Markov Processes Relat. Fields **27**, 245–313 (2021)

Markov Processes and **Related Fields**

Short Course on Fundamentals of Mathematics. Part I

V.A. Malyshev

Received January 12, 2020

Abstract. This is the first part of the shortest course which has various goals. Firstly, to confront somehow the growing fragmentation of mathematics. Also to create a feeling that mathematics is a very simple science, if one starts to study it not with very formal and complicated general theories but with simplest examples. Any mathematical mind has some picture of mathematics as the (Egyptian) pyramid, but it should not be brought in from outside, but to appear step-by-step from many intuitively evident assertions.

Keywords: short course of mathematics, algebra, analysis

AMS Subject Classification: 00-01

Contents

1. Introduction

This text has various goals. Firstly, to confront somehow the growing fragmentation of mathematics. Also to create a feeling that mathematics is a very simple science, if one starts to study it not with very formal and complicated general theories but with simplest examples. An idea of structure of the mathematics as the (Egyptian) pyramid should not be brought in from outside, but to appear step-by-step from many intuitively evident assertions. And if something is not quite evident then it should be brought out from the mind.

Mathematics is immense, you can develop it in depth, width, as well as in other, as yet unknown, directions. But it is difficult to get rid of the desire to cover more at a glance and understand what is more important and what is less. This desire is also pushed by the fact that mathematics becomes more and more fragmented on many domains, far apart from each other. And the picture of mathematical life resembles groups-clans of gold diggers, digging deeper and deeper, but do not know what neighboring clan is doing. Basically, new problems arise as natural (unfortunately, often formal) generalizations of already known

problems. And these deep caves can devour the digger for life. What is even worse – many areas remain untouched.

How to connect different areas of mathematics is a difficult question. It seems impossible to do this in the framework of pure mathematics, but surprisingly relation with other sciences, and primarily with physics, help a lot.First of all, because in physics there are global axioms. However, their rigorous formulations are absent. And if they exist in mathematical physics, their consistency is mostly not proved. Nevertheless, development of mathematical physics demands development of various areas of mathematics.

One of the goals of this text is to gather background information and more structured overview of very beginnings of mathematics, necessary for wide participation in development of mathematical physics. In particular, to involve students in this.

But, besides this, there are other goals:

1. This can raise interest to the mathematical structure of other sciences, including biology.

2. To understand what is needed to clearly imagine the mathematical structure of theoretical physics.

3. We tried to avoid excessive reliance on the school curriculum of mathematics, which often interferes with logical clarity. Therefore, the logical structure of the theory itself must always be visible before the eyes (partial order in the set of theorems). In other words, for any statement, the return path to the main axioms must be visible. On the other side, the intuition is always explicitly or implicitly present, which helps to see logical transitions.

4. Such text might look like a summary. That is why was our main problem was to combine brevity with accessibility. The thicker the book, even small logical steps are described in great detail. In our text logical transitions are described mostly without details but so that the reader could fill in the details without consulting with great monographs. But nevertheless, especially young mathematicians, should, looking through many mathematical books, choose themselves those that are more suitable for their perception.

5. Also this text can be useful for our Project [11] and be introduction to courses [9, 10].

2. Entering the language of mathematics

2.1. Sets

The basis of the mathematical language (i.e. the language in which mathematicians communicate and in which mathematical articles and books are written) are (besides natural numbers) two concepts: set and element of a set, sometimes referred to as undefinable (all the other concepts of mathematics can be defined through these):

1. **Element** – from the word elementary. Elements can be any objects. This term is used when we forget that element consists of something, has its own structure, but we know that he belongs to a particular set. Not to be confused: in physics there is the term "elementary particle" – but physicists mean that we do not know yet what does it consists of. The elements may be called points, objects, etc. and be labeled, for example, with Latin letters $x, y, \ldots$, or letters with indices $x_1, x_2, \ldots$, as if to emphasize that they are from the same set. It is important to keep in mind that the equality $x = y$ does not mean that elements x and y are similar but only that we used different letters to designate the same item. So. in the molecule or in the electric current there are many quite similar electrons, but they can be enumerated: to be able to follow trajectories of any of them.

2. **Set** (sometimes interchangeably used with other words – class, etc.). For example, the set of all trees in the wood, all letters of the Russian alphabet, all natural numbers greater than 5, etc. If the element x belongs to the set A, it is written $x \in A$. If, for example, the set A consists of three elements x_1, x_2, x_3, then it is written as

$$A = \{x_1, x_2, x_3\} = \{x_i : i = 1, 2, 3\}.$$

The simplest constructions of set theory A **subset** B of the set A is a set, elements of which can only be elements of A, but not necessarily all. More precisely, if $x \in B$. then necessarily $x \in A$. Write $B \subset A$ or $B \subseteq A$.

The **empty set**, denoted $\emptyset$, is a set which has no elements. It is a subset of any set, including itself.

Union $A \cup B$ of sets A and B is a set consisting of those and only those elements that belong to at least one of them. Similarly is defined the union of a greater number of sets.

Intersection $A \cap B$ of sets A and B is the set consisting of those and only those elements that belong to both, that is to A and B. Similarly can be defined the intersection of a larger number of sets.

The **difference** $A \setminus B$ of sets A and B is a set consisting of those and only those elements that belong to A but do not belong to B. If $B \subset A$, then $A \setminus B$ is sometimes called the complement of B in A.

Say that a set of pairwise disjoint subsets $A_1, \ldots, A_n$ of the set A (that is $A_i \cap A_j = \emptyset$ for all $i \neq j$) define a **partition** of A into subsets, if

$$A_1 \cup \ldots \cup A_n = A.$$

Cartesian product $X \times Y$ of two sets X and Y the set of all possible pairs (x, y) where $x \in X$ and $y \in Y$. In accordance with the notation $X \times Y$, the pairs are designated with the same order of elements, that is, in the first place (left) is element of X, and the second (right) – element of Y. Similarly is defined Cartesian product $X_1 \times X_2 \times \ldots \times X_N$ for any number of sets.

Any subset $A \subset X \times Y$ is called **relation**. If $(x, y) \in A$ one can imagine that x and y are related by some rule, denoted by A.

A **function** f from set X to Y (or a **mapping** of the set X to Y), often denoted as $f : X \to Y$, is defined as the subset $C_f \subset X \times Y$ such that any element $x \in X$ occurs in pairs of C_f exactly once. It would be more accurate to talk about **single-valued** function (or map), but the word single-valued is usually omitted. Then, if the pair (x, y) belongs to C_f, it is said that the function f maps $x \in X$ to $y = f(x) \in Y$, which is called value of the function f at point x.

Below we will see that in some exactly defined cases, the functions are called differently: operators, functionals, measures, etc.

Two sets A and B are called **isomorphic**, if there exists a subset F of the set $A \times B$ such that in its pairs each element of A and each element of B occur exactly once. In other words, a subset F defines a function $f : A \to B$ for which there exists a unique (called **inverse**) function $f^{(-1)}$ such that $f^{(-1)}(f(a)) = a$ for all $a \in A$. Such a function f is called **one-to-one**. One can say that isomorphic sets have the same **power** (generalization of the concept of number of elements). The set is called **finite** with number of elements N if it is isomorphic to the set $\{1, 2, \ldots, N\}$. The set is called **countable** if it is isomorphic to the set of all natural numbers $\mathcal{N} = \{1, 2, \ldots, N, \ldots\}$.

Note that the extended set of natural numbers $0, 1, 2, \ldots, N, \ldots$ is obtained by adding zero element and is denoted by $\mathcal{N}_0$.

2.2. Structures on sets

Structures on any set A often can be defined as special subsets of Cartesian products $A \times A, A \times A \times A, \ldots$. Examples:

1. Element of the N-fold Cartesian product $A \times A \times \ldots \times A$ is called an **ordered set** or a **sequence** $(a_1, a_2, \ldots, a_N) = a_1, a_2, \ldots, a_N$ of elements $a_i \in A$, or **words** $a_1 a_2 \ldots a_N$ from the **alphabet** A.

2. Any subset $G \subset A \times A$ defines a **graph**. The elements of set A are called **vertices**, while elements of the set G, that is the pairs $(a_1, a_2) \in G$ are called its **edges** connecting vertices a_1 and a_2. Geometrically one can imagine the edge as a (line) segment, connecting two points (vertices). The edge can be considered as directed from a_1 to a_2, and the graph is called **oriented graph** (without multiple directed edges, that is not more than one directed edge from a_1 to a_2).

3. A **partial order** on the set A is a subset $T \subset A \times A$ such that the following conditions hold:

3a) pairs with the same elements, i.e. of the form (a, a) cannot belong to T;

3b) if $(a, b) \in T$ then (b, a) does not belong to T. Then we say that a is greater (higher, more important, precedes) b (or b less a) and write $a > b$ (or $b < a$);

3c) if $a > b$ and $b > c$ then $a > c$.

If for every two elements we can say which one is greater, then this set is called **well ordered**. These are, for example, the sets $\mathcal{N}$ and $\mathcal{N}_0$.

If the set A is finite or countable, then with the partial order it is possible to associate a graph G as follows. The set (graph) G is obtained from the set T by deleting all pairs (a_1, a_2) such for which there is an "intermediate" element of a, that is such that $a_1 > a > a_2$. If each a, except one, called **root**, can occur in pairs (from G) on the right exactly once , then such partial order is called a **tree**. It is very intuitive to think that each element has at most one direct superior. Or to each node (except the root-source) there is only one incoming flow.

Algebraic operations. A set A is called a **group** if a function (operation) $f : A \times A \to A$ is defined, denoted by, for example, as multiplication

$$a = f(a_1, a_2) = a_1 a_2.$$

Under such notation, this operation is natural to call multiplication (a_1 multiplied from the right by a_2). It is required that the following conditions (axioms) hold:

1. there is a **unit** element $1 \in A$ such that $1a = a1 = a$;
2. for each $a \in A$, there exists its inverse element $a^{-1} \in A$ such that $aa^{-1} = a^{-1}a = 1$;
3. **associativity** – for all $a_1, a_2, a_3 \in A$

$$(a_1 a_2)a_3 = a_1(a_2 a_3).$$

A subset of a group is called a **subgroup** if it is a group relative to the same multiplication operation. Most of the groups are constructed as subgroups of the set $\Lambda = \Lambda_X$ of all one-to-one transformations f of some set X, which is the group under multiplication $f_1 f_2 = f_1(f_2)$. The unit of this group is the identical mapping I from X to itself, that is $I(x) = x$ for all $x \in X$.

If X is finite, then Λ_X is called the **permutation group**. A permutation is a bijective function $f : A \to A$, where $A = \{1, 2, \ldots, N\}$. Intuitively it determines on which place $f(k)$ will be placed the element which is on place k. There are $N!$ of such functions, and the product of two functions f_1 and f_2 is defined as $(f_1 f_2)(k) = f_1(f_2(k))$.

Problem 1. Elementary permutation is the permutation of two elements, that is, there exist $a_1, a_2 \in A$, such that $f(a_1) = a_2, f(a_2) = a_1$, and for the rest of a we have $f(a) = a$. Prove that any permutation f can be obtained as the product of such elementary permutations. The parity of the number of such permutations depends only on f. It is called the **parity** of permutation f and is denoted by $\sigma(f)$.

The group is called **commutative**, if $f_1 f_2 = f_2 f_1$ for all f_1, f_2. Often the operation in a commutative group is (or is called) the **addition** and is denoted by $f_1 + f_2$. For example, in the set of integers $Z = \{\dots, -2, -1, 0, 1, 2, \dots\}$, , which are obtained by adding zero and negative integers to the set of natural numbers. The integers form a group (under addition), where a unit is called zero, and inverse element to a is $-a$.

Another important algebraic object is a (commutative) **field**. A field is a commutative group with addition and zero element 0. Also $A \setminus \{0\}$ is a commutative group with multiplication and unit element 1. Moreover, also the **distributivity** axiom is assumed to hold:

$$a_1(a_2 + a_3) = a_1 a_2 + a_1 a_3.$$

The first example is the field of **rational numbers**, denoted by Q. If addition and multiplication exist for the integers, rational numbers arise from defining what is **division**. Intuitively we introduce one half, one third etc. of the whole, that is, numbers of the form $\frac{1}{n}$ (where n is an integer) as the number such that $n\frac{1}{n} = \frac{1}{n}n = 1$. **Formally**, we introduce the set of formal symbols of the form $\frac{p}{q}$, where p and $q \neq 0$ are integers, and define operations on them as we were taught at school

$$\frac{p_1}{q_1} + \frac{p_2}{q_2} = \frac{p_1 q_2 + q_1 p_2}{q_1 q_2}, \quad \frac{p_1}{q_1}\frac{p_2}{q_2} = \frac{p_1 p_2}{q_1 q_2}, \quad -\frac{p}{q} = \frac{-p}{q} = \frac{p}{-q}, \frac{p_1 q}{q p_2} = \frac{p_1}{p_2}.$$

Problem 2. Define when $\frac{p_1}{q_1} = \frac{p_2}{q_2}$ and prove that the set of all rational numbers is countable.

Remark 1. Many types of algebraic operations were invented. We are not going here to define them precisely, but only briefly describe the main ones (in addition to groups and fields):

linear space (commutative) – addition and multiplication by a number;
semigroup – only multiplication (no division);
ring – addition and multiplication;
algebra – addition, multiplication and multiplication by a number,
module is an abelian (i.e. commutative) group with the multiplication by elements of some ring.

Metrics and topology. Here all numbers (in particular, ε) are rational.

Metrics (or **distance**) on the set X is a symmetric function $\rho : X \times X \to Q$, that is, the function $\rho(x_1, x_2) = \rho(x_2, x_1)$ on $X \times X$ with values in the field of rational numbers, which is positive for $x_1 \neq x_2$ and is zero for $x_1 = x_2$. Moreover, it is required that the so-called triangle inequality holds: for all $x_1, x_2, x_3 \in X$

$$\rho(x_1, x_3) \leq \rho(x_1, x_2) + \rho(x_2, x_3).$$

A set with such metrics is called **metric space**.

Remark 2. It might seem strange that the metrics is introduced using the field of rational numbers. There are at least two reasons. The first is that we have not yet defined what is a real number. The second, more important, is that all existing physical theories are only approximations, and, for the practice, finite number of digits in the decimal representation of real number are sufficient. Although the field of real numbers is extremely convenient object, but doubtfully can be considered as physically real.

Neighborhood (more precisely, ε-neighborhood, where $\varepsilon > 0$) of the point x is the set $O_\varepsilon(x) = \{y : \rho(x, y) < \varepsilon\}$. It is a frequent notation which reads – the set of all $y \in X$, which are at a distance less than ε from x.

Now we come to two major concepts that permeate all mathematics.

A sequence of elements (points)

$$x_1.x_2, \ldots, x_k, \ldots \tag{2.1}$$

of some metric space X has the **limit** (limiting point) $x \in X$, if for any neighborhood O_ε of the point x there exists a natural number $N = N(\varepsilon)$ such that for all $k > N$ elements x_k belong to this neighborhood. It is said then that the sequence (2.1) converges to x (or is **convergent**).

A sequence of points (2.1) in a metric space X satisfies the **Cauchy condition** (or is called a **Cauchy sequence**) if for any $\varepsilon > 0$ there exists a natural number $N = N(\varepsilon)$ such that $\rho(x_k, x_m) < \varepsilon$ for all $k, m > N$. Figuratively speaking, the points cluster and group densely together.

Both of these definitions are automatically generalized from countable sequence to any well ordered set. So, assume that we have introduced full order on some subset $X' \subset X$. Then X' has the limit point $x \in X$ if for any neighborhood O_ε of the point x there is an element $x_\varepsilon \in X$' such that any $x' > x_\varepsilon, x' \in X'$, belongs to O_ε.

Metric space X is called **complete** if any Cauchy sequence has some limit $x \in X$.

Open set in the metric space X is defined to be either any neighborhood or any union of neighborhoods in any number.

Problem 3. Prove that the intersection of any finite number of open sets is an open set.

The **boundary** ∂A of open set $A \subset X$ is defined as the set of points $x \in X$ not belonging to A but such that any neighborhood of x has nonempty intersection with A. Subset A of the metric space X is called **closed** if its complement $X \setminus A$ is open.

A subset A of complete metric space is called compact set (or simply a **compact**) if any sequence of its points has a convergent subsequence. Equivalent definition (to prove !): from any coverage of set A by open sets, one can choose

a finite covering. The system of sets is called a **covering** of the set A. if their union contains A.

The system of open sets defines **topology** on the set X. Topology (i.e. open sets) can be introduced in more general way (without metrics) but we will not need this here.

Remark 3. (This is a general remark) There are two sources for development of mathematics:

1) Mathematical language with its structure of concepts. The number of people fluent in this mathematical language is very small, unlike for example Russian or English languages. However, the mathematical language is far from exhausting the mathematics.

2) Intuition in mathematics is based largely on non-mathematical disciplines (primarily physics), and life itself. Completely new concepts can arise from intuition. However, many deep developments, seemingly not associated with any life, appeared and gradually faded afterwards. Examples can be – the famous continuum hypothesis, analytic set theory. However, this does not mean that they will be forgotten or that they have nothing to do with our project. It is not excluded that someday they will form the basis of new theories. Moreover, such books as [1–3] will help to develop sense of logic in the life, and also to better understand the structure and depth of mathematics.

2.3. Real and complex numbers

We will give three definitions of real numbers. In each of them the most important thing is to state the conditions for equality of two objects, which we will call real numbers. So, real number is:

1. A Cauchy sequence $x_1, x_2, \dots, x_n, \dots$ of rational numbers. Two such sequences x_n and y_n define the same number iff for any (rational) $\varepsilon > 0$ there is such $N = N(\varepsilon)$ that for all $n > N, m > N$

$$|x_n - y_m| < \varepsilon$$

If Cauchy sequence has a limit in Q, then the so defined real number is rational. Thus, any rational number is real.

2. Cross section of the set of rational numbers, that is, a decomposition of the set of rational numbers Q into two subsets Q_1 and Q_2 such that, for all $x_1 \in Q_1, x_2 \in Q_2$ we have $x_1 < x_2$. All these real numbers are different and, moreover, are not rational except in the following cases: 1) there exists maximum of the set Q_1, that is a rational number $r \in Q_1$ such that $x < r$ for all other $x \in Q_1$; 2) there exists minimum of the set Q_2, that is a rational number $r \in Q_2$ such that $x > r$ for all other $x \in Q_2$. In these cases the section determines a rational number. Therefore, all rational numbers are real numbers.

3. Bilateral infinite sequence of ten digits, $a_n = 0, 1 \dots, 9$, of the form

$$\alpha = \dots a_{-m} \dots a_{-2} a_{-1}, a_0 a_1 a_2 \dots a_n \dots, \tag{2.2}$$

where $m, n > 0$ and only finite number of digits a_{-k}, with $k > 0$, are different from zero. Typically, the zeros $a_{-k}, k > 1$, to the left of all nonzero digits are not written, that is, decimal representation has the form

$$\alpha, a_0 a_1 \dots,$$

where $\alpha \geq 0$ is an integer. All numbers defined by these sequences are considered different, with one exception: any two numbers α and β like

$$\alpha = \dots a_{n-1} a_n a_{n+1} \dots, \quad \beta = \dots b_{n-1} b_n b_{n+1} \dots$$

are considered equal if for some n the following 5 conditions hold: 1) $a_k = b_k$ for all $k < n$; 2) $b_n \neq 9$; 3) $a_n = b_n + 1$; 4) $a_k = 0$ for all $k > n$; 5) $b_k = 9$ for all $k > n$. Speaking shortly, they are considered different except for the pairs of numbers of the form

$$\alpha, \beta 0000 \dots = \alpha, (\beta - 1) 9999 \dots$$

where α, β are natural numbers in decimal representation (in β the last digit is not zero). In particular,

$$1 = 0, 999 \dots$$

Otherwise, the real numbers form well-ordered set: $\alpha > \beta$ if there is (minimal) n such that $a_n > b_n$ and $a_k = b_k$ for all $k < n$.

Problem 4. Prove equivalence of all three definitions. That is, for every real number in the sense of one definition, specify the corresponding number in the sense of the other.

Problem 5. Define metrics $\rho(x.y) = |x - y|$ on the set of real numbers. Using each of these three definitions, prove that the unit interval $[0, 1]$ is compact in the sense of two definitions of compactness.

Real numbers form a field. To show this, for any real number (2.2) let $\alpha(N) = \dots b_n b_{n+1} \dots$ be the rational number such that $b_n = a_n$ when $n \leq N$, and $b_n = 0$ for $n > N$. Then addition, multiplication and division which are defined as the limits of operations on rational numbers of the form $\alpha(N)$ as $N \to \infty$.

Consider for example addition. We define the sum of two real numbers α and β:

$$\alpha + \beta = \lim_{N \to \infty} (\alpha_N + \beta_N),$$

where the symbol $\lim_{N\to\infty}$ is called the limit as $N \to \infty$ in the following sense. For any number γ and any $n \leq N$, denote $\gamma_n(N)$ the term with number n in the sequence $\gamma(N)$. Let α and β be given. Then for any n and for all $N > n+1$ the digits $\gamma_n(N)$ are the same.

The set of rational numbers Q can be considered as a subset of the set R of real numbers. For this it is necessary to use a decimal representation of rational numbers. For rational numbers $\frac{p}{q}$. performing standard division, it is easy to see that we get an infinite sequence of the form

$$\alpha, \beta\gamma\gamma\gamma\dots$$

where $\alpha, \beta > 0, \gamma > 0$ are integers in decimal representation. Vice versa, let us prove that any such sequence defines a rational number. Assume that the integer γ has exactly k digits:

$$0, \gamma\gamma\gamma\dots = \lim_{N\to\infty} \sum_{n=1}^{N} \frac{\gamma}{10^{kn}} = \frac{1}{1-10^{-n}} \frac{\gamma}{10^n}.$$

remembering algebraic formula from school

$$\frac{1-x^n}{1-x} = 1 + x + \dots + x^{n-1} \tag{2.3}$$

which in the limit gives

$$\lim_{n\to\infty} \frac{1-x^n}{1-x} = \frac{1}{1-x}. \tag{2.4}$$

Complex numbers. A complex number is written as an expression of the form $x + iy$, where x, y are real and i is a formal symbol called the **imaginary unit**.

Addition and multiplication of complex numbers are defined as follows:

$$(x_1 + iy_i) + (x_2 + iy_2) = (x_1 + x_2) + i(y_1 + y_2)$$

$$(x_1 + iy_i)(x_2 + iy_2) = (x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1)$$

(it is sufficient to remember that $i^2 = -1$). The set of all complex numbers is denoted C, and to check the (commutative) field axioms, it is sufficient to check that every element $a + bi$ has an inverse

$$\frac{1}{a^2+b^2}(a - bi)$$

and of course $0 = 0 + 0i, 1 = 1 + 0i$. Metrically, C is identified with the plane R^2: a complex number $x + iy$ is identified with the point $(x, y) \in R^2 = R \times R$.

The unit circle $S = \{a + bi : a^2 + b^2 = 1\}$, as a subset of the set of complex numbers, is invariant under multiplication and division but not addition. That is, it is a commutative group w.r.t. multiplication. For example, squaring gives

$$(a + bi)^2 = a^2 - b^2 + 2abi \Longrightarrow (a^2 - b^2)^2 + 4a^2b^2 = (a^2 + b^2)^2 = 1,$$

and also for multiplication and division

$$\begin{aligned}(a + bi)(c + di) &= ac - bd + (ad + bc)i \\ \Longrightarrow a^2c^2 + b^2d^2 - 2abcd + a^2d^2 + b^2c^2 + 2abcd &= \\ = a^2c^2 + (1 - a^2)d^2 + a^2d^2 + (1 - a^2)c^2 &= 1, \\ \frac{1}{a + bi} = \frac{a - bi}{(a + bi)(a - bi)} &= \frac{a - bi}{a^2 + b^2} = a - bi.\end{aligned}$$

The absolute value (module) of the complex number $z = x + iy$ is defined as, see the definition of the square root in (6),

$$r = |z| = \sqrt{x^2 + y^2}.$$

Metrics in C, or on two-dimensional plane, is defined as

$$\rho(z, y) - \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2},$$

where $x = x_1 + ix_2, y = y_1 + iy_2$.

3. Basics of analysis

3.1. Derivatives and integrals

One-dimensional real analysis deals with functions $f : A \to R$, and complex analysis – with functions $f : A \to C$, where A is some subset of R or C respectively. Most often, these subsets are open or closed.

Remind that in any metric space an open set is a union of neighborhoods of its points. And on the real line, every open set is the union of finite or countable set of open intervals, that is sets of the form $(a, b) = \{x : a < x < b\}$ for some $a < b$.

The first ideas on the function under consideration usually appear in the form of this function **graph**. If, for example, it is a function on R then the function graph is a curve in the plane along the real axis. The simplest functions are constant and piece-wise constant functions. This means that the domain is divided into finite or countable number of subsets, on each of which the function is constant. More complicated curves can have different degrees of smoothness. However, this requires more detailed definitions.

3.1.1. Scales of function behavior

Often, it is necessary to compare functions through the rate of growth or decrease. For example, it is clear that. x^{n+1} grows faster than x^n as $x \to \infty$.

But try, using only the definitions of addition and multiplication, to prove that $n^{1000}2^{-n} \to 0$ when $n \to \infty$. Or, for example, how to compare speed of growth of the following functions (with a constant $a > 0$), as $N \to \infty$:

$$a, \log_a N, N^a, 2^{aN}, N^{aN}, \dots$$

But of course, we should start with rigorous definition of these functions (see below in this Chapter). And the analysis develops methods to solve much more complicated problems.

There are three commonly used notations. For example, consider the behavior of two functions f, g, defined in some neighborhood of point a:

1. $f(x)$ is said to be O-large of $g(x)$ as $x \to a$, and write $f(x) = O(g(x))$ if there exists $C > 0$ such that for all x in some neighborhood of the point a

$$|f(x)| \le C|g(x)|.$$

2. Let $g \ne 0$ as $x \to a$. Then f is said to be o-small of g as $x \to a$, and write $f = o(g)$, if

$$\lim_{x \to a} \frac{f(x)}{g(x)} = 0.$$

3. f and $g \ne 0$ are asymptotically equivalent as $x \to a$ ($f(x) \sim g(x)$) if

$$\lim_{x \to a} \frac{f(x)}{g(x)} = 1.$$

But often you need a more accurate comparison, called an asymptotic expansion. For example, the formula

$$f(N) = N^2 + N + 1 + o(1)$$

for the **asymptotic expansion** of the function $f(N)$ up to o(1) as $N \to \infty$. This means that $f(N) - (N^2 + N + 1) \to 0$. From this it follows that $f(N) \sim N^2$.

Continuity and limits The function f, defined in some neighborhood of the point x, is called **continuous** at x if for any (sometimes it is said "arbitrarily small") $\delta > 0$ there exists $\varepsilon = \varepsilon(\delta, x) > 0$ such that for all $y \in O_{\varepsilon(\delta x)}(x)$

$$|f(x) - f(y)| < \delta.$$

Shortly one can say that

$$f(y) - f(x) \to_{y \to x} 0.$$

Function defined on open or closed set A is called continuous on A if it is continuous at each point $x \in A$. If for all $\delta > 0$ there exists $\varepsilon = \varepsilon(\delta)$, independent of x, then we say that f is **uniformly continuous** on the set A.

It is not difficult to prove, using the second definition of compactness, that continuous function on the segment $[a, b]$ is uniformly continuous on it.

More generally, right (left) limits on the real axis are defined respectively as

$$\lim_{y \to x+0} f(y) = f(x), \quad \lim_{y \to x-0} f(y) = f(x),$$

where $y \to x + 0$ ($y \to x - 0$) means that only sequences y_k such that $y_k > x$ ($y_k < x$) are considered. If these limits exist but are different, then their difference is called the **jump** at this point. There can be more complicated system of breaks:

1) at a given point it is possible that both right and left limits do not exist. For example, limits at zero the function $\sin \frac{1}{x}$ (see below the definition of trigonometric functions);

2) limits may not exist at each point. An example is a function equal 1 in every rational and 0 at every irrational point.

The set of all continuous functions on A is denoted by $C_0(A)$.

Function f is called **monotone,** if it is either **increasing** or **decreasing**, that is for all $x < y$ either $f(x) < f(y)$ or $f(x) > f(y)$.

Problem 6. Prove that the function x^2 maps the segment $[0, \infty)$ to itself is one-to-one. And that inverse function to it, denoted $\sqrt{x}$, on the same segment is continuous.

Smoothness. Significantly stronger constraint (than continuity) is the Lipschitz condition on the function f on some set Λ, where it is defined: there exists $C > 0$ such that for all $x, y \in \Lambda$ we have $|f(x) - f(y)| < C|x - y|$. Even stronger condition – the existence of the following limit (**differentiability** at x)

$$\lim_{\varepsilon \to 0} \frac{f(x+\varepsilon) - f(x)}{\varepsilon} = \lim_{\varepsilon = \Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta f(x)}{\Delta x}, \qquad (3.1)$$

where $\Delta x = \varepsilon$ is called the increment of the argument, and $\Delta f = f(x + \Delta x) - f(x)$ is the increment of the function (for a given increment of the argument). This limit is called the **derivative** of the function at the point x and is denoted by

$$\frac{df}{dx}(x) = f'(x) = f_x(x) = f^{(1)}(x),$$

where the notation dx, df is often called the infinitely small (as if in the limit) increments. Then the function f is called **differentiable** at the point x. The

condition of differentiability can be reformulated in the form of a **Taylor decomposition**

$$f(x+\Delta x) = f(x) + f'(x)\Delta x + o(\Delta x).$$

Indeed, if we denote the difference

$$g(\Delta x) = \frac{f(x+\Delta x) - f(x)}{\Delta x} - f'(x),$$

then $g(\Delta x) = \mathrm{o}(1)$, and hence $g(\Delta x)\Delta x = o(\Delta x)$.

A function is called differentiable (**smooth**) on some interval if it is differentiable at each point of this interval.

If the function $f'(x)$ also has a derivative, it is called the second derivative, and is denoted by

$$\frac{d^2 f}{dx^2}(x) = f"(x) = f_{xx}(x) = f^{(2)}(x).$$

Then we say that f is twice differentiable. By induction the derivatives of any order $n = 3, 4, \ldots$ can be defined, if they exist.

3.1.2. Calculation of derivatives

Polynomials

$$\frac{d(x^n)}{dx} = \lim \frac{(x+\varepsilon)^n - x^n}{\varepsilon} = \lim \frac{1}{\varepsilon}(\varepsilon n x^{n-1} + O(\varepsilon^2)) = n x^{n-1}$$

and then by induction in $k \le n$

$$\frac{d^k x^n}{dx^k} = n(n-1)\ldots(n-k+1)x^{n-k}.$$

Due to linearity of differentiation (derivative of sum of functions is equal to the sum of the derivatives), polynomials of degree n are differentiable arbitrary number of times, with all derivatives, since the degree $n+1$, equal zero.

Assume that the function $f(x)$ is defined in a neighborhood of point x_0. Then the following expression is called Taylor expansion of order n of $f(x)$ at x_0, if there exist constants $d_0, d_1, \ldots, d_n$ such that for $x - x_0 \to 0$

$$f(x) = \sum_{k=0}^{n} d_k (x - x_0)^k + o((x - x_0)^n. \tag{3.2}$$

Then the derivatives $f^{(k)}(x_0)$ exist for all $k \le n$, and are equal to $k! d_k$. Vice-versa, denote the difference

$$\xi(x) = f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k,$$

and differentiating, we get that all derivatives $\frac{d^k \xi(x)}{dx^k}$ are equal to zero for $k = 0, 1, \ldots, n$.

Derivative of the product Assume that f and g are both differentiable at x. Then

$$\frac{1}{\varepsilon}(f(x+\varepsilon)g(x+\varepsilon)-f(x)g(x))=$$

$$=\frac{1}{\varepsilon}(f(x+\varepsilon)g(x+\varepsilon)-f(x)g(x+\varepsilon))+\frac{1}{\varepsilon}(f(x)g(x+\varepsilon)-f(x)g(x))=$$

$$=\frac{1}{\varepsilon}g(x+\varepsilon)(f(x+\varepsilon)-f(x))+\frac{1}{\varepsilon}f(x)(g(x+\varepsilon)-g(x))\to f'(x)g(x)+f(x)g'(x).$$

As an example let us find the derivative of $\frac{1}{q(x)}$ at the point x, where the function $q(x)\neq 0$:

$$0=\Big(q\frac{1}{q}\Big)'=q'\frac{1}{q}+q\Big(\frac{1}{q}\Big)'\Longrightarrow\Big(\frac{1}{q}\Big)'=-\frac{q'}{q^2}.$$

The derivative of composite function Let f and g be differentiable respectively in the points $g(x)$ and x. Then

$$\frac{f(g(x+\varepsilon))-f(g(x))}{\varepsilon}=$$

$$=\frac{f(g(x)+g'(x)\varepsilon+o(\varepsilon)))-f(g(x))}{\varepsilon}=$$

$$=\frac{f(g(x))+f'(g(x))(g'(x)\varepsilon+o(\varepsilon))+o(g'(x)\varepsilon+o(\varepsilon)))-f(g(x))}{\varepsilon}$$

$$\to_{\varepsilon\to 0}f'(g(x))g'(x).$$

Derivative of inverse function. Let $y(x)$ maps the interval (a,b) one-to-one onto some other interval (a_1,b_1), and let $x(y)$ be the inverse the function on (a_1,b_1). Both functions are assumed to be **continuously differentiable**, that is, having continuous derivatives. Then $x(y(x))=x$ and, as above,

$$\frac{d(x(y(x)))}{dx}=\frac{dx}{dy}\frac{dy}{dx}=1\Longrightarrow x'(y)=\frac{1}{y'(x)}.$$

3.1.3. Sums, series, integrals

Series is a formal expression of the form $\sum_{k=0}^{\infty}a_k$, where a_k can be real (or complex) numbers or functions. In the latter case, the series called a **functional series**. Series is called **convergent**, if the sequence of partial sums

$$S_N=\sum_{k=0}^{N}a_k$$

converges.

Remark 4. If all a_k have the same sign, then the order of terms in the series does not play role for convergence. Otherwise, the order of terms of the series may be important. Moreover, sometimes you need to talk about the sum of a countable number of terms regardless of their enumeration. Then it is necessary first to sum up all a_k which are non-negative and secondly – all negative. If at least one of these two sums is finite, then the integral is defined as their sum. When both are infinite you should invent different way of summation. A series is called **absolutely convergent** if an increasing sequence of positive numbers $\sum_{k=0}^{N} |a_k|$ is bounded. In this case, the order of its terms does not matter.

In the theory of **cluster expansions** the series are more complicated – their terms are numerated by finite subsets of vertices of countable graphs.

Integral of piece-wise constant function. Consider an increasing sequence of real numbers $x_0 < x_1 < x_2 < \dots$ and the (piece-wise-constant) function $f(x)$ on $[x_0, \infty)$ equal to a_k on half-intervals $[x_k, x_{k+1})$, where $k = 0, 1, \dots$. Integrals of this function over the intervals $[x_0, x_N)$ (or the whole $[x_0, \infty)$) are by definition the following sums (or series)

$$\int_{x_0}^{x_N} f(x)dx = \sum_{k=0}^{N-1} a_k(x_{k+1} - x_k),$$

$$\int_{x_0}^{\infty} f(x)dx = \sum_{k=0}^{\infty} a_k(x_{k+1} - x_k).$$

The last integral is said to be convergent if the series converges.

The integral of a continuous function Now we want to extend the definition of the integral to functions that are not piece-wise constant. For example, consider the function $f(x) = x$ on the interval $[0, y]$. The natural desire is to find a sequence of piece-wise constant functions $f_N(x)$ on $[0, y], y \geq 0$, uniformly converging to $f(x)$ as $N \to \infty$. And define the integral of $f(x)$ as the limit of integrals of piece wise constant functions

$$\int_0^y f(x)dx = \lim_{N\to\infty} \int_0^y f_N(x)dx.$$

In this case it is very easy to prove the convergence, and even to calculate this limit. Put $f_N(x) = x_k = \frac{k}{N}$ if $x_k = \frac{k}{N} \leq x < x_{k+1} = \frac{k+1}{N}$. Then, for any $0 \leq y$ there exists m such that $0 \leq y - \frac{m}{N} \leq \frac{1}{N}$. Then

$$S_N(y) = \int_0^y f_N(x)dx = \sum_{k=0}^{m} \frac{k}{N}\frac{1}{N} + O\Big(\frac{1}{N}\Big) =$$

$$= \frac{1}{N^2}\frac{m^2}{2} + O\Big(\frac{1}{N}\Big) = \frac{y^2}{2} + O\Big(\frac{1}{N}\Big) \to \frac{y^2}{2}. \tag{3.3}$$

Similarly, for any continuous (say on the interval $[0,1]$) function $f(x)$ one can prove existence of the limit of integrals of piece wise continuous functions $f_N(x) = f(x_k)$, where x_k – point of the form $\frac{k}{N}$. nearest to x from the left.

Problem 7. Consider again a continuous function $f(x)$ on finite interval $[a,b]$. We call $R(f,B) = \sup_{x\in B} f(x) - \inf_{x\in B} f(x)$ the scope of function f on the set $B \subset [a,b]$. Here **supremum** $\sup_{x\in B} f(x)$ is defined as the minimal number, greater than any $f(x), x \in B$. Similarly, **infimum** $\inf_{x\in B} f(x)$ is the maximal number, less than any $f(x), x \in B$.

Prove that for any $\varepsilon > 0$, any segment A can be divided into finite number of half-intervals $A_k(\varepsilon), k = 1,\dots,N = N(\varepsilon)$ so that $R(f, A_k(\varepsilon)) < \varepsilon$ for all k . Use the uniform continuity of f on A.

However we cannot calculate this integral as above. Let us come back to the formula (3.3) and note that the function x is a derivative of its integral $\frac{x^2}{2}$. It turns out that this miracle is the rule in the general case. For this note two properties of additivity of the integral, which are derived from the definition:

$$\int_0^y (f_1(x) + f_2(x))dx = \int_0^y f_1(x)dx + \int_0^y f_2(x)dx,$$

$$\int_0^y f(x)dx = \int_0^{y_1} f(x)dx + \int_{y_1}^y f(x)dx, \;\; 0 \le y_1 \le y.$$

In the one-dimensional case, **anti-derivative** (or **indefinite integral**) of function $f(x), x \in A = [a,b] \subset R$, is called any function (if they exist) on A whose derivative is $f(x)$.

Problem 8. Prove the uniqueness of indefinite integral (up to addition of a constant), that is, if the function is differentiable and its derivative is identically zero on A, then the function is a constant.

In order to understand the connection of the integral with the anti-derivative we should always remember that the derivative corresponds to the difference and the integral – to the sum of the function values. Therefore, we divide the interval $[a,b)$ into N parts

$$a = x_0 < x_1 < \ldots < x_N = b,$$

where there exist constants $0 < C_1 < C_2$ such that for all k

$$\frac{C_1}{N} < x_{k+1} - x_k < \frac{C_2}{N}.$$

Take the function f_N equal to $f(x_k)$ on the interval $[x_k, x_{k+1}), k = 0,1,\dots,N-1$. Then its integral can be written as

$$\int_a^b f(x)dx = \lim \sum_{k=0}^{N-1} f(x_k)(x_{k+1} - x_k).$$

If $f(x)$ is continuous on $[a,b]$, the limit clearly exists. Put $b=x$ and find the derivative at x of the integral as the limit

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_x^{x+\varepsilon} f(t)dt = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_x^{x+\varepsilon}(f(x)+O(\varepsilon))dt = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}(\varepsilon f(x)+o(\varepsilon)) = f(x) \tag{3.4}$$

Therefore, this integral is anti derivative of $f(x)$. But the easiest way to remember the formula for the sum of the differences

$$f(x_N)-f(x_0)=\sum_{k=0}^{N-1}(f(x_{k+1})-f(x_k)) = \sum_{k=0}^{N-1}\frac{f(x_{k+1})-f(x_k)}{x_{k+1}-x_k}(x_{k+1}-x_k)$$

and its limit

$$f(b)-f(a)=\int_A f'(x)dx=\int_a^b f'(x)dx.$$

Due to this, often the evaluation of the integral will be reduced to the corresponding methods of calculating the derivatives. For example, the method of **integration by parts** allows to reduce the calculation of one of the integrals in the left part to another

$$\int_a^b fgdx+\int_a^b fg'dx=\int_a^b (fg)'dx=f(b)g(b)-f(a)g(a).$$

3.1.4. Length of a curve

We define d-dimensional space $R^d=\{(x_1,\ldots,x_n)\}$ as the set of all sequences $(x_1,\ldots,x_n)$ of real numbers $x_k\in R$. We can consider it as complete metric space with the euclidean distance $\rho(x,y)=\sqrt{\sum_{k=1}^d (y_k-x_k)^2}$.

Here, the **curve** in R^d=$\{(x_1,\ldots,x_n)\}$ is a map of the interval [a,b] to R^d. It can be defined by a continuous vector function $x(t)=(x_1(t),\ldots,x_n(t))$, that is all functions $x_k(t)$ are continuous, where $t\in[a,b]$. A curve is called a **straight** line if all the functions $x_k(t)$ are linear. A curve is called **broken** (or piece wise linear) if it consists of a finite number of straight lines, that is, if for some points $a=t_0^{(N)}<t_1^{(N)}<\ldots<t_N^{(N)}=b$ it is a straight on each of the segments $[t_{k-1},t_k]$ for $k=1,\ldots,N$. The length of a straight line is the distance between its start and end points. The length of broken line is the sum of the lengths of its straight segments. A curve is called **smooth** if the functions $x_k(t)$ are smooth. The length of such a curve is the limit of the lengths of piece wise linear curves defined by the functions $x^{(N)}(t)=(x_1^{(N)}(t),\ldots x_d^{(N)}(t))$ if for $N\to\infty$ we have $max_k|t_k^{(N)}-t_{k-1}^{(N)}|\to 0$ and uniformly in t and k

$$|x_k(t)-x_k^{(N)}(t)|\to 0.$$

We first consider the length L of the curve $f(x)$ in R^2 along the real axis x. Let $f(x)$ be given on the interval $[a.b]$ with the continuous derivative $f'(x)$. In our case of the curve defined by the function $f(x)$, when constructing the broken lines, divide the segment into segments of the same length $\Delta x - \frac{b-a}{N}$. Then the length L_N of the broken line defined by the points $(a+\frac{k}{N}(b-a), f(a+\frac{k}{N}(b-a))), k = 0, 1, \dots, N$ will be

$$L_N = \sum_k \sqrt{(\Delta x)^2 + (\Delta f)^2} = \sum \sqrt{(\Delta x)^2 + (\Delta f)^2}\frac{\Delta x}{\Delta x} =$$

$$= \sum \sqrt{1 + \frac{(\Delta f)^2}{(\Delta x)^2}}\Delta x \to L = \int_a^b \sqrt{1 + \Big(\frac{df}{dx}\Big)^2}dx.$$

Now let a curve be given on $R^2 = (x_1, x_2)\}$ by two equations $x_1(t), x_2(t)$, where the parameter varies within $0 \le t \le s$. Then its length is defined as the limit of the lengths of broken lines approximating this curve. And similarly,

$$L = \int_0^s \sqrt{\Big(\frac{dx_1}{dt}\Big)^2 + \Big(\frac{dx_2}{dt}\Big)^2}dt.$$

Generalizing to larger dimensions is simple.

3.1.5. Multiple integrals and measures

Algebra of sets and measures. Consider arbitrary set Ω with the elements $\omega \in \Omega$ and some set Σ of its subsets. The latter will be called **algebra of sets**, if $\Omega \in \Sigma$ and for any subsets $A, B \in \Sigma$ their union, intersection and difference also belong to Σ. Measure is defined, for given pair (Ω, Σ) as a real valued function $\mu(A), A \in \Sigma$, on Σ, if the following property (finite additivity) holds: $\mu(\cup A_k) = \sum \mu(A_k)$ for any finite number of non intersecting sets $A_k \in \Sigma$. Examples:

1. Let $\Omega = \Omega_N = [-N, N] \subset R$ with arbitrary N. Here Σ is the minimal algebra, containing all points and all intervals $(a, b) \subset \Omega$. Their measures will be correspondingly 0 and $b-a$. Then Σ will contain all segments, half-intervals and their finite unions. It is easy to prove that this measure can be uniquely defined on all Σ. This measure is called **Lebesgue measure** on Ω_N, and in the limit $N \to \infty$ Lebesgue measure on R.

2. Let $\Omega = \Omega_N \times \Omega_N = [-N, N] \times [-N, N] \subset R^2$. Remind that open set $O \subset R^d$ is called **connected**, if for any two points $x, y \in O$ there exists continuous line inside O, connecting these two points. Here Σ will be the minimal algebra, containing all, belonging to Ω, points, segments and connected open sets-polygons, the boundary of which is a closed piece-wise straight line. Their measure is defined to be 0 for points and segments, and the area in the latter case. This measure is also called Lebesgue measure (on R^2).

Now we can define the integral of continuous function $f(x,y)$ over any bounded open set $D \in \Sigma$. Consider any partition $\mathbb{Q}_\varepsilon$ of D on finite number sets $A_k \in \Sigma$ with diameter less than $\varepsilon > 0$ (prove that such partitions exist) and with measures $\mu_k = \mu(A_k)$. Define the **integral sum**

$$S(\mathbb{Q}_\varepsilon) = \sum_k f(x_k, y_k)\mu(A_k),$$

where $(x_k, y_k) \in A_k$ for all k. Then for any sequence ε_k, any partitions $\mathbb{Q}_{\varepsilon_k}$ and any choice of points $(x_k, y_k) \in A_k$, the sequence $S(Q_{\varepsilon_n})$ is the Cauchy sequence (the proof is completely the same as in one-dimensional case). It follows that the limit

$$\int\int_D f(x,y)dxdy = \lim_{n\to\infty} S(Q_{\varepsilon_n})$$

exists. It is called the integral of f in measure μ over the set D.

It is easy to generalize all these definitions to open domains with piece wise smooth boundary. Namely, it is sufficient to approximate such boundary by piece-wise linear curves.

Now very important question – how to calculate or estimate such two-dimensional integrals. Consider, for example, the rectangle $D = \{a_1 \le x \le b_1, a_2 \le y \le b_2\}$ in the plane and a continuous function $f(x,y)$ on it. Sets A_k here will be rectangles with each side parallel to one of the coordinate axes. Then the integral is defined as the limit, when $N_1, N_2 \to \infty$, of the sum

$$\sum_{k=0}^{N_1-1}\sum_{m=0}^{N_2-1} f\Big(a_1 + \frac{kL_1}{N_1}, a_2 + \frac{mL_2}{N_2}\Big)\frac{1}{N_1}\frac{1}{N_2}$$

where $L_i = b_i - a_i$, wherein the double limit $N_1, N_2 \to \infty$ can be understood whatever. For example, as $N_1 = N_2 \to \infty$.

But another approach is more useful. We saw above that we can fix sufficiently large N_2 an then choose N_1 even much larger, that is to consider very narrow rectangles. Even more, for given N_2, m we pass to the limit $N_1 \to \infty$. Otherwise, you can say that we compute one-dimensional integrals of $f(x, a_2 + \frac{mL_2}{N_2})$ in x, or even the integral of $f(x,y)$ for each $y \in [a_2, b_2]$. The result is a function $F(a_2 + \frac{mL_2}{N_2})$ or $F(y)$, and then we calculate the integral over y. And we get

$$\int\int_D f(x,y)dxdy = \int_{a_2}^{b_2}\left(\int_{a_1}^{b_1} f(x,y)dx\right)dy.$$

This method is called **iterated integration**. It is very important as to find integral it allows to use (much easier) calculation of derivatives.

Everywhere above our sets were bounded, and our partitions and sums were finite. It is possible to develop more general constructions. The algebra is called

σ-algebra if Σ contains Ω itself and is closed relatively to countable unions and intersections. In the latter case, the pair (Ω, Σ) is called a **measurable space**.

Structure of σ-algebras is very complicated, although its elements (called **measurable sets**) have interesting hierarchical structure, which was studied in the science called "Analytic set theory". Usually, σ-algebras are constructed as minimal σ-algebra, containing some class Σ_0 of subsets, called the **base**. If Σ_0 is all open sets, then this σ-algebra is called .

For given (Ω, Σ) **measure** is a real function $\mu(A), A \in \Sigma$ on Σ such that the property of **countable additive** holds: $\mu(\cup A_k) = \sum \mu(A_k)$ for countable number of non-intersecting sets $A_k \infty \Sigma$. Here of course one should avoid cases when there are two sets A_1, A_2 such that $\mu(A_1) = \infty, \mu(A_2) = -\infty$. Just because we do not know then the measure of their union.

3.2. Power series

Formal power series is the following expression

$$f(z) = \sum_{n=0}^{\infty} a_n z^n. \tag{3.5}$$

The simplest example of such a series is a geometric progression with the $a_n \equiv 1$. Below, we mainly consider the case when a_n and z are complex numbers.

The series (3.5) converges in some neighborhood of zero, if for some $\varepsilon > 0$ it converges for all z with $|z| < \varepsilon$. For example, a geometric progression converges for all z with |z|<1.

Suppose that we are given a series (3.5). Then the formal n-fold differentiation at zero yields

$$a_n = \frac{f^{(n)}(0)}{n!}.$$

Vice-versa, if we can calculate or estimate the modules of derivatives of f at point 0, it is possible to check the convergence of the **Taylor series**

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} z^n.$$

In the case of convergence we will get what is called **the expansion of f in a series** at the point 0.

Let us now consider the complex valued function $f(z)$ defined on some open set A of complex plane. If for the point $z_0 \in A$ there exists the limit of

$$\frac{f(z) - f(z_0)}{z - z_0}$$

as $z \to z_0$ (more exactly $|z - z_0| \to 0$), it is called the complex derivative and is denoted $f'(z_0) = \frac{df}{dz}(z_0)$.

Remark 5. The function $f(z)$ of complex variable z can be regarded as a complex function $f(x+iy)$ of two real variables x, y. Its partial derivatives in x and y are defined as

$$\frac{\partial f}{\partial x} = \lim_{x\to 0} \frac{f(z_0+x)-f(z_0)}{x}, \quad \frac{\partial f}{\partial y} = \lim_{y\to 0} \frac{f(z_0+iy)-f(z_0)}{y}.$$

If the complex derivative of $f(z)$ exists, then

$$\frac{\partial f}{\partial x} = \frac{df}{dz}, \quad \frac{\partial f}{\partial y} = i\frac{df}{dz},$$

and therefore the following condition holds

$$\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y} = 0.$$

Complex function $f(z), z \in C$, is called **analytic** at the point z_0, if it coincides with a power series in $z - z_0$, converging when $|z - z_0| < \varepsilon$ for some $\varepsilon > 0$. It turns out that then $f(z)$ is analytic at any point z of this neighborhood. But there is a much stronger statement.

Theorem 1. *If $f(z)$ is complex differentiable at every point of some open set A, it is analytic at each point z of the set A, and moreover, the corresponding power series converges in any circular neighborhood $O(z)$ such that $O(z) \subset A$.*

This statement will be proved below.

3.2.1. Main functions

Rational functions. These are functions of form $\frac{P(z)}{Q(z)}$, where $P = p_0 + p_1 z + \ldots + p_m z^m$ and $Q = q_0 + q_1 z + \ldots + q_n z^n$ are polynomials. If $Q(0) \neq 0$, that is if $q_0 \neq 0$, then Q^{-1} (and therefore also $\frac{P}{Q}$) is expanded in a power series in z in some neighborhood of the point 0

$$Q^{-1} = \frac{1}{q_0}\frac{1}{1-y} = \frac{1}{q_0}\sum_{k=0}^{\infty} y^n, \quad y = -\Big(\frac{q_1}{q_0}z + \ldots + \frac{q_n}{q_0}z^n\Big),$$

which converges for $|y| < 1$, and therefore also for sufficiently small $|z|$. Similarly, if $Q(z_0) \neq 0$ for some point z_0 then Q^{-1} can be expanded in a power series in $z - z_0$ in some neighborhood of z_0.

Remark 6. Note that any polynomial Q can be represented (as for $z_1 = 0$) as $Q = (z - z_1)Q_{n-1} + C$, where Q_{n-1} is some polynomial of degree $n-1$ and C is a constant. But if $Q(z_1) = 0$, then is divisible by $z - z_1$, that is $C = 0$. It

follows by induction (assuming also that each polynomial has at least one root) that Q can be represented in the form

$$Q = C(z - q_1) \dots (z - q_n) \tag{3.6}$$

for some complex numbers $C, q_1, \dots, q_n$, where q_k are not necessary different. It follows that Q cannot have more than n roots. Below we prove that Q has exactly n roots, that is, it can be represented in the form (3.6) (the main theorem of algebra).

The number e. There are two equivalent definitions of the number e:

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} \Longleftrightarrow e = \lim_{n\to\infty} \Big(1 + \frac{1}{n}\Big)^n. \tag{3.7}$$

Convergence of the series is proved majorizing it by geometric progression

$$1 + \sum_{n=0}^{\infty} \frac{1}{2^n} = 3.$$

For the proof of the formula (3.7) it is easy to estimate from above the modulus of the difference

$$\sum_{n=0}^{N} \frac{1}{n!} - \Big(1 + \frac{1}{N}\Big)^N.$$

To do this one should perform some $(\pm)$-cancellations.

The function e^z. Absolutely similarly we introduce the function $z(t) = e^t$

$$e^t = \sum_{n=0}^{\infty} \frac{t^n}{n!} \Longleftrightarrow e^t = \lim_{n\to\infty} \Big(1 + \frac{t}{n}\Big)^n$$

or as a function $z = z(t)$ satisfying the equation

$$\frac{dz}{dt} = z,$$

that is, as (the only, up to a constant multiplier) function which coincides with its derivative, and hence with all its derivatives.

The main property is

$$e^{x+y} = e^x e^y \tag{3.8}$$

that can be prove by multiplication of series.on the right side and using obvious combinatorial formulas

$$x^m y^n \frac{1}{(m+n)!} C_{m+n}^m = x^m y^n \frac{1}{m!n!}\,.$$

In particular, $e^z e^{-z} = 1$, which implies that e^z never vanishes, and $e^{-z} = \frac{1}{e^z}$. In addition, e^t is increasing on $[0, \infty)$ from 1 to ∞, and on $[0, -\infty)$ decreases from 1 to 0. And thus, it takes all positive values. Below we show that it takes all complex values except zero.

Trigonometric functions. Functions $\cos x, \sin x$ are introduced as, respectively, the real and imaginary parts of the series for e^{ix}

$$\cos x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}, \quad \sin x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!},$$

It follows that they are real for real x, and by differentiation of the series we get the formulas

$$(\sin x)' = \cos x, \ \ (\cos x)' = -\sin x. \tag{3.9}$$

Let us check the formula

$$\sin^2 x + \cos^2 x = 1, \tag{3.10}$$

which implies the boundedness of these functions for real x. To prove this one should square the two series, write the formula for the coefficient of x^{2n} and to show that they are canceled for any $n > 0$

$$2 \sum_{k=0}^{n} (-1)^k \frac{x^{2k}}{(2k)!} (-1)^{n-k} \frac{x^{2(n-k)}}{(2n-2k)!} = (-1)^n 2x^{2n} \frac{1}{(2n)!} \sum_{k=0}^{n} C_{2n}^{2k},$$

$$2 \sum_{k=0}^{n-1} (-1)^k \frac{x^{2k+1}}{(2k+1)!} (-1)^{n-k-1} \frac{x^{2(n-k)-1}}{(2n-2k-1)!} = (-1)^{n-1} 2x^{2n} \frac{1}{(2n)!} \sum_{k=0}^{n-1} C_{2n}^{2k+1}.$$

The last sum corresponds to the decomposition of $(1-1)^{2n}$, according to the binomial theorem.

Formula (3.10) shows that any real φ defines a point $(x, y) = (\cos\varphi, \sin\varphi)$ of the unit circle S in the plane, and this number formally can be called the angle. Then this angle is zero for the point (1,0). Usually the angle is measured from point (1,0) counterclockwise. Then the angle corresponds to the length of the corresponding arc of the circle from the point (1,0) to the point (x,y).

The length of the curve $y = f(x)$ (e.g. the length $\varphi(x)$ of the segment of the unit circle over the interval $[0, x] \subset R$, in the positive quarter plane) is equal to

$$\varphi(x) = \int_0^x \sqrt{1 + (\frac{dy}{dx})^2} dx = \int_0^x \frac{1}{\sqrt{1-x^2}} dx = \arcsin x, \ \ |x| < 1,$$

since it is possible to compute the derivative of the arc sine (the function inverse to the sine on the interval $(-\pi/2, \pi/2)$ as the derivative of the inverse function

$$\frac{d(\arcsin y)}{dy} = \frac{1}{\cos(\arcsin y)} = \frac{1}{\sqrt{1 - \sin^2(\arcsin y)}} = \frac{1}{\sqrt{1-y^2}}. \tag{3.11}$$

Then the first derivative

$$\frac{d\varphi}{dx} = \frac{1}{\sqrt{1-x^2}}, \quad |x| < 1,$$

and hence the rest derivatives look simple and you can write the series for $\varphi(x)$, that is actually series for solutions of the equation $\cos\varphi = x$ or the series for $\arccos\varphi$.

Note the invariance of the lengths on the circle relative to the rotation group of the circle.

The number π. There are various definitions:

1) The length of the circle of unit radius

$$x^2 + y^2 = 1 \Longleftrightarrow y = \sqrt{1-x^2}$$

is 2π. Or you can define the number π as the integral

$$\frac{\pi}{2} = \int_0^1 \frac{1}{\sqrt{1-x^2}} dx. \tag{3.12}$$

2) Quarter of the area of the unit circle

$$\frac{\pi}{4} = S = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sqrt{1 - \frac{n^2}{N^2}} = \lim_{N\to\infty} \frac{1}{N^2} \sum_{n=1}^{N} \sqrt{N^2 - n^2}$$

from where we get the Gregory series

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots$$

3) One more definition, see (4.10), known to any probabilist,

$$\sqrt{2\pi} = \int_{-\infty}^{\infty} \exp\Big\{-\frac{x^2}{2}\Big\}\, dx.$$

Periodicity. We prove the periodicity of trigonometric functions. Note that from their series it follows

$$\sin 0 = 0, \quad \cos 0 = 1$$

and from (3.11) and (3.12) it follows that

$$\sin\frac{\pi}{2} = 1, \ \cos\frac{\pi}{2} = 0, \ \dots, \ \cos 2\pi = 1, \ \sin 2\pi = 0.$$

Then

$$e^{2\pi i} = \cos 2\pi + i \sin 2\pi = 1,$$

and we obtain periodicity of the exponent

$$e^{2\pi i + x} = e^{2\pi i} e^x = e^x$$

and also that the function $f(x) = e^{x+i\pi}, x \in R$ takes all negative values. Similarly, considering the straight lines on the plane, passing through zero, we see that e^t takes all values except zero.

Logarithm and power. The function $\ln x$ is defined first for $x > 0$ as the function inverse to e^x. Then. from (3.8) we have

$$\ln(xy) = \ln x + \ln y$$

and

$$x = e^{\ln x} \Longrightarrow 1 = (e^{\ln x})' = e^{\ln x}(\ln x)' \Longrightarrow (\ln x)' = \frac{1}{x}.$$

Now it is easy to find all derivatives and the series for the logarithm at the point 1

$$\ln(! + x) = x + \sum_{k=2}^{\infty} \frac{x^k}{k}(-1)^{k-1}.$$

Because of periodicity of the exponent, at any point $z \neq 0$ logarithm is a multivalued function. In particular, on the circle $|z| = 1$ it takes imaginary values $\ln e^{i\varphi} = i(\varphi + 2\pi k)$ with integer k.

For any complex $x \neq 0$, the power function is defined as $x^a = e^{a \ln x}$. From this it is easy to derive all its properties, and in particular,

$$(x^a)' = (e^{a\ lnx})' = x^a \frac{a}{x} = ax^{a-1}.$$

3.2.2. Analytic functions

Complex integral along a curve in the complex plane. Let now be given a connected open set $\Lambda \subset C$ in the complex plane, a (complex) differentiable function $f(z)$ on Λ and the curve $\Gamma : z(t), t \in [0, T)$, in Λ, where $z(t)$ is differentiable in t. Choose $N + 1$ points $z_k = z(k\frac{T}{N}), k = 0, 1, \ldots, N$, on this curve, so that $z_0 = z(0), z_N = z(T) = z$ and

$$F_N(z) = \sum_{k=1}^{N} f(z_k)(z_k - z_{k-1}). \tag{3.13}$$

There exists the limit

$$F(z) = \lim_{N \to \infty} F_N(z) \tag{3.14}$$

as in fact we integrate separately the real and imaginary parts of $f(z)$ on the interval $[0,T]$.

However, we can even forget about parametrization and define this integral as the limit (3.14) under the condition that

$$\Delta(N) = \max_{k=1,\dots,N} |z_k - z_{k-1}| \to 0,$$

as $N \to \infty$. It helps to intuitively understand the integral as the sum of differences and how to calculate the integral. Simplest example is the constant function, for example function $f(z) \equiv 1$. Then in the sum (3.13) occurs great cancellation and the answer is $z_N - z_0$. In particular, this integral is zero for any closed contour Γ. In the general case we substitute

$$f(z_k)(z_k - z_{k-1}) = F(z_k) - F(z_{k-1}) + o(|z_k - z_{k-1}|),$$

where $F(z)$ is the anti-derivative of $f(z)$ "in a complex sense", and get the same cancellations, except the sum $\sum_k o(|z_k - z_{k-1}|)$, but it tends to 0 as $N \to \infty$. It is proved exactly as in (3.4), as a derivative "in the complex sense" exists and is the same in any direction (including the direction of the curve Γ at the point z). And so

$$F(z) = \int_\Gamma f(z)dz = \int_0^T f(z(t))z'(t)dt \Longrightarrow \frac{dF(z)}{dz} = f(z).$$

Let the anti derivative is known and single-valued on Λ. For example, the anti derivative of $z^n, n \neq -1$, is the function $\frac{1}{n+1}z^{n+1}$. Then, for any curve (contour) Γ with a start and end points z_0 and z, respectively, we have

$$\int_\Gamma z^n dz = \frac{1}{n+1}z^{n+1} - \frac{1}{n+1}z_0^{n+1},\ n \neq -1.$$

If the contour is closed, i.e. $z = z_0$, then the integral is equal to zero. Instead of a polynomial we could consider any function $f(z)$, single valued and differentiable at each point of some simply connected area, and then for any closed loop Γ

$$\int_\Gamma f(z)dz = 0,$$

where by a **simply connected** area Λ we mean that Λ itself and its complement are connected.

At the same time, if $n = -1$, and Γ is any circle with center at the point 0, then

$$\int_\Gamma \frac{1}{z}dz = \ln(ze^{2\pi i}) - \ln z = 2\pi i,$$

and anti derivative for $1/z$ will be $\ln z$. We agree, that further such integrals are taken counterclockwise.

Let $f(z)$ be analytic at the point zero. Denote $\Gamma(r) = \{z : |z| = r\}$ the circle of radius r. Then for sufficiently small r

$$f(0) = \frac{1}{2\pi i} \int_{\Gamma(r)} \frac{f(z)}{z} dz, \tag{3.15}$$

which follows from the previous formulas and term integration of the series. The last formula is called the **Cauchy formula**.

Another example. Let, in a small neighborhood of a point z_0, the function is represented in the form (with $m \leq -1$)

$$f(z) = \sum_{k=m}^{\infty} a_k (z - z_0)^k. \tag{3.16}$$

Then a_{-1} is called the **residue** at the point z_0, and if $\Gamma(r)$ is the circle of sufficiently small radius r around the point z_0, then

$$\frac{1}{2\pi i} \int_{\Gamma(r)} f(z) dz = a_{-1}.$$

Analyticity and differentiability.

Proposition 1. *Assume that the function f is complex differentiable in a neighborhood of zero of a certain radius $0 < R < \infty$. Then*

1) Cauchy's formula at the point $z = 0$ holds for any $r < R$,

2) the function f has all derivatives at the point 0, and its Taylor series converges absolutely for all $z : |z| < R$.

For the proof we need the following frequently used construction. Fix two numbers $0 < r_1 < r_2 < R$. Define two closed contours on C: 1) $\Gamma_+ = \Gamma_+(r_1, r_2)$ that starts at the point $(r_1, 0)$, goes along the real axis to the point $(r_2, 0)$, then along the circumference of radius r_2 counterclockwise to the point $(-r_2, 0)$, then along the real axis to the point $(-r_1, 0)$, and finally along the circle of radius r_1 clockwise it comes back to the point $(r_1, 0)$. 2) A symmetric (with respect to the real axis) contour $\Gamma_- = \{(r_2, 0) \to (r_1, 0) \to (-r_1, 0) \to (-r_2, 0) \to (r_2, 0)\}$. We want to prove that

$$\int_{\Gamma(r_1)} = \int_{\Gamma(r_2)}.$$

In fact, the integrals over the contours Γ_+ and Γ_- are equal to zero (no singular points inside), and therefore $0 = \int_{\Gamma_+} + \int_{\Gamma_-} = -\int_{\Gamma(r_1)} + \int_{\Gamma(r_2)}$, as in the left sum the integrals over both intervals of the real axis cancel each other.

To prove 1) note that when $|z| = r_1 \to 0$ one can write

$$f(z) = f(0) + f'(0)z + o(r_1).$$

Then the Cauchy integral from the second term of the sum will give 0, and the third will tend to zero. So in the limit there will be only $f(0)$. Since r_1 is arbitrary, the integral over the circumference of radius r_2 is equal to this limit.

2) Assume now that z_0 is an arbitrary point, f is differentiable in this point and Γ is a circle of sufficiently small radius $r < r$ around z_0. Then from 1) and (3.15) it follows

$$f(z_0) = \frac{1}{2\pi i} \oint_\Gamma \frac{f(z)}{z - z_0} dz.$$

Differentiating under the integral sign, we obtain

$$\begin{aligned} f'(z_0) &= \lim_{\varepsilon \to 0} \frac{f'(z_0 + \varepsilon) - f'(z_0)}{\varepsilon} = \\ &= \frac{1}{2\pi i} \lim_{\varepsilon \to 0} \int_\Gamma \frac{1}{\varepsilon} \left(\frac{1}{z - z_0 - \varepsilon} - \frac{1}{z - z_0} \right) f(z) dz = \\ &= \frac{1}{2\pi i} \lim_{\varepsilon \to 0} \int_\Gamma \frac{1}{(z - z_0 - \varepsilon)(z - z_0)} f(z) dz = \frac{1}{2\pi i} \int_\Gamma \frac{f(z)}{(z - z_0)^2} dz \end{aligned}$$

and similarly

$$f^{(n)}(z_0) = \frac{n!}{2\pi i} \int_{\Gamma(r)} \frac{f(z)}{(z - z_0)^{n+1}} dz.$$

Thus, if for example $z_0 = 0$, we have the estimate

$$|f^{(n)}(0)| \le \frac{n!}{r^{n+1}} \max_{|z|=r} |f(z)|.$$

Now, for arbitrary z with $|z| < R$ we choose r such that $|z| < r < R$. Therefore, the Taylor series for $f(z)$ at the point 0 absolutely converges when $|z| < r$.

The **residue theorem**: suppose that in a simply connected domain Λ a function is analytic everywhere except for a finite number of singular points, where it has the form (3.16). Then the integral over any closed loop without self-intersections is equal to the sum of the residues at the singular points inside this loop.

It is proved in the same way.

Note that such functions are called **meromorphic** in Λ, and the functions (like e^z) analytic at each point of the complex plane are called **entire** functions.

Main theorem of algebra Consider the polynomial $P(z) = z^n + p(z) = z^n + a_{n-1}z^{n-1} + \ldots + a_0$, as well as the function $f(z) = \frac{P'(z)}{P(z)}$. For example, if $P = z^n, n = 0, 1, \ldots$, then for any $r > 0$

$$\frac{1}{2\pi i} \int_{\Gamma(r)} f(z) dz = n \tag{3.17}$$

that is equal to the multiplicity of the root $z = 0$. Let z_0 be one of the roots of P. Then the same integral over a circle of small radius around z_0 will give the multiplicity of this root.

Due to the dominance of the term z^n for sufficiently large R, all roots will lie inside $\Gamma(R)$ for some R. But

$$f(z) = \frac{P'(z)}{P(z)} = \frac{nz^{n-1}+p'(z)}{z^n+p(z)} = \frac{nz^{-1}+p'(z)/z^n}{1+p(z)/z^n} = nz^{-1}\frac{1+p'(z)/(nz^{n-1})}{1+p(z)/z^n} =$$

$$= nz^{-1}\frac{1+\sum_{k=1}^{\infty} a_k z^{-k}}{1+\sum_{k=1}^{\infty} b_k z^{-k}} = \frac{n}{z} + \sum_{k=2}^{\infty} c_k z^{-k}.$$

The last series is convergent for sufficiently large R. Therefore, the integral of it is equal to zero. And so the formula (3.17) is correct.

But in addition, the integral of $f(z)$ over $\Gamma(R)$ is equal to the sum of the integrals of $f(z)$ over small circles around all zeros of P. This can be proved, as above, by constructing small circles around each zero and the contours connecting them together and with $\Gamma(R)$. So the sum of the multiplicities of all the roots of $P(z)$ is n.

3.3. Asymptotics of sums, products and integrals

Each of these methods is based on simple and clear ideas: 1). approximation of sums by integrals, 2) cancellations in frequent changes of sign, 3) neighborhood of exponentially large maximum allows to neglect everything else, 4) choice of the path of integration.

Sums. First we prove that

$$\sum_{k=1}^{n} \frac{1}{k} = \ln n + C + o(1), \tag{3.18}$$

where $C > 0$ is some constant. Growing members of the asymptotics (unlike constants) in such problems one can find easier, since they do not depend on a finite number of series terms. Consider the piece wise constant function $f(x)$ equal to $\frac{1}{k}$ on the interval $[k, k+1)$. Then our sum is equal to the area under it in the interval $[1, n+1)$. Since the difference $f(x) - \frac{1}{x} \geq 0$, then

$$\sum_{k=1}^{n} \frac{1}{n} > \int_1^{n+1} \frac{dx}{x} = \ln(n+1) = \ln n + \ln\Big(1+\frac{1}{n}\Big) = \ln n + O\Big(\frac{1}{n}\Big),$$

and the area between these two functions is equal to

$$\sum_{k=1}^{n}\left(\frac{1}{k} - \int_k^{k+1} \frac{dx}{x}\right) = C > 0,$$

which gives a convergent series with positive terms.

Products – Stirling formula. Similarly, in the asymptotics of the product, to find the multiplicative constant is significantly more difficult than increasing multipliers. An example is the Stirling formula

$$n! \sim \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n. \tag{3.19}$$

Everything comes down to finding the asymptotic expansion of the sum

$$\ln n! = \ln 1 + \ldots + \ln n = n\ln n - n + \frac{1}{2}\ln n + \frac{1}{2}\ln(2\pi) + C + o(1), \tag{3.20}$$

where everything is very simple, again with the exception of finding the constant. It seems to exist many links between the two famous numbers: π and e. For example, one given by the formula $e^{2\pi i} = 1$, and also Stirling formula (3.19) itself. It is difficult to join all these links to one simple structure.

Consider sum (3.20) as the area S_n under the graph of the piece wise constant function $f(x)$ on $(1, n]$, equal to $\ln k$ on half-interval $(k-1, k]$. The area under the function $\ln x$ on this interval is

$$\int_1^n \ln x dx = -\int_1^n x d(\ln x) + (x\ln x)|_1^n = -n + 1 + n\ln$$

gives the lower bound and the basic factor $(\frac{n}{e})^n$. The factor $\sqrt{n}$ is slightly more complicated because of the asymptotics of the difference of the two areas

$$S_n - \int_1^n \ln x dx = \sum_{k=2}^n \Big[\ln k - \int_{k-1}^k \ln x dx\Big] =$$

$$= \sum_{k=2}^n [\ln k - (k\ln k - (k-1)\ln(k-1) - 1)] = \frac{1}{2}\ln n + C + o(1) \tag{3.21}$$

which is easily calculated after decomposition (for large k)

$$\ln(k-1) = \ln k + \ln\Big(1 - \frac{1}{k}\Big) = \ln k - \frac{1}{k} - \frac{1}{2k^2} + O\Big(\frac{1}{k^3}\Big) \ln(k-1) =$$

$$= \ln k + \ln\Big(1 - \frac{1}{k}\Big) = \ln k - \frac{1}{k} - \frac{1}{2k^2} + O\Big(\frac{1}{k^3}\Big),$$

and use the last equality of (3.21) asymptotics of sums (3.18).

Integration by parts. A typical example is the following estimate of the integral for a function f differentiable on $[a, b]$: there exists a constant $A > 0$ such that

$$\Big|\int_a^b f(x)\cos(Nx)dx\Big| \le \frac{A}{N}. \tag{3.22}$$

Intuitively, this is due to fast cosine fluctuations on short intervals, where f is almost constant, which gives plus-minus cancellations. Formally, this looks like this:

$$f(x)\cos(Nx) = \frac{1}{N}f(x)(\sin(Nx))' = \frac{1}{N}[(f(x)\sin(Nx))' - f'(x)\sin(Nx)] \Longrightarrow$$

$$\Longrightarrow \int_a^b f(x)\cos Nxdx = \frac{1}{N}\left[f(b)\sin(Nb) - f(a)\sin(Na) - \int_a^b f'(x)\sin(Nx)dx\right].$$

If $f(b)\sin(Nb) = f(a)\sin(Na)$, then assuming twice differentiability of f, we can apply a similar procedure to the last integral, obtaining an estimate of the order N^{-2}, etc.

Laplace Method. Consider the integral

$$\int_a^b f(x)e^{NS(x)}dx,$$

where S has a unique non-degenerate maximum $S(x_0) > 0, S'(x_0) = 0, S''(x_0) < 0$, at the interior point $x_0 \in [a, b)$. First, let $f = 1$, put out $e^{NS(x_0)}$ of the integral

$$\int_a^b e^{NS(x)}dx = e^{NS(x_0)}\int_a^b e^{N(S(x)-S(x_0))}dx.$$

We show that in the integral on the right we can asymptotically neglect the integral outside the interval $I_0 = (x_0 - cN^{-1/2}, x_0 + cN^{-1/2})$ for an arbitrary constant c. Indeed, in the neighborhood of x_0, we have

$$S(x) - S(x_0) = \frac{1}{2}S''(x_0)(x - x_0)^2 + O((x - x_0)^3).$$

For sufficiently small $\delta > 0$ we introduce a smaller interval $I_\delta = (x_0 - cN^{-1/2-\delta}, x_0 + cN^{-1/2-\delta})$,

$$\int_{I_0} > \int_{I_\delta} \geq c_1 N^{-1/2-\delta}\exp\{N^\delta\}$$

At the same time, the integrand outside I_0 does not exceed a constant. Then

$$\begin{aligned}
&\int_{I_0}\exp\{N(S(x) - S(x_0))\}dx \\
&\quad\sim \int_{I_0}\exp\Big\{N\frac{1}{2}S''(x_0)(x - x_0)^2\Big\}dx \\
&\quad\sim \int_{-\infty}^{\infty}\exp\Big\{N\frac{1}{2}S''(x_0)(x - x_0)^2\Big\}dx = \sqrt{\frac{2\pi}{NS''(x_0)}}\,.
\end{aligned}$$

The last integral is called Gaussian and is calculated in (4.10).

In a curvilinear integral (real integral of vector field or complex of analytic function), it is often possible to deform a path so that it passes through a maximum point. Then they talk about the **saddle point method**.

Method of stationary phase. Generalization of integrals of type (3.22) are the integrals, like

$$\int f(x)\exp(iN\mathrm{S(x)})dx.$$

However there is a nuance – the ends of the interval of integration also can give essential contribution. Therefore, assume that $f(x)$ is smooth and finite (i.e. is zero outside a finite interval) function. If $S(x)$ has no stationary points, i.e. the points where $S'(x)=0$, then it is estimated by integration by parts as above. The existence of such points prevents cancellation due to sign changes.

Then in the neighborhood of stationary points the integral is treated like in Laplace method, and outside of small neighborhood of stationary points – like in integration by parts. Assume that at some point a we have $S'(a)=0$ and, for example, $S"(a)>0, f(a)\neq 0$. Then, like in Laplace method, it is necessary to choose a sufficiently small $\varepsilon=\varepsilon(N)$ so that

$$\int_{-\infty}^{\infty} f(x)\exp(iN\mathrm{S(x)})dx\sim$$

$$\sim\int_{a-\varepsilon}^{a+\varepsilon} f(x)\exp(iN\mathrm{S(x)})dx\sim$$

$$\sim f(a)\exp(iNS(a))\int_{a-\varepsilon}^{a+\varepsilon}\exp(iN(S(x)-S(a))dx\sim$$

$$\sim f(a)\exp(iNS(a))\int_{a-\varepsilon}^{a+\varepsilon}\exp\Big(\frac{iN}{2}S''(a)(x-a)^2\Big)dx\sim$$

$$\sim f(a)\exp(iNS(a))\int_{-\infty}^{\infty}\exp\Big(\frac{iN}{2}S''(a)(x-a)^2\Big)dx\sim$$

$$\sim C\frac{\exp(iNS(a))}{\sqrt{N}},$$

$$C=f(a)\exp\Big(\frac{i\pi}{4}\Big)\sqrt{\frac{2\pi}{S''(a)}}\,.$$

Thus, the connection with Laplace method is evident and also that in asymptotic expansions one always has to do a lot of calculations, although the methods (algorithms) can be predictable.

4. Linear algebra

4.1. Linear equations

4.1.1. Linear operators and matrices

Linear space L over the field C (or R) is the set of elements (called vectors), which is the commutative (addition) group with zero element 0, so that for any two elements x, y is defined their sum $x + y$. Moreover, for each vector x multiplication λx is defined for any complex (real) number λ that satisfies the following axioms:

$$0x = 0, \quad \lambda(x + y) = \lambda x + \lambda y, \quad (\lambda_1\lambda_2)x = \lambda_1(\lambda_2 x),$$

$$(\lambda_1 + \lambda_2)x = \lambda_1 x + \lambda_2 x.$$

Further on, the main field we are considering is C, but many definitions can be literally transferred to other fields.

N vectors $x_1, \ldots, x_N$ are called **linearly independent** if there are no such (not all zero) numbers λ_k that

$$\lambda_1 x_1 + \ldots + \lambda_N x_N = 0.$$

Maximal number of linearly independent vectors is called the **dimension** of L and is denoted by $\dim L$. Then (if the dimension is equal to N) any N such vectors can be selected as the **basis**. This means that any vector y can be represented as a linear combination of elements of this basis. In other words, there exist numbers $a \neq 0, a_1, \ldots, a_N$, such that

$$ay + a_1 x_1 + \ldots + a_N x_N = 0,$$

since the vectors $x_1, \ldots, x_N, y$ are linearly dependent. And of course, we can take $a = 1$.

Two linear spaces L_1, L_2 of the same dimension are (linearly) isomorphic, i.e. there exists a bijective mapping $\varphi : L_1 \to L_2$, consistent with addition and multiplication on numbers, i.e.

$$\varphi(\lambda_1 x_1 + \lambda_2 x_2) = \lambda_1\varphi(x_1) + \lambda_2\varphi(x_2).$$

For the proof it is sufficient to continue linearly some one-to-one correspondence between elements of any their bases.

Easy implementation (sometimes we will talk about the **standard representation**) of linear space of dimension N is the set of all vectors (sequences) $a = (a_1, \ldots, a_N)$, of length N, where $a_k \in C$ are called the coordinates (or components) of the vector a. In this representation for any two vectors a and

$b = (b_1, \dots, b_N)$ operations of addition and multiplication by number are as follows:

$$a + b = (a_1 + b_1, \dots, a_N + b_N), \quad \lambda a = (\lambda a_1, \dots, \lambda a_N).$$

Convenient basis here is $e_k, k = 1, \dots, N$, where all the coordinates of the vector e_k are equal to 0, except k-th coordinate, equal to 1.

Then the statement about the linear independence of the vectors $x_k = (x_{k1}, \dots, x_{kN}), k = 1, \dots, N$, can be reformulated as follows: one vector equation, or equivalent system of N linear equations

$$\sum_k x_k \lambda_k = y \Longleftrightarrow \sum_k x_{kl} \lambda_k = y_l, \;\; l = 1, 2, \dots, N,$$

with unknowns λ_k, has a unique solution $\{\lambda_k\}$ for any vector $y = (y_1, \dots, y_N)$.

Direct sum of N linear spaces $L_k, k = 1, \dots, N$, is defined as the linear space such that: 1) as the set it is the Cartesian product of the sets L_k, with elements $(x_1, \dots, x_N)$, where $x_k \in L_k$, 2) operations of addition and multiplication by a number are defined as

$$(x_1, \dots, x_N) + (y_1, \dots, y_N) = (x_1 + y_1, \dots, x_N + y_N),$$
$$\lambda(x_1, \dots, x_N) = (\lambda x_1, \dots, \lambda x_N).$$

In particular, any linear space of dimension N is a direct sum of N one-dimensional spaces.

A **linear functional** on L is a function $f : L \to C$, having the properties of linearity

$$f(x + y) = f(x) + f(y), \;\; f(\lambda x) = \lambda f(x). \tag{4.1}$$

It is obvious that in the standard representation any linear functional has the form

$$f(x) = x_1 f(e_1) + \dots + x_N f(e_N) = f_1 x_1 + \dots + f_N x_N$$

for some numbers $f_k = f(e_k)$.

Linear operator in C^N is a function $A : L \to L$ having the same properties of linearity (4.1). Linear functionals, linear operators (matrices, see below) in C^N also form a linear space of dimension respectively N and N^2. The linear space of operators is also an algebra with multiplication defined as $(L_2 L_1)x = L_2(L_1)x$. The **kernel** $KerA$ of the operator A, i.e. the set of vectors $x \in L$ such that $Ax = 0$, is a subspace of L. Similarly, the **image** ImA is also a linear space, its dimension is called the **rank** of the operator A.

Any linear operator can be written as (for some basis $e_1, \dots, e_N$ and some numbers $a_{kl}, k, l = 1, \dots, N$)

$$Ae_k = \sum_l a_{lk} e_l.$$

Then the vector $x = (x_1, \ldots, x_N)$ is transformed to

$$Ax = \sum_k x_k A e_k = \sum_k x_k \sum_l a_{lk} e_l = (\Sigma_{k=1}^N a_{1k} x_k, \ldots, \Sigma_{k=1}^N a_{Nk} x_k).$$

These numbers a_{ij} can be arranged in a matrix form (square matrix) $A = (a_{ij}), i, j = 1, \ldots, N$, where i enumerates the rows and j – the columns. The product of two operators $AB = D$ corresponds (in this basis) to multiplication of their matrices

$$d_{ij} = \sum_{k=1}^N a_{ik} b_{kj}.$$

Moreover, multiplication of a vector by a matrix can be understood as matrix multiplication, if the vector will be written as $(1 \times N)$-matrix, or as a $(N \times 1)$-matrix, that is

$$xA = (x_1, \ldots, x_N)A = y = (y_1, \ldots, y_N) = \Big(\sum_i x_i a_{i1}, \ldots, \sum_i x_i a_{iN}\Big),$$

$$Ax = A \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ x_N \end{pmatrix} = y = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_N \end{pmatrix} = \begin{pmatrix} \sum_j a_{1j} x_j \\ \cdot \\ \cdot \\ \sum_j a_{Nj} x_j \end{pmatrix}.$$

Note the summation over first indices of a_{ij} in xA and over second indices in Ax.

If in each of L_k of the direct sum $L = L_1 \times \ldots \times L_N$ some linear operator $A_k : L_k \to L_k$ is defined, then the direct sum of these operators is the operator $A : L \to L$ acting as

$$Ax = A(x_1, \ldots, x_N) = y = (y_1, \ldots, y_N) = (A_1 x_1, \ldots, A_N x_N).$$

Problem 9. Similarly, define linear operators from C^N to C^M and the corresponding matrices. For $M = 1$ this will be obviously linear functionals.

Multilinear forms. Let linear spaces $L_1, \ldots, L_n$ of dimensions N_k be given. The vectors of these spaces will be denoted by $x_k = (a_{k1}, \ldots, a_{kN_k})$.

A function $f : L_1 \times \ldots \times L_n \to C$ is called a **multilinear form** if for any index k, e.g. $k = 1$, the following holds:

1) addition of vector $y \in L_1$ to x_1 gives

$$f(x_1 + y, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n) + f(y, x_2, \ldots, x_n),$$

2) multiplication of x_1 by any number λ multiplies the result by this number

$$f(\lambda x_1, x_2, \ldots, x_n) = \lambda f(x_1, x_2, \ldots, x_n).$$

Further, we assume that all $L_k = L$ of dimension n, and will write every vector x_i of L in the standard basis $e_1, \dots, e_n$:

$$x_i = \sum_{j=1}^{n} a_{ij} e_j.$$

Then the form can be considered as a function of n^2 numbers a_{ij}.

The form is called **skew symmetric** if any permutation of two vectors changes sign. For example,

$$f(x_2, x_1, x_3, \dots, x_n) = -f(x_1, x_2, x_3, \dots, x_n).$$

We will show now that with normalization condition

$$f(e_1, e_2, \dots, e_n) = 1 \tag{4.2}$$

a skew symmetric form, as a function of a_{ij}, is unique and has the following form

$$f(\{a_{ij}\}) = \sum_{\sigma} a_{1j_1} a_{2j_2} \dots a_{nj_n} (-1)^{|\sigma|} \tag{4.3}$$

where the sum is taken over all permutations

$$\sigma = \begin{pmatrix} 1 & 2 & . & . & n \\ i_1 & i_2 & . & . & i_n \end{pmatrix}$$

and $|\sigma|$ is the parity of substitution σ. This number (4.3) is called the **determinant** $\det A$ of the matrix $A = (a_{ij})$.

For the proof of (4.3), it is sufficient first to open all brackets, using property 1), and put out all a_{ij} of f, using 2) :

$$f(x_1, \dots, x_n) = f\left(\sum_{j=1}^{n} a_{1j} e_j, \dots, \sum_{j=1}^{n} a_{nj} e_j\right) =$$

$$= \sum_{j_1, \dots, j_n = 1}^{n} a_{1j_1} \dots a_{nj_n} f(e_{j_1}, \dots, e_{j_n}).$$

After this it is necessary to use the normalization condition and the fact that $f(e_{j_1}, \dots, e_{j_n}) = (-1)^{|\sigma|}$, if all $e_{j_1}, \dots, e_{j_n}$ are different, and zero otherwise. Note that (4.3) can be rewritten in the form

$$\sum_{\sigma} a_{j_1 1} a_{j_2 2} \dots a_{j_n n} (-1)^{|\sigma|},$$

which implies that the determinants of matrices $A = (a_{ij})$ and the **transpose** of the matrix $A' = (a'_{ij} = a_{ji})$ are equal.

Problem 10. Prove by direct multiplication that $\det(AB) = \det(A) \det(B)$.

4.1.2. Geometry of determinant

The scalar product (a, b) of two vectors $a = (a_1, \ldots, a_N), b = (b_1, \ldots, b_N) \in R^N$ and the length $|a|$ of the vector a is defined as

$$(a, b) = a_1 b_1 + \ldots + a_N b_N, \quad |a| = \sqrt{(a, a)} \geq 0.$$

Cosine of the angle φ between the vectors a and b is defined (like on the plane) as

$$\cos \varphi = \frac{(a, b)}{|a||b|}.$$

It is easy to prove that this definition of the cosine does not contradict to the earlier definition in the section on power series.

We define n-dimensional parallelepipeds in R^n, generated by n vectors $a_1, \ldots, a_n$ as

$$P_n = P(a_1, \ldots, a_n) = \{s_1 a_1 + \ldots + s_n a_n : 0 \leq s_1, \ldots, s_n \leq 1\}.$$

We want also to define their (oriented) volumes $v(P(a_1, \ldots, a_n))$ by induction in n . Let M_{n-1} be the subspace generated by the vectors $a_1, \ldots, a_{n-1}$. Then the vector a_n can be written in the form $a_n = h + a$, where h is perpendicular to M_{n-1}, i.e. to each of the vectors $a_1, \ldots, a_{n-1}$, and a is a linear combination of the vectors $a_1, \ldots, a_{n-1}$.

This means that P_n is the union of straight line segments $\cup_{x \in P_{n-1}} [x, x + a_n]$. Then $|v(P_n)| = |v(P_{n-1})||h|$, as in the two-dimensional case. We have to show now that if $|v(P_{n-1})| = |\det A_{n-1}|$, then $|v(P_n)| = |\det A_n|$.

To do this, define the matrix $A = (a_{ij})$ with the vector-rows $a_i = (a_{i1}, \ldots, a_{iN})$, Then the matrix AA' can be written in the form of the **Gram matrix** $G = G(a_1, \ldots, a_N)$ of the system of vectors $a_1, \ldots, a_N$ with the elements $g_{ij} = (a_i, a_j)$. Then

$$\det G = \det(AA') = (\det A)^2.$$

Since $(a_k, h) = 0$ for $k = 1, \ldots, n-1$, we have

$$\det G(a_1, \ldots, a_{n-1}, a_n) = \det G(a_1, \ldots, a_{n-1}, h) = \det G(a_1, \ldots, a_{n-1})(h, h).$$

Remark 7. Throughout the previous we missed an important nuance: the volume could have minus sign, depending on the orientation – the sign of the determinant determines the orientation change. A simple example in two-dimensional case: the difference between $P_2(e_1, e_2)$ and $P_2(e_2, e_1)$ explains what is orientation.

We can now give a simple explanation of why

$$\det(A_1 A_2) = \det A_1 \det A_2.$$

Indeed, the determinant can be seen not only as a volume but also as the number showing how many times increases the volume of the unit cube generated by the basis vectors $e_1, e_2, \dots, e_n$, (and also of any volume) under the action of the matrix $A : R^n \to R^n$. Therefore, the determinant of the product of two matrices equals the product of their determinants.

4.1.3. Combinatorics of matrices

Solution of linear equations. To find vector $x = (x_1, \dots, x_n)$ from the equations $Ax = y = (y_q, \dots, y_n)$ with $(n \times n)$-matrix $A = (a_{ij})$ is a well-known natural trick. For example, suppose that $a_{11} \neq 0$, then in the system of linear equations

$$\sum_{j=1}^{n} a_{ij} x_x = y_i \tag{4.4}$$

we can eliminate the unknown x_1 (if it is present) from the equations with $i = 2, \dots, n$, by adding equation 1 to the equation i, multiplied by the corresponding number. If $a_{11} = 0$, then find the line i with $a_{i1} \neq 0$ and exchange it with the first row, and then similarly eliminate the variable x_1. After this we will have the equation 1 and $n-1$ equations with a matrix A_{n-1}, where there will not be unknown x_1. Suppose for example that the matrix A is invertible, then the matrix A_{n-1} is also invertible, since its rows will remain linearly independent, Next step – we eliminate by induction $x_2, x_{3,\dots}$. On the last step n we have one row with one unknown x_n, which can be easily found. Then the inverse procedure will find all the rest.

This procedure shows that the following 3 elementary operations leave the equation with the same set of solutions: 1) rearrange the equations, 2) multiply both parts of the same equation by a nonzero number, 3) adding one equation to another. Note that each of these operations does not change in the matrix the number of linearly independent rows, and the number of linearly independent columns.

Instead of doing these operations "by hands", one can use multiplication (like $A \to E(i,j)A$) by the corresponding matrix (called **elementary matrices**): 1) matrix $E(i,j) = E^{-1}(i,j)$ derived from unit matrix E by interchanging rows i and j, 2) $E(bi) = E^{-1}(b^{-1}i)$ obtained from E by multiplying its row i by the number b, 3) $E(j \to i) = E^{-1}(-j \to i)$, received from E by adding row j to row i. Of course, at the same time exactly the same matrix should be applied to the right side of the equations. This approach allows to better understand many well-known facts and formulas:

1) Any invertible matrix is a product of elementary ones – can be obtained from unit matrix by elementary operations.

2) For a square matrix A, the maximum number of its linearly independent rows and columns are equal. This number $r(A)$ coincides with the rank of the

matrix A. In fact, using elementary operations, one can make zero $n - r(A)$ rows and columns.

3) Also this makes clear famous Kramer formula for the solution

$$x_k = \frac{\det A_k}{\det A}, \tag{4.5}$$

of the linear system (4.4) with $\det A \neq 0$, where A_k is A, where the k-th column is replaced by right-side column y in (4.4). For example, assume that we transform A to triangular form (see below) with $x_1, \ldots, x_{n-1}, x_n$ on the diagonal, and simultaneously A_n (with exactly the same operations) obtaining triangular matrix with diagonal $x_1, \ldots, x_{n-1}, y_n$, then $\det A = x_1 \ldots x_{n-1} x_n$ and $\det A_n = x_1 \ldots x_{n-1} y_n$. Then $x_n = y_n$.

4) Formula for the inverse matrix (with nonzero determinant)

$$A^{-1} = \frac{1}{\det A}((-1)^{i+j} M_{ij}), \tag{4.6}$$

where (M_{ij}) is called minor of matrix A, that is the matrix A without i-th row and j-th column. In fact, calculation of the inverse matrix can be reduced to the subsequent solution of systems of equations. Thus, it is sufficient to prove this formula for simple (see below) matrices.

4.2. Spectrum and similarity

Spectrum of linear operator A is the set of λ such that the operator $A - \lambda E$ is not one-to-one (does not have inverse), that is where there exists at least one vector ψ, called **eigenvector**, and $\lambda \in C$, called **eigenvalue**, such that

$$A\psi = \lambda\psi. \tag{4.7}$$

The number of points of the spectrum does not exceed N and is not less than 1, that is there exists at least one λ and ψ, satisfying (4.7). To prove this define the **resolvent** of operator A in C^N as the following rational matrix function of λ

$$R(\lambda) = (A - \lambda E)^{-1}, \lambda \in C.$$

Resolvent exists for all $\lambda \in C$, except those when the operator $A - \lambda E$ is not invertible, that is except finite number of zeroes of the **characteristic polynomial** $P_A(\lambda) = \det(A - \lambda E)$, that is where

$$P(\lambda) = 0. \tag{4.8}$$

There can be several eigenvectors with eigenvalue λ. All of them generate the subspace called the **eigensubspace** $L_\lambda = L_\lambda(A)$. Let d_λ be their dimensions. Any number of eigenvectors ψ_k with different eigenvalues λ_k are linearly

independent. For two eigenvectors this is evident. For any number n this can be shown by induction in n. Thus, all L_λ are linearly independent but the problem is that it can be that $\sum_\lambda d_\lambda < N$, that is they do not generate all the space. Now we want to understand how A acts on other vectors.

Simple types of matrices. The matrix is called a **block** matrix, if the set of indices $\{1,\dots,n\}$ is divided into m consecutive blocks

$$\{1,\dots,k_1\},\{k_1+1,\dots,k_2\},\dots,\{k_{m-1},\dots,k_m=n\}$$

so that $a_{ij}=0$ if i,j belong to different blocks. Block matrix determines the partitioning of the basis into correspondent groups that, in turn, determines the partitioning of C^n is a direct sum of linear spaces of corresponding dimensions, where each block of the matrix defines a linear operator. And the whole matrix is the direct sum of these operators. Also it is easy to see that the determinant of a block matrix is equal to the product of determinants of blocks. A special case of a block matrix is a **diagonal** matrix where only the diagonal elements a_{ii} can be nonzero.

A matrix is called upper (lower) **triangular**, if $a_{ij}=0$ when $i>j$ ($i<j$). It is obvious that the determinant of a triangular matrix equals the product of diagonal elements. Indeed, for example for the upper triangular matrix in any product of (4.3) necessarily $j_n=n$, and then by induction. The matrix is called **block-triangular** if each of its block is a triangular matrix.

A matrix A of order n is called **Jordan block** (of order n) if for all $i=1,\dots,n-1$ we have $a_{i,i+1}=1, a_{ii}=\lambda$ for all i and some λ, and all the others are null. We shall denote such Jordan block as $J(N,\lambda)$. A block matrix (direct sum of matrices) is called a **Jordan** (having Jordan form) if each of its blocks is a Jordan block.

Similarity of matrices. Two matrices A and B are called **similar** (sometimes called also conjugate), if there is an invertible matrix C such that $B=CAC^{-1}$. We shall show that the matrices are similar, iff they represent the same linear operator in different bases. Suppose that there are two bases consisting of vectors-columns $e_1,\dots,e_n$ and $g_1,\dots,g_n$. Let the two bases be related by a linear operation C as $Ce_i=g_i$. and moreover in these bases the operators A,C are specified by the matrices

$$g_i=Ce_i=\sum_j c_{ij}e_j,\quad Ae_i=\sum_j a_{ij}e_j,\quad Ag_i=\sum_j b_{ij}g_j,$$

which we denote by the same letters C,A,B respectively. Then

$$C^{-1}Ag_i=\sum_j b_{ij}C^{-1}g_j=\sum_j b_{ij}e_j \Longrightarrow$$

$$\Longrightarrow C^{-1}ACC^{-1}g_j = C^{-1}ACe_j = \sum_j b_{ij}e_j \Longrightarrow B = CAC^{-1}.$$

It is easy to see that the similarity relation is symmetric, reflexive and transitive. This set of properties defines an **equivalence relation**, that is, partitioning the entire set of matrices into classes of "similarity". So, matrices of different classes may not be similar to each other.

If A and B are similar, then they have the same determinants, ranks, characteristic polynomials, spectrum, and also dimensions $\dim L_\lambda(A) = \dim L_\lambda(B)$ for any λ. This is simple to prove. However, all these conditions are not sufficient that A and B were similar. See example in [5].

A natural question is to find in each equivalence class the matrix of the simplest kind. But firstly, it is necessary understand what are simplest matrix representations, for which there is no diagonal representation. For example, to show that any matrix is similar to a triangular matrix consider any matrix A, select, as we did above, the elementary matrices D_k so that the matrix $D_m \dots D_1 A$ were triangular. For example, if the matrix has N linearly independent eigenvectors then it is similar to diagonal matrix with the diagonal representation, that is with the corresponding eigenvalues on the diagonal.

Theorem 2 (Basic statement). *Any matrix is similar to some Jordan matrix.*

Consider first an operator A in $L = C^N$ with the only eigenvector g_1, and assume that it has eigenvalue 0. We shall show that A is similar to the Jordan block $J(N, 0)$. More exactly, we shall show that A is **nilpotent** operator that is $L^n = 0$ for some n (in our case minimal such $n = N$), and there exists vector g_N such that the following sequence

$$g_N, g_{N-1} = Ag_N, \dots, g_1 = Ag_2$$

is a linear basis in L.

For example, let $N = 2$, then for any vector x, linear independent with g_1, we have $Ax = ag_1 + bx$. There could be 4 cases: 1) $a = b = 0$, then there is second eigenvector with zero eigenvalue, 2) $a = 0, b \neq 0$ then x is the second eigenvector with nonzero eigenvalue, 3) $a \neq 0, b \neq 0$, then the vector $y = \frac{a}{b}g_1 + x$ is the second eigenvector with eigenvalue b, 4) $a \neq 0, b = 0$, this is what we need. Thus $b = 0$ and $A(a^{-1}x) = g_1$.

For any N, assume that for some $n < N$ there exists sequence

$$g_n, g_{n-1} = Ag_n, \dots, g_1 = Ag_2.$$

Let L_n be the subspace having this sequence as a basis. Choose some vectors $x_1, \dots, x_{N-n}$ so that they, together with $g_1, \dots, g_n$ form the basis of L. Let L_x the subspace generated by the $x_1, \dots, x_{N-n}$. Then there are two possibilities:

1) there is no vector $x \in L_x$ such that $Ax \in L_n$, then L_n is invariant w.r.t. A and there exists at least one eigenvector inside it, that contradicts our assumption;

2) there is vector x such that $Ax = g = c_1 g_1 + \ldots + c_n g_n \in L_n$. If $c_n = 0$, then there are two vectors $x_1 = x$ and $x_2 = c_1 g_2 + \ldots + c_{n-1} g_n$ such that $x_1 - x_2$ is a second eigenvector. If $c_n \neq 0$ then $Ay = g_n$ for $y = c_n^{-1}(x - c_1 g_2 - \ldots - c_{n-1} g_n)$. And so on, we get $n = N$.

In case of several eigenvectors $\xi_1, \ldots, \xi_m$. each with eigenvalue 0 there will be m invariant subspaces, where A will be nilpotent with minimal degrees $n_1, \ldots, n_m$ such that for $k = 1, \ldots, m$ and some vectors $\xi_{k1} = \xi_k, \xi_{k2}, \ldots, \xi_{kn_k}$ (the basis of the subspace L_k)

$$A\xi_{k1} = 0, \quad A\xi_{k2} = \xi_{k1}, \ldots, A\xi_{kn_k} = \xi_{k,n_k-1}$$

and $L_1 \oplus \ldots \oplus L_m = L$.

For several eigenvectors $\xi_1, \ldots, \xi_m$ with any eigenvalues λ_k the situation is the same. The only difference is that

$$A\xi_{k1} = \lambda_k \xi_{k1}, \quad A\xi_{k2} = \xi_{k1} + \lambda_k \xi_{k2}, \ldots, A\xi_{kn_k} = \xi_{k,n_k-1} + \lambda_k \xi_{kn_k}.$$

The proof is just by reducing to the previous case. For example, let $N = 2$ and the only eigenvector g_1 has eigenvalue $\lambda \neq 0$. Consider the operator $B = A - \lambda E$ which has the only eigenvector g_1 with eigenvalue 0. Then B is similar to Jordan block $J(2, 0)$ and A – to $J(2, \lambda)$.such that $Ax = ag_1 + bx, a \neq 0$. If $b \neq 0$ then the vector $y = \frac{a}{b} g_1 + x$ is an eigenvector with eigenvalue b. Thus $b = 0$ and $A(a^{-1}x) = g_1$. If for $N = 2$ there is the only eigenvector g_1 with eigenvalue λ, we consider the operator $B = A - \lambda E$ which has the only eigenvector g_1 with eigenvalue 0. Then B is similar to Jordan block $J(2, 0)$ and A – to $J(2, \lambda)$.

4.3. Examples of local and infinite dimensional linearity

4.3.1. Local diffeomorphisms

Any smooth mapping is locally linear in some sense – the less the neighborhood the better the map of this neighborhood is approximated by linear map.

Mapping of the open region $\Lambda_1 \subset R^d$ to another open region $\Lambda_2 \subset R^n$ is called **diffeomorphism** if it is one-to-one and continuously differentiable.

First, let $d = 1$ and the function $f(x)$ is defined in a neighborhood of zero. When one can affirm that it maps some neighborhood of zero one-to-one to some neighborhood of the point $f(0)$? For this it is sufficient that $f(x)$ were continuously differentiable on some interval containing 0, and $f'(0) \neq 0$. Then on some interval $[x_1, x_2]$ that contains 0, its derivative has one sign and the answer is obvious from monotonicity

$$f(x_2) - f(x_1) = \int_{x_1}^{x_2} f(x)dx \neq 0. \tag{4.9}$$

Similarly for arbitrary d. Let the mapping of an open sets-polygons $O_1 \subset R^d$ to R^d is given by the system of real functions $f_i(x_1, \ldots, x_d), i = 1, \ldots, d$, which have all continuous partial derivatives $\frac{\partial f_i}{\partial x_j}, i, j = 1, \ldots, d$. Then, if the Jacobi matrix $(a_{ij} = \frac{\partial f_i}{\partial x_j})$, at each point of O_1, has a nonzero determinant (**Jacobian**), then some neighborhood of each point $(x_1^0, \ldots, x_d^0)$ is mapped one-to–one onto some neighborhood of the point $(f_1(x_1^0, \ldots, x_d^0), \ldots, f_d(x_1^0, \ldots, x_d^0))$.

Indeed, suppose there are two points $(x_1^0, \ldots, x_d^0)$ and $(x_1^0 + \delta_1, \ldots, x_d^0 + \delta_d)$ and direct path between them: $x_k^0 + t\delta_k, 0 \le t \le 1, k = 1, \ldots, d$. The derivative of a vector function $F = (f_1, \ldots, f_d)$ in the direction of this path is equal to

$$\frac{dF}{dt} = \left\{ \sum_j \delta_j \frac{\partial f_i}{\partial x_j}, \;\; i = 1, \ldots, d \right\}.$$

But none of these vectors can be identically equal to zero due to the linear independence of the rows and columns of the Jacobian matrix. Therefore, at least one component is different from zero. Then, similar to (4.9), vectors $f(x_1^0, \ldots, x_d^0)$ and $f(x_1^0 + \delta_1, \ldots, x_d^0 + \delta_d)$ are different.

Change of measure. Let $\mu = \mu_0$ be Lebesgue measure on R. In mathematical statistical physics there is a popular term "change of measure". It is easy to understand this term on simplest examples:

1. Let a function $f(x)$ on the segment $[a, b]$ be given. Then, as it is known, the Lebesgue measure of any subinterval $I = [c, d]$ is its length $d - c$. Then new measure is $\mu_1(I) = \int_I f(x)dx$, if of course all these integrals exist.

2. If this function is increasing, then another new measure is

$$\mu_2(I) = f(d) - f(c) = \int_I f'(x)dx = \mu(f(I)).$$

Very important, and related to change of measure, is change of variables $x \to y(x), y = x(y)$.

Change of variables. In one-dimensional case the change of variables under the integral

$$\int_a^b f(y)dy = \int_{x(a)}^{x(b)} f(y(x))y'(x)dx$$

is easy to understand as the limit of sums

$$\sum f(y_k)(y_{k+1} - y_k) \sim \sum f(y(x_k))y'_x(x_k)(x_{k+1} - x_k)$$

or

$$\sum f(y_k)\Delta_k(y) \sim \sum f(y(x_k))y'_x(x_k)\Delta_k(x).$$

In multi-dimensional case, if $\Lambda \subset R^d$ is an open set and $y : \Lambda \to y(\Lambda)$ is its diffeomorphism onto $y(\Lambda)$, then the corresponding formula is

$$\int_\Lambda f(y)dy = \int_{x(\Lambda)} f(y(x))j(x)dx,$$

where $J(x)$ is the Jacobian of the map $y(x)$.

As an example, consider the calculation of the gaussian integral

$$\int_{-\infty}^{\infty} \exp\Big\{-\frac{x^2}{2}\Big\}dx = \sqrt{2\pi} \tag{4.10}$$

is derived by passing in the integral to the polar coordinates.

Polar coordinates (r, φ) on the plane with coordinates (x, y) are given by

$$x = r\cos\varphi, y = r\sin\varphi,$$

It is enough to prove that the double integral (the square of our integral)

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\Big\{-\frac{x^2+y^2}{2}\Big\}dxdy$$

equals 2π. When passing to the polar coordinate system, the Jacobian is equal to r, which gives

$$\int_0^\infty \Big(\int_0^{2\pi} d\varphi\Big) \exp\Big\{-\frac{r^2}{2}\Big\}rdr = \int_0^\infty \int_0^{2\pi} d\varphi \exp\Big\{-\frac{r^2}{2}\Big\}d\Big(\frac{r^2}{2}\Big) =$$

$$= \int_0^\infty (2\pi)e^{-x}dx = 2\pi.$$

4.3.2. Fourier transform and generalized functions

Finite cyclic groups. Consider a cyclic group $I_N = \{0, 1, \dots, N-1\}$ in addition modulo N, and the Hilbert space $l_2(I_N)$ of complex functions $f(n), n \in I_N$ on it with the scalar product and norm

$$(f, g) = \sum_{n=0}^{N-1} f(n)g^*(n), \;\; ||f||^2 = \sum |f_n|^2 < \infty. \tag{4.11}$$

The **Cauchy inequality**

$$|(f, g)| \le ||f||||g|| \tag{4.12}$$

which easily follows from the inequality $2|ab| \le a^2 + b^2$, if you square both parts of (4.12) and open the brackets.

In $l_2(I_N)$ there are two remarkable bases: 1) $e_m(n) = \delta_{m,\mathrm{n}}$, 2) $\tilde{e}_m(n) = \frac{1}{\sqrt{N}} \exp(2\pi i \frac{m}{N} n)$, where the orthogonality and normalization are easily tested, if we use the formula for $n \neq 0$:

$$\sum_{m=0}^{N-1} \Bigl(\exp 2\pi i \frac{n}{N}\Bigr)^m = 0$$

for $n \neq 0$. The **Fourier transform** here is called the unitary operator U in $l_2(I_N)$ that maps e_m on $\tilde{e}_m$, and the inverse Fourier transform is the inverse operator $U = U^{-1} = U^*$. Thus, the direct and inverse Fourier transform with an arbitrary function $f(m)$ will be

$$\tilde{f}(n) = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} f(m) \exp\Bigl(2\pi i \frac{m}{N} n\Bigr), \quad f(m) = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} \tilde{f}(n) \exp\Bigl(-2\pi i \frac{n}{N} m\Bigr).$$

Hilbert and Banach spaces. Here, at the first time, we transcend from finite-dimensional linear algebra to **infinite dimensional**. The Hilbert space can be define as a vector space with a scalar product and orthonormal countable basis such that the property similar to (4.11) holds with $N = \infty$. However, sometimes the basis is not so easy to find, and it is necessary to introduce more general concepts. **Normed space** is a linear vector space over R or C with norm $0 \leq ||x|| < \infty$ such that: 1) $||x|| = 0 \Longleftrightarrow x = 0$, 2) $||ax|| = |a|||x||$ for any number a, 3) $||x + y|| \leq ||x|| + ||y||$. The metric in it is defined as $\rho(f, g) = ||f - g||$. The space is called **complete** if any Cauchy sequence, under this metrics, converges to some element of the space. A complete normed space is called **Banach space**.

Two examples of Banach spaces on finite or countable sets X: $l_1(X) \subset l_\infty(X)$, respectively with the norms

$$||f||_1 = \sum_{x \in X} |f(x)|, \quad ||f||_\infty = \sup_{x \in X} |f(x)|.$$

A Banach space where the norm is determined by the scalar product, that is, the bilinear form with the properties as above, is called a **Hilbert space**. It is easy to show that every Hilbert space has an orthonormal basis, and if it is countable, then the space is called **separable**.

The projection of vector x on vector y is defined as $(x, y)\frac{y}{||y||}$, and on the subspace with an orthonormal basis $\{g_k\}$ as $\sum_k (x, g_k) g_k$.

A classical example of Hilbert space is the set $L_2([0, 1])$ of complex or real functions on the interval $[0, 1]$, the square modulus of which is integrable. Since we, on purpose, do not consider measurable functions in all their generality, it is convenient to define $L_2([0, 1])$ and the like as the completion of any class of fairly simple functions in the metric of the norm, not identifying the limit

elements with concrete functions. For example, classes of piece-wise constant, continuous and differentiable functions. It is easy to show that in each of these three cases, we get the same space.

Group Z. Firstly, let us define two basic Hilbert spaces. The first one is $l_2(Z)$, where Z is the integer lattice. Let $e_n(k) = \delta_{nk}$ be the standard orthonormal basis in it. Any element (a function f on Z) of $l_2(Z)$, is, by definition, can be represented in the form

$$f = \sum_{n \in Z} f_n e_n$$

with complex coefficients $f_n = f(n)$, with scalar product and norm defined as above, and the Cauchy inequality also holds.

The second one is $L_2(S)$. This is a set of functions on the circle S of length 1 with finite norm and scalar product

$$||f||^2 = \int_0^1 |f(x)|^2 dx < \infty, \quad (f,g) = \int f(x) g^*(x) dx,$$

where the Cauchy inequality also holds. An orthonormal basis $g_n = e^{2\pi i n x}, x \in [0,1)$ determines the isomorphism between l_2 and $L_2(S)$:

$$f = f(n) = \sum_n f(n) e_n \Longleftrightarrow \varphi(x) = \sum_n f(n) e^{2\pi i n x}. \tag{4.13}$$

We will show the preservation of scalar products, i.e. the unitarity of the Fourier transform:

$$(\tilde{f}, \tilde{g}) = \int \tilde{f}(\tilde{g})^* du = \int_0^1 du \sum_n f(n) e^{2\pi i n u} \sum_m g^*(m) e^{-2\pi i m u} =$$

$$= \sum_n f(n) \sum_m g^*(m) \int_0^1 e^{2\pi i (n-m) u} du = \sum_n f(n) g^*(n) = (f,g). \tag{4.14}$$

In this case, the inverse transform will be

$$f(n) = \int \tilde{f}(u) e^{-2\pi i n u} du.$$

It remains to show that the Fourier transform is a map on all L_2, that is, for any function f in L_2 at least one coefficient (f, g_n) is not zero. In view of our definition of L_2, it is enough to prove it for example for any continuous function. We restrict ourselves to real functions – for example, let for some f, all Fourier coefficients are equal to zero. Note that then any sum

$$\sum_k (a_k \cos 2\pi k + b_k \sin 2\pi k) \tag{4.15}$$

will be orthogonal to this function. Suppose that at some point the function f is positive. For example, let $f(0) = 1, f(x) < C$ for all x and some $C > 0$. Let us prove that for sufficiently large N the function

$$\Big(\frac{1+\cos 2\pi x}{2}\Big)^N,$$

which is of course representable in the form (4.15), is not orthogonal to f. Divide the interval $[-\frac{1}{2}, \frac{1}{2}]$ into 3 parts: $I_1 = \{x : |x| < \varepsilon_1\}, I_1 = \{x : \varepsilon_1 \le |x| < \varepsilon_2\}, I_3 = \{x : \varepsilon_2 \le |x|\}$. For sufficiently small (but not dependent on N) $0 < \varepsilon_1 < \varepsilon_2$ there are small $0 < \delta_1 < \delta_2$ and $C_1 > 0$ (also independent of N) such that $f(x) > C_1, x \in I_1 \cup I_2$, and

$$1 - \delta_1 < \frac{1+\cos 2\pi x}{2} \le 1, \;\; x \in I_1,$$

$$\frac{1+\cos 2\pi x}{2} \le 1 - \delta_2, \;\; x \in I_1.$$

Divide the integral

$$\int_0^1 \Big(\frac{1+\cos 2\pi x}{2}\Big)^N f(x)dx = \int_{I_1} + \int_2 + \int_{I_3},$$

where the first term is greater than $2\varepsilon_1 C_1(1-\delta_1)^N$, the second is positive, and the third is less than $C(1-\delta_2)^N$. From here the result follows.

It also follows that the Fourier series (4.13) converges in the metrics of L_2. Fourier series were earlier an extremely popular area of mathematics, in which many technique were developed.

Remark 8. For non-commutative groups there is a similar science – the theory of characters and representations of groups, but its algebra is more complicated, and in many sections of classical mathematical physics is not necessary.

Fourier integrals. Denote by S the class of functions (the Schwartz space) on R which are infinitely smooth and all their derivatives decrease at infinity faster than any degree. Fourier transform $\tilde{f}$ of function f (direct and inverse) is usually determined as follows

$$\tilde{f}(u) = C_1 \int_R f(x) e^{-C_2 iux} dx, \tag{4.16}$$

it maps S to itself, that is proved by integration by parts. Historically there were used several types of real constants. For example, 1) when $C_1 = 1 = -C_2$ for direct and $C_1 = C_2 = 1$ for reverse, 2) $C_1 = \frac{1}{\sqrt{2\pi}}$, $C_2 = -1$, $C_1 = \frac{1}{\sqrt{2\pi}}$, $C_2 = 1$. Everywhere below we use option 1).

Main results:

1. All these transformations are one-to-one on S, it is enough to prove that there exists a constant $C \neq 0$ such that for any $f \in$S

$$\int_R \tilde{f}(u)e^{iux}du = Cf(x). \tag{4.17}$$

2. For functions from S they save (up to a multiplier depending only on C_1 and C_2) the scalar product in $L_2(R)$.

Let us prove first (4.17) for the Gaussian density, i.e. the Fourier transform of the function $\exp(-x^2/2)$ coincides with itself, multiplied by 2π:

$$\int_{-\infty}^{\infty} \exp\Big(-\frac{x^2}{2} - itx\Big)dx = \int_{-\infty}^{\infty} \exp\Big(-\Big(\frac{x}{\sqrt{2}} + \frac{it}{\sqrt{2}}\Big)^2 + \Big(\frac{it}{\sqrt{2}}\Big)^2\Big)dx =$$

$$= \exp\Big(-\frac{t^2}{2}\Big)\int_{-\infty}^{\infty} \exp\Big(-\Big(\frac{x}{\sqrt{2}} + \frac{it}{\sqrt{2}}\Big)^2\Big) = \exp\Big(-\frac{t^2}{2}\Big)\int_{-\infty+it}^{\infty+it} \exp\Big(-\frac{z^2}{2}\Big)dz,$$

where we made the change of variable $z = x + it$. But

$$\int_{-\infty+it}^{\infty+it} \exp\Big(-\frac{z^2}{2}\Big)dz = \int_{-\infty}^{\infty} \exp\Big\{-\frac{x^2}{2}\Big\}dx.$$

In fact, consider the complex integral of $\exp(-z^2/2)$ along closed contour consisting of 4 straight line segments on the complex plane

$$[(-N,0),(N,0)], \quad [(N,0),(N,N+it)],$$

$$[(N,N+it),(-N,-N+it)], \quad [(-N,-N+it),(-N,0)].$$

It is equal to zero. But when $N \to \infty$ the integrals on the second and fourth segment tend to zero. And then the first and third integrals have different signs. It remains to use the formula(4.10) below.

Now for any $f \in S$:

$$\int_R \tilde{f}(u)e^{iux}du = \int_R \Big(\int_R f(y)e^{-iuy}dy\Big)e^{iux}du = \int_R\int_R f(y)e^{iu(x-y)}dudy =$$

$$= \lim_{N\to\infty}\int_R f(y)dy\Big(\int_{-N}^{N} e^{iu(x-y)}du\Big) = \lim_{N\to\infty}\int_R f(y)\frac{2\sin N(x-y)}{x-y}dy =$$

$$= \Big(y = \frac{z}{N} + x\Big) = \lim_{N\to\infty} 2\int_R f\Big(\frac{z}{N} + x\Big)\frac{\sin z}{z}dz = f(x). \tag{4.18}$$

The last integral, as $N \to \infty$, tends to $f(x)\int_{-\infty}^{\infty}\frac{\sin z}{z}dz$, the latter integral is called the integral sine (Dirichlet integral), and it is well known that

$$\int_{-\infty}^{\infty} \frac{\sin z}{z}dz = \pi.$$

Thus the desired constant is equal to 2π. Which, incidentally, can be verified by ourselves if in (4.18) we take the Gaussian density for $f(x)$, for which we already know this.

And the fact that scalar products are preserved is verified by a similar calculation.

Remark 9. In case of double limits, as for example in $\int_{-\infty}^{\infty}$, there is a nuance – in what sense to understand it. In many cases, for simplicity, we can understand this as $\lim_{N\to\infty}\int_{-N}^{N}$.

It is very helpful, in solving various equations, to know how the Fourier transform changes after simplest transformations of the function f:

1) $f(x) \to g(x) = f(x+s) \Longrightarrow \tilde{f}(u) \to \tilde{g}(u) = e^{ius}\tilde{f}$;
2) $f(x) \to g(x) = \frac{df}{dx} \Longrightarrow \tilde{f}(u) \to \tilde{g}(u) = iu\tilde{f}$;
3) **Convolution** of two functions is defined as the function

$$(f * g)(x) = \int f(x-y)g(y)dy,$$

whose Fourier transform is equal to the product of their Fourier transforms

$$\int e^{-iux} f(x-y)g(y)dydx = \int [f(x-y)e^{-iu(x-y)}d(x-y)]e^{-iuy}g(y)dy = \tilde{f}\tilde{g}.$$

Generalized functions. It is often useful to expand the space C^∞, adding to it the "worst" functions (or even not functions) without reducing the freedom of the algebraic computations. For example, the function equal to 1 at one point and zero in the others, is very important on Z, but is actually zero on the real axis. And the function, equal ∞ at one point and zero in all other, it is not known what, but appeared to be very illustrative and important. Often one has to do with infinite integrals, and often needs to change the integrands to get finite results.

This can be done as follows: each function $f \in C^\infty$ defines also linear functional L_f on C^∞:

$$L_f(g) = \int fgdx, \;\; g \in C^\infty.$$

And we shall extend the class of such linear functionals.

Let us explain this on the example of the famous δ-function. Consider the function $D_N(x) = N$ on the interval $(-\frac{1}{2N}, \frac{1}{2N})$ and 0 otherwise. In the limit $N \to \infty$, it is equal to 0 everywhere except the point 0 where it is infinite. Moreover, its integral with any function from C^∞

$$\int f(x)D_N(x)dx \underset{N\to\infty}{\longrightarrow} f(0) = \int f(x)\delta(x)dx.$$

This is a linear functional on C^∞ defined by the so-called δ-function $\delta(x) = \lim D_N(x)$. As the physicists say, "the function equal to zero everywhere except for the point 0, where it is infinite, and its integral is equal to 1". Instead of simple function D_N one can use many others, such as the Gaussian density $\frac{1}{\sqrt{2\pi N}} \exp(-\frac{x^2}{N})$.

Another example of a linear functional on S is

$$L_1(f) = \int_R f(x)dx$$

generated by the function identically equal to 1.

In general, the Fourier transform $\tilde{L}$ of the linear functional L is defined as follows: $\tilde{L}\tilde{f} = Lf$. We will show in two ways that the Fourier transform of the δ-function is the function equal to 1. First way:

$$\tilde{D}_N(u) = N\int_{-1/(2N)}^{1/(2N)} e^{-iux}dx = N\int_{-1/(2N)}^{1/(2N)} \cos uxdx = \frac{N}{u}(\sin ux)\Big|_{-1/(2N)}^{1/(2N)} =$$

$$= \frac{2N}{u}\Big(\sin\frac{u}{2N}\Big) \to 1.$$

Second way:

$$\int_{-\infty}^{\infty} \tilde{f}du = \lim_{N\to\infty}\int_{-N}^{N} \tilde{f}du = \lim_{N\to\infty}\int f(x)\left(\int_{-N}^{N} e^{-iux}du\right)dx =$$

$$= \lim_{N\to\infty}\int f(x)\frac{2}{x}\sin Nxdx =$$

$$= \lim_{N\to\infty}\int (f(0) + f(0)x + O(x^2))\frac{2}{x}\sin Nxdx \sim \lim_{N\to\infty} f(0)\int \frac{2}{x}\sin Nxdx =$$

$$= \lim_{N\to\infty} f(0)\int \frac{2}{Nx}\sin Nxd(Nx) = \frac{1}{2\pi}f(0)\int \frac{2}{y}\sin ydy = f(0).$$

Here are other examples of generalized functions:

1) the shifts of δ-functions $\delta(x - x_0)$ give the linear functionals $L(f) = \int_R f(x)\delta(x - x_0)dx = f(x_0)$;

2) for generalized functions is defined their addition and multiplication by a number and by C_∞ function;

3) the derivative of a generalized function is defined as $L'(f) = -L(f')$. In particular, $L'_\delta(f) = -L_\delta(f') = -f'(0)$;

4) the regularization of a singular (ordinary) function x^{-1} – its transformation into a generalized one. Two ways: the **principal value integral**

$$Lf = \lim_{\varepsilon\to 0}\int_{-\varepsilon}^{\varepsilon} \frac{1}{x}f(x)dx,$$

and

$$Lf = \int_R \frac{1}{x}(f(x) - f(0))dx,$$

which is the simplest example of **subtracting divergences**, commonly used in quantum field theory.

5. First models of mathematical physics

5.1. Three finite-dimensional dynamics

All our life is in time and dynamics – evolution of the states of various systems in time. Mainly, three dynamics (and their mixtures) are considered. Here we give simplest examples of these 3 dynamics.

5.1.1. Deterministic dynamics

Iterations of finite set transformations. We consider the map $F : X \to X$ of finite set $X = \{1, \dots, N\}$ (called the the **set of states**, or **state space**) into itself. The graph of this mapping is called a directed graph with set of vertices X and N directed edges $l_{ij} = (i, j = F(i))$. If F is one-to-one, then the graph consists of several cycles, where cycle is a sequence of vertices $i_1, \dots, i_n$ such that $F(i_k) = i_{k+1}$ for $k = 1, \dots, n-1$ and $F(i_n) = i_1$. A cycle of length $n = 1$ is called a **fixed point**.

The initial state i_0 and the iterations $i_{0,}, i_1 = F(i_0), \dots, i_{k+1} = F(i_k), \dots$ of the map F determine the evolution of states in discrete time $t = k = 0, 1, 2, \dots$.

The map F induces a mapping $L_F : \Phi(X) \to \Phi(X)$ of the set $\Phi(X)$ of complex functions $f(x)$ on X to itself

$$(L_F f)(i) = f(Fi),$$

and also the mapping $M_F : M(X) \to M(X)$, of the set $M(X)$ of measures μ on X to itself

$$(M_F \mu)(A) = \mu(F^{-1}A). \tag{5.1}$$

A measure μ is called **invariant**, if $M_F \mu = \mu$. The map F is called **ergodic** if invariant measure is unique (up to a constant factor). In this case, for any (initial) measure μ

$$\frac{1}{T}(M_F(\mu) + M_F^2(\mu) + \ldots + M_F^T(\mu)) \tag{5.2}$$

tends to the invariant measure as $T \to \infty$. In the general case, for any initial measure μ (5.2) converges to some (dependent on μ) invariant measure.

All the following statements are almost obvious. The measure is called constant on the subset $A \subset X$ if it has the same value for any single point subset.

In the case of one-to-one F the invariant measure is constant on each cycle, and is unique (up to a constant multiplier) if there is exactly one cycle.

Similar convergence can be defined for functions f on X. As $T \to \infty$

$$\frac{1}{T}(L_F(f) + L_F^2(f) + \ldots + L_F^T(f))$$

converges to some function $\varphi(x)$ depending on f. It is of course a trivial case of the famous theorem of Birkhoff – Khinchin. If F is ergodic, the function $\varphi(x)$ is constant.

5.1.2. Stochastic dynamics

Matrix (operator) series. In the set of all matrices of given dimension N one can introduce the norm $||A||$, also the distance $||A - B||$, creating thus the metric space of all matrices. For example, the norm could be this:

$$||A|| = \sum_{i,j} |a_{ij}|.$$

Then for any square matrix A the matrix series (matrix exponent)

$$e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!} = E + A + \frac{A^2}{2} + \ldots$$

converges in the sense of this norm.

The dynamics of measures on X. Now the states are not only points of the set X, but all probability measures p on X, that is non-negative and normalized measures p, i.e. such that $p(A) \geq 0$ for all $A \subset X$ and $p(X) = 1$. Measures of single-point subsets $\{k\}$ (single-point measures) will be denoted by p_k. Thus, the state at time $t = 0, 1, \ldots$is defined by the row vector $p(t) = (p_1(t), \ldots, p_N(t))$.

General linear dynamics on the set of probability measures on X is defined by $(N \times N)$-matrix $P = (p_{ij})$ with non-negative elements, and moreover, $\sum_j p_{ij} = 1$ for all i. Such matrices are called **stochastic**. The dynamics is determined by multiplying the row vector

$$p(t) = p(0)P^t$$

where the components $p_i(t) : i \in X$ are called the probabilities for the system to be in state i at time t. If the components $p_i(0), i = 1, \ldots, N$ of the initial vector $p(0)$ are non-negative and normalized, i.e. $\sum_{i=1}^N p_i(0) = 1$, it is easy to prove that all vectors p(t) have these properties. Simple examples: 1) if the matrix P has identical rows (this is called **Bernoulli scheme**), all measures $p(t)$ will be the same for $t > 0$, 2) each row has exactly one unit, the rest are zeros, which corresponds to the induced deterministic dynamics (5.1).

This dynamics of measures is called a finite Markov chain with discrete time. Probability measure π is called invariant if $\pi P = \pi$. Markov chain is called **ergodic** if for any initial state $p(0)$ as $t \to \infty$

$$p(0)P^t \to \pi.$$

Let us prove that any matrix P, all elements of which are positive, defines ergodic Markov chain.

For this we need a **method of successive approximations**. We give now its formulation.

Let be given complete (that is where any Cauchy sequence has a limit) metric space M with a distance of ρ and finite diameter $D = D(M) = \sup_{x,y \in M} \rho(x, y)$, and the mapping $F : M \to M$ is defined such that there exists number $0 < \alpha < 1$ such that for any two points $x, y \in M$

$$\rho(Fx, Fy) < \alpha\rho(x, y). \tag{5.3}$$

Then there exists a unique fixed point of F, i.e. such that $Fx = x$.

Indeed, then for any point the sequence

$$x_0 0, x_1 = F(x_0), \ldots, x_{k+1} = F(x_k), \ldots$$

is a Cauchy sequence, since

$$\rho(x_k, x_{k+1}) < \alpha\rho(x_{k-1}, x_k) < \alpha^k D.$$

That is why the sequence x_k has unique limit $x = Fx$. Any other sequence y_k, starting from any y_0 has the same limit as the points x_k and y_k become closer when $k \to \infty$.

In our case define the norm $|\mu|$ of measure μ, and the distance between two measures μ_1, μ_2 (not necessarily probabilistic) as

$$|\mu| = \sum_{x \in X} |\mu(x)|, \quad \rho(\mu_1, \mu_2) = |\mu_1 - \mu_2| = \sum_{x \in X} |\mu_1(x) - \mu_2(x)|.$$

We call the carrier of measure μ the set of points where it is nonzero. Then, if two measures μ_1, μ_2 have disjoint carriers, then $\rho(\mu_1, \mu_2) = \sum_{x \in X} |\mu_1(x)| + |\mu_2(x)|$. Measure $\mu_1 - \mu_2$ and $(\mu_1 - \mu_2)P$ can be written as the difference of the two non-negative measures ν_1, ν_2 with the same norm and disjoint carriers. We want to prove that for any two such measures the norm of $(\mu_1 - \mu_2)P$ will decrease in comparison with the norm $(\mu_1 - \mu_2)$

$$||(\mu_1 - \mu_2)P|| \leq \alpha||\mu_1 - \mu_2||,$$

where

$$\alpha \leq 1 - \min_{i.j \in X} p_{ij}.$$

Indeed, consider the points x_1, x_2, where $\nu_m(x_m) \geq \frac{1}{N}|\nu_m|, m = 1, 2$. Then at each point $y \in X$ in the difference $\nu_1(x_1)p_{x_1y} - \nu_2(x_2)p_{x_2y}$ will occur plus-minus cancellation in

$$(\min \nu_1(x_1))(\min p_{xy}) \geq \frac{1}{N}|\nu_1|(\min p_{xy}).$$

Summation in y gives the result.

Finite Markov chains with discrete time. Now we shall look at the dynamics of the measures differently. Consider the set

$$\Omega = \Omega_\infty = \{\omega = (x_0, x_1, \dots, x_n, \dots\}$$

of infinite sequences of elements $x_n \in X$. Its elements ω we will call **elementary events** and present them as trajectories (paths) of a particle, jumping on the set X from point x_n to x_{n+1}. While the indices $0, 1, 2, \dots$ we consider as a sequence of discrete moments of time.

We call basic events $A(i_0, \dots, i_n)$ the set of all infinite sequences ω such that $x_k = i_k$ for $k = 0, 1, \dots, n$. Let us denote Σ the minimal σ-algebra, generated by all basic events. Now let be given measure $p_0(x)$ on X and stochastic matrix P. Define probability measure μ_n on events A(i_{0}, ...,i_{n}), which will define Markov chain

$$\mu_n(A(i_0, \dots, i_n)) = p_0(i_0)p_{i_0i_1}p_{i_1i_2} \dots p_{i_{n-1}i_n}.$$

These measures agree in the sense that

$$\mu_n(A(i_0, \dots, i_n)) = \sum_{i_{+1} \in X} \mu_{n+1}(A(i_0, \dots, i_{n+1})).$$

Then the Kolmogorov theorem states that on the σ-algebra Σ, generated by the base sets, there is unique measure μ, which coincides with the measures μ_n on base sets. Sets $A \in \Sigma$ are called **events**, and the measures $\mu(A)$ – the **probabilities** of these events. Real-valued functions $\xi = \xi(\omega)$ on Ω are called random variables, provided they satisfy the following measurability condition: the inverse image $\xi^{(-1)}(I)$ of any set $I \subset R$ belongs to Σ. For example, for any n, all functions depending only on the first n coordinates, are measurable.

Continuous time. We will start with formal definition using semigroup language: $(N \times N)$ matrix $H = (\lambda_{ij})$ with elements

$$\lambda_{ij} \geq 0, \ i \neq j, \quad \lambda_{ii} = -\sum_{j:j \neq i} \lambda_{ij}$$

is called the generator of the matrix semigroup

$$U^t = e^{Ht} = \sum_{n=0}^{\infty} \frac{(Ht)^n}{n!},$$

where the series converges for any t. Markov property on semigroup language is:

$$U^{t+s} = U^t U^s.$$

Firstly, we shall see that all matrices U^t are stochastic. It is sufficient to prove it for all t sufficiently small, then their products show that it is also true for all t. Assume for example that all $\lambda_{ij} > 0$ for $i \neq j$. Positivity of all elements follows as for small Δt:

$$U_{ij}^{\Delta t} = \lambda_{ij}\Delta t + o(\Delta t),\ i \neq j, \quad U_{ii}^{\Delta t} = 1 + \lambda_{ii}\Delta t + o(\Delta t). \tag{5.4}$$

The normalization property follows as any powers of H keep the property of H itself – zero sum of elements for every row.

If $p(0) = \{p_x(0), x \in X\}$ is the initial measure, then any component of the vector $p(t) - p(0)U^t$ is an analytic function of t. Now we will see how non analytic trajectories miraculously appear. Similar to discrete time introduce the so called random variables. We will consider them as functions on the set of ALL events (we do not introduce measurable functions on purpose) events, that is arbitrary functions $\omega = \omega(t) : [0, \infty) \to X$. For example, $\xi(t) = \xi(t, \omega) = x \in X$ is the event $\omega(t) = x$, that is the event that our system at time t will be at the state x. Elements U_{ij}^t will be understood as the probabilities to be at j at tome t, under the condition that at time 0 we were at i. More complicated events: fix finite number of time moments $0 \leq t_1 < \ldots < t_k$ and define probabilities (finite-dimensional distributions)

$$P(\omega(t_1) = i_1, \ldots, \omega(t_k) = i_k) = \sum_i p_i(0) U_{ii_1}^{t_1} U_{i_1 i_2}^{t_2 - t_1} \ldots U_{i_{k-1} i_k}^{t_k - t_{k-1}} \tag{5.5}$$

of the event where we were at these discrete moments.

It is a natural desire to define and find the probability of the event that during time t we always left at the state i, that is there was no jumps during time t. We define it as the limit of distributions (5.5)

$$\lim_{N\to\infty} P\Big(\omega\Big(\frac{kt}{N}\Big) = i,\ k = 1, 2, \ldots, N\Big) = \lim_{N\to\infty} \Big(1 + \lambda_{ii}\frac{t}{N} + o\Big(\frac{t}{N}\Big)\Big)^N = \exp\{\lambda_{ii} t\}.$$

And the probability that until time t we were at the state i, and in time interval $(t, t + \Delta t)$ already not, by Markov property, will be

$$e^{\lambda_{ii} t}(1 - e^{\lambda_{ii}\Delta t}) = -\lambda_{ii} e^{\lambda_{ii} t}\Delta t + o(\Delta t) = \int_t^{t+\Delta t} f(t)dt + + o(\Delta t),$$

where the function $f(t) = -\lambda_{ii} e^{\lambda_{ii} t}$ is called the exponential density of the distribution of the the moment of the first jump from state i. Note that the probability that such jump ever occurs is 1.

Example is the Poisson process, where $X = \{0, 1\}$ and

$$H = \begin{pmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{pmatrix},$$

which describes deterministic evolution in "random time moments". The process jumps from 0 to 1 and back with the same "intensity" λ, that is random lengths ξ_n of intervals between jumps have the same distribution. Their distribution function is

$$F(t) = P(\xi_n \le t) = 1 - e^{-\lambda t},$$

and the distribution density is $f(t) = \lambda e^{-\lambda t}$.

5.1.3. Quantum dynamics

Self-adjointness and unitarity. Any complex linear space L of dimension N, can be identified with the set of complex functions $\psi(i)$ on finite set $\{1, \dots, N\}$. Introducing scalar product and the norm

$$(\psi_1, \psi_2) = \sum_{i=1}^{N} \psi_1(i)\psi_2^*(i), \;\; ||\psi|| = \sqrt{(\psi, \psi)}$$

makes it a metric space with the distance $\rho(\psi_1, \psi_2) = ||\psi_1 - \psi_2||$, that gives it the name Hilbert space $l_2 = l_2(\{1, \dots, N\})$. The functions $e_k = e_k(j) = \delta_{kj} = 1$, if $k = j$, and 0 if $k \ne j$, form its basis. Moreover, they are orthogonal and normalized: $(e_k, e_l) = \delta_{kl}$. It is said that they form the **orthonormal basis**.

Spectral theorem for self-adjoint operators. Let A be linear operator in L, then linear operator A^* in L is called adjoint to A, if for any ψ_1, ψ_2

$$(A^*\psi_1, \psi_2) = (\psi_1, A\psi_2).$$

It is easy to show that it exists and unique for any A, and its matrix in the orthonormal basis will be (a_{ji}^*), is a_{ij} is the matrix for A in this basis.

Operator A is called **self-adjoint**, if $A = A^*$,that is $a_{ij} = a_{ji}^*$ for all i, j. As any linear operator, a self-adjoint operator A has some eigenvector ψ_1 (assume that its norm is 1) with some eigenvalue λ_1. This eigenvalue is real as $(\lambda_1\psi_1, \psi_1) = (A\psi_1, \psi_1) = (\psi_1, A\psi_1) = (\psi_1, \lambda_1\psi_1) \Longrightarrow \lambda_1 = \lambda_1^*$.

Denote $L_1^\perp$ the set of all ψ, orthogonal to ψ_1, that is $(\psi, \psi_1) = 0$. Then $L_1^\perp$ is a linear subspace of dimension $N - 1$. It is invariant with respect to A, as $(\psi_1, A\psi) = (\psi_1, \lambda_1\psi) = 0$, if $(\psi_1, \psi) = 0$.

Then we forget about ψ_1 and carry out the same reasoning in $L_1^\perp$, We construct λ_2, ψ_2, and similarly by induction construct eigenvectors $\psi_1, \dots, \psi_N$ – orthogonal and having real eigenvalues $\lambda_1, \dots, \lambda_N$. Then, in this orthonormal basis, A is diagonal, that is the multiplication on its eigenvalues in $l_2\{1, \dots, N\}$.

Self-adjoint operator is said to have **simple spectrum**, if the multiplicity of each eigenvalue is 1 (that is the number of eigenvalues is N).

Unitary operators. Linear operator U is called **unitary**, if one of two following equivalent conditions holds:

a) U is invertible and $U^{-1} = U^*$,

b) $(U\psi_1, U\psi_2) = (U^*U\psi_1, \psi_2) = (\psi_1, \psi_2)$.

Thus, it conserves scalar product, orthogonality and norm. Due to norm conservation, its eigenvalues belong to the unit circle, that is equal $e^{i\lambda}$ with real λ, and the eigenvectors are orthogonal. In fact, if ψ is orthogonal to the eigenvalue ψ_1, then $U\psi$ is also orthogonal to ψ_1, as $(\psi_1, U\psi) = (U^{-1}\psi_1, \psi) = (e^{-i\lambda}\psi_1, \psi) = 0$. That is why there exists $H = H^*$ such that

$$U = e^{iH}.$$

The matrix function $e^{itH}, t \in R$, is called a **unitary group**. In quantum physics (in infinite dimensional case) this group defines quantum dynamics (where t is time) with Hamiltonian H.

Operator (or matrix) algebra

Trace. The trace TrA of a square matrix A is defined as the sum of its diagonal elements. Simplest properties of the trace:

$$Tr(AB) = \sum_{i,j} a_{ij}b_{ji} = \sum_{i,j} b_{ij}a_{ji} = Tr(BA),$$

from where the invariance with respect to the choice of basis follows

$$Tr(CAC^{-1}) = Tr(C^{-1}CA) = TrA.$$

States. Consider the set of $(N \times N)$-matrices $\mathbf{M} = \mathbf{M}_N$. It is an **algebra**, that is a linear space with multiplication operation satisfying the **distributivity** axiom

$$\boldsymbol{A(B + C) = AB + AC},$$
$$\boldsymbol{(B + C)A = BA + CA.}$$

Any linear functional $L = L(M)$ on the matrix algebra can be presented as

$$L(M) = Tr(AM)$$

for some matrix A.

Namely, it is defined by its values λ_{ij} on the basis matrices $E_{ij} = (\delta_{ij})$, then defining matrix $\Lambda = (\lambda_{ij})$, for any matrix $M = (m_{ij})$:

$$L(M) = \sum_{i,j} \lambda_{ij}m_{ij} = \sum_{i,j} \lambda'_{ji}m_{ij} = \sum_{j}\sum_{i} \lambda'_{ji}m_{ij} = Tr\Lambda' M.$$

Also, any linear functional can be also presented as

$$L(M) = \sum_i (Mg_i, g_j)$$

where g_i are some vectors, and the sum is finite.

Self-adjoint operator is called **positive** (**positive definite**) $A > 0$ or **non-negative** $A \geq 0$, if all its eigenvalues are positive (non-negative). Or: $A \geq 0$ is equivalent to any of the following conditions

$$\exists B : A = BB^* \Leftrightarrow (A\psi, \psi) \geq 0, \forall \psi \in R^N.$$

State φ on $\mathbf{M}$ is a linear functional on $\mathbf{M}$ with the following properties: 1) positivity, that is $\varphi(AA^*) \geq 0$ for all $A \in \mathbf{M}$, 2) normalization – $\varphi(E) = 1$, where E is the unit matrix in $\mathbf{M}$. Examples: **pure** states – are defined by one vector $\psi \in C^N$

$$\varphi(A) = (A\psi, \psi),$$

where $||\psi||^2 = (\psi, \psi) = 1$. Other states are called **mixed states**.

Any state can be presented as

$$\varphi(A) = Tr(\rho A) \tag{5.6}$$

where $\rho \in \mathbf{M}$ and moreover $\rho \geq 0, Tr\rho = 1$. Let us show positivity

$$\rho = CC^* \to Tr(\rho BB^*) = Tr(CC^*BB^*) = Tr(C^*BB^*C) =$$

$$= Tr((B^*C)^*B^*C) \geq 0.$$

Vice-versa: if $Tr(\rho A)$ is a state, then $\rho \geq 0$. It is clear that ρ for a given state is unique.

For pure state $\rho = P_\psi$, where P_ψ is the orthogonal projector on the normalized vector ψ.

Any mixed state can be presented as

$$\rho(A) = \sum_i (A\psi_i, \psi_i),$$

where the sum is finite and

$$\sum_i ||\psi_i||^2 = \sum_i (\psi_i, \psi_i) = 1.$$

In fact, let us diagonalize

$$\rho = U^{-1}\Big(\sum c_i P_i\Big)U,$$

where P_i is the orthogonal projector on the unit vector f_i. Then putting $\varphi_i = \sqrt{c_i} f_i$, we have

$$Tr(\rho A) = Tr((\sum c_i P_i)UAU^{-1}) =$$

$$= \sum (\varphi_i, UAU^{-1}\varphi_i) = \sum (U^{-1}\varphi_i, AU^{-1}\varphi_i).$$

Quantum evolution. Pure quantum states are usually called normalized vectors ψ. And if they considered as complex functions $\psi(x)$ on finite (or countable) set X, then they are called **wave functions**. It is always assumed that $\psi(x)\psi^*(x) = |\psi(x)|^2$ may be considered as probabilities that the system is at the point x.

Observables (physical variables) are operators. Mean value of the observable A in the state ψ is called the number

$$\langle A\rangle_\psi = (A\psi, \psi).$$

Their time evolution can be defined by two equivalent presentations.

1) **Schroedinger representation**, where the states change in time, and the observables do not change. Let self-adjoint operator be given H (called **Hamiltonian**) and the unitary group $U^t = e^{itH}$, defining the evolution of the vector $\psi(0)$ (and thus of the pure state $\rho(0)$) as

$$\psi(t) = U^t\psi(0).$$

Then the equation holds (called **Schroedinger equation**)

$$\frac{d}{dt}\psi(t) = iH\psi(t), \quad \rho(t) = U^{-t}\rho(0)U^t.$$

2) **Heisenberg representation**, where the states do not change but the observables change. Any unitary group of operators U^t defines the automorphism group of the operator algebra

$$\alpha_t(A) = A_t = U^tAU^{-t}$$

and satisfies the **Heisenberg equation**

$$\frac{d}{dt}A_t = it[H, A_t].$$

Mean values are changed similarly in both presentations

$$(A\psi(t), \psi(t)) = (\alpha_t(A)\psi, \psi).$$

5.2. Linear Differential Equations

5.2.1. First and second order

It is natural firstly to understand the situation with general linear equations of the first order

$$\frac{dx}{dt} = a(t)x + f(t). \tag{5.7}$$

If $f(t) = 0$ and a does not depend on t, then the solution is Ce^{at}, where C is an arbitrary constant. It is uniquely determined by the initial condition

$x(0) = x_0 = C$. But will this solution be unique depends only on the class of functions in which we would like to prove the uniqueness. The proof of uniqueness is easy in the class of functions which, in some neighborhood, can be expanded in a power series. From (5.7) we have the recurrent relation

$$x(t) = \sum_{k=0}^{\infty} a_k t^k \Longrightarrow (k+1)a_{k+1} = a_k, \;\; k = 0, 1, \ldots,$$

and substituting $t = 0$, we obtain $a_0 = x(0)$. Whence we obtain the uniqueness.

With an arbitrary function $a(t)$ we also look for a solution in the form $x(t) = Ce^{g(t)}$, where $C = x(0)$. Substituting, we get a simple equation, from where

$$g(t) = \int_0^t a(t)dt. \tag{5.8}$$

Partial solution of the inhomogeneous equation we are looking for in the form $x = D(t)\exp(\int_0^t a(t)dt)$, whence we obtain the equation for D, and the solution is

$$\dot{D} = \frac{d}{dt}\left(x\exp\left(-\int_0^t adt\right)\right) = f(t)\exp\left(-\int_0^t adt\right) \Longrightarrow$$

$$\Longrightarrow D = \int_0^t f(t)\exp\left(-\int_0^t a(s)ds\right)dt.$$

General solution is the sum of this solution (it is zero for $t = 0$) and the solution of the **homogeneous** (with $f = 0$) equation with given initial condition.

Simplest second order equations. It is always useful to start with a large number of simple examples, what we will do now. The general linear equation of second order has the form

$$m\ddot{x} = F(x, \dot{x}, t) = -k(t)x - \alpha(t)\dot{x} + f(t) \tag{5.9}$$

with the initial conditions $x(0), v(0)$. Right part of it is the general class of forces for linear (in x and $\dot{x}$) **inhomogeneous** equations. If $f(t)$ is identically zero, the equation is called **homogeneous**. Firstly, we assume that $k(t) \geq 0$ and will denote it $k(t) = \omega_0^2(t)$.

This is the simplest of **Newton equations**, on which classical mechanics is based. Below, we shall consider the main special cases of this equation, where physical interpretation plays important role. Here $x(t) \in R$ – coordinate of the point particle at time t, an integer $m > 0$ is called **mass**, further on to simplify notation we take $m = 1$. The function F is called the **force** (acting on a particle), and in the general case it depends on the particle **position**, **velocity** and **time**, We only consider the cases when F is the sum of these three terms. In the order:

1) **Potential force** which keeps the particle in the potential well. This means that this term can be written in the form $-\frac{\partial U}{\partial x}$, where the function $U(x,t) = \omega_0^2(t)\frac{x^2}{2}$ is called **potential energy**. It has a minimum (bottom of the **potential well**) at the point $x = 0$. Full particle energy (or **Hamiltonian**) $H = T + U$ is the sum of potential and **kinetic energy** $T = m\frac{v^2}{2}$ of particle, where $v = \frac{dx}{dt}$ is called the **velocity** of the particle. If ω_0^2 is time-independent and $\alpha = f = 0$, then the energy conservation law is proved in one line: $\frac{dH}{dt} = 0$. In this case, it is necessary to use the fact that the equation (5.9) can be written in the form of **hamiltonian system** of two first-order equations $\dot{x} = v, \dot{v} = -\omega_0^2 x$.

2) **Dissipative force**, which reduces the absolute value of the velocity, and hence the kinetic energy.

3) $f(t)$ is usually called the **driving force**.

5.2.2. Only one force

Here ω^2 and α are not time-dependent.

Only potential force. The equation here

$$\ddot{x} = -\omega_0^2 x.$$

Its solution in the general form with constants C_1, C_2 determined from the initial data

$$x(t) = C_1 \cos\omega_0 t + C_2 \sin\omega_0 t, \quad C_1 = x(0),\ C_2 = \frac{v(0)}{\omega_0}.$$

Law of energy conservation says that on the **phase plane** $R^2 = \{(x,v)\}$ the point performs a periodic motion along the ellipse

$$\frac{v^2(t)}{2} + \omega_0^2\frac{x^2(t)}{2} = H(0).$$

Only dissipation. The equation

$$\frac{d^2x}{dt^2} = -\alpha\frac{dx}{dt}$$

has a solution

$$v(t) = \frac{dx}{dt} = C_1\exp(-\alpha t), \quad C_1 = v(0), \Longrightarrow$$

$$\Longrightarrow x(t) = C_2 + C_1\int_0^t \exp(-\alpha t)dt, C_2(0) = x(0),$$

it follows that the particle does not escape to infinity, and its coordinate even tends to a constant value.

Only time dependent external force. Consider the equation

$$\frac{d^2x}{dt^2} = \frac{dv}{dt} = f(t).$$

When there are bounded trajectories ? We have

$$\frac{dx}{dt} = v(t) = v(0) + \int_0^t f(s)ds.$$

The speed is bounded if and only if the last integral is bounded. For example, if f is periodic with zero integral over a period. But for

$$x(t) = x(0) + \int_0^t v(s)ds = x(0) + v(0)t + \int_0^t du \int_0^u f(s)ds$$

the situation is quite different. Let for example $v(0) = 0, f = \sin(t + A)$. Then

$$v(t) = -\cos(t + A) + \cos A \Longrightarrow$$

$$\Longrightarrow x(t) = x(0) + \int_0^t v(s)ds = x(0) + t\cos A - \sin(t + A) + \sin A,$$

and the trajectory is bounded if and only if $\cos A = 0$.

5.2.3. Two forces

External force and dissipation. This can be reduced to equation of the first order

$$\frac{d^2x}{dt^2} = \frac{dv}{dt} = -\alpha v + f(t),$$

and, as we saw above, has the explicit solution

$$x(t) = x(0) + \frac{v(0)}{\alpha} - \frac{v(0)}{\alpha}e^{-\alpha t} - \alpha^{-1}\int_0^t e^{\alpha(s-t)}f(s)ds + \alpha^{-1}\int_0^t f(s)ds,$$

and boundedness will be for all f when the integrals are bounded.

Potential and dissipation.

$$\ddot{x} = -\omega_0^2 x - \alpha v, \ \ \alpha > 0.$$

Here two linearly independent solutions are obtained by substitution $x = e^{\lambda t}$ in the equation, which gives the quadratic equation for λ with two roots

$$\lambda = -\frac{\alpha}{2} \pm \sqrt{\frac{\alpha^2}{4} - \omega_0^2}.$$

For example, if $\alpha^2/4 > \omega_0^2$, then both root λ_1 and λ_2 are negative, and the solution has the form

$$C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t},$$

where C_1, C_2 are determined from initial conditions. This implies exponential convergence to zero for all ω and all initial conditions. The reader will easily consider himself the cases $\alpha^2/4 < \omega_0^2$ and $\alpha^2/4 = \omega_0^2$.

Potential and constant external force. In the case of constant force $f = const$ and dissipation, we have the equation

$$\ddot{x} = -\omega^2 x + f - \alpha v - \ddot{x} = -\omega_0^2 x - \alpha\Big(v - \frac{f}{\alpha}\Big). \tag{5.10}$$

If there is no dissipation, the solution

$$x = \frac{f}{\omega_0^2} + C_1 \cos\omega_0 t + C_2 \sin\omega_0 t$$

just shifts the point around which the oscillations occur. If $\alpha \neq 0$, then the substitution $x = \alpha/\omega_0^2 + y$ shows that there will be an exponential convergence of $x(t)$ to the point f/ω_0^2.

Resonance. For $f = a\cos\omega t, \omega \neq \omega_0$, the general solution has the form of the sum of a particular solution of the inhomogeneous equation and the general solution of the homogeneous

$$\frac{a}{\omega_0^2 - \omega^2}\cos\omega t + C_1 \cos\omega_0 t + C_2 \sin\omega_0 t. \tag{5.11}$$

Moreover, the constants C_1, C_2 are selected so that (5.11) satisfies the initial conditions. The closer the frequency ω of the driving force to the natural frequency ω_0, the greater the amplitude of the oscillation (**resonance** phenomenon). When $\omega = \omega_0 > 0$, the particular solution of the inhomogeneous equation has the form

$$\frac{a}{2\omega_0} t \sin\omega_0 t, \tag{5.12}$$

that is, the amplitude maximum increases linearly with time. Such resonant terms (usually called **secular**) like (5.12) often appear and cause a lot of trouble in much more complicated problems, for example, in the case of general external forces $f = f(x, t)$.

Thus, the trajectory will be unbounded only at one frequency $\omega = \omega_0$. And it may seem that the unboundedness of the trajectory due to external disturbance is a rare phenomenon. To understand this, it is necessary to understand the resonance mechanism less formally. It seems to be due to some synchronisation

mechanism ? But if we shift the phase of the resonant external force $f(t) = \cos(\omega t + \varphi)$, then the solution will again be resonant

$$\frac{a}{2\omega_0} t \sin(\omega_0 t + \varphi).$$

There is no complete understanding of resonance phenomenon for general systems.

Non-periodic force. Consider a more complicated example: instead of ω_0 we consider arbitrary smooth function $a(\omega)$, which is zero outside some neighborhood of ω_0

$$f(t) = \int a(\omega) \cos \omega t d\omega.$$

Assume that $a(\omega)$ is smooth and has a carrier on the interval $(\omega_0 - \varepsilon, \omega_0 + \varepsilon)$, where $0 < \varepsilon < \omega_0$ (assuming $\omega_0 > 0$) is not necessarily small.

Proposition 2. *For any initial data the solution is uniformly bounded on the time interval* $[0, \infty)$.

Proof. The solution is (denoting $\omega = \omega_0 + x$)

$$\lim_{\varepsilon \to 0} \int \frac{a(\omega) \cos \omega t}{\omega_0^2 - \omega^2} d\omega = \lim_{\varepsilon \to 0} \int \int_{-\varepsilon}^{\varepsilon} \frac{a(\omega_0 + x) \cos(\omega_0 + x)t}{\omega_0^2 - (\omega_0 + x)^2} dx =$$

$$= \lim_{\varepsilon \to 0} \int \int_{-\varepsilon}^{\varepsilon} \frac{a(\omega_0 + x)}{x} \frac{1}{-2\omega_0 - x} \cos(\omega_0 + x)t dx =$$

$$= \lim_{\varepsilon \to 0} \int \frac{a(\omega_0)}{-2\omega_0} \int_{-\varepsilon}^{\varepsilon} \frac{\cos(\omega_0 t + xt)}{x} dx =$$

$$= \lim_{\varepsilon \to 0} \int \frac{a(\omega_0)}{-2\omega_0} \int_{-\varepsilon}^{\varepsilon} \Big(\frac{\cos \omega_0 t \cos xt}{x} - \frac{\sin \omega_0 t \sin xt}{x} \Big) dx.$$

It is sufficient to consider only the case $a(\omega_0) \neq 0$. We see that unboundedness in time can only arise when integrating in a small neighborhood of ω_0 The first term in the neighborhood of the point $x = 0$ is zero because $(\pm)$-symmetry. At the same time

$$\int_{-\varepsilon}^{\varepsilon} \frac{\sin xt}{x} dx = \int_{-t\varepsilon}^{t\varepsilon} \frac{\sin x}{x} dx$$

is bounded uniformly in t. Indeed, on arbitrary period $(N, N + 2\pi)$ put $x = N + y$, then

$$\frac{1}{x} = \frac{1}{N} \frac{1}{1 + y/N} = \frac{1}{N} - \frac{x}{N^2} + \dots$$

The first term gives 0 in the integrals for such periods, and the rest will give a convergent sum.

Remark 10. There are many other examples of bounded solutions for other driving forces $f(t)$, including random, but it would be necessary to have a complete overview of all available results.

5.2.4. More general linear systems

Consider a system of the form

$$\frac{d\varphi}{dt} = A\varphi + f, \;\; \varphi(o) = \varphi_0 \tag{5.13}$$

where $\psi = \psi(t) = \{\psi_k(t) : k = 1, \dots, n\}$ is unknown column vector, $f = f(t) = \{f_k(t) : k = 1, \dots, n\}$ is a known column vector, and A is $(n \times n)$-matrix with constant coefficients. This system can easily be solved explicitly by replacing the unknown vector by the vector $C(t)$

$$\psi(t) = e^{At}C(t) \Longleftrightarrow C(t) = e^{-At}\psi(t), \;\; \psi(0) = C(0),$$

whence we get the equation for the new vector and its solution

$$\frac{dC}{dt} = e^{-At}f \Longrightarrow C(t) - C(0) = \int_0^t e^{-As}f(s)ds$$

from where the solution to the equation (5.13), with the initial condition $\psi(0) = C(0)$, is

$$\frac{dC}{dt} = e^{-At}f \Longrightarrow C(t) - C(0) = \int_0^t e^{-As}f(s)ds.$$

In case of a homogeneous equation, that is if $f = 0$, for example if possible, we bring matrix A to diagonal form

$$A = DBD^{-1},$$

where B is diagonal. Then for the vector $\xi(t) = C^{-1}\varphi(t)$ we obtain the equation

$$\frac{d\xi}{dt} = B\xi,$$

$$W = F_1\frac{dF_2}{dt} - \frac{dF_1}{dt}F_2,$$

and the problem reduces to one-dimensional equations.

In case when A depends on the time, a trick similar to this only takes place in the one-dimensional case, see equation (5.13) above. That is why, the theory, even for the simplest system of two equations, to which reduces, for example, the Hill equation

$$\ddot{x} = -k(t)x, \tag{5.14}$$

is much more difficult, see below. The main problem is to find linearly independent solutions of the homogeneous equation. However, if these solutions are known, then the corresponding inhomogeneous equation is not a problem. So, for the equation

$$\frac{d^2x}{dt^2} + g_1(t)\frac{dx}{dt} + g_0(t)x = f(t)$$

with arbitrary g_1, g_0, f the general solution is

$$c_1F_1(t)+c_2F_2(t)+F_2(t)\int_0^t F_1(s)f(s)W^{-1}(s)ds - F_1(t)\int_0^t F_2(s)f(s)W^{-1}(s)ds,$$

where F_1, F_2 are linearly independent solutions of the homogeneous equation and

$$W = F_1\frac{dF_2}{dt} - \frac{dF_1}{dt}F_2.$$

Note that even in the case $g_1 = 0, f = 0$, in the matrix case it is impossible to search for a solution as in (5.8).

For the equation

$$\frac{d^2x}{dt^2} = -\omega_0^2 x + f(t)$$

$F_1 = \cos\omega t, F_2 == \sin\omega t$, and $W(t) = \omega$. The the general solution has the form

$$x(t) = c_1\cos\omega_0 t + c_2\sin\omega_0 t + \sin\omega_0 t\int_0^t \cos\omega_0 s f(s)ds - \cos\omega_0 t\int_0^t \sin\omega_0 s f(s)ds.$$

From this, in many cases it is easy to find the conditions for boundedness.

Hill equation and the multiplicative integral. We discussed above two simple cases of Hill equation when $k(t) = k$ is a constant. If $k < 0$ then, as we have seen,the solution is bounded. If $k > 0$, then the solution is

$$x(t) = \frac{1}{2}\Big(x(0) + \frac{v(0)}{\sqrt{k}}\Big)e^{\sqrt{k}} + \frac{1}{2}\Big(x(0) - \frac{v(0)}{\sqrt{k}}\Big)e^{-\sqrt{k}}$$

and grows exponentially, except for the case when $x(0) = -\frac{v(0)}{\sqrt{k}}$. In case of irregular or random $k(t)$ we can expect quite different behavior of solutions. There is a general approach, see [6–8], based on the concept of the **multiplicative integral**, which is a direct generalization of the usual (additive) integral. If the latter is the limit of sums of many small terms, any of which tends to zero, the first is the limit of products, where each factor tends to 1. Possibly, some mathematicians did not hear about the science, but the ideas of it obviously by various examples were understood by many of them. For example, there is a natural approximation (belonging apparently to Euler) for solving linear vector

equations $\dot{x} = A(t)x$ on the interval $[0, 1]$ with initial condition $x(0)$. The segment is subdivided by points $t_k = k/N$ on N intervals of length $1/N$, and we consider linear recurrence equation:

$$x_{k+1} = x_k + \dot{A}(t_k)(t_{k+1} - t_k)x_k = (E + \dot{A}(t_k)(t_{k+1} - t_k))x_k,$$
$$k = 0, 1, \ldots, N-1.$$

We get then the system of piece wise constant solutions on $t_k \leq t < t_{k+1}$

$$x^{(N)}(t) = x_k^{(N)} = B(k(N))x_0, \;\; B(k(N)) = \prod_{k=0}^{k(N)/N} (E + \dot{A}(t_k)).$$

Moreover, under certain conditions it is possible to prove that there exist limiting matrices $B(t)$ such that $B(t)x_0 = x(t)$, and for $k(N)/N \to t, N \to \infty$, will be $B(k(N)) \to B(t)$.

However, it could bring us rather to the field of computational mathematics and computational methods in physics. Moreover, it is unclear whether all types of solution behavior can be covered by these methods.

References

[1] F. Hausdorff (1962) *Set Theory*. 2nd edition.

[2] E. Landau (1951) *Foundations of Analysis*. Chelsey Publ.

[3] N. Bourbaki (1982) *A Panorama of Pure Mathematics* (*as Seen by N. Bourbaki*).

[4] E.B. Vinberg (2003) *A Course in Algebra*.

[5] R. Brualdi and D. Cvetkovic (2009) *A Combinatorial Approach to Matrix Theory and Its Applications*. Cambridge.

[6] F.R. Gantmakher (2004) *Theory of Matrices* (Chapter 15). Moscow.

[7] J. Dollard and Ch. Friedman (1979) *Product Integration with Application to Differential Equations*. Cambridge.

[8] A. Slavik (2007) *Product Integration, Its History and Applications*. Prague.

[9] V.A. Malyshev (2018) Non-relativistic classical mechanics of point particles: shortest elementary introduction. *Structure of Mathematical Physics* (1), 121–157.

[10] V.A. Malyshev (2018) Classical Microscopic Electrodynamics: short introduction. *Structure of Mathematical Physics* (1), 159–184.

[11] (2018) Projects. *Structure of Mathematical Physics* (1), 3–32.