I. A GENERAL VIEW OF MATHEMATICS

mathematics which distinguishes it from the mathematics of preceding ages.*

Suggested Reading

Preliminary remark. The original Russian text of Mathematics: its content, methods, and meaning contains a list of recommended books at the end of each of its twenty chapters. In the present translation these books have been retained only if they have been translated into English. In compensation, the lists given here contain many other, readily available, works in the English language.

Books dealing with mathematics in general

G. H. Hardy, A mathematician's apology, Macmillan, New York, 1940.

Books of a historical character

R. C. Archibald, Outline of the history of mathematics, 5th ed., Mathematical Association of America, Oberlin, Ohio, 1941.

§1. Introduction

The rise at the end of the Middle Ages of new conditions of manufacture in Europe, namely the birth of capitalism, which at this time was replacing the feudal system, was accompanied by important geographical discoveries and explorations. In 1492, relying on the idea that the earth is spherical, Columbus discovered the New World. The discovery by Columbus greatly extended the boundaries of the known world and produced a revolution in the minds of men. The end of the 15th century and the beginning of the 16th saw the creative activity of the great artist-humanists Leonardo da Vinci, Raphael, and Michelangelo, which gave new meaning to art. In 1543 Copernicus published his work "On the revolution of the heavenly bodies," which completely changed the face of astronomy; in 1609 appeared the "New astronomy" of Kepler, containing his first and second laws for the motion of the planets around the sun, and in 1618 his book "Harmony of the world," containing the third law. Galileo, on the basis of his study of the works of Archimedes and his own bold experiments, laid the foundations for the new mechanics, an indispensable science for the newly arising technology. In 1609 Galileo directed his recently constructed telescope, though still small and imperfect, toward the night sky; the first glance in a telescope was enough to destroy the ideal celestial spheres of Aristotle and the dogma of the perfect form of celestial bodies. The surface of the moon was seen to be covered with mountains and pitted with craters. Venus displayed phases like the Moon, Jupiter was surrounded by four satellites and provided a miniature visual model of the solar system. The Milky Way fell apart into separate stars, and for the first time men felt the staggeringly immense distance...
II. ANALYSIS

of the stars. No other scientific discovery has ever made such an impression on the civilized world.*

The further development of navigation, and consequently of astronomy, and also the new development of technology and mechanics necessitated the study of many new mathematical problems. The novelty of these problems consisted chiefly in the fact that they required mathematical study of the laws of motion in a broad sense of the word.

The state of rest and motionlessness is unknown in nature. The whole of nature, from the smallest particles up to the most massive bodies, is in a state of eternal creation and annihilation, in a perpetual flux, in unceasing motion and charge. In the final analysis, every natural science studies some aspect of this motion. Mathematical analysis is that branch of mathematics that provides methods for the quantitative investigation of various processes of change, motion, and dependence of one magnitude on another. So it naturally arose in a period when the development of mechanics and astronomy, brought to life by questions of technology and navigation, had already produced a considerable accumulation of observations, measurements, and hypotheses and was leading science straight toward quantitative investigation of the simplest forms of motion.

The name "infinitesimal analysis" says nothing about the subject matter under discussion but emphasizes the method. We are dealing here with the special mathematical method of infinitesimals, or in its modern form, of limits. We now give some typical examples of arguments which make use of the method of limits and in one of the later sections we will define the necessary concepts.

**Example 1.** As was established experimentally by Galileo, the distance $s$ covered in the time $t$ by a body falling freely in a vacuum is expressed by the formula

$$ s = \frac{gt^2}{2} $$

(1)

($g$ is a constant equal to 9.81 m/sec$^2$).† What is the velocity of the falling body at each point in its path?

Let the body be passing through the point $A$ at the time $t$ and consider what happens in the short interval of time of length $\Delta t$; that is, in the time from $t$ to $t + \Delta t$. The distance covered will be increased by a certain

increment $\Delta s$. The original distance is $s_1 = \frac{gt^2}{2}$; the increased distance is

$$ s_2 = \frac{g(t + \Delta t)^2}{2} = \frac{gt^2}{2} + \frac{g}{2}(2t\Delta t + \Delta t^2). $$

From this we find the increment

$$ \Delta s = s_2 - s_1 = \frac{g}{2}(2t\Delta t + \Delta t^2). $$

This represents the distance covered in the time from $t$ to $t + \Delta t$. To find the average velocity over the section of the path $\Delta s$, we divide $\Delta s$ by $\Delta t$:

$$ v_{av} = \frac{\Delta s}{\Delta t} = gt + \frac{g}{2} \Delta t. $$

Letting $\Delta t$ approach zero we obtain an average velocity which approaches as close as we like to the true velocity at the point $A$. On the other hand, we see that the second summand on the right-hand side of the equation becomes vanishingly small with decreasing $\Delta t$, so that the average $v_{av}$ approaches the value $gt$, a fact which it is convenient to write as follows:

$$ v = \lim_{\Delta t \to 0} v_{av} = \lim_{\Delta t \to 0} \frac{\Delta s}{\Delta t} = \lim_{\Delta t \to 0} \left( gt + \frac{g}{2} \Delta t \right) = gt. $$

Consequently, $gt$ is the true velocity at the time $t$.

**Example 2.** A reservoir with a square base of side $a$ and vertical walls of height $h$ is full to the top with water (figure 1). With what force is the water acting on one of the walls of the reservoir?

We divide the surface of the wall into $n$ horizontal strips of height $h/n$. The pressure exerted at each point of the vessel is equal, by a well-known law, to the weight of the column of water lying above it. So at the lower edge of each of the strips the pressure, expressed in suitable units, will be equal respectively to

$$ \frac{h}{n}, \frac{2h}{n}, \frac{3h}{n}, \ldots, \frac{(n-1)h}{n}, h. $$
We obtain an approximate expression for the desired force \( P \), if we assume that the pressure is constant over each strip. Thus the approximate value of \( P \) is equal to

\[
P \approx \frac{ah}{n} \cdot \frac{h}{n} + \frac{ah}{n} \cdot \frac{2h}{n} + \cdots + \frac{ah}{n} \cdot \frac{n-1}{n} + \frac{ah}{n} \cdot \frac{h}{n}
\]

\[
= \frac{ah^2}{n^2} \cdot (1 + 2 + \cdots + n) = \frac{ah^2}{n^2} \cdot \frac{n(n + 1)}{2} = \frac{ah^2}{2} \left(1 + \frac{1}{n}\right).
\]

To find the true value of the force, we divide the side into narrower and narrower strips, increasing \( n \) without limit. With increasing \( n \) the magnitude \( 1/n \) in the above formula will become smaller and smaller and in the limit we obtain the exact formula

\[
P = \frac{ah^2}{2}.
\]

The idea of the method of limits is simple and amounts to the following. In order to determine the exact value of a certain magnitude, we first determine not the magnitude itself but some approximation to it. However, we make not one approximation but a whole series of them, each more accurate than the last. Then from examination of this chain of approximations, that is from examination of the process of approximation itself, we uniquely determine the exact value of the magnitude. By this method, which is in essence a profoundly dialectical one, we obtain a fixed constant as the result of a process or motion.

The mathematical method of limits was worked as the result of the persistent labor of many generations on problems that could not be solved by the simple methods of arithmetic, algebra, and elementary geometry.

What were the problems whose solution led to the fundamental concepts of analysis, and what were the methods of solution that were set up for these problems? Let us examine some of them.

The mathematicians of the 17th century gradually discovered that a large number of problems arising from various kinds of motion with consequent dependence of certain variables on others, and also from geometric problems which had not yielded to former methods, could be reduced to two types. Simple examples of problems of the first type are: find the velocity at any time of a given nonuniform motion (or more generally, find the rate of change of a given magnitude), and draw a tangent to a given curve. These problems (our first example is one of them) led to a branch of analysis that received the name “differential calculus.” The simplest examples of the second type of problem are:

find the area of a curvilinear figure (the problem of quadrature), or the distance traversed in a nonuniform motion, or more generally the total effect of the action of a continuously changing magnitude (compare the second of our two examples). This group of problems led to another branch of analysis, the “integral calculus.” Thus two fundamental problems were singled out: the problem of tangents and the problem of quadratures.

In this chapter we will describe in detail the underlying ideas of the solution of these two problems. Particularly important here is the theorem of Newton and Leibnitz to the effect that the problem of quadratures is the inverse, in a well-known sense, of the problem of tangents. For solving the problem of tangents, and problems that can be reduced to it, there was worked out a suitable algorithm, a completely general method leading directly to the solution, namely the method of derivatives or of differentiation.

The history of the creation and development of analysis and of the role played in its growth by the analytic geometry of Descartes has already been described in Chapter I. We see that in the second half of the 17th century and the first half of the 18th a complete change took place in the whole of mathematics. To the divisions that already existed, arithmetic, elementary geometry, and the rudiments of algebra and trigonometry, were added such general methods as analytic geometry, differential and integral calculus, and the theory of the simplest differential equations. It was now possible to solve problems whose solutions up to now had been quite inaccessible.

It turned out that if the law for the formation of a given curve is not too complicated, then it is always possible to construct a tangent to it at an arbitrary point; it is only necessary to calculate, with the help of the rules of differential calculus, the so-called derivative, which in most cases requires a very short time. Up till then it had been possible to draw tangents only to the circle and to one or two other curves, and no one had suspected the existence of a general solution of the problem.

If we know the distance traversed by a moving point up to any desired instant of time, then by the same method we can at once find the velocity of the point at a given moment, and also its acceleration. Conversely, from the acceleration it is possible to find the velocity and the distance, by making use of the inverse of differentiation, namely integration. As a result, it was not very difficult, for example, to prove from the Newtonian laws of motion and the law of universal gravitation that the planets must move around the sun in ellipses according to the laws of Kepler.

Of the greatest importance in practical life is the problem of the greatest and least values of a magnitude, the so-called problem of maxima and
II. ANALYSIS

minima. Let us take an example: From a log of wood with circular cross section of given radius we wish to cut a beam of rectangular cross section such that it will offer the greatest resistance to bending. What should be the ratio of the sides? A short argument on the stiffness of beams of rectangular cross section (applying simple concepts from the integral calculus), followed by the solving of a maximum problem (which involves calculating a derivative) provides the answer that the greatest stiffness is produced for a rectangular cross section whose height is in the ratio to its base of $\sqrt{2} : 1$. The problems of maxima and minima are solved as simply as those of drawing tangents.

At various points of a curved line, if it is not a straight line or a circle, the curvature is in general different. How can we calculate the radius of a circle with the same curvature as the given line at the given point, the so-called radius of curvature of the curve at the point? It turns out that this is equally simple; it is only necessary to apply the operation of differentiation twice. The radius of curvature plays a great role in many questions of mechanics.

Before the invention of the new methods of calculation, it had been possible to find the area only of polygons, of the circle, of a sector or a segment of the circle, and of two or three other figures. In addition, Archimedes had already invented a way to calculate the area of a segment of a parabola. The extremely ingenious method which he used in this problem was based on special properties of the parabola and consequently gave rise to the idea that every new problem in the calculation of area would very likely require its own methods of investigation, even more ingenious and difficult than those of Archimedes. So mathematicians were greatly pleased when it turned out that the theorem of Newton and Leibnitz, to the effect that the inversion of the problem of tangents would solve the problem of quadrature, at one provided a method of calculating the areas bounded by curves of widely different kinds. It became clear that a general method exists, which is suitable for an infinite number of the most different figures. The same remark is true for the calculation of volumes, surfaces, the lengths of curves, the mass of inhomogeneous bodies, and so forth.

The new method accomplished even more in mechanics. It seemed that there was no problem in mechanics that the new calculations would not clarify and solve.

Not long before, Pascal had explained the increase in the size of the Torricelli vacuum with increasing altitude as a consequence of the decrease in atmospheric pressure. But exactly what is the law governing this decrease? The question is answered immediately by the investigation of a simple differential equation.

§1. INTRODUCTION

It is well known to sailors that they should take two or three turns of the mooring cable around the capstan if one man is to be able to keep a large vessel at its mooring. Why is this? It turned out that from a mathematical point of view the problem is almost completely identical with the preceding one and can be solved at once.

Thus, after the creation of analysis, there followed a period of tempestuous development of its applications to the most varied branches of technology and natural science. Since it is founded on abstraction from the special features of particular problems, mathematical analysis reflects the actual deep-lying properties of the material world, and this is the reason why it provides the means for investigation of such a wide range of practical questions. The mechanical motion of solid bodies, the motion of liquids and gases of their particular particles, their laws of flow in the mass, the conduction of heat and electricity, the course of chemical reactions, all these phenomena are studied in the corresponding sciences by means of mathematical analysis.

At the same time as its applications were being extended, the subject of analysis itself was being immeasurably enriched by the creation and development of various new branches, such as the theory of series, applications of geometry to analysis, and the theory of differential equations.

Among mathematicians of the 18th century, there was a widespread opinion that any problem of the natural sciences, provided only that one could find a correct mathematical description of it, could be solved by means of analytic geometry and the differential and integral calculus.

Mathematicians proceeded gradually to more complicated problems of natural science and technology, which demanded further development of their methods. For the solution of such problems it became necessary to create further branches of mathematics: the calculus of variations, the theory of functions of a complex variable, field theory, integral equations, and functional analysis. But all these new methods of calculation were essentially immediate extensions and generalizations of the remarkable methods discovered in the 17th century. The greatest mathematicians of the 18th century, David Bernoulli (1700–1782), Leonhard Euler (1707–1783) and Lagrange (1736–1813), who blazed new paths in science, constantly took as their starting point the fundamental problems of the exact sciences. This energetic development of analysis was continued into the 19th century by such famous mathematicians as Gauss (1777–1855), Cauchy (1789–1857), M. V. Ostrogradskii (1801–1861), P. L. Čebyšev (1821–1894), Riemann (1826–1866), Abel (1802–1829), Weierstrass (1815–1897), all of whom made truly remarkable contributions to the development of mathematical analysis.

The Russian mathematical genius, N. I. Lobačevskii, had an influence
on the development of certain questions of mathematical analysis, and we should also mention the leading mathematicians who were active at the turn of the 20th century: A. A. Markov (1856–1922), A. M. Lyapunov (1857–1918), H. Poincaré (1854–1912), F. Klein (1849–1925), D. Hilbert (1862–1943).

The second half of the 19th century witnessed a profound critical examination and clarification of the foundations of analysis. The various powerful methods that had accumulated were now put on a uniform systematic basis, corresponding to the advanced level of mathematical rigor. All these methods are the means by which, along with arithmetic, algebra, geometry and trigonometry, we give a mathematical interpretation to the world around us, describe the course of actual events, and solve the important practical problems connected with them.

Analytic geometry, differential and integral calculus, and the theory of differential equations are studied at all technical institutes, so that these branches of mathematics are known to millions of citizens; the elements of these sciences are also taught at many technical schools; there is also some question of their being introduced into the secondary schools.

In most recent times the general use of rapid calculating machines has introduced a new era in mathematics. These machines, in conjunction with the branches of mathematics just mentioned, open up strange new possibilities for mankind.

At the present time, analysis and the branches arising from it represent a widely diversified mathematical science, consisting of several broad independent disciplines closely connected with one another; each of these disciplines is being developed and perfected.

More than ever before, a significant role is being played in analysis by the requirements of daily life, by problems connected with the imposing development of technology. Of great importance are the aerodynamical problems of supersonic velocities, which are being solved with constant success. The most difficult problems of mathematical physics have now reached the stage where they can be solved in practical numerical form.

In contemporary physics such theories as quantum mechanics (which studies the problems peculiar to the microcosm of the atom) not only require the most advanced branches of contemporary mathematical analysis for solving their problems but could not even describe their fundamental concepts without the use of analysis.

The purpose of the present chapter is to give a popular presentation, suitable for a reader acquainted only with elementary mathematics, of the growth and the simplest applications of such basic concepts of analysis as function, limit, derivative, and integral. Since the various special branches of analysis will be dealt with in other chapters of the

§2. FUNCTION

book, the present chapter has a more elementary character and a reader who has already studied a usual first course in analysis may omit it without harm to his understanding of the rest of the book.

§2. Function

The concept of a function. The various objects or phenomena that we observe in nature are organically connected with one another; they are interdependent. The simplest relations of this sort have long been known to mankind and information about them has been accumulated and formulated as physical laws. These laws indicate that the various magnitudes characterizing a given phenomenon are so closely related to one another that some of them are completely determined by the values of others. For example, the length of the sides of a rectangle completely determine its area, the volume of a given amount of gas at a given temperature is determined by the pressure, and the elongation of a given metallic rod is determined by its temperature. It was uniformities of this sort that served as the origin of the concept of function.

Consider an algebraic formula which, corresponding to each value of the literal magnitudes occurring in it, allows us to find the value of the magnitude expressed by the formula; the basic idea here is that of a function. Let us consider some examples of functions expressed by such formulas.

1. Let us suppose that at the beginning of a certain period of time a material point was at rest and that subsequently it began to fall as the result of gravity. Then the distance s traced out by the point up to time t is expressed by the formula

$$s = \frac{gt^2}{2},$$

where g is the acceleration of gravity.

![Fig. 2.](image-url)
2. From a square of side \( a \) we construct an open rectangular box of height \( x \) (figure 2). The volume \( V \) of the box is calculated from the formula
\[
V = x(a - 2x)^2. \tag{2}
\]
Formula (2) allows us, for every height \( x \) under the obvious restriction \( 0 \leq x \leq a/2 \), to find the volume of the box.

3. Let a pillar (figure 3) be erected at the center of a circular skating rink with a light at height \( h \). The illumination \( T \) at the edge of the circle may be expressed by the formula
\[
T = \frac{A \sin \alpha}{h^2 + r^2}, \tag{3}
\]
where \( r \) is the radius of the circle, \( \tan \alpha = h/r \), and \( A \) is a certain magnitude characterizing the power of the light. If we know the height \( h \) we can calculate \( T \) from formula (3).

4. The roots of the quadratic equation
\[
x^2 + px - 1 = 0 \tag{4}
\]
are given by the formula
\[
x = -\frac{p}{2} \pm \sqrt{1 + \frac{p^2}{4}}. \tag{5}
\]

The characteristic feature of a formula in general, and of the examples just given in particular, is that the formula enables us, for any given value of one of the variables (the time \( t \), the height \( x \) of the box, the height \( h \) of the pillar, the coefficient \( p \) of the quadratic equation), which is called the independent variable, to calculate the value of the other variable (the distance \( s \), the volume \( V \), the illumination \( T \), the root \( x \) of the equation), which is called a dependent variable or a function of the first variable.

Each of the formulas introduced provides an example of a function: the distance \( s \) traced by the point is a function to the time \( t \); the volume \( V \) of the box is a function of height \( x \); the illumination \( T \) of the edge of the rink is a function of the height \( h \) of the pillar; the two roots of the quadratic equation (4) are functions of the coefficient \( p \).

It should be remarked that in some cases the independent variable may assume any desired numerical value, as in example 4 where the coefficient \( p \) of the quadratic equation (4) may be an arbitrary number. In others the independent variable may take an arbitrary value from some set (or collection) of numbers determined in advance; as in example 2, where the volume of the box is a function of its height \( x \), which can take any value from the set of numbers \( x \) satisfying the inequality \( 0 < x < a/2 \). Similarly, in example 3 the illumination \( T \) at the edge of the rink is a function of the height \( h \) of the pillar, which theoretically can take any value satisfying the inequality \( h > 0 \), but in practice \( h \) must satisfy the inequalities \( 0 < h < H \), where the magnitude \( H \) is determined by the technical facilities at the disposal of the administration of the rink.

Let us introduce other examples of this kind. The formula
\[
\nu = \sqrt{1 - x^2}
\]
determines a real function (expressing a relationship between the real numbers \( x \) and \( \nu \)) only for those values of \( x \) which satisfy the inequalities \(-1 \leq x \leq 1\), and the formula \( \nu = \log (1 - x^2) \) only for those \( x \) which satisfy the inequalities \(-1 < x < 1\).

So it is necessary to take account of the fact that actual functions may not be defined for all numerical values of the independent variable but only for those values which belong to a certain set, which most often fills out an interval on the \( x \)-axis, with or without the end points.

We are now in the position to give the definition of a function accepted in present-day mathematics.

The (dependent) magnitude \( \nu \) is a function of the (independent) magnitude \( x \) if there exists a rule whereby to each value of \( x \) belonging to a certain set of numbers there corresponds a definite value of \( \nu \).

The set of values \( x \) appearing in this definition is called the domain of the function.

Every new concept gives rise to a new symbolim. The transition from arithmetic to algebra was made possible by the construction of formulas which were valid for arbitrary numbers, and the search for general solutions gave rise to the literal symbolism of algebra.

The problem of analysis is the study of functions, that is of the dependence of one variable on another. Consequently, just as in algebra a transition took place from concrete numbers to arbitrary numbers, denoted by letters, so in analysis there was the corresponding transition from
concrete formulas to arbitrary formulas. The phrase "y is a function of x" is conventionally written as

\[ y = f(x) .\]

Just as in algebra different letters are used for different numbers, so in analysis different notations are used for different types of dependence, that is for different functions: thus we write \( y = f(x), \ y = \phi(x), \ldots .\)

**Graphs of functions.** One of the most fruitful and brilliant ideas of the second half of the 17th century was the idea of the connection between the concept of a function and the geometric representation of a line. This connection can be realized, for example, by means of a rectangular Cartesian system of coordinates, with which the reader is certainly familiar in a general way from his secondary school mathematics.

Let us set up on the plane a rectangular Cartesian system of coordinates. This means that on the plane we choose two mutually perpendicular lines (the axis of abscissas and the axis of ordinates), on each of which we fix a positive direction. Then to each point \( M \) of the plane we may assign two numbers \((x, y)\), which are its coordinates, expressing in the given system of measurement the distance, taken with the proper sign,* of the point \( M \) from the axis of ordinates and the axis of abscissas respectively.

With such a system of coordinates we may represent functions graphically in the form of certain lines. Suppose we are given a function

\[ y = f(x) .\]  

This means, as we know, that for every value of \( x \) belonging to the domain of definition of the given function, it is possible to determine by some means, for example by calculation, a corresponding value \( y \). Let us give to \( x \) all possible numerical values, for each \( x \) determine \( y \) according to our rule (6), and construct on the plane the point with coordinates \( x \) and \( y \). In this way, for every point \( M' \) on the \( x \)-axis (figure 4) there will correspond

![Fig. 4.](image)

![Fig. 5.](image)

* The number \( x \) is the abscissa and \( y \) is the ordinate of the point \( M \).

§2. **FUNCTION**

A point \( M \) with coordinates \( x \) and \( y = f(x) \). The set of all points \( M \) forms a certain line, which we call the graph of the function \( y = f(x) \).

Thus, the graph of the function \( f(x) \) is the geometric locus of the points whose coordinates satisfy equation (6).

In school we became acquainted with the graphs of the simplest functions. Thus the reader probably knows that the function \( y = kx + b \), where \( k \) and \( b \) are constants, is the graph (figure 5) of a straight line forming the angle \( \alpha \) with the positive direction of the \( x \)-axis, where \( \tan \alpha = k \), and intersecting the \( y \)-axis at the point \((0, b)\). This function is called a linear function.

Linear functions occur very frequently in the applications. Let us recall that many physical laws are represented, with considerable accuracy, by linear functions. For example, the length \( l \) of a body may be considered with good approximation as a linear function of its temperature

\[ i = l_0 + \alpha l \, t, \]

where \( \alpha \) is the coefficient of linear expansion, and \( l_0 \) is the length of the body for \( t = 0 \). If \( x \) is the time and \( y \) is the distance covered by a moving point, then the linear function \( y = kx + b \) obviously expresses the fact that the point is moving with uniform velocity \( k \); and the number \( b \) denotes the distance, at time \( x_0 = 0 \), of the moving point from the fixed zero-point from which we measure our distances. Linear functions are extremely useful because of their simplicity and because it is possible to consider nonuniform changes as being approximately linear, even if only for small intervals.

But in many cases it is necessary to make use of nonlinear functional dependence. Let us recall for example the law of Boyle-Mariotte

\[ v = \frac{c}{p}, \]

where the magnitudes \( p \) and \( v \) are inversely proportional. The graph of such a relation represents a hyperbola (figure 6).

The physical law of Boyle-Mariotte corresponds actually to the case that \( p \) and \( v \) are positive; it represents a branch of the hyperbola lying in the first quadrant.

The general class of oscillatory processes includes periodic motions, which are usually described by the familiar trigonometric functions. For example, if we extend a hanging spring from its position of equilibrium, then, so long as we stay within the elastic limits of the spring, the point \( A \) will perform vertical oscillations which are quite accurately expressed by the law

\[ x = a \cos(pt + \beta), \]
where \( x \) is the displacement of the point \( A \) from its position of equilibrium, \( t \) is the time, and the numbers \( a, p \) and \( \alpha \) are certain constants determined by the material, the dimensions, and the initial extension of the spring.

It should be kept in mind that a function may be defined in various domains by various formulas, determined by the circumstances of the case. For example, the relation \( Q = f(t) \) between the temperature \( t \) of a gram of water (or ice) and the quantity of heat \( Q \) in it, as \( t \) varies between \(-10^\circ\) and \(+10^\circ\), is a completely determined function which it is difficult to express in a single formula, but it is easy to represent this function by two formulas. Since the specific heat of ice is equal to 0.5 and that of water is equal to 1, this function, if we agree that \( Q = 0 \) at \(-10^\circ\), is represented by the formula

\[
Q = 0.5t + 5,
\]

as \( t \) varies in the interval \(-10^\circ \leq t < 0^\circ\) and by another formula

\[
Q = t + 85,
\]

as \( t \) varies in the interval \( 0^\circ < t \leq 10^\circ\). For \( t = 0 \) this function is indefinite or multiple-valued; for convenience, we may agree that at \( t = 0 \) it takes some well-defined value, for example \( f(0) = 45 \). The graph of the function \( Q = f(t) \) is given in figure 7.

---

We have introduced many examples of functions given by formulas. The possibility of representing a function by means of formulas is extremely important from the mathematical point of view, since such formulas provide very favorable conditions for investigating the properties of the functions by mathematical methods.

But one must not think that a formula is the only method of defining a function. There are many other methods; for example, the graph of the function, which gives a visual geometric picture of it. The following example gives a good illustration of another method.

To record variation of the temperature of the air during the course of 24 hours, meteorological stations make use of an instrument called the thermograph. A thermograph consists of a drum rotated about its axis by a clockwork mechanism, and of a curved brass framework that is extremely sensitive to changes of temperature. As a result, a pen fastened to the framework by a system of levers rises with rising temperature; and conversely, a fall in the temperature lowers the pen. On the drum is wound a ribbon of graph paper, on which the pen draws a continuous line, forming the graph of the function \( T = f(t) \), which expresses the interdependence of the time and the temperature of the air. From this graph we may determine, without calculation, the value of the temperature at any moment of time \( t \).

This example shows that a graph in itself determines a function independently of whether the function is given by a formula or not.

Incidentally, we shall return to this question (see Chapter XII) and shall prove the following important assertion: Every continuous graph can be represented by a formula, or, as it is still customary to say, by an analytic expression. This statement is also true for many discontinuous graphs.

We remark that the truth of this statement, which is of great theoretical importance, was completely realized in mathematics only in the middle of the past century. Up to that time mathematicians understood by the term "function" only an analytic expression (formula). But they were under the mistaken impression that many discontinuous graphs did not correspond to any analytic expression, since they assumed that if a function was given by a formula, then its graph must possess certain particularly desirable properties in comparison with the other graphs.

But in the 19th century, it was discovered that every continuous graph may be represented by a more or less complicated formula. Thus the exceptional role of the analytic expression as a means of definition of

---

* Of course, the above statement will be completely clear to the reader only after we have given a precise definition of exactly what is meant in mathematics by the term "formula" and "analytic expression."
functions was weakened and there came into existence the new, more flexible definition given above for the concept of a function. By this definition a variable \( y \) is called a function of a variable \( x \) if there exists a rule whereby to every value of \( x \) in the domain of definition of the function there corresponds a completely determined value \( y \), independent of the way in which this rule is given: by a formula, a graph, a table or in any other way.

We may remark here that in the mathematical literature the above definition of a function is often associated with the name of Dirichlet, but it is worth emphasizing that this definition was given simultaneously and independently by N. I. Lobachevskii. Finally we suggest as an exercise that the reader sketch the graphs of the functions \( x^2, \sqrt{x}, \sin x, \sin 2x, \sin (x + \pi/4), \ln x, \ln(1 + x), |x - 3|, (x + |x|)/2 \).

We should also note that the graph of a function which for all values of \( x \) satisfies the relation

\[
f(-x) = f(x)
\]

is symmetric with respect to the \( y \)-axis and in the case

\[
f(-x) = -f(x)
\]

the graph is symmetric with respect to the origin of coordinates. Consider also how to obtain the graph of a function \( f(a + x) \), when \( a \) is a constant, from the graph of \( f(x) \). Finally, consider how, using the graphs of the functions \( f(x) \) and \( g(x) \), it is possible to find the values of the composite function \( y = f[\phi(x)] \).

§3. Limits

In §1 it was stated that modern mathematical analysis uses a special method, which was worked out in the course of many centuries and serves now as its basic instrument. We are speaking here of the method of infinitesimals, or, as is essentially the same, of limits. We shall try to give some idea of these concepts. For this purpose we consider the following example.

We wish to calculate the area bounded by the parabola with equation \( y = x^2 \), by the \( x \)-axis and by the straight line \( x = 1 \) (figure 8). Elementary mathematics will not furnish us with a means for solving this problem. But here is how we may proceed.

We divide the interval \([0, 1]\) along the \( x \)-axis into \( n \) equal parts at the points

\[
0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1
\]

and on each of these parts construct the rectangle whose left side extends up to the parabola. As a result we obtain the system of rectangles shaded in figure 8, the sum \( S_n \) of whose areas is given by

\[
S_n = 0 \cdot \frac{1}{n} + 1 \cdot \frac{1}{n} + 2 \cdot \frac{1}{n} + \cdots + \frac{n-1}{n} \cdot \frac{1}{n}
\]

\[
= \frac{1^2 + 2^2 + \cdots + (n-1)^2}{n^2} = \frac{(n-1)(2n-1)}{6n^2}.
\]

Let us express \( S_n \) in the following form:

\[
S_n = \frac{1}{3} + \frac{1}{6n^2} - \frac{1}{2n} = \frac{1}{3} + a_n.
\]

The quantity \( a_n \), which depends on \( n \), is admittedly rather unwieldy in appearance, but it possesses a certain remarkable property: If \( n \) is increased beyond all bounds, then \( a_n \) approaches 0. This property may also be expressed as follows: If we are given an arbitrary positive number \( \epsilon \), then it is possible to choose an integer \( N \) sufficiently large that for all \( n \) greater than \( N \) number \( a_n \) will be less than the given \( \epsilon \) in absolute value.\footnote{If in the obvious equalities \((k+1)^2 - k^2 = 3k^2 + 3k + 1\), for the different values \( k = 1, 2, \ldots, n - 1 \), we add the left and right sides separately, we obtain the equation \( n^2 - 1 = 3n + \frac{3(n-1)n}{2} + n - 1 \) where \( a_n = 1^2 + 2^2 + \cdots + (n-1)^2 \). Solving this equation for \( a_n \), we get

\[
a_n = \frac{(n-1)n(2n-1)}{6}.
\]

\footnote{For example, if \( \epsilon = 0.001 \), we may take \( N = 500 \). In fact, since

\[
\frac{1}{6n^2} < \frac{1}{2n}
\]

for positive integral \( n \), therefore

\[
\|x\| = \left| 1 - \frac{1}{6n^2} - \frac{1}{2n} \right| = \frac{1}{2n} \quad 0 < \frac{1}{2n} < 0.001
\]}

\[
\frac{1}{6n^2} \quad < \quad \frac{1}{2n} \quad < \quad 0.001
\]
II. ANALYSIS

The magnitude \( \alpha_n \) is an example of an infinitesimal in the sense in which that word is used in modern mathematics.

In Figure 8 we see that if we increase the number \( n \) beyond all bounds, the sum \( S_n \) of the areas of the shaded rectangles will approach the desired area of the curvilinear figure. On the other hand, equation (7), in view of the fact that \( \alpha_n \) approaches zero as \( n \) increases beyond all bounds, shows that the sum \( S_n \) at the same time approaches \( 1/3 \). From this it follows that the desired area \( S \) of the figure is equal to \( 1/3 \), and we have solved our problem.

So the method under discussion amounts to this, that in order to find a certain magnitude \( S \) we introduce another magnitude \( S_n \), a variable magnitude which approaches zero through particular values \( S_1, S_2, S_3, \ldots \), which depend according to some law on the natural numbers \( n = 1, 2, \ldots \). Then, from the fact that the variable \( S_n \) may be represented as the sum of a constant \( \frac{1}{3} \) and an infinitesimal \( \alpha_n \), we conclude that \( S_n \) approaches \( \frac{1}{3} \) and so \( S = \frac{1}{3} \). In the language of the modern theory of limits we may say that for increasing \( n \) the variable magnitude \( S_n \) approaches a limit, which is equal to \( \frac{1}{3} \).

Now let us give a precise definition of the concepts introduced here. If a variable magnitude \( \alpha_n (n = 1, 2, \ldots) \) has the property that for every arbitrarily small positive number \( \epsilon \) it is possible to choose an integer \( N \) so large that for all \( n > N \) we have \( |\alpha_n| < \epsilon \), then we say that \( \alpha_n \) is an infinitesimal and we write

\[ \lim_{n \to \infty} \alpha_n = 0 \text{ or } \alpha_n \to 0. \]

On the other hand, if a variable \( x_n \) may be represented as a sum

\[ x_n = a + \alpha_n, \]

where \( a \) is constant and \( \alpha_n \) is an infinitesimal, then we say that the variable \( x_n \), for \( n \) increasing beyond all bounds, approaches the number \( a \) and we write

\[ \lim x_n = a \text{ or } x_n \to a. \]

The number \( a \) is called the limit of \( x_n \). In particular the limit of an infinitesimal is obviously zero.

---

§3. LIMITS

Let us consider the following examples of variable magnitudes

\[ x_n = \frac{1}{n}, \quad y_n = -\frac{1}{n^2}, \quad z_n = \frac{(-1)^n}{n}, \quad u_n = \frac{n - 1}{n}, \quad v_n = (-1)^n (n = 1, 2, \ldots). \]

It is clear that \( x_n, y_n, \) and \( z_n \) are infinitesimals, the first of them approaching zero through decreasing values, the second through increasing negative values, while the third takes on values which oscillate around zero. Further, \( u_n \to 1 \), while \( v_n \) does not have a limit at all, since with increasing \( n \) it does not approach any constant number but continually oscillates, taking on the values 1 and \(-1\).

Another important concept in analysis is that of an infinitely large magnitude, which is defined as a variable \( x_n (n = 1, 2, \ldots) \), with the property that after choice of an arbitrarily large positive number \( M \) it is possible to find a number \( N \) such that for all \( n > N \)

\[ |x_n| > M. \]

The fact that the magnitude \( x_n \) is infinitely large is written thus

\[ \lim x_n = \infty \text{ or } x_n \to \infty. \]

Such a magnitude \( x_n \) is said to approach infinity. If it is positive (negative) from some value on, this fact is expressed thus: \( x_n \to +\infty \) (\( x_n \to -\infty \)). For example, for \( n = 1, 2, \ldots \)

\[ \lim n^2 = +\infty, \quad \lim (-n^n) = -\infty; \]

\[ \lim \log \frac{1}{n} = -\infty, \quad \lim \tan \left( \frac{\pi}{2} + \frac{1}{n} \right) = -\infty. \]

It is easy to see that if a magnitude \( \alpha_n \) is infinitely large, then \( \beta_n = 1/\alpha_n \) is infinitely small, and conversely.

Two variable magnitudes \( x_n \) and \( y_n \) may be added, subtracted, multiplied, and divided the one by the other so as to produce new magnitudes that are in general also variable: namely their sum \( x_n + y_n \), their difference \( x_n - y_n \), their product \( x_n y_n \), and their quotient \( x_n / y_n \). Correspondingly their particular values will be

\[ x_1 \pm y_1, x_2 \pm y_2, x_3 \pm y_3, \ldots \]

\[ X_1 Y_1, X_2 Y_2, X_3 Y_3, \ldots \]

\[ X_1, x_2, X_3, \frac{X_1}{Y_1}, \frac{x_2}{Y_2}, \frac{X_3}{Y_3}, \ldots. \]

---

for arbitrary \( n > 500 \). In the same way it would be possible to assign arbitrarily small values \( \epsilon \), or example:

\[ \varepsilon_1 = 0.0001, \quad \varepsilon_2 = 0.00001, \quad \ldots, \]

and for each of them to choose, as above, appropriate values \( N = N_1, N_2, \ldots \).
It is also possible to prove, as is fairly evident, that if the variables \( x_n \) and \( y_n \) approach finite limits, then their sum, difference, product, and quotient also approach limits which are correspondingly equal to the sum, difference, product, and quotient of these limits. This fact may be expressed thus:

\[
\lim (x_n \pm y_n) = \lim x_n \pm \lim y_n; \quad \lim (x_n y_n) = \lim x_n \lim y_n;
\]

\[
\lim \frac{x_n}{y_n} = \frac{\lim x_n}{\lim y_n}.
\]

However, in the case of the quotient it is necessary to assume that the limit of the denominator (\( \lim y_n \)) is not equal to zero. If \( \lim y_n = 0 \) and \( \lim x_n \neq 0 \), then the ratio of \( x_n \) to \( y_n \) will not have a finite limit but will approach infinity.

Especially interesting, and at the same time important, is the case when the numerator and the denominator simultaneously approach zero. Here it is impossible to state in advance whether the ratio \( x_n/y_n \) will approach a limit, and if it does, what that limit will be, since the answer to this question depends entirely on the character of the approach of \( x_n \) and \( y_n \) to zero. For example, if

\[
x_n = \frac{1}{n}, \quad y_n = \frac{1}{n^2}, \quad z_n = \frac{(-1)^n}{n} \quad (n = 1, 2, \ldots),
\]

then

\[
\frac{y_n}{x_n} = n \to 0, \quad \frac{x_n}{y_n} = n \to \infty.
\]

On the other hand, the magnitude

\[
\frac{x_n}{z_n} = (-1)^n
\]

evidently does not approach any limit.

Thus the case when the numerator and the denominator of the fraction both approach zero cannot be dealt with in advance by general theorems, and for each particular fraction of this kind it is necessary to make a special investigation.

We shall see later that the fundamental problem of the differential calculus, which may be considered as the problem of determining the velocity of a nonuniform motion at a given moment, reduces to determining the limit of the ratio of two infinitesimal magnitudes, namely the increase of the distance covered and the increase in the time.

So far we have considered variables \( x_n \) which take on a sequence of numerical values \( x_1, x_2, x_3, \ldots, x_n, \ldots \), while the index \( n \) runs through the sequence of natural numbers \( n = 1, 2, 3, \ldots \). But it is also possible to consider the case that \( n \) varies continuously, like the time for example, and here also to determine the limit of the variable \( x_n \). The properties of such limits are completely analogous to those formulated earlier for discrete (that is, discontinuous) variables. We also note that there is no special significance in the fact that \( n \) increases beyond all bounds. It is equally possible to consider the case that, while varying continuously, \( n \) approaches a given value \( n_0 \).

As an example let us investigate the variation in the magnitude of \( (\sin x)/x \) as \( x \) approaches zero. Table 1 shows the values of this magnitude for certain values of \( x \):

<table>
<thead>
<tr>
<th>( x )</th>
<th>( \sin x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>0.9589 ...</td>
</tr>
<tr>
<td>0.10</td>
<td>0.9983 ...</td>
</tr>
<tr>
<td>0.05</td>
<td>0.9996 ...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

(it is assumed that the values of \( x \) are given in radian measure).

It is obvious that as \( x \) approaches zero the magnitude \( (\sin x)/x \) approaches 1, but of course we must still give a rigorous proof of this fact. The proof may be obtained, for example, from the following inequality, which is valid for all nonzero angles in the first quadrant:

\[
\sin x < x < \tan x.
\]

If we divide both sides of this inequality by \( \sin x \), we obtain

\[
1 < \frac{x}{\sin x} < \frac{1}{\cos x},
\]

from which follows

\[
\cos x < \frac{\sin x}{x} < 1.
\]

But as \( x \) decreases to zero \( \cos x \) approaches 1, so that the magnitude \( (\sin x)/x \), being contained in the interval between \( \cos x \) and 1, also approaches 1, that is

\[
\lim_{x \to 0} \frac{\sin x}{x} = 1.
\]
II. ANALYSIS

We shall have occasion below to make use of this fact.

Our equation has been proved for the case that \( x \) approaches zero through positive values. But by changing the proof in an obvious way, it is possible to obtain the same result when \( x \) approaches zero through negative values.

Let us now discuss for a moment the following question. A variable magnitude may or may not have a limit and the question arises whether it is possible to give a criterion for determining the existence of a limit for a variable. We will confine ourselves to an important and sufficiently general case, for which such a criterion can be given. Let us suppose that the variable magnitude \( x_1 \) increases or at least does not decrease; that is, it satisfies the inequalities:

\[
x_1 \leq x_2 \leq x_3 \leq \cdots,
\]

and let us also suppose we have determined that none of its values exceeds a certain fixed number \( M \); that is, \( x_n \leq M \) \((n = 1, 2, \ldots)\). If we mark the values of \( x_n \) and the number \( M \) on the \( x \)-axis, we can see that the variable point \( x_n \) moves along the axis to the right but constantly remains to the left of the point \( M \). It is rather obvious that the variable point \( x_n \) must inevitably approach a certain limit point \( a \), situated to the left of \( M \) or at most coinciding with \( M \).

So, in the case under consideration, the limit

\[
\lim_{n \to \infty} x_n = a
\]

of our variable exists.

The above argument has an intuitive character but we may consider it as a proof. In a course in modern analysis a complete proof of this fact is given on the basis of the theory of real numbers.

As an example let us consider the variable

\[
u_n = \left(1 + \frac{1}{n}\right)^n (n = 1, 2, 3, \ldots),\]

The first few values are \( u_1 = 2, u_2 = 2.25, u_3 \approx 2.37, u_4 \approx 2.44, \ldots \), which are seen to increase. From the binomial theorem of Newton it is possible to prove that this increase holds for arbitrary \( n \). Moreover, it is also easy to prove that for all \( n \) the inequality \( u_n < 3 \) is valid. Consequently, our variable must have a limit which is not greater than 3. We shall see that this limit plays a very important role in mathematical physics and in a certain sense is the most natural base for logarithms of numbers.

§3. LIMITS

It is customary to denote this limit by the letter \( e \). It is equal to

\[
e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = 2.718281828459045 \ldots
\]

A more detailed analysis shows that the number \( e \) is not rational.*

It is also possible to show that the limit under consideration exists and is equal to \( e \) not only when \( n \to \infty \) but also when \( n \to -\infty \). In both cases \( n \) may also take on noninteger values.

Let us mention an important application to physics of the concept of a limit. It consists of the remarkable fact that only by using the concept of a limit (passage to the limit) it is possible for us to give a complete definition of many of the concrete magnitudes encountered in physics.

Let us also consider for the moment the following geometric example. In elementary geometry the figures considered first are those bounded by straight line segments. But later there arises the more difficult task of finding the length of the circumference of a circle with given radius.

If we analyze the difficulties connected with the solution of this problem, we find that they reduce to the following.

We must give an answer to the question, what is meant by the length of the circumference; that is, we must give a precise definition of this length. It is essential that the definition should be expressible in terms of the lengths of straight-line segments and also that it should provide us with the possibility of effectively calculating the length of the circumference.

It is understood, of course, that the result of this calculation should be in agreement with practical experience. For example, if we consider a circumference consisting of an actual thread, then, if we cut the thread and stretch it out, we must obtain a segment whose length, within the limits of accuracy of measurement, coincides with our computed length.

As is known from elementary geometry, the solution of this problem reduces to the following definition. The length of a circumference is defined to be the limit approached by the perimeter of a regular\(^1\) polygon inscribed in it as the number of sides of the polygon increases beyond all bounds. Thus the solution of the problem is based essentially on the concept of a limit.

The length of an arbitrary smooth curve is defined in the same way.

* In this connection we should remark that addition, subtraction, multiplication, and division (excluding division by zero) of rational numbers, that is numbers of the form \( p/q \), where \( p \) and \( q \) are integers, ends to rational numbers. But this is not necessarily the case for the operation of taking a limit. The limit of a sequence of rational numbers may be irrational number.

\(^1\) It is not important that the polygon should be regular. The only essential feature is that the greatest side of the variable inscribed polygon should approach zero.
In the paragraphs just following, we will meet with a number of examples of geometric and physical magnitudes that can be defined only with the concept of a limit.

The concepts of limit and infinitesimal were given a definitive formulation at the beginning of the last century. The definitions introduced here are connected with the name of Cauchy, before whose time mathematicians operated with concepts that were less clear. The present-day concepts of a limit, of an infinitesimal as a variable magnitude, and of a real number, resulted from the development of mathematical analysis and were at the same time the means of stating and clarifying its many achievements.

§4. Continuous Functions

Continuous functions form the basic class of functions for the operations of mathematical analysis. The general idea of a continuous function may be obtained from the fact that its graph is continuous; that is, its curve may be drawn without lifting the pencil from the paper.

A continuous function gives the mathematical expression of a situation often encountered in practical life, namely that to a small increase in an independent variable there corresponds a small increase in the dependent variable, or function. Excellent examples of a continuous function are given by the various rules governing the motion of bodies \(s = f(t)\), expressing the dependence of the distance \(s\) on the time \(t\). Since the time and the distance are continuous, a law of motion of the body \(s = f(t)\) sets up between them a definite relation, characterized by the fact that to a small increase in the time corresponds a small increase in the distance.

Mankind arrived at the abstraction of continuity by observing the surrounding so-called dense media, namely solids, liquids, and gases; for example, metals, water, and air. In actual fact, as is well known now, every physical medium represents the accumulation of a large number of separate particles in motion. But these particles and the distances between them are so small in comparison with the dimensions of the media in which the phenomena of microscopic physics take place that many of these phenomena may be studied with sufficient accuracy if we consider the medium as being approximately without interstices, that is as continuously distributed over the occupied space. It is on such an assumption that many of the physical sciences are based, for example, hydrodynamics, aerodynamics, and the theory of elasticity. The mathematical concept of continuity naturally plays a large role in these sciences, and in many others as well.

§4. CONTINUOUS FUNCTIONS

Let us consider an arbitrary function \(y = f(x)\) and some specific value of the independent variable \(x_0\). If our function reflects a continuous process, then to values \(x\) which differ only slightly from \(x_0\) will correspond values of the function \(f(x)\) differing only slightly from the value \(f(x_0)\) at the point \(x_0\). Thus if the increment \(x - x_0\) of the independent variable is small, then the corresponding increment \(f(x) - f(x_0)\) of the function will also be small. In other words if the increment of the independent variable \(x - x_0\) approaches zero, then the increment \(f(x) - f(x_0)\) of the function must also approach zero, a fact which may be expressed in the following way:

\[
\lim_{x \to x_0} [f(x) - f(x_0)] = 0.
\] (8)

This relation constitutes the mathematical definition of continuity of the function at the point \(x_0\), namely, the function \(f(x)\) is said to be continuous at the point \(x_0\), if equality (8) holds.

Finally, we give the following definition. A function is said to be continuous in a given interval, if it is continuous at every point \(x_0\) of this interval; that is, if at every such point equality (8) is fulfilled.

Thus, in order to introduce a mathematical definition of the property of a function reflected in the fact that its graph is continuous (in the everyday sense of this word), it was necessary first to define local continuity (continuity at the point \(x_0\)) and then on this basis to define continuity of the function in the whole interval.

This definition, first introduced at the beginning of the last century by Cauchy, is now generally adopted in contemporary mathematical analysis. The test of many concrete examples has shown that it corresponds very well to the practical notion we have of a continuous function, for instance, as represented by its continuous graph.

As examples of continuous functions, the reader may consider the elementary functions well known to him from school mathematics \(x^n, \sin x, \cos x, a^x, \log x, \arcsin x, \arccos x\). All these functions are continuous in the intervals for which they are defined.

If continuous functions are added, subtracted, multiplied, or divided (except for division by zero), the result is also a continuous function. But in the case of division the continuity is usually destroyed for those values \(x_0\) for which the function in the denominator vanishes. The result of the division in that case is a function which is discontinuous at the point \(x_0\).

The function \(y = 1/x\) may serve as an example of a function which is discontinuous at the point \(x = 0\). Other discontinuous functions are represented by the graphs in figure 9.
§4. CONTINUOUS FUNCTIONS

We recommend that the reader examine these graphs carefully. He will notice that the breaks in the functions are different kinds: In some cases a limit \( f(x) \) exists as \( x \) approaches the point \( x_0 \) where the function suffers a discontinuity, but this limit is different from \( f(x_0) \). In other cases, as in figure 9c, the limit simply does not exist. It may also happen that as \( x \) approaches \( x_0 \) from one side \( f(x) - f(x_0) \to 0 \), but as \( x \to x_0 \) from the other side, \( f(x) - f(x_0) \) does not approach zero. In this case, of course, the function has a discontinuity, but we may say that at such a point it is "continuous from one side." All these cases are represented in the graphs of figure 9.

As an exercise we recommend to the reader to consider the question, what value must be given to the functions

\[
\frac{\sin x}{x}, \frac{1 - \cos x}{x}, \frac{x^2 - 1}{x^2}, \frac{\tan x}{x}, \frac{1}{x-1}, \frac{x}{(x^2 - 4)}
\]

at those points where they are not defined (that is, at the points where the denominator is equal to zero), in order that they may be continuous at these points. Also, is it possible to find such numbers for the functions

\[
\tan x, \frac{1}{x-1}, \frac{x-2}{(x^2 - 4)}
\]

These discontinuous functions in mathematics represent the numerous jumplike processes to be met with in nature. In the case of a sudden blow, for example, the value of the velocity of a body changes in such a jumplike fashion. Many qualitative transitions take place with such jumps. In §2 we introduced the function \( Q = f(t) \), expressing the way in which the quantity of heat in a given quantity of water (or ice) depends on the temperature. In the neighborhood of the melting point of ice the quantity of heat \( Q = f(t) \) changes in a jumplike fashion with changing \( t \).

Functions with isolated discontinuities are encountered quite often in analysis, along with the continuous functions. But as an example of a more complicated function, where the number of discontinuities is infinite, let us consider the so-called Riemann function, which is equal to zero at all irrational points and equal to \( 1/q \) at rational points of the form \( x = p/q \) (where \( p/q \) is a fraction in its lowest terms). This function is discontinuous at all rational points and continuous at irrational points. By altering it slightly we may easily obtain an example of a function which is discontinuous at all points.\(^*\) Let us remark by the way that even for such complicated functions modern analysis has discovered many in-

\(^*\) It is sufficient to set the function equal to unity at the irrational points.
II. ANALYSIS

Interesting laws, which are investigated in one of the independent branches of analysis, the theory of functions of a real variable. This theory has developed with extraordinary rapidity during the past 50 years.

§5. DERIVATIVE

The next fundamental concept of analysis is the concept of derivative. Let us consider two problems from which it arose historically.

Velocity. At the beginning of the present chapter we defined the velocity of a freely falling body. To do so we made use of a passage to the limit from the average velocity over short distances to the velocity at the given point and the given time. The same procedure may be used to define the instantaneous velocity for an arbitrary nonuniform motion. In fact, let the function

\[ s = f(t) \]

express the dependence of the distance \( s \) covered by the material point in the time \( t \). To find the velocity at the moment \( t = t_0 \), let us consider the interval of time from \( t_0 \) to \( t_0 + h \) (\( h \neq 0 \)). During this time the point will cover the distance

\[ \Delta s = f(t_0 + h) - f(t_0). \]

The average velocity \( v_{av} \) over this part of the path will depend on \( h \)

\[ v_{av} = \frac{\Delta s}{h} = \frac{1}{h} [f(t_0 + h) - f(t_0)], \]

and will represent the actual velocity at the point \( t_0 \) with greater and greater accuracy as \( h \) becomes smaller. It follows that the true velocity at the time \( t_0 \) is equal to the limit

\[ v = \lim_{h \to 0} \frac{f(t_0 + h) - f(t_0)}{h} \]

of the ratio of the increase in the distance to the increase in the time, as the latter approaches zero without ever being actually equal to zero. In order to calculate the velocity for different forms of motion, we must discover how to find this limit for various functions \( f(t) \).

Tangent. We are led to investigate a precisely analogous limit by another problem, this time a geometric one, namely the problem of drawing a tangent to an arbitrary plane curve.

Let the curve \( C \) be the graph of a function \( y = f(x) \), and let \( A \) be the point on the curve \( C \) with abscissa \( x_0 \) (figure 10). Which straight line shall we call the tangent to \( C \) at the point \( A \)? In elementary geometry this question does not arise. The only curve studied there, namely the circumference of a circle, allows us to define the tangent as a straight line which has only one point in common with the curve. But for other curves such a definition will clearly not correspond to our intuitive picture of "tangency." Thus, of the two straight lines \( L \) and \( M \) in figure 11, the first is obviously not tangent to the curve drawn there (a sinusoidal curve), although it has only one point in common with it; while the second straight line has many points in common with the curve, and yet it is tangent to the curve at each of these points.

To define the tangent, let us consider on the curve \( C \) (figure 10) another point \( A' \), distinct from \( A \), with abscissa \( x_0 + h \). Let us draw the secant \( AA' \) and denote the angle which it forms with the \( x \)-axis by \( \beta \). We now allow the point \( A' \) to approach \( A \) along the curve \( C \). If the secant \( AA' \) correspondingly approaches a limiting position, then the straight line \( T \) which has this limiting position is called the tangent at the point \( A \). Evidently the angle \( \alpha \) formed by the straight line \( T \) with the \( x \)-axis, must be equal to the limiting value of the variable angle \( \beta \).

The value of \( \tan \beta \) is easily determined from the triangle \( ABA' \) (figure 10):

\[ \tan \beta = \frac{BA'}{AB} = \frac{f(x_0 + h) - f(x_0)}{h}. \]

For the limiting position we must have

\[ \tan \alpha = \lim_{h \to 0} \tan \beta = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}. \]

![Fig. 10.](image1.png)

![Fig. 11.](image2.png)
II. ANALYSIS

that is, the trigonometric tangent of the angle of inclination of the tangent line is equal to the limit of the ratio of the increase in the function \( f(x) \) at the point \( x_0 \) to the corresponding increase in the independent variable, as the latter approaches zero without ever being actually equal to zero.

Let us give still another example leading to the calculation of an analogous limit. Let us suppose that a variable electric current is flowing through a conductor. Let us assume that we know the function \( Q = f(t) \) expressing the quantity of electricity that has passed through a fixed cross section of the conductor up to time \( t \). In the period from \( t_0 \) to \( t_0 + h \), there will flow through this cross section a quantity of electricity \( \Delta Q \) equal to \( f(t_0 + h) - f(t_0) \). The average value of the current will therefore be equal to

\[
I_{av} = \frac{\Delta Q}{h} = \frac{f(t_0 + h) - f(t_0)}{h}.
\]

The limit of this ratio as \( h \to 0 \) will give us the value of the current at the time \( t_0 \)

\[
I = \lim_{h \to 0} \frac{f(t_0 + h) - f(t_0)}{h}.
\]

All the three problems discussed, in spite of the fact that they refer to different branches of science, namely mechanics, geometry, and the theory of electricity, have led to one and the same mathematical operation to be performed on a given function, namely to find the limit of the ratio of the increase of the function to the corresponding increase \( h \) of the independent variable as \( h \to 0 \). The number of such widely different problems could be increased at will, and their solution would lead to the same operation. To it we are led, for example, by the question of the rate of a chemical reaction, or of the density of a nonhomogeneous mass and so forth. In view of the exceptional role played by this operation on functions, it has received a special name, differentiation, and the result of the operation is called the derivative of the function.

Thus, the derivative of the function \( y = f(x) \), or more precisely, the value of the derivative at the given point \( x \) is the limit* approached by the ratio of the increase \( f(x + h) - f(x) \) of the function to the increase \( h \) of the independent variable, as the latter approaches zero. We often write \( h = \Delta x \), and \( f(x + \Delta x) - f(x) = \Delta y \), in which case the definition of the derivative is written in the concise form:

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

* It is understood that we are speaking here of the case where the limit in question actually exists. If this limit does not exist, then we say that at the point \( x \) the function does not have a derivative.

§5. DERIVATIVE

The value of the derivative obviously depends on the point \( x \) at which it is found. Thus the derivative of a function \( y = f(x) \) is itself a function of \( x \). It is customary to denote the derivative thus

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

Certain other notations are also customary for the derivative:

\[
\frac{df(x)}{dx}, \text{ or } \frac{dy}{dx}, \text{ or } y', \text{ or } y'.
\]

We should also remark that the notation \( \frac{dy}{dx} \) looks like a fraction, although it is read as a single symbol for the derivative. In the following sections the numerator and the denominator of this “fraction” will take on independent meaning, in such a way that their ratio will coincide with the derivative so that this manner of writing is completely justified.

The results of these examples may now be formulated as follows.

The velocity of a point for which the distance \( s \) is a given function of the time \( s = f(t) \) is equal to the derivative of this function

\[
v = s' = f'(t).
\]

More concisely, the velocity is the derivative of the distance with respect to time.

The trigonometric tangent of the angle of inclination of the tangent line to the curve \( y = f(x) \) at the point with abscissa \( x \) is equal to the derivative of the function \( f(x) \) at this point:

\[
\tan \alpha = y' = f'(x).
\]

The strength of the current \( I \) at the time \( t \), if \( Q = f(t) \) is the quantity of electricity which up to time \( t \) has passed through a cross section of the conductor, is equal to the derivative

\[
I = Q' = f'(t).
\]

Let us make the following remark. The velocity of a nonuniform motion at a given time is a purely physical concept, arising from practical experience. Mankind arrived at it as the result of numerous observations on different concrete motions. The study of nonuniform motion of a body on different parts of its path, the comparison of different motions of this sort taking place simultaneously, and in particular the study of the phenomena of collisions of bodies, all represented an accumulation of
practical experience that led to the setting up of the physical concept of the velocity of a nonuniform motion at a given time. But the exact definition of velocity necessarily depended upon the method of defining its numerical value, and to define this value was possible only with the concept of the derivative.

In mechanics the velocity of a body moving according to the rule \( s = f(t) \) at the time \( t \) is defined as the derivative of the function \( f(t) \) for this value of \( t \).

The discussion at the beginning of the present section has shown, on the one hand, the advantages of introducing the operation of finding the derivative, and on the other has given a reasonable justification for the above formulated definition of the velocity at any given moment.

Thus, when we raised the question of finding the velocity of a point in nonuniform motion we had, properly speaking, only an empirical notion of its value but no exact definition. But now, as a result of our analysis, we have reached an exact definition of the value of the velocity at a given moment, namely the derivative of the distance with respect to the time. This result is extremely important from a practical point of view, since our empirical knowledge of the velocity has been greatly enriched by the fact that we can now make an exact numerical calculation.

What has just been said refers equally well, of course, to the strength of a current and to many other concepts expressing the rate of some process, physical, chemical, and so forth.

This situation may serve as an example for numerous others of a similar nature, where practical experience has led to the formation of a concept relating to the external world (velocity, work, density, area, and so forth) and then mathematics has enabled us to define this concept precisely, whereupon we can make use of the concept in practical calculations.

We have already noted at the beginning of the chapter that the concept of a derivative arose chiefly as the result of many centuries of effort directed towards the solving of two problems: drawing a tangent to a curve and finding the velocity of a nonuniform motion. These problems, and also the calculation of areas discussed later, interested mathematicians in ancient times. But until the 16th century the statement and the method of solution for each of these was not found in the work of Newton and Leibnitz. An important contribution to the foundations of present-day analysis was also made by Euler.

But it must be said that Newton and Leibnitz and their contemporaries provided very little logical basis for their great mathematical discoveries;

§5. DERIVATIVE

in their methods of reasoning and in the concepts with which they operated, there was much that is unclear from our point of view. Even at that time, the mathematicians themselves were quite conscious of this, as is shown by the embittered discussions to be found in their correspondence with one another. However, these mathematicians of the 17th and 18th centuries carried on their purely mathematical activities in very close association with the research of other investigators, in the various branches of natural science (physics, mechanics, chemistry, technology). The statement of a mathematical problem usually arose from practical needs or from a wish to understand some phenomenon of nature, and as soon as the problem was solved, the solution was submitted in one way or another to practical test. Consequently, in spite of a certain lack of logical basis, mathematics was able to advance in extremely useful directions.

Examples for the calculation of derivatives. The definition of the derivative as the limit

\[
f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}
\]

allows us to calculate the derivative of any given concrete function.

Of course, it must be admitted that cases are possible where the function at one point or another or even at many points simply does not have a derivative; in other words, the ratio

\[
\frac{f(x + h) - f(x)}{h}
\]

as \( h \to 0 \) does not approach a finite limit. This case obviously occurs at every point of discontinuity of the function \( f(x) \), since here the ratio

\[
\frac{f(x + h) - f(x)}{h}
\]

has a numerator which does not approach zero while the denominator decreases without bound. The derivative may also fail to exist at a point where the function is continuous. A simple example is given by any...
II. ANALYSIS

point where the graph of the function forms an angle (figure 12). At such a point the curve of the graph has no definite tangent, and consequently the function has no derivative. Often at such points the expression (10) approaches different values, depending on whether \( h \) approaches zero from the right or from the left, so that if \( h \) approaches zero in an arbitrary manner, the ratio (10) simply has no limit. An example of a more complicated function without a derivative is given by

\[
y = \begin{cases} 
  \frac{1}{x} \sin \frac{1}{x} & \text{for } x \neq 0, \\
  0 & \text{for } x = 0.
\end{cases}
\]

The graph of this function is drawn in figure 13. At the point \( x = 0 \) it has no derivative because, as is evident from the graph, the secant \( OA \) does not approach any definite position even when \( A \to 0 \) from one side. In fact, the secant \( OA \) oscillates endlessly back and forth between the straight line \( OM \) and the straight line \( OL \). The corresponding ratio (10) in this case has no limit, even if \( h \) preserves the same sign as it approaches zero.

\[
\frac{f(x + h) - f(x)}{h} = \frac{(x + h)^n - x^n}{h} = 2x + h.
\]

As \( h \to 0 \) we obtain* in the limit \( 2x \); consequently

\[
y' = (x^n)' = 2x.
\]

3. \( y = x^n \) (\( n \) a positive integer).

\[
\frac{f(x + h) - f(x)}{h} = \frac{(x - h)^n - x^n}{h} = \frac{x^n - nx^{n-1}h + \frac{n(n-1)}{2!}x^{n-2}h^2 + \cdots + h^n - x^n}{h} = \frac{nx^{n-1} + \frac{n(n-1)}{2!}x^{n-2}h + \cdots + h^{n-1}}{h}.
\]

Every addend on the right side, beginning with the second, approaches zero as \( h \to 0 \); consequently

\[
y' = (x^n)' = nx^{n-1}.
\]

This formula remains true for arbitrary \( n \) positive or negative, fractional.

§5. DERIVATIVE

Let us remark finally that it is possible to define, in a purely analytical way by means of a formula, a continuous function which does not have a derivative at any point. An example of such a function was first given by the outstanding German mathematician of the last century, Weierstrass.

Consequently the class of differentiable functions is considerably narrower than that of continuous functions.

Let us pass now to the actual calculation of the derivatives of the simplest functions.

1. \( y = c \), where \( c \) is a constant. A constant may be considered as a special case of a function that remains equal to the same number for arbitrary \( x \). Its graph is a straight line parallel to the \( x \)-axis at a distance equal to \( c \). This straight line forms with the \( x \)-axis an angle \( \alpha = 0 \), and obviously the derivative of a constant is identically equal to zero. \( y' = (c)' = 0 \). From the point of view of mechanics, this equation means that the velocity of a fixed point is equal to zero.

2. \( y = x^n \)

\[
\frac{f(x + h) - f(x)}{h} = \frac{(x + h)^n - x^n}{h} = 2x + h.
\]

As \( h \to 0 \) we obtain* in the limit \( 2x \); consequently

\[
y' = (x^n)' = 2x.
\]

3. \( y = x^n \) (\( n \) a positive integer).

\[
\frac{f(x + h) - f(x)}{h} = \frac{(x - h)^n - x^n}{h} = \frac{x^n - nx^{n-1}h + \frac{n(n-1)}{2!}x^{n-2}h^2 + \cdots + h^n - x^n}{h} = \frac{nx^{n-1} + \frac{n(n-1)}{2!}x^{n-2}h + \cdots + h^{n-1}}{h}.
\]

Every addend on the right side, beginning with the second, approaches zero as \( h \to 0 \); consequently

\[
y' = (x^n)' = nx^{n-1}.
\]

This formula remains true for arbitrary \( n \) positive or negative, fractional.

* We always assume here that \( h \neq 0 \).
or even irrational, although the proof must then be different. We will make use of this fact without proving it. Thus for example

\[(\sqrt{x})' = (x^2)' = \frac{1}{2} x^{-\frac{1}{2}} = \frac{1}{2 \sqrt{x}}, \quad (x > 0);\]

\[(\sqrt[3]{x})' = (x^3)' = \frac{1}{3} x^{-\frac{2}{3}} = \frac{1}{3 \sqrt[3]{x^2}}, \quad (x \neq 0);\]

\[\left(\frac{1}{x}\right)' = (x^{-1})' = -1 \cdot x^{-2} = -\frac{1}{x^2}, \quad (x \neq 0);\]

\[(x^n)' = nx^{n-1}, \quad (x > 0).\]

4. \(y = \sin x.\)

\[
\frac{\sin (x + h) - \sin x}{h} = \frac{2 \sin \frac{h}{2} \cos \left(\frac{x + h}{2}\right)}{h} = \frac{\sin \frac{h}{2}}{h} \cdot \cos \left(\frac{x + h}{2}\right).
\]

As explained earlier, the first fraction approaches unity as \(h \to 0,\) and \(\cos (x + h/2)\) obviously approaches \(\cos x.\) Thus the derivative of the sine is equal to the cosine

\[y' = (\sin x)' = \cos x.\]

We suggest to the reader that by the same sort of argument he prove that

\[\cos x)' = -\sin x.\]

5. Earlier (Chapter II, §2) we have already noted the existence of the limit

\[
\lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = e = 2.71828 \ldots.
\]

We also remarked that for the calculation of this limit no essential role is played by the fact that \(x\) took on only positive integral values. It is important only that the infinitesimal \(1/n,\) which is being added to unity, and the exponent \(n,\) which is increasing beyond all bounds, should be reciprocal to each other.

Making use of this assertion, we can easily find the derivative of the logarithm \(y = \log_a x\)

\[\log_a (x + h) - \log_a x = \frac{1}{h} \log_a \frac{x + h}{x} = \frac{1}{x} \log_a \left(1 + \frac{h}{x}\right)^{x/h}.\]

\[\frac{(\log_a x)' = 1}{x}\]

or again

\[\frac{(\log_a x)' = 1}{x}.\]

\[\frac{(\log_a x)' = 1}{x}.\]

§6. RULES FOR DIFFERENTIATION

The continuity of the logarithm allows us to replace the quantity under the log sign by its limit, which is equal to \(e;\) thus

\[
\lim_{h \to 0} \left(1 + \frac{h}{x}\right)^{x/h} = e.
\]

(in this case the role of \(n \to \infty\) is played by the increasing quantity \(x/h).\)

As a result, we obtain the rule for differentiating a logarithm

\[(\log_a x)' = \frac{1}{x} \log_a e.\]

This rule becomes particularly simple if as the base of our logarithms we choose the number \(e.\) Logarithms taken to this base are called natural logarithms and are denoted by \(\ln x.\) We may write

\[\frac{(\log_e x)' = 1}{x}\]

or again

\[\frac{(\ln x)' = 1}{x}.\]

§6. Rules for Differentiation

From the examples given earlier it may appear that the calculation of the derivative of every new function demands the invention of new methods. This is not the case. The development of analysis was made possible to no small extent by the discovery of a simple unified method for finding the derivative of an arbitrary "elementary" function (that is, a function which may be expressed by a formula consisting of a finite combination of the fundamental algebraic operations, the trigonometric functions, the operation of raising to a power, and the taking of logarithms). At the basis of this method are the so-called rules of differentiation. They consist of a number of theorems that allow us to reduce more complicated problems to simpler ones.

We will explain here the rules of differentiation and will try to be very brief in deducing them. If the reader wishes to form merely a general idea of analysis, he may omit the present section, remembering only that there exists a means of actually finding the derivative of any elementary function. In this case it will be necessary, of course, for him to take on faith some of the calculations in our later examples.

**Derivative of a sum.** Assume that \(y\) is given as a function of \(x\) by the expression

\[y = \phi(x) + \psi(x),\]
where \( u = \phi(x) \) and \( v = \psi(x) \) are known functions of \( x \). We assume moreover that we can find the derivatives of the functions \( u \) and \( v \). How then are we to find the derivative of the function \( y \)? The answer is simple

\[
y' = (u + v)' = u' + v'.
\]

(11)

In fact, let us give \( x \) an increment \( \Delta x \); then \( u, v, \) and \( y \) will each receive an increment \( \Delta u, \Delta v, \) and \( \Delta y, \) connected by the equation

\[
\Delta y = \Delta u + \Delta v.
\]

Thus*

\[
\frac{\Delta y}{\Delta x} = \frac{\Delta u}{\Delta x} + \frac{\Delta v}{\Delta x},
\]

and after the passage to the limit for \( \Delta x \to 0 \) we at once get formula (11), if, of course, the functions \( u \) and \( v \) have derivatives.

Analogously we may derive the formula for differentiating the difference of two functions

\[
(u - v)' = u' - v'.
\]

(12)

**Derivative of a product.** The rule for the differentiation of a product is somewhat more complicated. The derivative of the product of two functions, each of which has a derivative, exists, and is equal to the sum of the product of the first function by the derivative of the second and the product of the second by the derivative of the first; that is

\[
(uv)' = uv' + vu'.
\]

(13)

In fact, let us give \( x \) an increment \( \Delta x \). Then the functions \( u, v \) and \( y = uv \) will receive the increments \( \Delta u, \Delta v, \Delta y, \) satisfying the relation

\[
\Delta y = (u + \Delta u)(v + \Delta v) - uv = u \Delta v + v \Delta u + \Delta u \Delta v,
\]

from which

\[
\frac{\Delta y}{\Delta x} = u \frac{\Delta v}{\Delta x} + v \frac{\Delta u}{\Delta x} + \Delta u \frac{\Delta v}{\Delta x}.
\]

After passage to the limit for \( \Delta x \to 0 \) the first two summands on the right side produce the right side of formula (13) while the third summand vanishes.† Consequently, in the limit we obtain the rule (13).

---

* Here \( \Delta x \) is never equal to zero.

† The final summand here approaches zero for \( \Delta x \to 0 \), since \( \Delta u/\Delta x \) approaches a finite number, equal to the derivative \( v' \), which was assumed from the beginning to exist, and \( \Delta u \to 0 \), since the function \( u \), assumed to have a derivative, is continuous.

§6. RULES FOR DIFFERENTIATION

In the particular case \( v = c = \text{const} \), we have

\[
(cu)' = cu' + uc' = cu',
\]

since the derivative of a constant is equal to zero.

**Derivative of a quotient.** Let \( y = u/v \), where \( u \) and \( v \) have a derivative for a given \( x \), with \( v \neq 0 \) for that value of \( x \). Obviously

\[
\frac{\Delta y}{\Delta x} = \frac{\frac{\Delta u}{v} - \frac{u \Delta v}{v^2}}{1 + \Delta u v},
\]

from which

\[
\frac{\Delta y}{\Delta x} = \frac{\frac{\Delta u}{v} - \frac{u \Delta v}{v^2}}{1 + \Delta u v} \to \frac{uu' - vv'}{v^2} (\Delta x \to 0).
\]

Here we have again made use of the fact that for a function \( v \) which has a derivative we necessarily have \( \Delta v \to 0 \), when \( \Delta x \to 0 \). Thus

\[
\left( \frac{u}{v} \right)' = \frac{vu' - u v'}{v^2}.
\]

(15)

Let us give some examples of the application of these rules

\[
(2x^3 - 5)' = 2(3x^2), \quad (5)' = 0 = 6x^2; \quad (x^3 \sin x)' = (x^3)' \sin x + (x^3) \sin x = 2x^2 \cos x + 2x \sin x; \quad (\tan x)' = \left( \frac{\sin x}{\cos x} \right)' = \frac{\cos x \sin x - \sin x \cos x}{\cos^2 x} = \frac{\cos x \cdot \cos x - \sin x \cdot (\sin x)}{\cos^2 x} = \frac{1}{\cos^2 x} = \sec^2 x.
\]

We recommend to the reader to prove for himself the formula

\[
(\cot x)' = -\csc^2 x.
\]

**Derivative of the inverse function.** Let us consider a function \( y = f(x) \), which is continuous and increasing (decreasing) on the interval \([a, b]\). By increasing (decreasing) we mean that to a greater value of \( x \) in the interval \([a, b]\) corresponds a greater (smaller) value of \( y \) (figure 14).

Let \( c = f(a) \) and \( d = f(b) \). In figure 14 it is evident that for each value of \( y \) from the interval \([c, d]\) (or \([d, c]\), respectively) there corresponds exactly one value of \( x \) from the interval \([a, b]\) such that \( y = f(x) \). Thus
on the interval \([c, d]\) (or \([a, c]\)) we have a completely determined function \(x = \psi(y)\), which is called the inverse function of \(y = f(x)\). In figure 14 it is clear that the function \(\psi(y)\) is continuous, a fact which is proved in modern analysis by strictly analytical methods. Now let \(dx\) and \(dy\)

![Graph showing inverse function](image)

Fig. 14.

correspond respectively to the increments in \(x\) and \(y\). It is evident that

\[
\frac{\Delta y}{\Delta x} = \frac{1}{\frac{\Delta x}{\Delta y}}, \text{ if } \Delta y \neq 0.
\]

In the limit this gives us a simple relation between derivatives of the direct and inverse functions

\[
y'_{x} = \frac{1}{x'_{y}}.
\]

(16)

Let us make use of this relation to find the derivative of the function \(y = a^x\). The inverse function is \(x = \log_a y\), which we are already able to differentiate, and so we may write

\[
(a^x)' = \frac{1}{(\log_a y)'y} = \frac{1}{\frac{1}{\log_a e}y} = y \log a = a^x \ln a.
\]

(17)

In particular \((e^x)' = e^x\).

As another example let us take \(y = \arcsin x\). The inverse function is \(x = \sin y\). Thus

\[
(\arcsin x)' = \frac{1}{\cos y} = \frac{1}{\sqrt{1 - (\sin y)^2}} = \frac{1}{\sqrt{1 - x^2}}.
\]

\section{6. RULES FOR DIFFERENTIATION}

Table of derivatives. Let us tabulate the derivatives of the simplest elementary functions (Table 2).

<table>
<thead>
<tr>
<th>(y)</th>
<th>(y')</th>
<th>(y)</th>
<th>(y')</th>
<th>(y)</th>
<th>(y')</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c)</td>
<td>0</td>
<td>(\ln x)</td>
<td>(\frac{1}{x})</td>
<td>(\tan x)</td>
<td>(\sec^2 x)</td>
</tr>
<tr>
<td>(x^a)</td>
<td>(ax^{a-1})</td>
<td>(\log_a x)</td>
<td>(\frac{1}{x})</td>
<td>(\log a)</td>
<td>(\frac{1}{\sqrt{1 - x^2}})</td>
</tr>
<tr>
<td>(e^x)</td>
<td>(e^x)</td>
<td>(\sin x)</td>
<td>(\cos x)</td>
<td>(\arccos x)</td>
<td>(\frac{1}{\sqrt{1 - x^2}})</td>
</tr>
<tr>
<td>(a^x)</td>
<td>(a^x \ln a)</td>
<td>(\cos x)</td>
<td>(-\sin x)</td>
<td>(\arcsin x)</td>
<td>(\frac{1}{\sqrt{1 + x^2}})</td>
</tr>
</tbody>
</table>

These formulas have been calculated and explained earlier, with the exception of the last two which the reader may, if he wishes, easily derive for himself by using the rule for differentiation of an inverse function.

Calculation of the derivative of a function of a function. It remains to consider the last and most difficult rule for differentiation. The reader in possession of this rule and of a set of tables may with perfect right consider that he is able to differentiate any elementary function.

In order to apply the rule we are about to give, it is necessary to be completely clear about how the function we wish to differentiate is constructed; that is, which operations must we perform on the independent variable \(x\), and in which order, to produce the value of the dependent variable \(y\).

For example, to calculate the function

\[ y = \sin x^3, \]

it is necessary first of all to raise \(x\) to the second power and then to take the sine of the magnitude so obtained, a procedure which may be described in the following way: \(y = \sin u\), where \(u = x^3\).

On the other hand, in order to calculate the function

\[ y = \sin^2 x, \]

it is necessary first of all to find the sine of \(x\), and then to raise the value so found to the second power, a procedure which may be written thus: \(y = u^2\), where \(u = \sin x\).
II. ANALYSIS

Here are some examples:

1. \( y = (3x + 4)^3, y = u^3, u = 3x + 4. \)
2. \( y = \sqrt{1 - x^2}, y = u^2, u = 1 - x^2. \)
3. \( y = e^{ex}; y = \epsilon^x, u = kx. \)

In more complicated cases we have a chain of simple relations, which may have several links. For example,

4. \( y = \cos^3 x^2; y = u^3, u = \cos v; v = x^2. \)

If \( y \) is a function of the variable \( u \)

\[ y = f(u), \tag{18} \]

and \( u \) in its turn is a function of the variable \( x \)

\[ u = \phi(x), \tag{19} \]

then \( y \), being a function of \( u \), is also a certain function of \( x \), which may denote as follows

\[ y = F(x) = f(\phi(x)). \tag{20} \]

By considering more complicated cases we may form, for example, the function

\[ y = \Phi(x) = f(\phi(\psi(x))), \]

which is equivalent to the equations

\[ y = f(u), \quad u = \phi(x), \quad v = \psi(x), \]

and we could form still longer chains.

We now show how to calculate the derivative of the function \( F(x) \) defined by equation (20) if we know the derivative of \( f(u) \) with respect to \( u \) and the derivative of \( \phi(x) \) with respect to \( x \).

Let us give to \( x \) the increment \( \Delta x \); then by (19) \( u \) will receive a certain increment \( \Delta u \) and by (18) \( y \) will receive an increment \( \Delta y \). Thus we may write

\[ \frac{\Delta y}{\Delta x} = \frac{\Delta y}{\Delta u} \cdot \frac{\Delta u}{\Delta x}. \]

Now let \( \Delta x \) approach zero. Then \( \Delta u/\Delta x \to u'_x \). Furthermore, from the continuity of \( u \), the increase \( \Delta u \to 0 \), and therefore \( \Delta y/\Delta u \to y'_u \) (the existence of the derivatives \( y'_u \) and \( u'_x \) was assumed).

\[ y'_x = y'_u u'_x. \tag{21} \]

Thus we have proved the important formula for the derivative of a function of a function:

Let us calculate, from formula (21) and the fundamental table of derivatives given, the derivatives of the functions we have been considering:

1. \( y = (3x + 4)^3 = u^3, y'_u = (u^3)'u(3x + 4)' = 3u^2 \cdot 3 = 9(3x + 4)^2. \)
2. \( y = \sqrt{1 - x^2} = u^2, y'_u = (u^2)'(1 - x^2)' = \frac{1}{2}u^{-1}(-2x) = -\frac{x}{\sqrt{1 - x^2}}. \)
3. \( y = e^{ex} = e^x, y'_e = (e^x)'u'_x = e^x \cdot k = ke^{ex}. \)

If \( y = f(u), u = \phi(v), v = \psi(x) \), then

\[ y'_x = y'_u u'_x = y'_u u'_e v'_e = y'_e u'_v v'_v . \]

It is clear how to generalize this formula for the case of an arbitrary (finite) number of functions in the chain. For example,

4. \( y = \cos^3 x^2; y'_x = (u^3)'(\cos v)'(x^2)' = 3u^2(-\sin v) \cdot 2x = -6x \cos^2 x^2 \sin x^2. \)

In our explanation of how to calculate the derivative of a function of a function, we have introduced intermediate variables \( u, v, \ldots \). But in fact, after a little practice one may dispense with them, simply keeping in mind the functions they denote.

The elementary functions. To close the present section let us remark that the functions whose derivatives were listed in tabular form (Table 2) may be used to define the so-called elementary functions. These elementary functions are defined as those functions that may be obtained from the preceding simple functions by the four arithmetical operations and the operation of taking a function of a function, each of these operations being performed a finite number of times.

For example, the polynomial \( x^4 - 2x^2 + 3x - 5 \) is an elementary function since it is obtained by arithmetic operations from a number of functions to the form \( x^4 \). The function \( \ln \sqrt{1 - x^2} \) is also elementary,
since it is obtained from the polynomial \( u = 1 - x^2 \) by the operation \( v = u^{1/2} \), and subsequently the operation in \( v \).

The rules for differentiation discussed earlier are sufficient to obtain the derivative of any elementary function, as soon as we know the derivatives of the simplest elementary functions.

§7. Maximum and Minimum; Investigation of the Graphs of Functions

One of the simplest and most important applications of the derivative is in the theory of maxima and minima. Let us suppose that on a certain interval \( a < x < b \) we are given a function \( y = f(x) \) which is not only continuous but also has a derivative at every point. Our ability to calculate the derivative enables us to form a clear picture of the graph of the function. On an interval on which the derivative is always positive the tangent to the graph will be directed upward. On such an interval the function will increase; that is, to a greater value of \( x \) will correspond a greater value of \( f(x) \). On the other hand, on an interval where the derivative is always negative, the function will decrease; the graph will run downward.

Maximum and minimum. In figure 15 we have drawn the graph of a function \( y = f(x) \) defined on the interval \( (a, b) \). Of a special interest are the points of this graph whose abscissas are \( x_0, x_1, x_3 \).

At the point \( x_0 \) the function \( f(x) \) is said to have a local maximum; by this we mean that at this point \( f(x) \) is greater than at neighboring points; more precisely \( f(x_0) > f(x) \) for every \( x \) in a certain interval around the point \( x_0 \).

\[
\text{Fig. 15.}
\]

§7. Maximum and Minimum

A local minimum is defined analogously.

For our function a local maximum occurs at the points \( x_0 \) and \( x_3 \), and a local minimum at the point \( x_1 \).

At every maximum or minimum point, if it is inside the interval \([a, b]\), i.e., if it does not coincide with one of the end points \( a \) or \( b \), the derivative must be equal to zero.

This last statement, a very important one, follows immediately from the definition of the derivative as the limit of the ratio \( \frac{\Delta y}{\Delta x} \). In fact, if we move a short distance from the maximum point, then \( \Delta y \leq 0 \). Thus for positive \( \Delta x \) the ratio \( \frac{\Delta y}{\Delta x} \) is nonpositive, and for negative \( \Delta x \) the ratio \( \frac{\Delta y}{\Delta x} \) is nonnegative. The limit of this ratio, which exists by hypothesis, can therefore be neither positive nor negative and there remains only the possibility that it is zero. By inspection of the diagram it is seen that this means that at maximum or minimum points (it is customary to leave out the word "local," although it is understood) the tangent to the graph is horizontal. In figure 15 we should remark that at the points \( x_2 \) and \( x_1 \) also the tangent is horizontal, just as it is at the points \( x_0, x_1, x_3 \), although at these points the function has neither maximum nor minimum. In general, there may be more points at which the derivative of the function is equal to zero (stationary points) than there are maximum or minimum points.

Determination of the greatest and least values of a function. In numerous technical questions it is necessary to find the point \( x \) at which a given function \( f(x) \) attains its greatest or its least value on a given interval.

In case we are interested in the greatest value, we must find \( x_0 \) on the interval \([a, b]\) for which among all \( x \) on \([a, b]\) the inequality \( f(x_0) \geq f(x) \) is fulfilled.

But now the fundamental question arises, whether in general there exists such a point. By the methods of modern analysis it is possible to prove the following existence theorem: If the function \( f(x) \) is continuous on a finite interval, then there exists at least one point on the interval for which the function attains its maximum (minimum) value on the interval \([a, b]\).

From what has been said already, it follows that these maximum or minimum points must be sought among the "stationary" points. This fact is the basis for the following well-known method for finding maxima and minima.

First we find the derivative of \( f(x) \) and then solve the equation obtained by setting it equal to zero

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

If \( x_1, x_2, \ldots, x_n \) are the roots of this equation, we then compare the numbers \( f(x_1), f(x_2), \ldots, f(x_n) \) with one another. Of course, it is necessary to take into account that the maximum or minimum of the function may be found not within the interval but at the end (as is the case with the minimum in figure 15) or at a point where the function has no derivative (as in figure 12). Thus to the points \( x_1, x_2, \ldots, x_n \) we must add the ends \( a \) and \( b \) of the interval and also those points, if they exist, at which there is no derivative. It only remains to compare the values of the function at all these points and choose among them the greatest or the least.

With respect to the stated existence theorem, it is important to add that this theorem ceases, in general, to hold in the case that the function \( f(x) \) is continuous only on the interval \((a, b)\); that is, on the set of points \( x \) satisfying the inequalities \( a < x < b \). We leave it to the reader to consider the fact that the function \( 1/x \) has neither a maximum nor a minimum on the interval \((0, 1)\).

Let us look at some examples.

From a square piece of tin of side \( a \) it is required to make a rectangular open box of maximum volume. If from the corners of the original square we take away squares of side \( x \) (see §2, example 2) we get a box with the volume

\[
V = x(a - 2x)^2.
\]

Our problem then becomes to find the value of \( x \) for which the function \( V(x) \) attains its greatest value on the interval \( 0 \leq x \leq a/2 \). In accordance with the rule, we find the derivative and set it equal to zero

\[
V'(x) = (a - 2x)^2 - 4x(a - 2x) = 0.
\]

Solving this equation, we find the two roots

\[
x_1 = \frac{a}{2}, x_2 = \frac{a}{6}.
\]

To these we adjoin the left end of the interval (the right end is identical with \( x_1 \)) and compare the values of the function at these points

\[
V(0) = 0; V\left(\frac{a}{6}\right) = \frac{2}{27}a^3, V\left(\frac{a}{2}\right) = 0.
\]

Thus the box will have the greatest volume, equal to \( 2/27 a^3 \), for the height \( x = a/6 \).

As a second example, let us examine the problem of the lamp at the skating rink (see §2, example 3). At what height \( h \) should we place the lamp in order that the edge of the rink may receive the greatest illumination?

§7. MAXIMUM AND MINIMUM

For formula (3) §2, our problem reduces to determining the value of \( h \) for \( T = A \sin \alpha/r^2 + r^2 \) takes on its greatest value. Instead of \( h \) it is more convenient here to find the angle \( \alpha \) (figure 3, Chapter I). We have

\[
h = r \tan \alpha,
\]

so that

\[
T = \frac{A}{r^2} \sin \frac{\alpha}{1 + \tan^2 \alpha} = \frac{A}{r^2} \sin \alpha \cos^2 \alpha.
\]

Then it is required to find the maximum of the function \( T(\alpha) \) among those values of \( \alpha \) which satisfy the inequality \( 0 < \alpha < \pi/2 \). To do this, we find the derivative and set it equal to zero

\[
T'(\alpha) = \frac{A}{r^2} (\cos^2 \alpha - 2 \sin^2 \alpha \cos \alpha) = 0.
\]

This equation splits into two

\[
\cos \alpha = 0, \cos^2 \alpha - 2 \sin^2 \alpha = 0.
\]

The first equation has the root \( \alpha = \pi/2 \), which coincides with the end of the interval \((0, \pi/2)\). The second equation may be put in the form

\[
\tan^2 \alpha = \frac{1}{2}.
\]

But since \( 0 < \alpha < \pi/2 \), we have the result \( \alpha \approx 35^\circ 15' \). So this is the value for which the function \( T(\alpha) \) attains its maximum (at the ends of the interval, \( T = 0 \)). The desired height \( h \) is thus equal to

\[
h = r \tan \alpha = \frac{r}{\sqrt{2}} \approx 0.7r.
\]

For best illumination of the edge of the rink the lamp should be placed at a height equal to about 0.7 times the radius.

But now let us suppose that the facilities at our disposal do not allow us to raise the lamp to a height greater than a certain \( H \). Then the angle \( \alpha \) may vary not from 0 to \( \pi/2 \) but only within the narrower limits \( 0 < \alpha \leq \arctan (H/r) \). For example, let \( r = 12 \) meters and \( H = 9 \) meters. In this case, it is in fact possible to raise the lamp to the height \( h = r/\sqrt{2} \), which amounts to somewhat more than 8 meters, so that this is what we ought to do. But if \( H \) is less than 8 meters (for example, if we have at our disposal only a pole of length 6 meters), then it turns out that the derivative of the function \( T(\alpha) \) in the interval \([0, \arctan (H/r)]\) is nowhere equal to zero. In this case the maximum is attained at the end of the interval, and the lamp should be raised to the greatest possible height \( H = 6 \) meters.
II. ANALYSIS

Up to now we have considered a function on a finite interval. If the interval is infinite in length, then even a continuous function may fail to attain its greatest or least value but may, for example, continue to grow or to decrease as \( x \) approaches infinity.

Thus the functions \( y = kx + b \) (see figure 5, Chapter I), \( y = \arctan x \) (figure 16a), \( y = \ln x \) (figure 16b) nowhere attain either a maximum or a minimum. The function \( y = e^{-x^2} \) (figure 16c) attains its maximum at the point \( x = 1 \), but nowhere attains a minimum. As for the function \( y = x/(1 + x^2) \) (figure 16d), it reaches its minimum at the point \( x = -1 \) and its maximum at the point \( x = 1 \).

\[ y = \frac{x}{1 + x^2} \]

**Fig. 16d.**

\[ y = e^{-x^2} \]

**Fig. 16c.**

\[ y = \arctan x \]

**Fig. 16a.**

\[ y = \ln x \]

**Fig. 16b.**

§7. MAXIMUM AND MINIMUM

In the case of an interval of infinite length the investigation may be reduced to the ordinary rules. It is only necessary to consider in place of \( d(a) \) and \( f(b) \) the limits

\[ A = \lim_{x \to -\infty} f(x), \quad B = \lim_{x \to +\infty} f(x). \]

Derivatives of higher orders. We have just seen how, for closer study of the graph of a function, we must examine the changes in its derivative \( f'(x) \). This derivative is a function of \( x \), so that we may in turn find its derivative.

The derivative of the derivative is called the second derivative and is denoted by

\[ [y']' = y'' \quad \text{or} \quad [f'(x)]' = f''(x). \]

Analogously, we may calculate the third derivative

\[ [y']' = y''' \quad \text{or} \quad [f''(x)]' = f'''(x) \]

and more generally the \( n \)th derivative or, as it is also called, the derivative of \( n \)th order

\[ y^{(n)} = f^{(n)}(x). \]

Of course, it must be kept in mind that, for a certain value of \( x \) (or even for all values of \( x \)) this sequence may break off at the derivative of some order, say the \( k \)th; it may happen that \( f^{(k)}(x) \) exists but not \( f^{(k+1)}(x) \). Derivatives of arbitrary order will appear later in §9 in connection with the Taylor formula. For the moment we confine ourselves to the second derivative.

Significance of the second derivative; convexity and concavity. The second derivative has a simple significance in mechanics. Let \( s = f(t) \) be a law of motion along a straight line; then \( s ' \) is the velocity and \( s '' \) is the “velocity of the change in the velocity” or more simply the “acceleration” of the point at time \( t \). For example, for a falling body under the force of gravity

\[ s = \frac{gt^2}{2} + vt + v_0, \]

\[ s ' = gt + v, \]

\[ s '' = g, \]

that is, the acceleration of falling bodies is constant.

The second derivative also has a simple geometric meaning. Just as the sign of the first derivative determines whether the function is increasing
or decreasing, so the sign of the second derivative determines the side
ward which the graph of the function will be curved.
Suppose, for example, that on a given interval the second derivative
is everywhere positive. Then the first derivative increases and therefore

\[ f'(x) = \tan \alpha \text{ increases and the angle } \alpha \text{ of inclination of the tangent line itself increases (figure 17). Thus as we move along the curve it keeps turning constantly to the same side, namely upward, and is thus, as they say, "convex downward."}

On the other hand, in a part of a curve where the second derivative is constantly negative (figure 18) the graph of the function is "convex upward." *

---

* Strictly defined, the "convexity upward" is that property of the curve that consists of its lying above (more precisely "not below") the chord joining any two of its points; analogously, for "convexity downward" (which is also simply called "concavity"), the curve does not lie above its chords.

---

\[ f(x) = \frac{x^3}{3} - \frac{5x^2}{2} + 6x - 2. \]

We take its first derivative and set it equal to zero,

\[ f'(x) = x^2 - 5x + 6 = 0. \]

The roots of the equation obtained in this way are \( x_1 = 2, x_2 = 3 \). The corresponding values of the function are

\[ f(2) = 2\frac{2}{3}, f(3) = 2\frac{1}{2}. \]

We then mark these two points on the diagram. Along with these we may also mark the point with coordinates \( x = 0, y = f(0) = -2 \) where the graph intersects the \( y \)-axis. The second derivative is \( f''(x) = 2x - 5 \). This reduces to zero for \( x = \frac{5}{2} \), so that

\[ f''(x) > 0 \text{ for } x > \frac{5}{2}, \]

\[ f''(x) < 0 \text{ for } x < \frac{5}{2}. \]

The point

\[ x = \frac{5}{2}, y = f\left(\frac{5}{2}\right) = 2\frac{7}{2} \]

is a point of inflection of the graph. To the left of this point the curve is convex upward, and to the right it is convex downward.

It is now evident that the point \( x = 2 \) is a maximum point and the point \( x = 3 \) is a minimum point for the function.

---

* In more complicated cases, where the second derivative itself changes sign, the problem of determining the character of the stationary point is solved by means of the Taylor formula (§9).
II. ANALYSIS

On the basis of these results we conclude that the graph of the function $y = f(x)$ has the appearance sketched in figure 19. To the right of the point $(0, -2)$ the curve rises with increasing $x$, is convex upwards, and attains its maximum at the point $(2, 2\frac{2}{3})$, after which it begins to fall. At the point $(2\frac{1}{2}, 2\frac{7}{12})$, where $f''(x) = 0$, the convexity changes to concavity. Then at the point $(3, 2\frac{1}{2})$ the function attains its minimum and from there on rises to infinity. The final statement comes from the fact that the first term of the function, the one containing the highest (third) power of $x$, approaches infinity faster than the second and third terms. For the same reason the graph of the function approaches $-\infty$ as $x$ assumes numerically larger negative values.

Example 2. We shall prove the inequality $e^x \geq 1 + x$ for arbitrary $x$. For this purpose we consider the function $f(x) = e^x - x - 1$. Its first derivative is $f'(x) = e^x - 1$, which reduces to zero only for $x = 0$. The second derivative $f''(x) = e^x > 0$ for all $x$. Consequently the graph of the function $f(x)$ is convex downward. The number $f(0) = 0$ is a minimum for the function and $e^x - x - 1 \geq 0$ for all $x$.

The study of graphs has many different purposes. They often show very clearly, for example, the number of real roots of a given equation. Thus, in order to demonstrate that the equation

$$xe^x = 2$$

has a single real root, we may study the graphs of the functions $y = e^x$ and $y = 2/x$ (as sketched in figure 20). It is easy to see that these graphs intersect at only one point, so that the equation $e^x = 2/x$ has exactly one root.

§8. INCREMENT AND DIFFERENTIAL OF A FUNCTION

The methods of analysis are extensively applied to questions of approximate calculation of the roots of an equation. On this subject, see Chapter IV, §5.

§8. Increment and Differential of a Function

The differential of a function. Let us consider a function $y = f(x)$ that has a derivative. The increment of this function

$$\Delta y = f(x + \Delta x) - f(x),$$

corresponding to the increment $\Delta x$, has the property that the ratio $\Delta y/\Delta x$, as $\Delta x \to 0$, approaches a finite limit, equal to the derivative

$$\frac{\Delta y}{\Delta x} \to f'(x).$$

This fact may be written as an equality

$$\frac{\Delta y}{\Delta x} = f'(x) + \alpha,$$

where the value of $\alpha$ depends on $\Delta x$ in such a way that as $\Delta x \to 0$, $\alpha$ also approaches zero. Thus the increment of a function may be represented in the form

$$\Delta y = f'(x) \Delta x + \alpha \Delta x,$$

where $\alpha \to 0$, if $\Delta x \to 0$.

The first summand on the right side of this equality depends on $\Delta x$ in a very simple way, namely it is proportional to $\Delta x$. It is called the differential of the function, at the point $x$, corresponding to the given increment $\Delta x$, and is denoted by

$$dy = f'(x) \Delta x.$$

The second summand has the characteristic property that, as $\Delta x \to 0$, it approaches zero more rapidly than $\Delta x$, as a result of the presence of the
factor $\alpha$. It is therefore said to be an infinitesimal of higher order than $\Delta x$ and, in case $f'(x) \neq 0$, it is also of higher order than the first summand. By this we mean that for sufficiently small $\Delta x$ the second summand is small in itself and its ratio to $\Delta x$ is also arbitrarily small.

The decomposition of $\Delta y$ into two summands, of which the first (the principal part) depends linearly on $\Delta x$ and the second is negligible for small $\Delta x$, may be illustrated by figure 21. The segment $BC = \Delta y$, where $BC = BD + DC$, $BD = \tan \beta \cdot \Delta x = f'(x) \Delta x = dy$, and $DC$ is an infinitesimal of higher order than $\Delta x$.

In practical problems the differential is often used as an approximate value for the increment in the function. For example, suppose we have the problem of determining the volume of the walls of a closed cubical box whose interior dimensions are $10 \times 10 \times 10$ cm and the thickness of whose walls is 0.05 cm. If great accuracy is not required, we may argue as follows. The volume of all the walls of the box represents the increment $\Delta y$ of the function $y = x^2$ for $x = 10$ and $\Delta x = 0.1$. So we find approximately

$$\Delta y \approx dy = (x^2)\cdot 2x \cdot dx = 3 \cdot 10^2 \cdot 0.1 = 30 \text{ cm}^3.$$ 

For symmetry in the notation it is customary to denote the increment of the independent variable by $dx$ and to call it also a differential. With this notation the differential of the function may be written thus:

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

Then the derivative is the ratio $f'(x) = dy/dx$ of the differential of the function to the differential of the independent variable.

The differential of a function originated historically in the concept of an "indivisible." This concept, which from a modern point of view was never very clearly defined, was in its time, in the 18th century, a fundamental one in mathematical analysis. The ideas concerning it have undergone essential changes in the course of several centuries. The indivisible, and later the differential of a function, were represented as actual infinitesimals, as something in the nature of an extremely small constant magnitude, which however was not zero. The definition given in this section is the one accepted in present-day analysis. According to this definition the differential is a finite magnitude for each increment $\Delta x$ and is at the same time proportional to $\Delta x$. The other fundamental property of the differential, the character of its difference from $dy$, may be recognized only in motion, so to speak: if we consider an increment $\Delta x$ which is approaching zero (which is infinitesimal), then the difference between $dy$ and $\Delta y$ will be arbitrarily small even in comparison with $\Delta x$.

This substitution of the differential in place of small increments of the function forms the basis of most of the applications of infinitesimal analysis to the study of nature. The reader will see this in a particularly clear way in the case of differential equations, dealt in this book in Chapters V and VI.

Thus, in order to determine the function that represents a given physical process, we try first of all to set up an equation that connects this function in some definite way with its derivatives of various orders. The method of obtaining such an equation, which is called a differential equation, often amounts to replacing increments of the desired functions by their corresponding differentials.

As an example let us solve the following problem. In a rectangular system of coordinates $Oxyz$, we consider the surface obtained by rotation of the parabola whose equation (in the $Oyz$ plane) is $z = y^2$. This surface is called a paraboloid of revolution (figure 22). Let $v$ denote the volume of the body bounded by the paraboloid and the plane parallel to the $Oxy$ plane at a distance $z$ from it. It is evident that $v$ is a function of $z$ ($z > 0$).

To determine the function $v$, we attempt to find its differential $dv$. The increment $\Delta v$ of the function $v$ at the point $z$ is equal to the volume bounded by the paraboloid and by two planes parallel to the $Oxy$ plane at distances $z$ and $z + \Delta z$ from it.

It is easy to see that the magnitude of $\Delta v$ is greater than the volume of the circular cylinder of radius $\sqrt{z}$ and height $\Delta z$ but less than that of the circular cylinder with radius $\sqrt{z + \Delta z}$ and height $\Delta z$. 

FIG. 21.

FIG. 22.
II. ANALYSIS

Thus

\[ \pi z \Delta z < \Delta v < \pi(z + \Delta z) \Delta z \]

and so

\[ \Delta v = \pi(z + \theta \Delta z) \Delta z = \pi z \Delta z + \pi \theta \Delta z^2, \]

where \( \theta \) is some number depending on \( \Delta z \) and satisfying the inequality \( 0 < \theta < 1 \).

So we have succeeded in representing the increment \( \Delta v \) in the form of a sum, the first summand of which is proportional to \( \Delta z \), while the second is an infinitesimal of higher order than \( \Delta z \) (as \( \Delta z \to 0 \)). It follows that the first summand is the differential of the function \( v \)

\[ dv = \pi z \Delta z, \]

or

\[ dv = \pi z dz, \]

since \( \Delta z = dz \) for the independent variable \( z \).

The equation so obtained relates the differentials \( dv \) and \( dz \) (of the variables \( v \) and \( z \)) to each other and thus is called a differential equation.

If we take into account that

\[ \frac{dv}{dz} = v', \]

where \( v' \) is the derivative of \( v \) with respect to the variable \( z \), our differential equation may also be written in the form

\[ v' = \pi z. \]

To solve this very simple differential equation we must find a function of \( z \) whose derivative is equal to \( \pi z \). Problems of this sort are treated in a general way in §§10 and 11, but for the moment we urge the reader to verify that a solution of our equation is given by \( v = \pi z^2/2 + C \), where for \( C \) we may choose an arbitrary number.* In our case the volume of the body is obviously zero for \( z = 0 \) (see figure 22), so that \( C = 0 \). Thus our function is given by \( v = \pi z^2/2 \).

The mean value theorem and examples of its application. The differential expresses the approximate value of the increment of the function in terms of the increment of the independent variable and of the derivative at the initial point. So for the increment from \( x = a \) to \( x = b \), we have

\[ f(b) - f(a) \approx f'(a)(b - a). \]

* This formula gives all the solutions.

§8. INCREMENT AND DIFFERENTIAL OF A FUNCTION

It is possible to obtain an exact equation of this sort if we replace the derivative \( f'(a) \) at the initial point by the derivative at some intermediate point, suitably chosen in the interval \((a, b)\). More precisely: If \( y = f(x) \) is a function which is differentiable on the interval \( a \leq x \leq b \), then there exists a point \( \xi \), strictly within this interval, such that the following exact equality holds

\[ f(b) - f(a) = f'(\xi)(b - a). \]  

(22)

The geometric interpretation of this "mean-value theorem" (also called Lagrange's formula or the finite-difference formula) is extraordinarily simple. Let \( A, B \) be the points on the graph of the function \( f(x) \) which correspond to \( x = a \) and \( x = b \), and let us join \( A \) and \( B \) by the chord \( AB \) (figure 23). Now let us move the straight line \( AB \), keeping it constant parallel to itself, up or down. At the moment when this straight line cuts the graph for the last time, it will be tangent to the graph at a certain point \( C \). At this point (let the corresponding abscissa be \( x = \xi \)), the tangent line will form the same angle of inclination \( \alpha \) as the chord \( AB \). But for the chord we have

\[ \tan \alpha = \frac{f(b) - f(a)}{b - a}. \]

On the other hand at the point \( C \)

\[ \tan \alpha = f'(\xi). \]
This equation
\[
\frac{f(b) - f(a)}{b - a} = f'(\xi)
\]
is exactly the mean-value theorem. *

Formula (22) has the peculiar feature that the point \( \xi \) appearing in it is unknown to us; we know only that it lies "somewhere in the interval \((a, b)\)." But in spite of this indeterminacy, the formula has great theoretical significance and is part of the proof of many theorems in analysis. The immediate practical importance of this formula is also very great, since it enables us to estimate the increase in a function when we know the limits between which its derivative can vary. For example,
\[
|\sin b - \sin a| = |\cos \xi| (b - a) \leq b - a.
\]

Here \(a, b\) and \(\xi\) are angles, expressed in radian measure; \(\xi\) is some value between \(a\) and \(b\); \(\xi\) itself is unknown, but we know that \(|\cos \xi| \leq 1\).

From formula (22) it is clear that a function whose derivative is everywhere equal to zero must be a constant; at no part of the interval can it receive an increment different from zero. Analogously, the reader will easily prove that a function whose derivative is everywhere positive must everywhere increase, and if its derivative is negative, the function must decrease. We give here without proof one of the many generalizations of the mean-value theorem.

For arbitrary functions \(\phi(x)\) and \(\psi(x)\) differentiable in the interval \([a, b]\), provided only that \(\psi'(x) \neq 0\) in \((a, b)\), the following equation\(^*\) holds
\[
\frac{\phi(b) - \phi(a)}{\psi(b) - \psi(a)} = \frac{\phi'(\xi)}{\psi'(\xi)}, \tag{23}
\]
where \(\xi\) is some point in the interval \((a, b)\). \(^\dagger\)

From this theorem we can derive a general method for calculating the limits of an expression like
\[
\lim_{x \to a} \frac{\phi(x)}{\psi(x)}, \tag{24}
\]

\(^*\) Of course these arguments only give a geometric interpretation of the theorem and by no means form a rigorous proof.

\(^\dagger\) Formula (23) can be derived by a simple application of the mean-value theorem to the function
\[
f(z) = \phi(x) - \frac{\phi(b) - \phi(a)}{\phi(b) - \phi(a)} \psi(x).
\]

\(^2\) By the symbols \([a, b]\) and \((a, b)\) we denote the sets of values of \(x\) satisfying the inequalities \(a \leq x \leq b\) and \(a < x < b\) respectively.

§8. Taylor's Formula

If \(\phi(0) = \psi(0) = 0\). From formula (23) we have
\[
\frac{\phi(x)}{\psi(x)} = \frac{\phi(x) - \phi(0)}{\psi(x) - \psi(0)} = \frac{\phi'(\xi)}{\psi'(\xi)},
\]
where \(\xi\) is between 0 and \(x\), and therefore \(\xi \to 0\) together with \(x\). This allows us to calculate the limit
\[
\lim_{x \to 0} \frac{\phi(x)}{\psi(x)},
\]
instead of the limit (24), which is in many cases very much easier. *

Example. Let us find the limit \(\lim_{x \to 0} \frac{x - \sin x}{x^3}\). By making use of the rule three times, we have successively
\[
\lim_{x \to 0} \frac{x - \sin x}{x^3} = \lim_{x \to 0} \frac{1 - \cos x}{3x^2} = \lim_{x \to 0} \frac{\sin x}{6x} = \lim_{x \to 0} \frac{\cos x}{6} = \frac{1}{6}.
\]

§9. Taylor's Formula

The function
\[
p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n,
\]
where the coefficients \(a_k\) are constants, is called a polynomial of degree \(n\). In particular, \(y = ax + b\) is a polynomial of the first degree and \(y = ax^2 + bx + c\) is a polynomial of the second degree. Polynomials may be considered as the simplest of all functions. In order to calculate their value for a given \(x\), we require only the operations of addition, subtraction, and multiplication; not even division is needed. Polynomials are continuous for all \(x\) and have derivatives of arbitrary order. Also, the derivative of a polynomial is again a polynomial, of degree lower by one, and the derivatives of order \(n + 1\) and higher of a polynomial of degree \(n\) are equal to zero.

If to the polynomials we adjoint functions of the form
\[
y = \frac{a_0 + a_1 x + \cdots + a_n x^n}{b_0 + b_1 x + \cdots + b_m x^m},
\]
* The same rule is valid for finding the limit of a fractional expression in which the numerator and the denominator both approach infinity. This method, which is very convenient for finding such limits (or, as we say, for the removal of indeterminacies), will be used, for example, in §5 of Chapter XII.
for the calculation of which we also need division, and also the functions \( \sqrt{x} \) and \( \sqrt[3]{x} \), and, finally, arithmetical combinations of these functions, we obtain essentially all the functions whose values can be calculated by methods learned in the secondary school.

While we were still in school, we formed some notion of a number of other functions, like

\[ \sqrt{x}, \log x, \sin x, \arctan x, \ldots. \]

But though we became acquainted with the most important properties of these functions, we found no answer in elementary mathematics to the question: How can we calculate them? What sort of operations, for example, is it necessary to perform on \( x \) in order to obtain \( \log x \) or \( \sin x \)? The answer to this question is given by methods that have been worked out in analysis. Let us examine one of these methods.

**Taylor’s formula.** On an interval containing the point \( a \), let there be given a function \( f(x) \) with derivatives of every order. The polynomial of first degree

\[ p_1(x) = f(a) + f'(a)(x - a) \]

has the same value as \( f(x) \) at the point \( x = a \) and also, as is easily verified, has the same derivative as \( f(x) \) at this point. Its graph is a straight line, which is tangent to the graph of \( f(x) \) at the point \( a \). It is possible to choose a polynomial of the second degree, namely

\[ p_2(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2, \]

which at the point of \( x = a \) has with \( f(x) \) a common value and a common first and second derivative. Its graph at the point \( a \) will follow that of \( f(x) \) even more closely. It is natural to expect that if we construct a polynomial which at \( x = a \) has the same first \( n \) derivatives as \( f(x) \) at the same point, then this polynomial will be a still better approximation to \( f(x) \) at points \( x \) near \( a \). Thus we obtain the following approximate equality, which is Taylor’s formula

\[ f(x) \approx f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n. \tag{25} \]

The right side of this formula is a polynomial of degree \( n \) in \( (x - a) \). For each \( x \) the value of this polynomial can be calculated if we know the values of \( f(a), f'(a), \ldots, f^{(n)}(a) \).

§9. **Taylor’s formula**

For functions which have an \((n + 1)\)th derivative, the right side of this formula, as is easy to show, differs from the left side by a small quantity which approaches zero more rapidly than \((x - a)^n \). Moreover, it is the only possible polynomial of degree \( n \) that differs from \( f(x) \), for \( x \) close to \( a \), by a quantity that approaches zero, as \( x \to a \), more rapidly than \((x - a)^n \). If \( f(x) \) itself is an algebraic polynomial of degree \( n \), then the approximate equality (25) becomes an exact one.

Finally, and this is particularly important, we can give a simple expression for the difference between the right side of formula (25) and the actual value of \( f(x) \). To make the approximate equality (25) exact, we must add to the right side a further term, called the “remainder term”

\[ f(x) = f(a) + f'(a)(x - a) + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n + \frac{f^{(n+1)}(\xi)}{(n + 1)!}(x - a)^{n+1} \tag{26} \]

This final supplementary term*

\[ R_{n+1}(x) = \frac{f^{(n+1)}(\xi)}{(n + 1)!}(x - a)^{n+1} \]

has the peculiarity that the derivative appearing in it is to be calculated in each case not at the point \( a \) but at a suitably chosen point \( \xi \), which is unknown but lies somewhere in the interval between \( a \) and \( x \).

The proof of equality (26) is rather cumbersome but quite simple in essence. We shall give here a somewhat artificial version of the proof which has the merit of being concise.

In order to find out by how much the left side in the approximate formula (25) differs from the right, let us consider the ratio of the difference between the two sides in equality (25) to the quantity \(-(x - a)^{n+1}\)

\[ \frac{f(x) - \left[ f(a) + f'(a)(x - a) + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n \right]}{-(x - a)^{n+1}} \tag{27} \]

We also introduce the function

\[ \phi(u) = f(u) + f'(u)(x - u) + \cdots + \frac{f^{(n)}(u)}{n!}(x - u)^n \]

of a variable \( u \), taking \( x \) to be fixed (constant). Then the numerator in (27) will represent the increase of this function as we pass from \( u = a \) to \( u = x \), and the denominator will be the increase over the same interval of the function

\[ \phi(u) = (x - u)^{n+1}. \]

* This is only one of the possible forms for the remainder term \( R_{n+1}(x) \).
II. ANALYSIS

We now make use of the generalized mean-value theorem quoted earlier

\[ \frac{\phi(x) - \phi(a)}{\psi(x) - \psi(a)} = \frac{\phi'()}{\psi'()} \]

Differentiating the functions \(\phi(u)\) and \(\psi(u)\) with respect to \(u\) (it must be recalled that the value of \(x\) has been fixed) we find that

\[ \frac{\phi''(\xi)}{\psi''(\xi)} = -\frac{f^{(n+1)}(\xi)}{(n + 1)!} \]

The equality of this last expression with the original quantity \((27)\) gives Taylor's formula in the form \((26)\).

In the form \((26)\) Taylor's formula not only provides a means of approximate calculation of \(f(x)\) but also allows us to estimate the error. Let us consider the simple example

\[ y = \sin x \]

The values of the function \(\sin x\) and of its derivatives of arbitrary order are known for \(x = 0\). Let us make use of these values to write Taylor's formula for \(\sin x\), choosing \(a = 0\) and limiting ourselves to the case \(n = 4\). We find successively

\[
\begin{align*}
    f(x) &= \sin x, \\
    f'(x) &= \cos x, \\
    f''(x) &= -\sin x, \\
    f'''(x) &= -\cos x, \\
    f''''(x) &= \sin x, \\
    f''''(0) &= 1, \\
    f''''(0) &= -1.
\end{align*}
\]

Therefore

\[ \sin x = x - \frac{x^3}{6} + R_5, \quad \text{where} \quad R_5 = \frac{x^5}{120} \cos \xi. \]

Although the exact value \(R_5\) is unknown, still we can easily estimate it from the fact that \(|\cos \xi| \leq 1\). For all values of \(x\) between 0 and \(\pi/4\) we have

\[ |R_5| < \frac{\pi^5}{120} < \frac{1}{400}. \]

Consequently, on the interval \([0, \pi/4]\) the function \(\sin x\) may be considered, with accuracy up to \(\frac{1}{400}\), as equal to the polynomial of third degree

\[ \sin x = x - \frac{1}{6} x^3. \]

§9. TAYLOR'S FORMULA

If we were to take more terms in Taylor's expansion for \(\sin x\), we would obtain a polynomial of higher degree which would approximate \(\sin x\) still more closely.

The tables for trigonometric and other functions are calculated by similar methods.

The laws of nature, as a rule, can be expressed with good approximation by functions that may be differentiated as often as we like and that in their turn may be approximated by polynomials, the degree of the polynomial being determined by the accuracy desired.

Taylor's series. If in formula \((25)\) we take a larger and larger number of terms, then the difference between the right side and \(f(x)\), expressed by the remainder term \(R_n(x)\), may tend to zero. Of course this will not always occur; neither for all functions nor for all values of \(x\). But there exists a broad class of functions (the so-called analytic functions) for which the remainder term \(R_n(x)\) does in fact approach zero as \(n \to \infty\) at least for all values of \(x\) within a certain interval around the point \(a\).

For these functions the Taylor formula allows us to calculate \(f(x)\) with any desired degree of accuracy. Let us examine such functions more closely.

If \(R_{n+1}(x) \to 0\) as \(n \to \infty\), then from \((25)\) it follows that

\[ f(x) = \lim_{n \to \infty} \left[ f(a) + f'(a)(x - a) + \cdots + \frac{f^{(n)}(a)}{n!} (x - a)^n \right] \]

In this case we say that \(f(x)\) has been expanded in a convergent infinite series

\[ f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!} (x - a)^2 + \cdots, \]

in increasing powers of \((x - a)\). This series is called a Taylor series, and \(f(x)\) is said to be the sum of the series. Let us consider some examples (with \(a = 0\)):

1. \((1 + x)^n = 1 + nx + \frac{n(n - 1)}{2!} x^2 + \frac{n(n - 1)(n - 2)}{3!} x^3 + \cdots \quad \text{(valid for } |x| < 1 \text{ and for arbitrary real } n)\).

2. \[ \sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \quad \text{(valid for all } x). \]
II. ANALYSIS

3. \[ \cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \] (valid for all \( x \)).

4. \[ e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \] (valid for all \( x \)).

5. \[ \arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots \] (valid for \( |x| \leq 1 \)).

The first of these examples is the famous binomial theorem of Newton, which was obtained by Newton for all \( n \) but completely proved in his time only for integral \( n \). This example served as a model for the establishment of the general Taylor formula. The last two formulas allow us, for \( x = 1 \), to calculate with arbitrarily good approximation the numbers \( e \) and \( \pi \).

The Taylor formula, which opens the way for most of the calculations in applied analysis, is extremely important from the practical point of view.

Many of the laws of nature, physical and chemical processes, the motion of bodies, and the like, are expressed with great accuracy by functions which may be expanded in a Taylor series. The theory of such functions can be formulated in a clearer and more complete way if we consider them as functions of a complex variable (see Chapter IX).

The idea of approximating a function by polynomials or of representing it as the sum of an infinite number of simpler functions underwent far-reaching developments in analysis, where it now forms an independent branch, the theory of approximation of functions (see Chapter XII).

§10. Integral

From Chapter I and from §1 of the present chapter the reader already knows that the concept of the integral, and more generally of the integral calculus, had its historical origin in the need for solving concrete problems, a characteristic example of which is the calculation of the area of a curvilinear figure. The present section is devoted to these questions. In it we will also discuss the aforementioned connection between the problems of the differential and the integral calculus, which was not fully cleared up until the 18th century.

Area. Let us suppose that a curve above the \( x \)-axis forms the graph of the function \( y = f(x) \). We attempt to find the area \( S \) of the segment bounded by the line \( y = f(x) \), by the \( x \)-axis and by the straight lines drawn through the points \( x = a \) and \( x = b \) parallel to the \( y \)-axis.

To solve this problem we divide the interval \([a, b]\) into \( n \) parts, not necessarily equal. We denote the length of the first part by \( \Delta x_1 \), of the second by \( \Delta x_2 \), and so forth up to the final part \( \Delta x_n \). In each segment we choose points \( \xi_1, \xi_2, \ldots, \xi_n \) and set up the sum

\[ S_n = f(\xi_1) \Delta x_1 + f(\xi_2) \Delta x_2 + \cdots + f(\xi_n) \Delta x_n \]  \( \cdots \) (28)

The magnitude \( S_n \) is obviously equal to the sum of the areas of the rectangles shaded in figure 24.

The finer we make the subdivision of the segment \([a, b]\), the closer \( S_n \) will be to the area \( S \). If we carry out a sequence of such constructions, dividing the interval \([a, b]\) into successively smaller and smaller parts, then the sums \( S_n \) will approach \( S \).

The possibility of dividing \([a, b]\) into unequal parts makes it necessary for us to define what we mean by "successively smaller" subdivisions. We assume not only that \( n \) increases beyond all bounds but also that the length of the greatest \( \Delta x_i \), in the \( n \)th subdivision approaches zero. Thus

\[ S = \lim_{\max \Delta x_i \to 0} \left[ f(\xi_1) \Delta x_1 + f(\xi_2) \Delta x_2 + \cdots + f(\xi_n) \Delta x_n \right] \]

\[ = \lim_{\max \Delta x_i \to 0} \sum_{i=1}^{n} f(\xi_i) \Delta x_i \]  \( \cdots \) (29)

The calculation of the desired area has in this way been reduced to finding the limit (29).

We note that when we first set up the problem, we had only an empirical idea of what we mean by the area of our curvilinear figure, but we had no precise definition. But now we have obtained an exact definition of the concept of area: It is the limit (29). We now have not only an intuitive notion of area but also a mathematical definition, on the
basis of which we can calculate the area numerically (compare the remarks at the end of §3, concerning velocity and the length of a circumference).

We have assumed that \( f(x) \geq 0 \). If \( f(x) \) changes sign, then in figure 25, the limit (29) will give us the algebraic sum of the areas of the segments lying between the curve \( y = f(x) \) and the \( x \)-axis, where the segments above the \( x \)-axis are taken with a plus sign and those below with a minus sign.

**Definite integral.** The need to calculate the limit (29) arises in many other problems. For example, suppose that a point is moving along a straight line with variable velocity \( v = f(t) \). How are we to determine the distance \( s \) covered by the point in the time from \( t = a \) to \( t = b \)?

Let us assume that the function \( f(t) \) is continuous; that is, in small intervals of time the velocity changes only slightly. We divide the interval \([a, b]\) into \( n \) parts, of length \( \Delta t_1, \Delta t_2, \ldots, \Delta t_n \). To calculate an approximate value for the distance covered in each interval \( \Delta t_i \), we will suppose that the velocity in this period of time is constant, equal throughout to its actual value at some intermediate point \( \xi_i \). The whole distance covered will then be expressed approximately by the sum

\[
\tilde{s}_n = \sum_{i=1}^{n} f(\xi_i) \Delta t_i,
\]

and the exact value of the distance \( s \) covered in the time from \( a \) to \( b \), will be the limit of such sums for finer and finer subdivisions; that is, it will be the limit (29)

\[
s = \lim_{n \to \infty} \sum_{i=1}^{n} f(\xi_i) \Delta t_i.
\]

It would be easy to give many examples of practical problems leading to the calculation of such a limit. We will discuss some of them later, but for the moment the examples already given will sufficiently indicate the importance of this idea. The limit (29) is called the **definite integral** of the function \( f(x) \) taken over the interval \([a, b]\), and it is denoted by

\[
\int_{a}^{b} f(x) \, dx.
\]

The expression \( f(x) \, dx \) is called the **integrand**, \( a \) and \( b \) are the limits of integration; \( a \) is the lower limit, \( b \) is the upper limit.

---

**§10. INTEGRAL**

The connection between differential and integral calculus. As an example of the direct calculation of a definite integral, we may take example 2, §1. We may now say that the problem considered there reduces to calculation of the definite integral

\[
\int_{0}^{8} ax \, dx.
\]

Another example was considered in §3, where we solved the problem of finding the area bounded by the parabola \( y = x^2 \). Here the problem reduces to calculation of the integral

\[
\int_{0}^{1} x^2 \, dx.
\]

We were able to calculate both these integrals directly, because we have simple formulas for the sum of the first \( n \) natural numbers and for the sum of their squares. But for an arbitrary function \( f(x) \), we are far from being able to add up the sum (28) (that is, to express the result in a simple formula) if the points \( \xi_i \) and the increments \( \Delta x_i \) are given to some particular problem. Moreover, even when such a summation is possible, there is no general method for carrying it out; various methods, each of a quite special character, must be used in the various cases.

So we are confronted by the problem of finding a general method for the calculation of definite integrals. Historically this question interested mathematicians for a long period of time, since there were many practical aspects involved in a general method for finding the area of curvilinear figures, the volume of bodies bounded by a curved surface, and so forth.

We have already noted that Archimedes was able to calculate the area of a segment and of certain other figures. The number of special problems that could be solved, involving areas, volumes, centers of gravity of solids, and so forth, gradually increased, but progress in finding a general method was at first extremely slow. The general method could not be discovered until sufficient theoretical and computational material had been accumulated through the demands of practical life. The work of gathering and generalizing this material proceeded very gradually until the end of the Middle Ages, and its subsequent energetic development was a direct consequence of the rapid growth in the productive powers of Europe resulting from the breakup of the former (feudal) methods of manufacturing and the creation of new ones (capitalistic).

The accumulation of facts connected with definite integrals proceeded alongside of the corresponding investigations of problems related to the derivative of a function. The reader already knows from §1 that this
immense preparatory labor was crowned with success in the 17th century by the work of Newton and Leibnitz. It is in this sense that Newton and Leibnitz are the creators of the differential and integral calculus.

One of the fundamental contributions of Newton and Leibnitz consists of the fact that they finally cleared up the profound connection between differential and integral calculus, which provides us, in particular, with a general method of calculating definite integrals for an extremely wide class of functions.

To explain this connection, we turn to an example from mechanics.

We suppose that a material point is moving along a straight line with velocity \( v = f(t) \), where \( t \) is the time. We already know that the distance \( \sigma \) covered by our point in the time between \( t = t_1 \) and \( t = t_2 \) is given by the definite integral

\[
\sigma = \int_{t_1}^{t_2} f(t) \, dt.
\]

Now let us assume that the law of motion of the point is known to us; that is, we know the function \( s = F(t) \) expressing the dependence on the time \( t \) of the distance \( s \) calculated from some initial point \( A \) on the straight line. The distance \( \sigma \) covered in the interval of time \([t_1, t_2]\) is obviously equal to the difference

\[
\sigma = F(t_2) - F(t_1).
\]

In this way we are led by physical considerations to the equality

\[
\int_{t_1}^{t_2} f(t) \, dt = F(t_2) - F(t_1),
\]

which expresses the connection between the law of motion of our point and its velocity.

From a mathematical point of view the function \( F(t) \), as we already know from §5, may be defined as a function whose derivative for all values of \( t \) in the given interval is equal to \( f(t) \), that is

\[
F'(t) = f(t).
\]

Such a function is called a primitive for \( f(t) \).

We must keep in mind that if the function \( f(t) \) has at least one primitive, then along with this one it will have an infinite number of others; for if \( F(t) \) is a primitive for \( f(t) \), then \( F(t) + C \), where \( C \) is an arbitrary constant, is also a primitive. Moreover, in this way we exhaust the whole set of primitives for \( f(t) \), since if \( F_1(t) \) and \( F_2(t) \) are primitives for the same function \( f(t) \), then their difference \( \phi(t) = F_1(t) - F_2(t) \) has a derivative

\[
\phi(t) = \frac{d}{dt} \phi(t) = \frac{d}{dt} [F_1(t) - F_2(t)] = f(t) - f(t) = 0.
\]

From this formula of Newton and Leibnitz, which reduces the problem of calculating the definite integral of a function to finding a primitive for the function and in this way forms a link between the differential and the integral calculus.

Many particular problems that were studied by the greatest mathematicians are automatically solved by this formula, stating that the definite integral of the function \( f(x) \) on the interval \([a, b]\) is equal to the difference between the values of any primitive at the left and right ends of the interval. It is customary to write the difference (30) thus:

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

where \( F(x) \) is an arbitrary primitive for \( f(x) \).

Example 1. The equality

\[
\left( \frac{x^3}{3} \right)' = x^2
\]

shows that the function \( x^3/3 \) is a primitive for the function \( x^2 \). Thus, by the formula of Newton and Leibnitz,

\[
\int_{0}^{a} x^2 \, dx = \left. \frac{x^3}{3} \right|_{0}^{a} = \frac{a^3}{3} - 0 = \frac{a^3}{3}.
\]

* By the mean value theorem

\[
\phi(t) = \phi(t_0) + \phi'(t)(t - t_0) = 0,
\]

when \( v \) lies between \( t \) and \( t_0 \). Thus \( \phi(t) = \phi(t_0) \) is constant for all \( t \).

† It is possible to prove mathematically, without recourse to examples from mechanics, that if the function \( f(x) \) is continuous (and even if it is discontinuous but Lebesgue-summable; see Chapter XV) on the interval \([a, b]\), then there exists a primitive \( F(x) \) satisfying equality (30).

‡ This formula has been generalized in various ways (see for example §13, the formula of Ostrogradskii).
had not yet completely "broken away" from their physical and geometric
origins, such as velocity and area. In fact, they were half mathematical
in character and half physical. The conditions existing at that time were
not yet suitable for producing a purely mathematical definition of
these concepts. Consequently, the investigator could handle them cor-
crctly in complicated situations only if he remained in close contact
with the practical aspects of his problem even during the intermediate
(mathematical) stages of his argument.

From this point of view the creative work of Newton was different in
character from that of Leibnitz.* Newton was guided at all stages by a
physical way of looking at the problem. But the investigations of Leibnitz
do not have such an immediate connection with physics, a fact that in the
absence of clear-cut mathematical definitions sometimes led him to mis-
taken conclusions. On the other hand, the most characteristic feature
of the creative activity of Leibnitz was his striving for generality, his
efforts to find the most general methods for the problems of mathematical
analysis.

The greatest merit of Leibnitz was his creation of a mathematical
symbolism expressing the essence of the matter. The notations for such
fundamental concepts of mathematical analysis as the differential dx,
the second differential $d^2x$, the integral $\int f \, dx$, and the derivative $\frac{d}{dx}$
were proposed by Leibnitz. The fact that these notations are still used
shows how well they were chosen.

One advantage of a well-chosen symbolism is that it makes our proofs
and calculations shorter and easier; also, it sometimes protects us against
mistaken conclusions. Leibnitz, who was well aware of this, paid especial
attention in all his work to the choice of notation.

The evolution of the concepts of mathematical analysis (derivative,
integral, and so forth) continued, of course, after Newton and Leibnitz
and is still continuing in our day; but there is one stage in this evolution
that should be mentioned especially. It took place at the beginning of the
last century and is related particularly to the work of Cauchy.

Cauchy gave a clear-cut formal definition of the concept of a limit and
used it as the basis for his definitions of continuity, derivative, differential,
and integral. These definitions have been introduced at the corre-
spinding places in the present chapter. They are widely used in present-day
analysis.

The great importance of these achievements lies in the fact that it is
now possible to operate in a purely formal way not only in arithmetic,

* The discoveries of Newton and Leibnitz were made independently.
algebra, and elementary geometry, but also in this new and very extensive branch of mathematics, in mathematical analysis, and to obtain correct results in so doing.

Regarding practical application of the results of mathematical analysis, it is now possible to say: If the original data are verified in the actual world, then the results of our mathematical arguments will also be verified there. If we are properly assured of the accuracy of the original data, then there is no need to make a practical check of the correctness of the mathematical results; it is sufficient to check only the correctness of the formal arguments.

This statement naturally requires the following limitation. In mathematical arguments the original data, which we take from the actual world, are true only up to a certain accuracy. This means that at every step of our mathematical argument the results obtained will contain certain errors, which may accumulate as the number of steps in the argument increases.*

Returning now to the definite integral, let us consider a question of fundamental importance. For what functions \( f(x) \), defined on the interval \([a, b]\), is it possible to guarantee the existence of the definite integral \( \int_a^b f(x) \, dx \), namely a number to which the sum \( \sum_{i=1}^{n} f(\xi_i) \Delta x_i \) tends as limit as \( \Delta x_i \to 0 \)? It must be kept in view that this number is to be the same for all subdivisions of the interval \([a, b]\) and all choices of the points \( \xi_i \).

Functions for which the definite integral, namely the limit (29), exists are said to be integrable on the interval \([a, b]\). Investigations carried out in the last century show that all continuous functions are integrable.

But there are also discontinuous functions which are integrable. Among them, for example, are those functions which are bounded and either increasing or decreasing on the interval \([a, b]\).

The function that is equal to zero at the rational points in \([a, b]\) and equal to unity at the irrational points, may serve as an example of a non-integrable function, since for an arbitrary subdivision the integral sum \( s_n \) will be equal to zero or unity, depending on whether we choose the points \( \xi_i \) as rational numbers or irrational.

Let us note that in many cases the formula of Newton and Leibnitz provides an answer to the practical question of calculating a definite integral. But here arises the problem of finding a primitive for a given function; that is, of finding a function that has the given function for its derivative. We now proceed to discuss this problem. Let us note by the way that the problem of finding a primitive has great importance in other branches of mathematics also, particularly in the solution of differential equations.

§11. Indefinite Integrals; the Technique of Integration

An arbitrary primitive of a given function \( f(x) \) is usually called an indefinite integral of \( f(x) \) and is written in the form

\[
\int f(x) \, dx.
\]

In this way, if \( F(x) \) is a completely determined primitive of \( f(x) \), then the indefinite integral of \( f(x) \) is given by

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

where \( C \) is an arbitrary constant.

Let us also note that if the function \( f(x) \) is given on the interval \([a, b]\) and, if \( F(x) \) is a primitive for \( f(x) \) and \( x \) is a point in the interval \([a, b]\), then by the formula of Newton and Leibnitz we may write

\[
F(x) = F(a) + \int_a^x f(t) \, dt.
\]

Here the integral on the right side differs from the primitive \( F(x) \) only by the constant \( F(a) \). In such a case this integral, if we consider it as a function of its upper limit \( x \) (for variable \( x \)), is a completely determined primitive of \( f(x) \). Consequently, an indefinite integral of \( f(x) \) may also be written as follows:

\[
\int f(x) \, dx = \int_a^x f(t) \, dt + C,
\]

where \( C \) is an arbitrary constant.

Let us set up a fundamental table of indefinite integrals, which can be obtained directly from the corresponding table of derivatives (see §6):
\[ \int x^a \, dx = \frac{x^{a+1}}{a+1} + C (a \neq -1), \]
\[ \int \frac{dx}{x} = \ln |x| + C, ^* \]
\[ \int a^x \, dx = \frac{a^x}{\ln a} + C, \]
\[ \int e^x \, dx = e^x + C, \]
\[ \int \sin x \, dx = -\cos x + C, \]
\[ \int \cos x \, dx = \sin x + C, \]
\[ \int \sec^2 x \, dx = \tan x + C, \]
\[ \int \frac{dx}{\sqrt{1 - x^2}} = \arcsin x + C \]
\[ = -\arccos x + C_1 \left( C_1 - C = \frac{\pi}{2} \right), \]
\[ \int \frac{dx}{1 + x^2} = \arctan x + C. \]

The general properties of indefinite integrals may also be deduced from the corresponding properties of derivatives. For example, from the rule for the differentiation of a sum we obtain the formula

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

and from the corresponding rule expressing the fact that a constant factor \( k \) may be taken outside the sign of differentiation we get

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

For example,

\[ \int \left( 3x^2 + 2x - \frac{3}{\sqrt{x}} + \frac{4}{x} - 1 \right) \, dx \]
\[ = \frac{3x^3}{3} + \frac{2x^2}{2} - 3 \frac{x^{-1/2} + 1}{-\frac{1}{2} + 1} + 4 \ln |x| - x + C. \]

\(^*\) For \( x > 0 \), \( (\ln |x|)' = (\ln x)' = 1/x; \) for \( x < 0 \), \( (\ln |x|)' = [\ln(-x)]' = 1/(-x(-1)) = 1/x. \)

§11. INDEFINITE INTEGRALS

There are a number of methods for calculating indefinite integrals. Let us consider one of them, namely the method of substitution or change of variable, which is based on the following equality

\[ \int f(x) \, dx = \int [\phi(t)] \phi'(t) \, dt + C, \tag{33} \]

where \( x = \phi(t) \) is a differentiable function. The relation (33) is to be understood in the sense that if in the function

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

on the left side of equality (33), we set \( x = \phi(t) \), we thereby obtain a function \( F[\phi(t)] \) whose derivative with respect to \( t \) is equal to the expression under the sign of integration on the right side of equality (33). This fact follows immediately from the theorem on the derivative of a function of a function.

Let us give some examples of this method of substitution

\[ \int e^{tx} \, dx = \int e^{t} \frac{1}{k} \, dt = \frac{1}{k} \int e^{t} \, dt = \frac{1}{k} e^{t} + C = \frac{e^{tx}}{k} + C \]

(substitution of \( kx = t \), from which \( k \, dx = dt \)).

\[ \int \frac{x \, dx}{\sqrt{a^2 - x^2}} = -\int dt = -t + C = -\sqrt{a^2 - x^2} + C \]

(substitution of \( t = \sqrt{a^2 - x^2} \), from which \( dt = -\frac{x \, dx}{\sqrt{a^2 - x^2}} \)).

\[ \int \sqrt{a^2 - x^2} \, dx = \int \sqrt{a^2 - a^2 \sin^2 u} \, a \cos u \, du = a^2 \int \cos^2 u \, du \]
\[ = a^2 \int \frac{1 + \cos 2u}{2} \, du = a^2 \left( \frac{u + \sin 2u}{2} \right) + C \]
\[ = a^2 \left( \arcsin \frac{x}{a} + \frac{x}{a^2} \sqrt{a^2 - x^2} \right) + C \]

(substitution of \( x = a \sin u \)).

As can be seen from these examples, the method of substitution or change of variables greatly extends the class of elementary functions that we are able to integrate; that is, for which we can find primitives.
that are themselves elementary functions. But it must be noted that from
the point of view of actually calculating the result, we are in a much worse
position, generally speaking, with respect to integration than for differ-
entiation.

From §6 we know that the derivative of an arbitrary elementary function
is itself an elementary function, which we may effectively calculate by
making use of the rules of differentiation. But the converse statement
is in general untrue, since there exist elementary functions whose indefinite
integrals are not elementary functions. Examples are \( e^{-x^2} \), \( 1/(\ln x) \), \( (\sin x)/x \)
and so forth. To obtain integrals of these functions we must make use of
approximative methods and also introduce new functions which cannot be
reduced to elementary ones. We can not spend more time here on this
question but must simply note that even in elementary mathematics it is
possible to find many examples in which a direct operation can be carried
out on a certain class of numbers, while the inverse operation can not be
carried out on the same class; thus, a square of an arbitrary rational
number is again a rational number, but the square root of a rational
number is by no means always rational. Analogously, differentiation of
elementary functions produces a function that is again elementary, but
integration may lead us outside the class of elementary functions.

Some of the integrals that cannot be expressed in terms of elementary
functions have great importance in mathematics and its applications.
An example is

\[ \int_0^\pi e^{-t^2} \, dt, \]

which plays a very important role in the theory of probability (see
Chapter XI). Other examples are the integrals

\[ \int_0^\phi \frac{d\theta}{\sqrt{1 - k^2 \sin^2 \theta}}, \quad \text{and} \quad \int_0^\phi \sqrt{1 - k^2 \sin^2 \theta} \, d\theta \quad (k^2 < 1), \]

which are called elliptic integrals of the first and second kind respectively.
We are led to the calculation of these integrals by a large number of
problems in physics (see Chapter V, §1, example 3). Detailed tables of
these integrals for various values of the arguments \( x \) and \( \phi \) have been
calculated by approximate methods but with great accuracy.

It must be emphasized that the proof of the very fact that a given ele-
mentary function cannot be integrated in terms of elementary functions is in
each case quite difficult. Such questions occupied the attention of out-
standing mathematicians in the last century and have played an important
role in the development of analysis. Fundamental results were obtained

\[ \int x^m(a + bx^r)^p \, dx, \]

where \( m, s, \) and \( p \) are rational numbers. Up to his time three relations,
obtained by Newton, were known for the exponents \( m, s, \) and \( p \), which
implied the integrability of this integral in terms of elementary functions.
Čebyšev proved that in all other cases the integral cannot be expressed
in terms of elementary functions.

We introduce here another method of integration, namely integration
by parts. It is based on the formula we already know

\[ (uv)' = uv' + u'v, \]

for the derivative of the product of the functions \( u \) and \( v \). This formula
may also be written

\[ uv' = (uv)' - u'v. \]

Let us now integrate the left and right sides, keeping in mind that

\[ \int (uv)' \, dx = uv + C. \]

We now finally obtain the equality

\[ \int uv' \, dx = uv - \int u'v \, dx, \]

which is also called the formula of integration by parts. We have not written
the constant \( C \) since we may consider that it is included in one of the
indefinite integrals occurring in this equation.

Let us introduce some applications of this formula. Suppose we have
to calculate \( \int xe^x \, dx \). Here we will take \( u = x \) and \( v' = e^x \), and thus
\( u' = 1, \, v = e^x \), and consequently

\[ \int xe^x \, dx = xe^x - \int 1 \cdot e^x \, dx = xe^x - e^x + C. \]

In the integral \( \int \ln x \, dx \) it is convenient to take \( u = \ln x, \, v' = 1 \), so
that \( u' = 1/x, \, v = \ln x \) and

\[ \int \ln x \, dx = x \ln x - \int dx = x \ln x - x + C. \]
In the following characteristic example it is necessary to integrate twice by parts and then to find the desired integral from the equations so obtained:

\[ \int e^x \sin x \, dx = e^x \sin x - \int e^x \cos x \, dx \]

\[ = e^x \sin x - e^x \cos x - \int e^x \sin x \, dx, \]

from which

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

We end this section here; from it the reader will have obtained only a superficial idea of the theory of integration. We have not given any attention to many different methods in this theory. In particular we have not touched here on the very interesting question of the integration of rational fractions, a theory in which an important contribution was made by the well-known mathematician and mechanician, Ostrogradskii.

§12. Functions of Several Variables

Up to now we have spoken only of functions of one variable, but in practice it is often necessary to deal also with functions depending on two, three, or in general many variables. For example, the area of a rectangle is a function

\[ S = xy \]

of its base x and its height y. The volume of a rectangular parallelepiped is a function

\[ v = xyz \]

of its three dimensions. The distance between two points A and B is a function

\[ r = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \]

of the six coordinates of these points. The well-known formula

\[ pv = RT \]

expresses the dependence of the volume v of a definite amount of gas on the pressure p and absolute temperature T.

§12. Functions of Several Variables

Functions of several variables, like functions of one variable, are in many cases defined only on a certain region of values of the variables themselves. For example, the function

\[ u = \ln(1 - x^2 - y^2 - z^2) \]

is defined only for values of x, y and z that satisfy the condition

\[ x^2 + y^2 + z^2 < 1. \]

(For other x, y, z its values are not real numbers.) The set of points of space whose coordinates satisfy the inequality (35) obviously fills up a sphere of unit radius with its center at the origin of coordinates. The points on the boundary are not included in this sphere; the surface of the sphere has been so to speak "peeled off." Such a sphere is said to be open. The function (34) is defined only for such sets of three numbers (x, y, z) as are coordinates of points in the open sphere G. It is customary to state this fact concisely by saying that the function (34) is defined on the sphere G.

Let us give another example. The temperature of a nonuniformly heated body \( V \) is a function of the coordinates x, y, z of the points of the body. This function is not defined for all sets of three numbers x, y, z but only for such sets as are coordinates of points of the body \( V \).

Finally, as a third example, let us consider the function

\[ u = \phi(x) + \phi(y) + \phi(z), \]

where \( \phi \) is a function of one variable defined on the interval [0, 1]. Obviously the function u is defined only for sets of three numbers (x, y, z) which are coordinates of points in the cube:

\[ 0 \leq x \leq 1, \quad 0 \leq y \leq 1, \quad 0 \leq z \leq 1. \]

We now give a formal definition of a function of three variables. Suppose that we are given a set \( E \) of triples of numbers (x, y, z) (points of space). If to each of these triples of numbers (points) of \( E \) there corresponds a definite number u in accordance with some law, then u is said to be a function of x, y, z (of the point), defined on the set of triples of numbers (on the points) E, a fact which is written thus:

\[ v = F(x, y, z). \]

In place of \( F \) we may also write other letters: \( f, \phi, \psi \).
II. **ANALYSIS**

In practice the set $E$ will usually be a set of points, filling out some geometrical body or surface: sphere, cube, annulus, and so forth, and then we simply say that the function is defined on this body or surface. Functions of two, four, and so forth, variables are defined analogously.

**Implicit definition of a function.** Let us note that functions of two variables may serve, under certain circumstances, as a useful means for the definition of functions of one variable. Given a function $F(x, y)$ of two variables let us set up the equation

$$F(x, y) = 0. \quad (36)$$

In general, this equation will define a certain set of points $(x, y)$ of the surface on which our function is equal to zero. Such sets of points usually represent curves that may be considered as the graphs of one or several one-valued functions $y = \phi(x)$ or $x = \psi(y)$ of one variable. In such a case these one-valued functions are said to be defined implicitly by the equation (36). For example, the equation

$$x^2 + y^2 - r^2 = 0$$

gives an implicit definition of two functions of one variable

$$y = +\sqrt{r^2 - x^2} \quad \text{and} \quad y = -\sqrt{r^2 - x^2}.$$ 

But it is necessary to keep in mind that an equation of the form (36) may fail to define any function at all. For example, the equation

$$x^2 + y^2 + 1 = 0$$

obviously does not define any real function, since no pair of real numbers satisfies it.

**Geometric representation.** Functions of two variables may always be visualized as surfaces by means of a system of space coordinates. Thus the function

$$z = f(x, y) \quad (37)$$

is represented in a three-dimensional rectangular coordinate system by a surface, which is the geometric locus of points $M$ whose coordinates $x, y, z$ satisfy equation (37) (figure 26).

There is another, extremely useful method, of representing the function (37), which has found wide application in practice. Let us choose a sequence of numbers $z_1, z_2, \ldots$, and then draw on one and the same plane $Oxy$ the curves

$$z_1 = f(x, y), \quad z_2 = f(x, y), \quad \ldots$$

which are the so-called level lines of the function $f(x, y)$. From a set of level lines, if they correspond to values of $z$ that are sufficiently close to one another, it is possible to form a very good opinion of the variation of the function $f(x, y)$, just as from the level lines of a topographical map one may judge the variation in altitude of the locality.

Figure 27 shows a map of the level lines of the function $z = x^2 + y^2$, the diagram at the right indicating how the function is built up from its level lines. In Chapter III, figure 50, a similar map is drawn for the level lines of the function $z = xy$.

**Partial derivatives and differential.** Let us make some remarks about the differentiation of the functions of several variables. As an example we take the arbitrary function

$$z = f(x, y)$$
II. ANALYSIS

of two variables. If we fix the value of \( y \), that is if we consider it as not varying, then our function of two variables becomes a function of the one variable \( x \). The derivative of this function with respect to \( x \), if it exists, is called the partial derivative with respect to \( x \) and is denoted thus:

\[
\frac{\partial z}{\partial x}, \quad \text{or} \quad \frac{\partial f}{\partial x}, \quad \text{or} \quad f'_x(x, y).
\]

The last of these three notations indicates clearly that the partial derivative with respect to \( x \) is in general a function of \( x \) and \( y \). The partial derivative with respect to \( y \) is defined similarly.

Geometrically the function \( f(x, y) \) represents a surface in a rectangular three-dimensional system of coordinates. The corresponding function of \( x \) for fixed \( y \) represents a plane curve (figure 28) obtained from the intersection of the surface with a plane parallel to the plane \( Oxy \) and at a distance \( y \) from it. The partial derivative \( \partial z/\partial x \) is obviously equal to the trigonometric tangent of the angle between the tangent to the curve at the point \((x, y)\) and the positive direction of the \( x \)-axis.

More generally, if we consider a function \( z = f(x_1, x_2, \ldots, x_n) \) of the \( n \) variables \( x_1, x_2, \ldots, x_n \), the partial derivative \( \partial z/\partial x_i \) is defined as the derivative of this function with respect to \( x_i \), calculated for fixed values of the other variables:

\[x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n.\]

We may say that the partial derivative of a function with respect to the variable \( x_i \) is the rate of change of this function in the direction of the change in \( x_i \). It would also be possible to define a derivative in an arbitrary assigned direction, not necessarily coinciding with any of the coordinate axis, but we will not take the time to do this.

Examples.

1. \( z = \frac{x}{y} \cdot \frac{\partial z}{\partial x} = \frac{1}{y} \cdot \frac{\partial z}{\partial y} = -\frac{x}{y^2} \).

§12. FUNCTIONS OF SEVERAL VARIABLES

2. \( u = \frac{1}{\sqrt{x^2 + y^2 + z^2}} \).

\[
\frac{\partial u}{\partial x} = -\frac{1}{x^2 + y^2 + z^2} - \frac{2x}{2\sqrt{x^2 + y^2 + z^2}} = -\frac{x}{(x^2 + y^2 + z^2)^{3/2}}.
\]

It is sometimes necessary to form the partial derivatives of these partial derivatives; that is, the so-called partial derivatives of second order. For functions of two variables there are four of them

\[
\frac{\partial^2 u}{\partial x^2}, \quad \frac{\partial^2 u}{\partial x \partial y}, \quad \frac{\partial^2 u}{\partial y \partial x}, \quad \frac{\partial^2 u}{\partial y^2}.
\]

However, if these derivatives are continuous, then it is not hard to prove that the second and third of these four (the so-called mixed derivatives) coincide:

\[
\frac{\partial^2 u}{\partial x \partial y} = \frac{\partial^2 u}{\partial y \partial x}.
\]

For example, in the case of first function considered,

\[
\frac{\partial^2 z}{\partial x^2} = 0, \quad \frac{\partial^2 z}{\partial x \partial y} = -\frac{1}{y^2}, \quad \frac{\partial^2 z}{\partial y^2} = -\frac{1}{y^2}, \quad \frac{\partial^2 z}{\partial y^2} = \frac{2x}{y^3},
\]

the two mixed derivatives are seen to coincide.

For functions of several variables, just as was done for functions of one variable, we may introduce the concept of a differential. For definiteness let us consider a function

\[z = f(x, y)\]

of two variables. If it has continuous partial derivatives, we can prove that its increment

\[\Delta z = f(x + \Delta x, y + \Delta y) - f(x, y),\]

corresponding to the increments \( \Delta x \) and \( \Delta y \) of its arguments, may be put in the form

\[\Delta z = \frac{\partial f}{\partial x} \Delta x + \frac{\partial f}{\partial y} \Delta y + \alpha \Delta x^2 + \beta \Delta y^2,\]

where \( \partial f/\partial x \) and \( \partial f/\partial y \) are the partial derivatives of the function at the point \((x, y)\) and the magnitude \( \alpha \) depends on \( \Delta x \) and \( \Delta y \) in such a way that \( \alpha \to 0 \) as \( \Delta x \to 0 \) and \( \Delta y \to 0 \).
The sum of the first two components
\[ dz = \frac{\partial f}{\partial x} \Delta x + \frac{\partial f}{\partial y} \Delta y \]
is linearly dependent* on \( \Delta x \) and \( \Delta y \) and is called the differential of the function. The third summand, because of the presence of the factor \( \alpha \), tending to zero with \( \Delta x \) and \( \Delta y \), is an infinitesimal of higher order than the magnitude
\[ \rho = \sqrt{\Delta x^2 + \Delta y^2}, \]
describing the change in \( x \) and \( y \).

Let us give an application of the concept of differential. The period of oscillation of a pendulum is calculated from the formula
\[ T = 2\pi \sqrt{\frac{l}{g}}, \]
where \( l \) is its length and \( g \) is the acceleration of gravity. Let us suppose that \( l \) and \( g \) are known with errors respectively equal to \( \Delta l \) and \( \Delta g \). Then the error in the calculation of \( T \) will be equal to the increment \( \Delta T \) corresponding to the increments of the arguments \( \Delta l \) and \( \Delta g \). Replacing \( \Delta T \) approximately by \( dT \), we will have
\[ \Delta T \approx dT = \pi \left( \frac{\Delta l}{\sqrt{lg}} - \frac{\sqrt{l} \Delta g}{\sqrt{g^3}} \right). \]

The signs of \( \Delta l \) and \( \Delta g \) are unknown, but we may obviously estimate \( \Delta T \) by the inequality
\[ |\Delta T| < \pi \left( \frac{|\Delta l|}{\sqrt{lg}} + \frac{\sqrt{l} |\Delta g|}{\sqrt{g^3}} \right), \]
from which after division by \( T \) we get
\[ \frac{|\Delta T|}{T} < \left( \frac{|\Delta l|}{l} + \frac{|\Delta g|}{g} \right). \]
Thus we may consider in practice that the relative error for \( T \) is equal to the sum of the relative errors for \( l \) and \( g \).

§12. FUNCTIONS OF SEVERAL VARIABLES

For symmetry of notation, the increments of the independent variables \( \Delta x \) and \( \Delta y \) are usually denoted by the symbols \( dx \) and \( dy \) and are also called differentials. With this notation the differential of the function \( u = f(x, y, z) \) may be written thus:
\[ du = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz. \]

Partial derivatives play a large role whenever we have to do with functions of several variables, as happens in many of the applications of analysis to technology and physics. We shall be dealing in Chapter VI with the problem of reconstructing a function from the properties of its partial derivatives.

In the following paragraphs, we give some simple examples of applications of partial derivatives in analysis.

**Differentiation of implicit functions.** Suppose we wish to find the derivative of \( y \), where \( y \) is a function of \( x \) defined implicitly by the relation
\[ F(x, y) = 0 \]
between these variables. If \( x \) and \( y \) satisfy the relation \( (38) \) and we give \( x \) the increment \( \Delta x \), then \( y \) will receive an increment \( \Delta y \) such that \( x + \Delta x \) and \( y + \Delta y \) again satisfy \( (38) \). Consequently* \( F(x + \Delta x, y + \Delta y) - F(x, y) = \frac{\partial F}{\partial x} \Delta x + \frac{\partial F}{\partial y} \Delta y + \alpha \sqrt{\Delta x^2 + \Delta y^2} = 0. \)

Thus, provided \( \frac{\partial F}{\partial y} \neq 0 \), it follows that
\[ \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = y'_x = \frac{\frac{\partial F}{\partial x}}{\frac{\partial F}{\partial y}}. \]
In this way we have obtained a method for finding the derivative of an implicit function \( y \) without first solving the equation \( (38) \) for \( y \).

**Maximum and minimum problems.** If a function, let us say of two variables \( z = f(x, y) \), attains its maximum at the point \((x_0, y_0)\), that is if \( f(x_0, y_0) \geq f(x, y) \) for all points \((x, y)\) close to \((x_0, y_0)\), then this point must also be the point of maximum altitude for any line formed by the

* In general a function \( Ax + By + C \), where \( A, B, C \) are constants, is called a linear function of \( x \) and \( y \). If \( C = 0 \), it is called a homogeneous linear function. Here we omit the word "homogeneous."*

* We assume that \( F(x, y) \) has continuous derivatives with respect to \( x \) and \( y \).*
intersection of the surface \( z = f(x, y) \) with a plane parallel to \( Oxz \) or \( Oyz \). So at such a point we must have

\[
j'_x(x, y) = 0, j'_y(x, y) = 0. \tag{39}\]

The same equations must also hold for a point of local minimum. Consequently, the greatest or least values of the function are to be sought first of all at points where the conditions (39) are satisfied, but we must also not forget about points on the boundary of the domain of definition of the function and points where the function fails to have a derivative, if such points exist.

To establish whether a point \((x, y)\) satisfying (39) is actually a maximum or minimum point, use is frequently made of various indirect arguments. For example, if for any reason it is clear that the function is differentiable and attains its minimum inside the region and that there is only one point where the conditions (39) are fulfilled, then obviously the minimum must be attained at this point.

For example, let it be required to make a rectangular tin box (without lid) with assigned volume \( V \), using the smallest possible amount of material. If the sides of the base of this box are denoted by \( x \) and \( y \), then its height \( h \) will be equal to \( V/xy \), and consequently the surface \( S \) will be given by the function

\[
S = xy + \frac{V}{xy} (2x + 2y) = xy + 2V \left(\frac{1}{x} + \frac{1}{y}\right) \tag{40}\]

of \( x \) and \( y \). Since \( x \) and \( y \) by the terms of the problem must be positive, the question has been reduced to finding the minimum of the function \( S(x, y) \) for all possible points \((x, y)\) in the first quadrant of the plane \((x, y)\), which we will denote by the letter \( G \).

If the minimum is attained at some point of the region \( G \), then the partial derivatives must be equal to zero

\[
\frac{\partial S}{\partial x} = y - 2V\frac{1}{x^2} = 0,
\]

\[
\frac{\partial S}{\partial y} = x - 2V\frac{1}{y^2} = 0,
\]

that is \( xy^2 = 2V \), \( xy^2 = 2V \), from which we find as the dimensions of the box:

\[
x = y = \sqrt[3]{2V} \quad \text{and} \quad h = \frac{\sqrt{2V}}{\sqrt{4}}. \tag{41}\]

We have solved the problem but have not altogether proved that our solution is correct. A rigorous mathematician will say to us: "You have supposed from the very beginning that under the given conditions the box with minimum surface actually exists and, proceeding from this assumption, you have found its dimensions. So you have really obtained only the following result: If there exists a point \((x, y)\) in \( G \) for which the function \( S \) attains its minimum, then the coordinates of this point must necessarily be determined by the equation (41). But now you must show that the minimum of \( S \) does exist for some point in \( G \) and then \( I \) will admit the correctness of your result." This remark is a very reasonable one, since, for example, our function \( S \), as we shall soon see, does not possess any maximum in the region \( G \). But let us show how it is possible to convince ourselves that in the given case the function actually does attain its minimum at a certain point \((x, y)\) of the region \( G \).

The fundamental theorem on which we shall base our argument is one that is proved in analysis with complete rigor; it amounts to the following. If the function \( f \) of one or several variables is everywhere continuous in a certain finite region \( H \) which is bounded and includes its boundary, then there always exists in \( H \) at least one point at which the function attains its minimum (maximum). With this theorem we can easily complete our analysis of the problem.

Let us consider an arbitrary point \((x_0, y_0)\) of the region \( G \); at this point let \( S(x_0, y_0) = N \). Let us also choose a number \( R \) satisfying the two inequalities \( R > N \), \( 2VR > N \) and construct a square \( \Omega_R \) with side \( R \), as in figure 29, where \( AB = CD = 1/R \).

We now give a lower bound for the values of our function \( S(x, y) \) at points of the region \( G \) lying outside the square \( \Omega_R \). If the point of the region \( G \) has abscissa \( x < 1/R \), then

\[
S(x, y) = xy + 2V \left(\frac{1}{x} + \frac{1}{y}\right) > 2V \frac{1}{x} > 2VR > N.
\]

Analogously, if the point of the region \( G \) has its ordinate \( y < 1/R \), then also \( S > N \). Also, if the point of the region \( G \) has its abscissa \( x > 1/R \)
\textbf{II. ANALYSIS}

and if it lies above the straight line $AF$ or has its ordinate $y > 1/R$ and lies to the right of the straight line $CE$, then

$$S(x, y) > xy > \frac{1}{R} R^2 = R > N.$$ 

Thus, for all points $(x, y)$ of the region $G$ lying outside the square $Q_2$, the inequality $S(x, y) > N$ holds, and since $S(x_0, y_0) = N$, the point $(x_0, y_0)$ must belong to the square and consequently the minimum of our function on $G$ is equal to its minimum on the square.

But the function $S(x, y)$ is continuous in this square and on its boundary, so that by the theorem stated earlier there exists in the square a point $(x, y)$ where our function assumes its minimum for points in the square and consequently for the entire region $G$. Thus the existence of a minimum has been proved.

This argument may serve as an example of the way that it is possible to discuss the existence of a maximum or a minimum for a function defined on an unbounded domain.

\textbf{The Taylor formula.} Like functions of one variable, functions of several variables may be represented by a Taylor formula. For example, an expansion of the function

$$u = f(x, y)$$

in the neighborhood of the point $(x_0, y_0)$ has the following form, if we confine ourselves to the first and second powers of $x - x_0$ and $y - y_0$:

$$f(x, y) = f(x_0, y_0) + [f'_x(x_0, y_0)(x - x_0) + f'_y(x_0, y_0)(y - y_0)] + \frac{1}{2!} \left[ f''_{xx}(x_0, y_0)(x - x_0)^2 + 2f''_{xy}(x_0, y_0)(x - x_0)(y - y_0) + f''_{yy}(x_0, y_0)(y - y_0)^2 \right] + R_3.$$

If the function $f(x, y)$ has continuous partial derivatives of the second order, the remainder term here will approach zero faster than

$$r^2 - (x - x_0)^2 + (y - y_0)^2,$$

that is, faster than the square of the distance between the points $(x, y)$ and $(x_0, y_0)$, as $r \to 0$. The Taylor formula provides a widely used method of defining and approximately calculating the values of various functions.

Let us note that with the help of this formula we can also answer the question asked earlier, whether a given function actually has a maximum or minimum at a point where \( \partial f / \partial x = \partial f / \partial y = 0 \). In fact, if these conditions are satisfied at the point $(x_0, y_0)$, then for points $(x, y)$ close to $(x_0, y_0)$, the value of the function will, by the Taylor formula, differ from $f(x_0, y_0)$ by the amount

$$f(x, y) - f(x_0, y_0) = \frac{1}{2!} \left[ A(x - x_0)^2 + 2B(x - x_0)(y - y_0) + C(y - y_0)^2 \right] + R_3,$$

where $A$, $B$, and $C$ denote respectively the second partial derivatives $f''_{xx}, f''_{xy}, f''_{yy}$ at the point $(x_0, y_0)$.

If it turns out that the function

$$\Phi(x, y) = A(x - x_0)^2 + 2B(x - x_0)(y - y_0) + C(y - y_0)^2$$

is positive for arbitrary values of $(x - x_0)$ and $(y - y_0)$ not both equal to zero, then the right side of equation (42) will also be positive for small values of $(x - x_0)$ and $(y - y_0)$, since for sufficiently small $(x - x_0)$ and $(y - y_0)$ the quantity $R_3$ is known to be less in absolute value than $\frac{1}{2} \Phi(x, y)$. Thus it will follow that at the point $(x_0, y_0)$ the function $f$ attains its minimum. On the other hand, if the function $\Phi(x, y)$ is negative for arbitrary $(x - x_0)$ and $(y - y_0)$ the right side of (42) will be negative for $(x - x_0)$ and $(y - y_0)$, so that at the point $(x_0, y_0)$ the function will have a maximum. In more complicated cases it is necessary to consider the succeeding terms in the Taylor formula.

Problems concerning the maximum or the minimum of functions of three or more variables may be treated in a completely analogous fashion. As an exercise the reader may prove that if given masses

$$m_1, m_2, \ldots, m_n$$

are arranged in space at given points

$$P_1(x_1, y_1, z_1), P_2(x_2, y_2, z_2), \ldots, P_n(x_n, y_n, z_n),$$

the moment (of inertia) $M$ of this system of masses about the point $P(x, y, z)$, defined as the sum of the products of the masses and the squares of their distances from the point $P$,

$$M(x, y, z) = \sum_{i=1}^{n} m_i (x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2,$$
will be a minimum if the point P is at the so-called center of gravity of the system, with the coordinates

\[
x = \frac{\sum_{i=1}^{n} m_i x_i}{\sum_{i=1}^{n} m_i}, \quad y = \frac{\sum_{i=1}^{n} m_i y_i}{\sum_{i=1}^{n} m_i}, \quad z = \frac{\sum_{i=1}^{n} m_i z_i}{\sum_{i=1}^{n} m_i}.
\]

**Maxima and minima with subsidiary conditions.** For functions of several variables we may set up various problems concerning maximum and minimum. Let us illustrate with a simple example. Suppose that among all rectangles inscribed in a circle of radius \( R \), we wish to find the one with greatest area. The area of a rectangle is equal to the product \( xy \) of its sides, where \( x \) and \( y \) are positive numbers connected in this case by the relation \( x^2 + y^2 = (2R)^2 \), as is clear from figure 30. Thus we are required to find the maximum of the function \( f(x, y) = xy \) for all \( x \) and \( y \) satisfying the relation \( x^2 + y^2 = 4R^2 \).

Problems of this sort, where it is necessary to find the maximum (or minimum) of a function \( f(x, y) \) for those values only of \( x \) and \( y \) that satisfy a certain relation that \( \phi(x, y) = 0 \) are very common in practice.

Of course, it would be possible to solve the equation \( \phi(x, y) = 0 \) for \( y \), to substitute the solution into the function \( f(x, y) \) and in this way to seek the ordinary maximum for a function of one variable \( x \). But this method is usually complicated and sometimes impossible.

For the solution of such problems in analysis, a much more convenient procedure called the method of Lagrange multipliers, has been worked out. The idea behind it is extremely simple. Let us consider the function

\[
F(x, y) = f(x, y) + \lambda \phi(x, y),
\]

where \( \lambda \) is an arbitrary positive number. Obviously, for \( x, y \) satisfying the condition \( \phi(x, y) = 0 \), the values of \( F(x, y) \) coincide with those of \( f(x, y) \).

For function \( F(x, y) \) let us seek a maximum without conditions of any kind on \( x \) and \( y \). At the maximum point the conditions \( \partial F/\partial x = \partial F/\partial y = 0 \) must hold, in other words

\[
\frac{\partial f}{\partial x} + \lambda \frac{\partial \phi}{\partial x} = 0;
\]

\[
\frac{\partial f}{\partial y} + \lambda \frac{\partial \phi}{\partial y} = 0.
\]

*We are speaking here, of course, of a maximum attained in the domain of definition of the function \( F(x, y) \). The functions \( f(x, y) \) and \( \phi(x, y) \) are assumed to be differentiable.

\[\text{§12. FUNCTIONS OF SEVERAL VARIABLES}\]

The values of \( x \) and \( y \) at the maximum point for \( F(x, y) \), being a solution of the system (43) and (44), depend on the coefficient \( \lambda \) in these equations. Let us now suppose that we have succeeded in choosing the number \( \lambda \) in such a way that the coordinates of the maximum point satisfy the condition

\[
\phi(x, y) = 0.
\]

Then this point will be an exact local maximum for the original problem.

In fact, we may consider the problem geometrically as follows. The function \( f(x, y) \) is defined on a certain region \( G \) (figure 31). The condition \( \phi(x, y) = 0 \) will ordinarily be satisfied by the points of some curve \( \Gamma \). We are required to find the greatest value of \( x \) and \( y \) on points of the line \( \Gamma \). If \( F(x, y) \) attains its maximum on the curve \( \Gamma \), then \( F(x, y) \) does not increase for small shifts in an arbitrary direction from this point, and in particular for shifts along the curve \( \Gamma \). But for shifts along \( \Gamma \), the values of \( F(x, y) \), coincide with those of \( f(x, y) \) which means that for small shifts along the curve the function \( f(x, y) \) does not increase, or in other words it has the local maximum at the point.

These arguments indicate a simple method of solving the problem. We solve equations (43), (44), (45) for the unknowns \( x, y, \) and \( \lambda \), obtaining one or more solutions

\[
(x_1, y_1, \lambda_1), \quad (x_2, y_2, \lambda_2), \quad \ldots.
\]

To the points \( (x_1, y_1), (x_2, y_2), \ldots \) so determined we adjoin those points of the boundary of \( G \) where the curve \( \Gamma \) leaves the region \( G \). Then from all these points we choose that one at which \( f(x, y) \) takes on its greatest (or smallest) value.

Of course, the arguments here are far from proving the correctness of the method. In fact, we have not yet even proved that the points of local maximum for \( f(x, y) \) on the curve \( \Gamma \) can be obtained as maximum points for the function \( F(x, y) \) for some value of \( \lambda \). However, it is possible to prove, as is done in the textbooks in analysis, that every point \( (x_0, y_0) \) where \( f(x, y) \) has a local maximum on the curve will be obtained by the
method indicated, provided only that at this point the partial derivatives
\( \phi'_x(x_0, y_0) \) and \( \phi'_y(x_0, y_0) \) are not both equal to zero.

Let us use the method of Lagrange to solve the problem at the beginning
of the present section. In this case \( f(x, y) = xy' \); \( \phi(x, y) = x^2 + y^2 - 4R^2 \).
We set up the equations (43), (44), (45)
\[
\begin{align*}
y + 2\lambda x &= 0, \\
x + 2\lambda y &= 0, \\
x^2 + y^2 &= 4R^2,
\end{align*}
\]
for which, taking into account that \( x \) and \( y \) are positive, we find the unique solution
\[
x = y = R \sqrt{2} \left( \lambda = -\frac{1}{2} \right).
\]
For these values of \( x \) and \( y \), which are equal to one another so that the
inscribed rectangle is a square, the area is in fact a maximum.

The method of Lagrange may be extended to deal with functions of
three or more variables. There may be any number of subsidiary conditions
(smaller than the number of variables) of the type of condition (45), and
we will introduce the corresponding number of auxiliary multipliers.

Let us give some examples of problems involving maxima or minima
with subsidiary conditions.

**Example 1.** For what height \( h \) and radius \( r \) will an open cylindrical
tank of given volume \( V \) require the least amount of sheet metal for its
manufacture; that is, the area of its sides and circular base will be a
minimum?

The problem obviously reduces to finding the minimum of the function
of the variables \( r \) and \( h \)
\[
f(r, h) = 2\pi rh + \pi r^2
\]
under the condition \( \pi r^2 h = V \), which may be written in the form
\[
\phi(r, h) = \pi r^2 h - V = 0.
\]

**Example 2.** A moving point is required to pass from \( A \) to \( B \) (figure 32).
On the path \( AM \) it moves with the velocity \( v_1 \), and on \( MB \) with the
velocity \( v_2 \). Where should the point \( M \) be placed on the line \( DD' \) so that
the entire path from \( A \) to \( B \) may be covered as quickly as possible?

Let us take as unknowns the angles \( \alpha \) and \( \beta \) marked in figure 32. The
lengths \( a \) and \( b \) of the perpendiculars from the points \( A \) and \( B \) to the
straight line \( DD' \) and the distance \( c \) between them are known. The time
required for covering the entire path is represented as can easily be seen,
by the formula
\[
f(\alpha, \beta) = \frac{a}{v_1 \cos \alpha} + \frac{b}{v_2 \cos \beta}.
\]
It is required to find the minimum of this expression, taking into account
the fact that \( \alpha \) and \( \beta \) are connected by the relation
\[
a \tan \alpha + b \tan \beta = c.
\]

The reader may solve these examples by the Lagrange method. In the
second example he will find that the best position for \( M \) is given by the condition
\[
\frac{\sin \alpha}{\sin \beta} = \frac{v_1}{v_2}.
\]
This is the well-known law for the refraction of light. Consequently, a ray
of light will be refracted in its passage from one medium to another in
such a way that the time for its passage from a point in one medium
to a point in the other is a minimum. Conclusions of this sort are interesting
not only for computational purposes but also from a general philosophical
point of view; they have inspired researchers in the exact sciences to
penetrate further and further into the profound and general laws of
nature.

Finally let us note that the multipliers \( \lambda \), introduced in the solution of
problems by the method of Lagrange, are not merely auxiliary numbers.
In each case they are closely connected with the essential nature of the
particular problem and have a concrete interpretation.
§13. Generalizations of the Concept of Integral

In §10 we defined the definite integral of the function \( f(x) \) on the interval \([a, b]\) as the limit of the sum

\[
\sum_{i=1}^{n} f(\xi_i) \Delta x_i
\]

when the length of the greatest segment \( \Delta x_i \) in the subdivision of \([a, b]\) approaches zero. In spite of the fact that the class of functions \( f(x) \) for which this limit actually exists (the class of integrable functions) is a very wide one, and in particular includes all continuous and even many discontinuous functions, this class of functions has a serious shortcoming. If we add, subtract, or multiply, or under certain conditions divide the values of two integrable functions \( f(x) \) and \( g(x) \), we obtain functions which, as may easily be proved, are again integrable. For \( f(x)/g(x) \) this will be true in all cases in which \( 1/|g(x)| \) remains bounded on \([a, b]\). But if a function is obtained as a result of a limiting process from a sequence of approximating integrable functions \( f_1(x), f_2(x), f_3(x), \ldots \) such that for all values of \( x \) in the interval \([a, b]\)

\[
f(x) = \lim_{n \to \infty} f_n(x),
\]

then the limit function \( f(x) \) is not necessarily integrable.

In many cases this and other circumstances give rise to considerable complication, since the process of passing to a limit is widely used.

A way out of the difficulty was discovered by making further generalizations of the concept of an integral. The most important of these is the integral of Lebesgue, with which the reader will become acquainted in Chapter XV on the theory of functions of a real variable. But here we will confine ourselves to generalizations of the integral in other directions, which are also of the greatest importance in practice.

Multiple integrals. We have already studied the process of integration for functions of one variable defined on a one-dimensional region, namely an interval. But the analogous process may be extended to functions of two, three, or more variables, defined on corresponding regions.

For example, let us consider a surface

\[
z = f(x, y)
\]

defined in a rectangular system of coordinates, and on the plane \( Ox \) let there be given a region \( G \) bounded by a closed curve \( \Gamma \). It is required to find the volume bounded by the surface, by the plane \( Ox \) and by the cylindrical surface passing through the curve \( \Gamma \) with generators parallel to the \( Oz \) axis (figure 33). To solve this problem we divide the plane region \( G \) into subregions by a network of straight lines parallel to the axes \( Ox \) and \( Oy \) and denote by

\[
G_1, G_2, \ldots, G_n
\]

those subregions which consist of complete rectangles. If the net is sufficiently fine, then practically the whole of the region \( G \) will be covered by the enumerated rectangles. In each of them we choose at will a point

\[
(\xi_1, \eta_1), (\xi_2, \eta_2), \ldots, (\xi_n, \eta_n)
\]

and, assuming for simplicity that \( G_i \) denotes not only the rectangle but also its area, we set up the sum

\[
S_n = f(\xi_1, \eta_1) G_1 + f(\xi_2, \eta_2) G_2 + \cdots + f(\xi_n, \eta_n) G_n = \sum_{i=1}^{n} f(\xi_i, \eta_i) G_i.
\]

(47)

It is clear that, if the surface is continuous and the net is sufficiently fine, this sum may be brought as near as we like to the desired volume \( V \). We will obtain the desired volume exactly if we take the limit of the sum (47) for finer and finer sub divisions (that is, for subdivisions such that the greatest of the diagonals of our rectangles approaches zero)

\[
\lim_{\max_{\Delta (G_i) \to 0}} \sum_{i=1}^{n} f(\xi_i, \eta_i) G_i = V.
\]

(48)

From the point of view of analysis it is therefore necessary, in order to determine the volume \( V \), to carry out a certain mathematical operation on the function \( f(x, y) \) and its domain of definition \( G \), an operation indicated by the left side of equality (48). This operation is called the integration of the function \( f \) over the region \( G \), and its result is the integral of \( f \) over \( G \). It is customary to denote this result in the following way:

\[
\iint_{G} f(x, y) \, dx \, dy = \lim_{\max_{\Delta (G_i) \to 0}} \sum_{i=1}^{n} f(\xi_i, \eta_i) G_i.
\]

(49)
Similarly, we may define the integral of a function of three variables over a three-dimensional region $G$, representing a certain body in space. Again we divide the region $G$ into parts, this time by planes parallel to the coordinate planes. Among these parts we choose the ones which represent complete parallelepipeds and enumerate them

$$G_1, G_2, \ldots, G_n.$$

In each of these we choose an arbitrary point

$$(\xi_1, \eta_1, \zeta_1), (\xi_2, \eta_2, \zeta_2), \ldots, (\xi_n, \eta_n, \zeta_n)$$

and set up the sum

$$S = \sum_{i=1}^{n} f(\xi_i, \eta_i, \zeta_i) G_i,$$  \hspace{1cm} (50)

where $G_i$ denotes the volume of the parallelepiped $G_i$. Finally we define the integral of $f(x, y, z)$ over the region $G$ as the limit

$$\lim_{\max d(G_i) \to 0} \sum_{i=1}^{n} f(\xi_i, \eta_i, \zeta_i) G_i = \iint_{G} f(x, y, z) \, dx \, dy \, dz,$$  \hspace{1cm} (51)

to which the sum (50) tends when the greatest diagonal $d(G_i)$ approaches zero.

Let us consider an example. We imagine the region $G$ is filled with a nonhomogeneous mass whose density at each point in $G$ is given by a known function $\rho(x, y, z)$. The density $\rho(x, y, z)$ of the mass at the point $(x, y, z)$ is defined as the limit approached by the ratio of the mass of an arbitrary small region containing the point $(x, y, z)$ to the volume of this region as its diameter approaches zero.* To determine the mass of the body $G$ it is natural to proceed as follows. We divide the region $G$ into parts by planes parallel to the coordinate planes and enumerate the complete parallelepipeds formed in this way

$$G_1, G_2, \ldots, G_n.$$

Assuming that the dividing planes are sufficiently close to one another, we will make only a small error if we neglect the irregular regions of the body and define the mass of each of the regular regions $G_i$ (the complete parallelepipeds) as the product

$$\rho(\xi_i, \eta_i, \zeta_i) G_i.$$

* The diameter of a region is defined as the least upper bound of the distance between two points of the region.

§13. GENERALIZATIONS OF CONCEPT OF INTEGRAL

where $(\xi_i, \eta_i, \zeta_i)$ is an arbitrary point $G_i$. As a result the approximate value of the mass $M$ will be expressed by the sum

$$S_n = \sum_{i=1}^{n} \rho(\xi_i, \eta_i, \zeta_i) G_i,$$

and its exact value will clearly be the limit of this sum as the greatest diagonal $G_i$ approaches zero; that is

$$M = \iiint_{G} \rho(x, y, z) \, dx \, dy \, dz = \lim_{\max d(G_i) \to 0} \sum_{i=1}^{n} \rho(\xi_i, \eta_i, \zeta_i) G_i.$$

The integrals (49) and (51) are called double and triple integrals respectively.

Let us examine a problem which leads to a double integral. We imagine that water is flowing over a plane surface. Also, on this surface the underground water is seeping through (or soaking back into the ground) with an intensity $f(x, y)$ which is different at different points. We consider a region $G$ bounded by a closed contour (figure 34) and assume that at every point of $G$ we know the intensity $f(x, y)$, namely the amount of underground water seeping through per minute per cm² of surface; we will have $f(x, y) > 0$ where the water is seeping through and $f(x, y) < 0$ where it is soaking into the ground. How much water will accumulate on the surface $G$ per minute?

If we divide $G$ into small parts, consider the rate of seepage as approximately constant in each part and then pass to the limit for finer and finer subdivisions, we will obtain an expression for the whole amount of accumulated water in the form of an integral

$$\iint_{G} f(x, y) \, dx \, dy.$$

Double (two-fold) integrals were first introduced by Euler. Multiple integrals form an instrument which is used everyday in calculations and investigations of the most varied kind.
II. ANALYSIS

It would also be possible to show, though we will not do it here, that calculation of multiple integrals may be reduced, as a rule, to iterated calculation of ordinary one-dimensional integrals.

Contour and surface integrals. Finally, we must mention that still other generalizations of the integral are possible. For example, the problem of defining the work done by a variable force applied to a material point, as the latter moves along a given curve, naturally leads to a so-called curvilinear integral, and the problem of finding the general charge on a surface on which electricity is continuously distributed with a given surface density leads to another new concept, an integral over a curved surface.

For example, suppose that a liquid is flowing through space (figure 35) and that the velocity of a particle of the liquid at the point \((x, y)\) is given by a function \(P(x, y)\), not depending on \(z\). If we wish to determine the amount of liquid flowing per minute through the contour \(\Gamma\), we may reason in the following way. Let us divide \(\Gamma\) up into segments \(\Delta s_i\). The amount of water flowing through one segment \(\Delta s_i\) is approximately equal to the column of liquid shaded in figure 35; this column may be considered as the amount of liquid forcing its way per minute through that segment of the contour. But the area of the shaded parallelogram is equal to

\[ P(x, y) \cdot \Delta s_i \cdot \cos \alpha_i, \]

where \(\alpha_i\) is the angle between the direction \(\vec{x}\) of the \(x\)-axis and the outward normal of the surface bounded by the contour \(\Gamma\); this normal is the perpendicular \(\vec{n}\) to the tangent, which we may consider as defining the direction of the segment \(\Delta s_i\). By summing up the areas of such parallelograms and passing to the limit for finer and finer subdivisions

\[ \int_\Gamma P(x, y) \cos (\vec{n}, \vec{x}) \, ds \]

and is called a curvilinear integral. If the flow is not everywhere parallel, then its velocity at each point \((x, y)\) will have a component \(P(x, y)\) along the \(x\)-axis and a component \(Q(x, y)\) along the \(y\)-axis. In this case we can show by an analogous argument that the quantity of water flowing through the contour will be equal to

\[ \int_\Gamma [P(x, y) \cos (\vec{n}, \vec{x}) + Q(x, y) \cos (\vec{n}, \vec{y})] \, ds. \]

When we speak of an integral over a curved surface \(G\) for a function \(f(M)\) of its points \(M(x, y, z)\), we mean the limit of sums of the form

\[ \lim_{n} \sum_{i=1}^{n} f(M_i) \Delta \sigma_i = \int_G f(x, y, z) \, d\sigma \]

for finer and finer subdivisions of the region \(G\) into other forms and for calculating their values, either exactly or approximately.

† Formula of Ostrogradskii. Several important and very general formulas relating an integral over a volume to an integral over its surface (and also an integral over a surface, curved or plane, to an integral around its boundary) were discovered in the middle of the past century by Ostrogradskii.

We shall not try to give here a proof of the general formula of Ostrogradskii, which has very wide application, but will merely illustrate it by an example of its simplest particular case.

Let us imagine, as we did before, that over a plane surface there is a horizontal flow of water that is also soaking into the ground or seeping out again from it. We mark off a region \(G\), bounded by a curve \(\Gamma\), and

\[ \int_\Gamma \left[ P(x, y) \, dy - Q(x, y) \, dx \right]. \]

* More precisely, through a cylindrical surface with the contour for its base and with height equal to unity.
II. ANALYSIS

assume that for each point of the region we know the components \( P(x, y) \) and \( Q(x, y) \) of the velocity of the water in the direction of the \( x \)-axis and of the \( y \)-axis respectively.

Let us calculate the rate at which the water is seeping from the ground at a point with coordinates \( (x, y) \). For this purpose we consider a small rectangle with sides \( \Delta x \) and \( \Delta y \) situated at the point \( (x, y) \).

As a result of the velocity \( P(x, y) \) through the left vertical edge of this rectangle, there will flow approximately \( P(x, y) \Delta y \) units of water per minute into the rectangle, and through the right side in the same time will flow out approximately \( P(x + \Delta x, y) \Delta y \) units. In general, the net amount of water leaving a square unit of surface as a result of the flow through its left and right vertical sides will be approximately

\[
\frac{[P(x + \Delta x, y) - P(x, y)] \Delta y}{\Delta x \Delta y}.
\]

If we let \( \Delta x \) approach zero, we obtain in the limit

\[
\frac{\partial P}{\partial x}.
\]

Correspondingly, the net rate of flow of water per unit area in the direction of the \( y \)-axis will be given by

\[
\frac{\partial Q}{\partial y}.
\]

This means that the intensity of the seepage of ground water at the point with coordinates \( (x, y) \) will be equal to

\[
\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y}.
\]

But in general, as we saw earlier, the quantity of water coming out from the ground will be given by the double integral of the function expressing the intensity of the seepage of ground water at each point, namely

\[
\iint_{\Omega} \left( \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} \right) dx \, dy.
\] (52)

But, since the water is incompressible, this entire quantity must flow out during the same time through the boundaries of the contour \( \Gamma \). The quantity of water flowing out through the contour \( \Gamma \) is expressed, as we saw earlier, by the curvilinear integral over \( \Gamma \)

\[
\int_{\Gamma} [P(x, y) \cos (\bar{n}, \bar{x}) + Q(x, y) \cos (\bar{n}, \bar{y})] \, ds.
\] (53)

§13. GENERALIZATIONS OF CONCEPT OF INTEGRAL

The equality of the magnitudes (52) and (53) expresses the formula of Ostrogradski in its simplest two-dimensional case

\[
\iint_{\Omega} \left( \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} \right) dx \, dy = \int_{\Gamma} \left[ P(x, y) \cos (\bar{n}, \bar{x}) + Q(x, y) \cos (\bar{n}, \bar{y}) \right] \, ds.
\]

We have merely explained the meaning of this formula by a physical example, but it can be proved mathematically.

In this way the mathematical theorem of Ostrogradski reflects a widespread phenomenon in the external world, which in our example we interpreted in a readily visualized way as preservation of the volume of an incompressible fluid.

Ostrogradski established a considerably more general formula expressing the connection between an integral over a multidimensional volume and an integral over its surface. In particular, for a three-dimensional body \( G \), bounded by the surface \( \Gamma \), his formula is

\[
\iiint_{G} \left( \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z} \right) dx \, dy \, dz = \int_{\Gamma} \left[ P \cos (\bar{n}, \bar{x}) + Q \cos (\bar{n}, \bar{y}) + R \cos (\bar{n}, \bar{z}) \right] \, ds,
\]

where \( ds \) is the element of surface.

It is interesting to note that the fundamental formula of the integral calculus

\[
\int_{a}^{b} f(x) \, dx = F(b) - F(a)
\] (54)

may be considered as a one-dimensional case of the formula of Ostrogradski. The equation (54) connects the integral over an interval with the "integral" over its "null-dimensional" boundary, consisting of the two end points.

Formula (54) may be illustrated by the following analogy. Let us imagine that in a straight pipe with constant cross section \( s = 1 \) water is flowing with velocity \( F(x) \), which is different for different cross sections (figure 36). Through the porous walls of the pipe, water is seeping into it.
II. ANALYSIS

(or out of it) at a rate which is also different for different cross sections.

If we consider a segment of the pipe from \( x \) to \( x + \Delta x \), the quantity of water seeping into it in unit time must be compensated by the difference \( F(x + \Delta x) - F(x) \) between the quantity flowing out of this segment and the quantity flowing into it along the pipe. So the quantity seeping into the segment is equal to the difference \( F(x + \Delta x) - F(x) \), and consequently the rate of seepage per unit length of pipe (the ratio of the seepage over an infinitesimal segment to the length of the segment) will be equal to

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

More generally, the quantity of water seeping into the pipe over the whole section \([a, b]\) must be equal to the amount lost by flow through the ends of the pipe. But the amount seeping through the walls is equal to \( \int_a^b f(x) \, dx \) and the amount lost by flow through the ends is \( F(b) - F(a) \). The equality of these two magnitudes produces formula (54).

§14. Series

Concept of a series. A series in mathematics is an expression of the form

\[ u_0 + u_1 + u_2 + \cdots. \]

The numbers \( u_k \) are called the terms of the series. There is an infinite number of them, and they are arranged in a definite order, so that to each natural number \( k = 0, 1, 2, \ldots \) there corresponds a definite value \( u_k \).

The reader must keep in mind that we have not said whether it is possible to calculate a value for such expressions or, in case it is possible, how to do it. The presence of a plus sign between the terms \( u_k \) in our expression seems to indicate that in some way all the terms should be added. But there are infinitely many of them and addition of numbers is defined only for a finite number of terms.

Let us denote by \( S_n \) the sum of the first \( n \) terms of the series; we will call it the \( n \)th partial sum. As a result we obtain a sequence of numbers

\[
S_1 = u_0,
S_2 = u_0 + u_1,
S_3 = u_0 + u_1 + u_2,
S_4 = u_0 + u_1 + \cdots + u_{n-1},
\]

and we may speak of a variable quantity \( S_n \), where \( n = 1, 2, \ldots \).

§14. SERIES

The series is said to be **convergent** if, as \( n \to \infty \), the variable \( S_n \) approaches a definite finite limit

\[
\lim_{n \to \infty} S_n = S.
\]

This limit is called the **sum of the series**, and in this case we write

\[
S = u_0 + u_1 + u_2 + \cdots.
\]

But if, as \( n \to \infty \), the limit \( S_n \) does not exist, then the series is said to be **divergent** and in this case there is no sense in speaking of its sum.* But if all the \( u_n \) have the same sign, then it is customary to say that the sum of the series is equal to infinity with the corresponding sign.

As an example, let us consider the series

\[
1 + x + x^2 + \cdots,
\]

whose terms form a geometric progression with common ratio \( x \).

The sum of the first \( n \) terms is equal to

\[
S_n(x) = \frac{1 - x^n}{1 - x} \quad (x \neq 1); \tag{55}
\]

if \( |x| < 1 \) this sum has a limit

\[
\lim_{n \to \infty} S_n(x) = \frac{1}{1 - x},
\]

and so for \( |x| < 1 \) we may write

\[
\frac{1}{1 - x} = 1 + x + x^2 + \cdots.
\]

If \( |x| > 1 \), then obviously

\[
\lim_{n \to \infty} S_n(x) = \infty,
\]

and the series diverges. The same situation holds also for \( x = 1 \), as may be seen immediately without use of formula (55), which for \( x = 1 \) has no meaning. Finally, if \( x = -1 \) the partial sums take the values \( +1 \) and \( 0 \) alternately, so that this series also is divergent.

* Let us note that it is also possible to give generalized definitions of the sum of a series, by virtue of which it is possible to assign to certain divergent series a more or less natural concept of "generalized sum." Such series are said to be summable. Operations with generalized sums of divergent series are sometimes useful.
II. ANALYSIS

To each series there corresponds a definite sequence of values of its partial sums $S_1, S_2, S_3, \ldots$ such that the convergence of the series depends on the fact that the sums approach a limit. Conversely, to an arbitrary sequence of numbers $S_1, S_2, S_3, \ldots$ corresponds a series

$$S_1 + (S_2 - S_1) + (S_3 - S_2) + \ldots,$$

the partial sums of which will be the numbers of the sequence. Thus the theory of variables ranging over a sequence may be reduced to the theory of the corresponding series, and conversely. Yet each of these theories has independent significance. In some cases it is more convenient to study the variable directly and in others to consider the equivalent series.

Let us note that series have long served as an important method of representing various entities (above all, functions) and of calculating their value. Of course, the views of mathematicians concerning series have changed with the passage of time, corresponding to the changes in their ideas about infinitesimals. The above clear-cut definition of convergence and divergence of a series was formulated at the beginning of the last century at the same time as the closely associated concept of a limit.

If the series converges, then its general term approaches zero with increasing $n$, since

$$\lim_{n \to \infty} u_n = \lim_{n \to \infty} (S_{n+1} - S_n) = S - S = 0.$$

From examples given in the following paragraphs, it will be clear that the converse statement is in general false. But the criterion is still a useful one, since it provides a necessary condition for the convergence of a series. For example, the divergence of a geometric progression with common ratio $x > 1$ follows immediately from the fact its general term does not approach zero.

If the series consists of positive terms, then its partial sum $S_n$ increases with increasing $n$ and only two cases can exist: Either the variable $S_n$ becomes and remains greater than any preassigned number $A$ for sufficiently large $n$, in which case $\lim_{n \to \infty} S_n = \infty$, so that the series diverges; or else there exists a number $A$ such that for all $n$ the value of $S_n$ does not exceed $A$; but then the variable $S_n$ necessarily approaches a definite finite limit not greater than $A$ and the series is convergent.

Convergence of a series. The question whether a given series converges or diverges may often be settled by comparing it with another series. Here it is customary to make use of the following criterion.

§14. SERIES

If we are given two series

$$u_0 + u_1 + u_2 + \ldots,$$
$$v_0 + v_1 + v_2 + \ldots$$

with positive terms such that for all values of $n$, beginning with a certain one, we have the inequality

$$u_n \leq v_n,$$

then the convergence of the second series implies the convergence of the first, and the divergence of the first implies the divergence of the second.

For example, let us consider the so-called harmonic series

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \ldots.$$

Its terms are correspondingly not less than the terms of the series

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{16} + \ldots,$$

in which the sum of the underlined terms in each case is equal to $\frac{1}{2}$.

It is clear that the sum $S_n$ of the second series approaches infinity with increasing $n$, and consequently that the harmonic series diverges. The series

$$1 + \frac{1}{2^x} + \frac{1}{3^x} + \frac{1}{4^x} + \ldots,$$

where $\alpha$ is a positive number less than unity, also obviously diverges, since for arbitrary $n$

$$\frac{1}{n^x} > \frac{1}{n} (0 < \alpha < 1).$$

On the other hand, it is possible to prove that series (56) for $\alpha > 1$ is convergent. We will prove this here only for the case $\alpha \geq 2$; for this purpose we consider the series

$$(1 - \frac{1}{2}) + (\frac{1}{2} - \frac{1}{3}) + \ldots + (\frac{1}{n-1} - \frac{1}{n}) + \ldots$$

with positive terms. It converges to unity as its sum, since its partial sums $S_n$ are equal to

$$S_n = 1 - \frac{1}{n+1} \to 1 \ (n \to \infty).$$
II. ANALYSIS

On the other hand, the general term of this series satisfies the inequality
\[ \frac{1}{n-1} - \frac{1}{n} = \frac{1}{n(n-1)} \geq \frac{1}{n^2}, \]
from which it follows that the series
\[ 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots \]
converges. All the more then will the series (56) converge with \( x > 2 \).

Let us give here without proof another useful criterion for convergence and divergence of series with positive terms, the so-called criterion of d’Alembert.

Let us suppose that, as \( n \) approaches infinity, the ratio \( (u_n + 1)/u_n \) has a limit \( q \). Then for \( q < 1 \) the sequence will certainly converge, while for \( q > 1 \) it will diverge. But for \( q = 1 \) the question of its convergence remains open.

The sum of a finite number of summands does not change if we permute the summands. But in general this is no longer true for infinite series. There exist convergent series for which it is possible to permute the terms in such a way as to change their sum and even to turn them into divergent series. Series with unstable sums of this sort fail to possess one of the fundamental properties of ordinary sums, permutability of the summands. So it is important to distinguish those series which preserve this property. It turns out that they are the so-called absolutely convergent series. The series
\[ u_0 + u_1 + u_2 + u_3 + \cdots \]
is said to be absolutely convergent if the series
\[ |u_0| + |u_1| + |u_2| + |u_3| + \cdots \]
of absolute values of its terms is also convergent. It is possible to prove that an absolutely convergent series is always convergent; in other words, that its partial sums \( S_n \) approach a finite limit. It is obvious that every convergent series with terms of one sign is absolutely convergent.

The series
\[ \frac{\sin x}{1^2} + \frac{\sin 2x}{2^2} + \frac{\sin 3x}{3^2} + \cdots \]
is an example of an absolutely convergent series, since the terms of the series
\[ \left| \frac{\sin x}{1^2} \right| + \left| \frac{\sin 2x}{2^2} \right| + \left| \frac{\sin 3x}{3^2} \right| + \cdots \]
are not greater than the corresponding terms of the convergent series
\[ 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots. \]

An example of a series which is convergent, but not absolutely convergent, is the following
\[ 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots \]
as the reader may prove for himself.

§14. SERIES

Series of functions; uniformly convergent series. In analysis we often have to do with series whose terms are functions of \( x \). In the preceding paragraphs we have already had examples of this sort, for instance, the series
\[ 1 + x + x^2 + x^3 + \cdots. \]
For some values of \( x \) this series converges, but for others it diverges. Particularly important in applications are series of functions convergent for all values of \( x \) belonging to a certain interval, which may in particular be the whole of the real axis or the positive half of it and so forth. Then the necessity arises for differentiating such series term by term, integrating them, deciding whether their sum is continuous, and so forth. For the familiar ease of the sum of a finite number of terms, there are simple general rules. We know that the derivative of a sum of differentiable functions is equal to the sum of their derivatives, the integral of a sum of continuous functions is the sum of their integrals, and a sum of continuous functions is itself a continuous function: All this holds for the sum of a finite number of terms.

But for infinite series these simple rules are in general no longer true. We could give many examples of convergent series of functions for which the rules of termwise integration and differentiation are false. In the same way a series of continuous functions may turn out to have a discontinuous sum. On the other hand many infinite series behave like finite sums with respect to these rules.

Profound investigations of this question have shown that these rules may still be applied if the infinite series in question are not only convergent at each separate point of the interval of definition (the domain over which \( x \) varies) but if they are uniformly convergent over the whole interval. In this way there was crystallized in mathematical analysis, in the middle of the 19th century, the important concept of the uniform convergence of a series.

Let us consider the series
\[ S(x) = u_0(x) + u_1(x) + u_2(x) + \cdots, \]
II. ANALYSIS

whose terms are functions defined on the interval \([a, b]\). We suppose that for each separate value of \(x\) in the interval this series converges to a certain sum \(S(x)\). The sum of the first \(n\) terms of the series

\[ S_n(x) = u_0(x) + u_1(x) + \cdots + u_{n-1}(x) \]

is also a certain function of \(x\), defined on \([a, b]\).

We now introduce a magnitude \(\eta_n\), which is equal to the least upper bound of the values* \(|S(x) - S_n(x)|\), as \(x\) varies on the interval \([a, b]\). This magnitude is written as follows:

\[ \eta_n = \sup_{a < x < b} |S(x) - S_n(x)|. \]

In case the quantity \(S(x) - S_n(x)\) attains its maximum value, which will certainly occur for example, when \(S(x)\) and \(S_n(x)\) are continuous, then \(\eta_n\) is simply the maximum of \(|S(x) - S_n(x)|\) on \([a, b]\).

From the assumed convergence of our series, we have for every individual value of \(x\) in the interval \([a, b]\)

\[ \lim_{n \to \infty} |S(x) - S_n(x)| = 0. \]

But the magnitude \(\eta_n\) may approach zero or it may not. If \(\eta_n\) approaches zero as \(n \to \infty\), then the series is said to be uniformly convergent, and in the opposite case nonuniformly convergent. In the same sense it is possible to speak of the uniform or nonuniform convergence of a sequence of functions \(S_n(x)\) without necessarily interpreting them as partial sums of a series.

Example 1. The series of functions

\[ \frac{1}{x + 1} - \frac{1}{(x - 1)(x + 2)} + \frac{1}{(x + 2)(x + 3)} - \cdots, \]

which we take to be defined only for nonnegative values of \(x\), namely on the half line \([0, \infty)\), may be written in the form

\[ \frac{1}{x + 1} + \left( \frac{1}{x + 2} - \frac{1}{x + 1} \right) + \left( \frac{1}{x + 3} - \frac{1}{x + 2} \right) + \cdots, \]

from which we see that its partial sums are equal to

\[ S_n(x) = \frac{1}{x + n} \]

and

\[ \lim_{n \to \infty} S_n(x) = 0. \]

§14. SERIES

Thus the series is convergent for all nonnegative \(x\) and has the sum \(S(x) = 0\). Furthermore,

\[ \eta_n = \sup_{0 < x < \infty} |S_n(x) - S(x)| = \sup_{0 < x < \infty} \frac{1}{x + n} = \frac{1}{n} \to 0 \quad (n \to \infty), \]

so that the series is uniformly convergent to zero on the half axis \([0, \infty)\). Figure 37 shows the graphs of some of the partial sums \(S_n(x)\).

Example 2. The series

\[ x + x(x - 1) + x^2(x - 1) + \cdots \]

may be written in the form

\[ x + (x^2 - x) + (x^3 - x^2) + \cdots, \]

from which

\[ S_n(x) = x^n, \]

and therefore

\[ \lim_{n \to \infty} S_n(x) = \begin{cases} 0, & \text{if } 0 < x < 1; \\ 1, & \text{if } x = 1. \end{cases} \]

Thus the sum of the series is discontinuous on the interval \([0, 1]\) with a discontinuity at the point \(x = 1\). The quantity \(|S_n(x) - S(x)|\) is less than unity for every \(x\) in \([0, 1]\), but for \(x\) close to \(x = 1\) it is arbitrarily close to unity. So,

\[ \eta_n = \sup_{0 < x < 1} |S_n(x) - S(x)| = 1 \]

for all \(n = 1, 2, \ldots\). Thus the series is nonuniformly convergent on the interval \([0, 1]\). Figure 38 shows some of the graphs of the function \(S_n(x)\). The graph of the sum of the series consists of the segment \(0 < x < 1\) of the \(x\)-axis omitting the right end point and of the point \((1, 1)\).
II. ANALYSIS

This example shows that the sum of a nonuniformly convergent series of continuous functions may in fact be a discontinuous function.

On the other hand, if we consider the series on the interval $0 \leq x \leq q$ with $q < 1$, then

$$\eta_n = \sup_{0 \leq x \leq q} |S_n(x) - S(x)| = \max_{0 \leq x \leq q} x^n = q^n \to 0,$$

so that on this interval the series converges uniformly and its sum is continuous. The fact that the sum of a uniformly convergent series of continuous functions is itself a continuous function is a general rule, as was pointed out earlier, which can be rigorously proved.

Example 3. The sum of the first $n$ terms of the series $S_n(x)$ has the graph represented by the heavy broken line in figure 39. Obviously, for all $n$ we have $S_n(0) = 0$, but if $0 < x \leq 1$, then for all $n \geq 1/x$, we will have $S_n(x) = 0$, and consequently for arbitrary $x$ in the interval $[0, 1]$,

$$S(x) = \lim_{n \to \infty} S_n(x) = 0.$$

On the other hand,

$$\eta_n = \sup_{0 \leq x \leq 1} |S_n(x) - S(x)| = \sup |S_n(x)| = n^n.$$

So the quantity $\eta_n$ does not approach zero but even approaches infinity. We now note that the series corresponding to this sequence $S_n(x)$ cannot be integrated term by term on the interval $[0, 1]$, since

$$\int_0^1 S(x) \, dx = 0, \quad \int_0^1 S_n(x) \, dx = \frac{1}{2} n^2 \frac{1}{n} = \frac{n}{2},$$

so that the series

$$\int_0^1 S_1(x) \, dx + \int_0^1 [S_2(x) - S_1(x)] \, dx + \int_0^1 [S_3(x) - S_2(x)] \, dx + \cdots$$

reduces to the divergent series

$$\frac{1}{2} + \left( \frac{3}{2} - \frac{1}{2} \right) + \left( \frac{5}{2} - \frac{3}{2} \right) + \left( \frac{7}{2} - \frac{5}{2} \right) + \cdots.$$

§14. SERIES

Let us state without proof the fundamental properties of uniformly convergent series:

1. The sum of a series of continuous functions which is uniformly convergent on the interval $[a, b]$ is a continuous function on this interval.
2. If the series of continuous functions

$$S(x) = u_0(x) + u_1(x) + u_2(x) + \cdots$$

(57)

converges uniformly on the interval $[a, b]$, then it may be integrated term by term on this interval; that is, for all $x_1, x_2$ in $[a, b]$ we have the equality

$$\int_{x_1}^{x_2} S(t) \, dt = \int_{x_1}^{x_2} u_0(t) \, dt + \int_{x_1}^{x_2} u_1(t) \, dt + \cdots.$$

3. If on the interval $[a, b]$ the series (57) converges and the functions $u_n(x)$ have continuous derivatives, then the equality

$$S'(x) = u_0'(x) + u_1'(x) + u_2'(x) + \cdots,$$

(58)

obtained by termwise differentiation of (57) will be valid on the interval $[a, b]$ if the series on the right in (58) converges uniformly.

Power series. In §9, a function $f(x)$ defined on an interval $[a, b]$ was called analytic, if on this interval it has derivatives of arbitrary order and if in a sufficiently small neighborhood of any point $x_0$ of the interval $[a, b]$ it may be expanded in a convergent Taylor series

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!} (x - x_0) + \frac{f''(x_0)}{2!} (x - x_0)^2 + \cdots.$$

(59)

If we introduce the notation

$$a_n = \frac{f^{(n)}(x_0)}{n!},$$

this series may be written in the following form

$$f(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots.$$
II. ANALYSIS

A series of this sort, where the numbers \( a_1, a_2, \ldots \) are constants independent of \( x \), is called a power series.

As an example let us consider the power series

\[
1 + x + x^2 + x^3 + \cdots, \tag{61}
\]

whose terms form a geometric progression.

We know that for all values of \( x \) in the interval \(-1 < x < 1\) this series converges and its sum is equal to

\[
S(x) = \frac{1}{1 - x}.
\]

For other values of \( x \) the series diverges.

It is also easy to see that the difference between the sum of the series and the sum of its first \( n \) terms is given by the formula

\[
S(x) - S_n(x) = \frac{x^n}{1 - x}, \tag{62}
\]

and if \(-q \leq x \leq q\), where \( q \) is a positive number less than unity, then

\[
\eta_n = \max |S(x) - S_n(x)| = \frac{q^n}{1 - q}.
\]

From this it is clear that \( \eta_n \) approaches zero with increasing \( n \) so that the series is uniformly convergent on the interval \(-q \leq x \leq q\), for all positive values of \( q < 1 \).

It is easy to verify that the function

\[
S(x) = \frac{1}{1 - x}
\]

has a derivative of \( n \)th order, which is equal to

\[
S^{(n)}(x) = \frac{n!}{(1 - x)^{n+1}},
\]

from which

\[
S^{(n)}(0) = n!
\]

and the sum of the first \( n \) terms of the Taylor series for the function \( S(x) \) exactly coincides for \( x_0 = 0 \) with the sum of the first \( n \) terms of the series (59). Moreover, we know that the remainder term of the formula, given by the equality (62), approaches zero with increasing \( n \), for all \( x \)

\[\text{§14. SERIES}\]

on the interval \(-1 < x < 1\). Thus we have shown that the series (61) is the Taylor series of its sum \( S(x) \).

Let us note one further fact. From the interval of convergence \(-1 < x < 1\) of our series, let us choose an arbitrary point \( x_0 \). It is easy to see that for all \( x \) sufficiently close to \( x_0 \), namely for all \( x \) satisfying the inequality

\[
\left| x - x_0 \right| < 1,
\]

we have the equality

\[
S(x) = \frac{1}{1 - x} = \frac{1}{1 - x_0} \cdot \frac{1}{1 - \left( \frac{x - x_0}{1 - x_0} \right)}
\]

\[
= \frac{1}{1 - x_0} \left[ 1 + \frac{x - x_0}{1 - x_0} + \frac{(x - x_0)^2}{1 - x_0^2} + \cdots \right]
\]

\[
= \frac{1}{1 - x_0} + \frac{x - x_0}{(1 - x_0)^2} + \frac{(x - x_0)^2}{(1 - x_0)^2} + \cdots. \tag{63}
\]

The reader may prove without difficulty that

\[
S^{(n)}(x_0) = \frac{n!}{(1 - x_0)^{n+1}}.
\]

Consequently, series (63) is the Taylor series of its sum \( S(x) \) and converges to it in a sufficiently small neighborhood of any point \( x_0 \) belonging to the interval of convergence of (61). Since the point \( x_0 \) is arbitrary, this means that the function \( S(x) \) is analytic on the interval.

All these facts that we have observed for the particular power series (61) hold for arbitrary power series.* Namely, for every power series of the form (60) where the constants \( a_n \) are chosen by any given law, there exists a certain nonnegative number \( R \) (which may also be infinite), called the radius of convergence of the series (60), with the following properties:

1. For all values of \( x \) from the interval \( x_0 - R < x < x_0 + R \), which is called its interval of convergence, the series converges and its sum \( S(x) \) is an analytic function of \( x \) in its interval. Here the convergence is uniform for every interval \([a, b]\) lying completely within the interval of convergence. The series itself is the Taylor series of its sum.

2. At the end points of the interval of convergence, the series may converge or diverge, depending on its individual character. But it will certainly diverge outside the closed interval \( x_0 - R \leq x \leq x_0 + R \).

* For more detailed information on this point see Chapter IX.
II. ANALYSIS

We suggest to the reader that he consider the power series

\[ 1 + \frac{x}{1} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots, \]
\[ 1 + x + 2!x^2 + 3!x^3 + \cdots, \]
\[ 1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots \]

and convince himself that their radii of convergence are respectively infinite, zero, and unity.

By the definition given earlier every analytic function may be expanded, in a sufficiently small neighborhood of an arbitrary point where it is defined, into a power series which converges to the function. Conversely, from what has been said it follows that the sum of every power series whose radius of convergence is not zero is an analytic function in its interval of convergence.

So we see that power series are organically connected with analytic functions. We could even say that on their interval of convergence power series are the natural means of representing analytic functions, and consequently they are also the natural means of approximating analytic functions by algebraic polynomials.*

For example, from the fact that the function \(1/(1-x)\) may be expanded in the power series

\[ \frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots, \]

which is convergent on the interval \(-1 < x < 1\), it follows that the power series is uniformly convergent on an arbitrary interval \(-a \leq x \leq a\) with \(a < 1\), and this implies the possibility of approximating the function on the whole interval \([-a, a]\) by the partial sums of the series with any preassigned degree of accuracy.

Let us suppose that we are required to approximate the function \(1/(1-x)\) by polynomials on the interval \([-\frac{1}{2}, \frac{1}{2}]\) with an accuracy of 0.01. We note that for all \(x\) in this interval we have the inequality

\[ \left| \frac{1}{1-x} - 1 - x - \cdots - x^n \right| = \left| x^{n+1} + x^{n+2} + \cdots \right| \]
\[ \leq \left| x \right|^{n+1} + \left| x \right|^{n+2} + \cdots \leq \frac{1}{2^n+1} + \frac{1}{2^n+2} + \cdots = \frac{1}{2^n}, \]

and since \(2^6 = 64\) and \(2^7 = 128\), the desired polynomial, approximating the function on the whole interval \([-\frac{1}{2}, \frac{1}{2}]\) with an accuracy of 0.01, will have the form

\[ \frac{1}{1-x} \approx 1 + x + x^2 + \cdots + x^7. \]

Let us note one further extremely valuable property of power series: They may be differentiated termwise everywhere in the interval of convergence. This property finds extremely wide application in the solution of various problems in mathematics.

For example, let it be required to find the solution of the differential equation \(y' = y\) under the auxiliary condition \(y(0) = 1\). We will seek the solution in the form of a power series,

\[ y = a_0 + a_1x + a_2x^2 + \cdots. \]

Because of the auxiliary condition, we must set \(a_0 = 1\). Assuming that this series converges, we may differentiate it termwise; as a result we obtain

\[ y' = a_1 + 2a_2x + 3a_3x^2 + \cdots. \]

If we substitute these two series into the differential equation and equate coefficients for each of the powers of \(x\), we obtain

\[ a_k = \frac{1}{k!} (k = 1, 2, \cdots) \]

and the desired solution has the form

\[ y = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots. \]

It is well known that this series converges for all values of \(x\) and that its sum is equal to \(y = e^x\).

In this case we have obtained a series whose sum is a well-known elementary function. But this does not always happen; it may turn out that a convergent power series so obtained has a sum that is not an elementary function. An example is the series

\[ y_p(x) = x^p \left[ 1 - \frac{x^2}{2(2p+2)} + \frac{x^4}{2 \cdot 4(2p+2)(2p+4)} - \cdots \right], \]

obtained as a solution of Bessel's differential equation, which is of great importance in applications. In this way power series may serve to define functions.

* Approximations going beyond the limits of the interval of convergence of a power series require other methods. (See Chapter XII.)
Suggested Reading


VOLUME ONE

MATHEMATICS
Its Content, Methods, and Meaning

EDITED BY
A. D. Aleksandrov, A. N. Kolmogorov, M. A. Lavrent’ev

TRANSLATED BY
S. H. Gould and T. Barthe

THE M.I.T. PRESS
Massachusetts Institute of Technology
Cambridge, Massachusetts