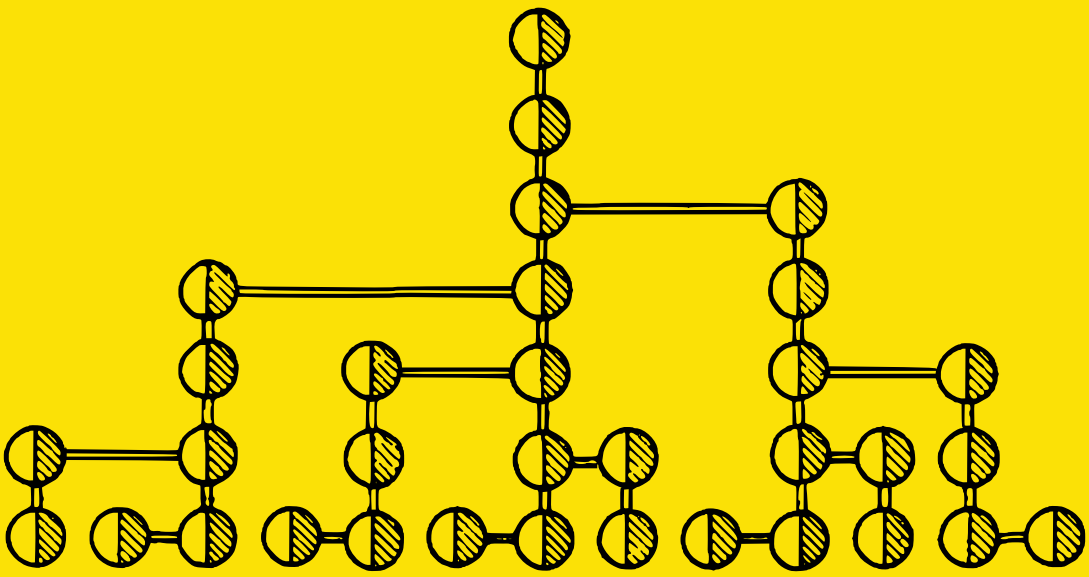
LEV TARASOV

CALCULUS

BASIC CONCEPTS FOR HIGH SCHOOLS



MIR PUBLISHERS MOSCOW

Lev Tarasov

Calculus

Mir Publishers Moscow

Translated from the Russian by *V. Kisin* and *A. Zilberman*

First published 1988

Revised from the 1984 Russian edition

This electronic version typeset using XeLaTeX by *Damitr Mazanov*.

Released on the web by http://mirtitles.org in 2018.

Contents

Preface

> Many objects are obscure to us not because our perceptions are poor, but simply because these objects are outside of the realm of our conceptions.
>
> Kosma Prutkov

Confession Of The Author My first acquaintance with calculus (or mathematical analysis) dates back to nearly a quarter of a century. This happened in the Moscow Engineering Physics Institute during splendid lectures given at that time by Professor D. A. Vasilkov. Even now I remember that feeling of delight and almost happiness. In the discussions with my classmates I rather heatedly insisted on a simile of higher mathematics to literature, which at that time was to me the most admired subject. Sure enough, these comparisons of mine lacked in objectivity. Nevertheless, my arguments were to a certain extent justified. The presence of an inner logic, coherence, dynamics, as well as the use of the most precise words to express a way of thinking, these were the characteristics of the prominent pieces of literature. They were present, in a different form of course, in higher mathematics as well. I remember that all of a sudden elementary mathematics which until that moment had seemed to me very dull and stagnant, turned to be brimming with life and inner motion governed by an impeccable logic.

Years have passed. The elapsed period of time has inevitably erased that highly emotional perception of calculus which has become a working tool for me. However, my memory keeps intact that unusual happy feeling which I experienced at the time of my initiation to this extraordinarily beautiful world of ideas which we call higher mathematics.

Confession Of The Reader Recently our professor of mathematics told us that we begin to study a new subject which he called calculus. He said that this subject is a foundation of higher mathematics and that it is going to be very difficult. We have already studied real numbers, the real line, infinite numerical sequences, and limits of sequences. The professor was indeed right saying that com.. prehension of the subject would present difficulties. I listen very carefully to his explanations and during the same day study the relevant pages of my textbook. I seem to understand everything, but at the same time have a feeling of a certain dissatisfaction. It is difficult for me to construct a consistent picture out of the pieces obtained in the classroom. It is equally difficult to remember exact wordings and definitions, for example, the definition of the limit of sequence. In other words, I fail to grasp something very important. Perhaps, all things will become clearer in the future, but so far calculus has not become an open book for me. Moreover, I do not see any substantial difference between calculus and algebra. It seems that everything has become rather difficult to perceive and even more difficult to keep in my memory.

Comments Of The Author These two confessions provide an opportunity to get acquainted with the two interlocutors in this book. In fact, the whole book is presented as a relatively free-flowing dialogue between the AUTHOR and the READER. From one discussion to another the AUTHOR will lead the inquisitive and receptive READER to different notions, ideas, and theorems of calculus, emphasizing especially complicated

or delicate aspects, stressing the inner logic of proofs, and attracting the reader's attention to special points. I hope that this form of presentation will help a reader of the book in learning new definitions such as those of *derivative, antiderivative, definite integral, differential equation*, etc. I also expect that it will lead the reader to better understanding of such concepts as *numerical sequence, limit of sequence, and function*. Briefly, these discussions are intended to assist pupils entering a novel world of calculus.And if in the long run the reader of the book gets a feeling of the intrinsic beauty and integrity of higher mathematics or even is appealed to it the author will consider his mission as successfully completed.

Working on this book, the author consulted the existing manuals and textbooks such as *Algebra and Elements of Analysis* edited by A. N. Kolmogorov, as well as the specialized textbook by N. Ya. Vilenkin and S. I. Shvartsburd *Calculus*. Appreciable help was given to the author in the form of comments and recommendations by N. Ya. Vilenkin, B. M. Ivlev, A. M. Kisin, S. N. Krachkovsky, and N. Ch. Krutitskaya, who read the first version of the manuscript. I wish to express gratitude for their advice and interest in my work. I am especially grateful to A. N. Tarasova for her help in preparing the manuscript.

Dialogue 1

INFINITE NUMERICAL SEQUENCE

AUTHOR: Let us start our discussions of calculus by considering the definition of an *infinite numerical sequence* or simply a *sequence*.

We shall consider the following examples of sequences:

$$1, 2, 4, 8, 16, 32, 64, 128, \ldots \tag{1}$$

$$5, 7, 9, 11, 13, 15, 17, 19, \ldots \tag{2}$$

$$1, 4, 9, 16, 25, 36, 49, 64, \ldots \tag{3}$$

$$1, \sqrt{2}, \sqrt{3}, 2, \sqrt{5}, \sqrt{6}, \sqrt{7}, 2\sqrt{2}, \ldots \tag{4}$$

$$\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \frac{6}{7}, \frac{7}{8}, \frac{8}{9}, \cdots \tag{5}$$

$$2, 0, -2, -4, -6, -8, -10, -12, \ldots \tag{6}$$

$$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{8}, \cdots \tag{7}$$

$$1, \frac{1}{2}, 3, \frac{1}{4}, 5, \frac{1}{6}, 7, \frac{1}{8}, \ldots \tag{8}$$

$$1, -1, \frac{1}{3}, -\frac{1}{3}, \frac{1}{5}, -\frac{1}{5}, \frac{1}{7}, -\frac{1}{7}, \dots \tag{9}$$

$$1, \frac{2}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{6}{7}, \frac{1}{7}, \frac{8}{9}, \dots \tag{10}$$

Have a closer look at these examples. What do they have in common?

READER: It is assumed that in each example there must be an infinite number of terms in a sequence. But in general, they are all different.

AUTHOR: In each example we have eight terms of a sequence. Could you write, say, the ninth term?

READER: Sure, in the first example the ninth term must be 256, while in the second example it must be 21.

AUTHOR: Correct. It means that in all the examples there is a *certain law*, which makes it possible to write down the ninth, tenth, and other terms of the sequences. Note, though, that if there is a *finite number* of terms in a sequence, one may fail to discover the law which governs the infinite sequence.

READER: Yes, but in our case these laws are easily recognizable. In example (1) we have the terms of an infinite geometric progression with common ratio 2. In example (2) we notice a sequence of odd numbers starting from 5. In example (3) we recognize a sequence of squares of natural numbers.

AUTHOR: Now let us look at the situation more rigorously. Let us enumerate all the terms of the sequence in sequential order, i.e. $1, 2, 3, \dots, n, \dots$ There is a certain law (a rule) by which each of these natural numbers is *assigned* to a certain number (the corresponding term of the sequence). In example (1) this arrangement is as follows:

1	2	4	8	16	32	…	2^{n-1}	…	(terms of the sequence)
↑	↑	↑	↑	↑	↑		↑		
1	2	3	4	5	6	…	n	…	(position numbers of the terms)

In order to describe a sequence it is sufficient to indicate the term of the sequence corresponding to the number n, i.e. to write down the term of the sequence occupying the nth position. Thus, we can formulate the following definition of a sequence.

Definition

We say that there is an infinite numerical sequence if every natural number (position numbers is unambiguously placed in correspondence with a definite number (term of the sequence) by a specific rule.

This relationship may be presented in the following general form:

y_1	y_2	y_3	y_4	y_5	…	y_n	…
↑	↑	↑	↑	↑		↑	
1	2	3	4	5	…	n	…

'The number y_n is the nth term of the sequence, and the whole sequence is sometimes denoted by a symbol (y_n).

READER: We have been given a somewhat different definition of a sequence: a sequence is a function defined on a set of natural numbers (integers).

AUTHOR: Well, actually the two definitions are equivalent. However, I am not inclined to use the term "function" too early. First, because the discussion of a function will come later. Second, you will normally deal with somewhat different functions, namely those defined not on a set of

integers but on the real line or within its segment. Anyway, the above definition of a sequence is quite correct.

Getting back to our examples of sequences, let us look in each case for an *analytical expression (formula)* for the nth term. Go ahead.

READER: Oh, this is not difficult. In example (1) it is $y_n = 2^n$. In (2) it is $y_n = 2n + 3$. In (3) it is $y_n = n^2$. In (4) it is $y_n = \sqrt{n}$. In (5) it is $y_n = 1 - \dfrac{1}{n+1} = \dfrac{n}{n+1}$. In (6) it is $y_n = 4 - 2n$. In (7) it is $y_n = \dfrac{1}{n}$.
In the remaining three examples I just do not know.

AUTHOR: Let us look at example (8). One can easily see that if n is an even integer, then $y_n = \dfrac{1}{n}$, but if n is odd, then $y_n = n$, It means that

$$y_n = \begin{cases} \dfrac{1}{n} \text{ if } n = 2k \\ n \text{ if } n = 2k-1 \end{cases}$$

READER: Can I, in this particular case, find a single analytical expression for y_n?

AUTHOR: Yes, you can. Though I think you needn't. Let us present y_n in a different form:

$$y_n = a_n n + b_n \frac{1}{n}$$

and demand that the coefficient a_n be equal to unity if n is odd, and to zero if n is even; the coefficient b_n should behave in quite an opposite manner. In this particular case these coefficients can be determined as follows:

$$a_n = \frac{1}{2}[1 - (-1)^n], \quad b_n = \frac{1}{2}[1 + (-1)^n]$$

Consequently

$$y_n = \frac{n}{2}[1 - (-1)^n] + \frac{1}{2n}[1 + (-1)^n]$$

Do in the same manner in the other two examples.

READER: For sequence (9) I can write

$$y_n = \frac{1}{2n}[1-(-1)^n] + \frac{1}{2(n-1)}[1+(-1)^n]$$

and for sequence (10)

$$y_n = \frac{1}{2n}[1-(-1)^n] + \frac{n}{2(n+1)}[1+(-1)^n]$$

AUTHOR: It is important to note that an analytical expression for the nth term of a given sequence is not necessarily a unique method of defining a sequence. A sequence can be defined, for example, by *recursion* (or the *recurrence method*) (Latin word *recurrere* means to run back). In this case, in order to define a sequence one should describe the first term (or the first several terms) of the sequence and a recurrence (or a recursion) relation, which is an expression for the nth term of the sequence via the preceding one (or several preceding terms).

Using the recurrence method, let us present sequence (1), as follows

$$y_1 = 1, \quad y_n = 2y_{n-1}$$

READER: It's clear. Sequence (2) can be apparently represented by formulas

$$y_1 = 5, \quad y_n = y_{n-1} + 2$$

AUTHOR: That's right. Using recursion, let us try to determine one interesting sequence

$$y_1 = 1, \quad y_2 = 1, \quad y_n = y_{n-2} + y_{n-1}$$

$$1, 1, 2, 3, 5, 8, 13, 21, \ldots \tag{11}$$

This sequence is known as the *Fibonacci sequence* (or *numbers*).

READER: I understand, I have heard something about. the problem of Fibonacci rabbits.

AUTHOR: Yes, it was this problem, formulated by Fibonacci, the 13th century Italian mathematician, that gave the name to this sequence (11). The problem reads as follows. A man places a pair of newly born rabbits into a warren and wants to know how many rabbits he would have over a certain period of time. A pair of rabbits will start producing offspring two months after they were born and every following month one new pair of rabbits will appear. At the beginning (during the first month) the man will have in his warren only one pair of rabbits ($y_1 = 1$); during the second month he will have the same pair of rabbits ($y_2 = 1$); during the third month the offspring will appear, and therefore the number of the pairs of rabbits in the warren will grow to two ($y_3 = 3$); during the fourth month there will be one more reproduction of the first pair ($y_4 = 3$); during the fifth month there will be offspring both from the first and second couples of rabbits ($y_5 = 5$), etc. An increase of the number of pairs in the warren from month to month is plotted in Figure 1. One can see that the numbers of pairs of rabbits counted at the end of each month form sequence (11), i.e. the Fibonacci, sequence.

READER: But in reality the rabbits do not multiply in accordance with such an idealized pattern. Furthermore, as time goes on, the first pairs of rabbits should obviously stop proliferating.

AUTHOR: The Fibonacci sequence is interesting not because it describes a simplified growth pattern of rabbits' population. It so happens that

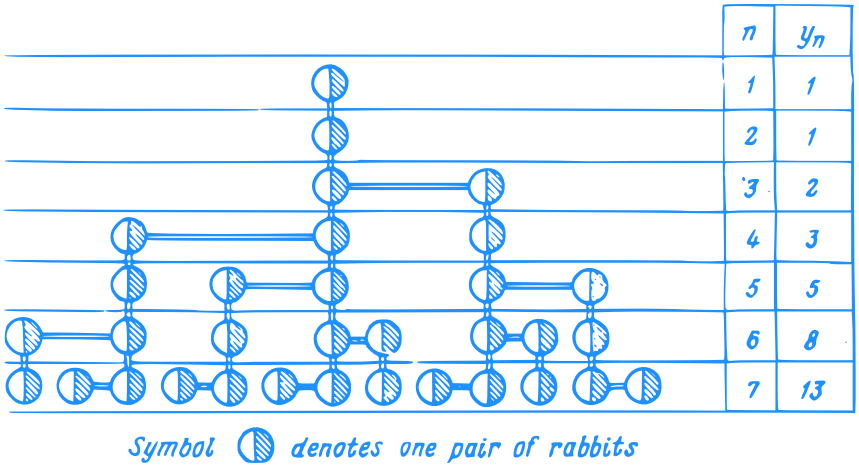


Figure 1: Visualising Fibonacci rabbits.

this sequence appears, as if by magic, in quite unexpected situations. For example, the Fibonacci numbers are used to process information by computers and to optimize programming for computers. However, this is a digression from our main topic.

Getting back to the ways of describing sequences, I would like to point out that the *very method chosen to describe a sequence is not of principal importance*. One sequence may be described, for the sake of convenience, by a formula for the nth term, and another (as, for example, the Fibonacci sequence), by the recurrence method. What is important, however, is the method used to describe the *law of correspondence*, i.e. the law by which any natural number is placed in correspondence with a certain term of the sequence. In a number of cases such a law can be formulated only by words. The examples of such cases are shown below:

$$2, 3, 5, 7, 11, 13, 17, 19, 23, \ldots \tag{12}$$

$$3, 3.1, 3.14, 3.141, 3.1415, 3.14159, \ldots \tag{13}$$

In both cases we cannot indicate either the formula for the- nth term or the recurrence relation. Nevertheless, you can without great difficulties identify specific laws of correspondence and put them in words.

READER: Wait a minute. Sequence (12) is a sequence of *prime numbers* arranged in an increasing order, while (13) is, apparently, a sequence composed of decimal approximations, with deficit, for π.

AUTHOR: You are absolutely right.

READER: It may seem that a numerical sequence differs from a random set of numbers by a presence of an intrinsic *degree of order* that is reflected either by the formula for tho nth term or by the recurrence relation. However, the last two examples show that such a degree of order needn't. be present.

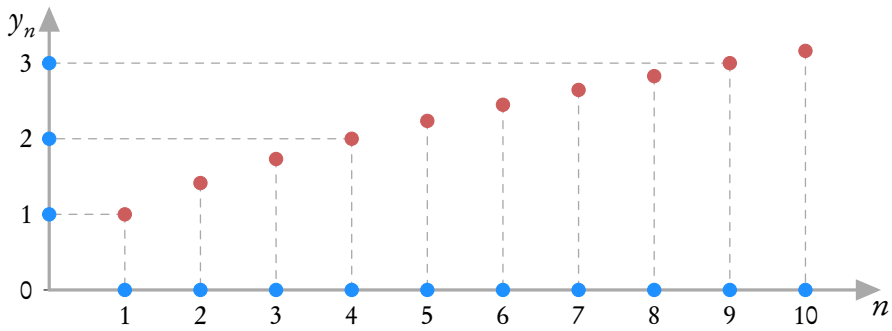


Figure 2: Graph of the series (4).

AUTHOR: Actually, a degree of order determined by a formula (an analytical expression) is not mandatory. It is important, however, to have a law (a rule, a characteristic) of correspondence, which enables one to relate any natural number to a certain term of a sequence. In examples (12) and (13) such laws of correspondence are obvious. Therefore, (12)

and (13) are not inferior (and not superior) to sequences (1)-(11) which permit an analytical description. Later we shall talk about the *geometric*

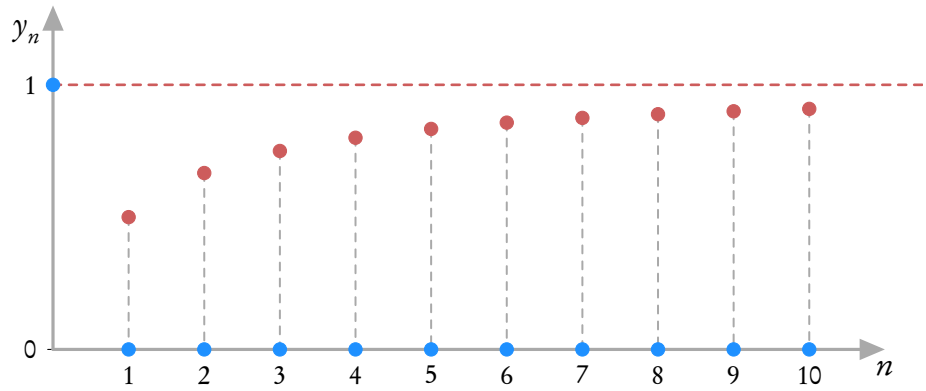


Figure 3: Image of the series (5).

image (or *map*) of a numerical sequence. Let us take two coordinate axes, x and y. We shall mark on the first axis integers $1, 2, 3, \ldots, n, \ldots$ and on the second axis, the corresponding terms of a sequence, i.e. the numbers $y_1, y_2, y_3, \ldots y_n, \ldots$. Then the sequence can be represented by a set of points $M(n, y_n)$ on the coordinate plane. For example Figure 2 images sequence (4), Figure 3 images sequence (5), Figure 4 images sequence (9), and Figure 5 images sequence (10).

As a matter of fact, there are other types of geometry images of a numerical sequence. Let us retain, for example only one coordinate y-axis and plot on it points $y_1, y_2, y_3, \ldots y_n, \ldots$ which map the terms of a sequence. In Figure 6 this method of mapping is illustrated for the sequences that have been shown in Figure 2-Figure 5. One has to admit that the latter method is less descriptive in comparison with the former method.

READER: But in the case of sequences (4) and (5) the seond method looks rather obvious.

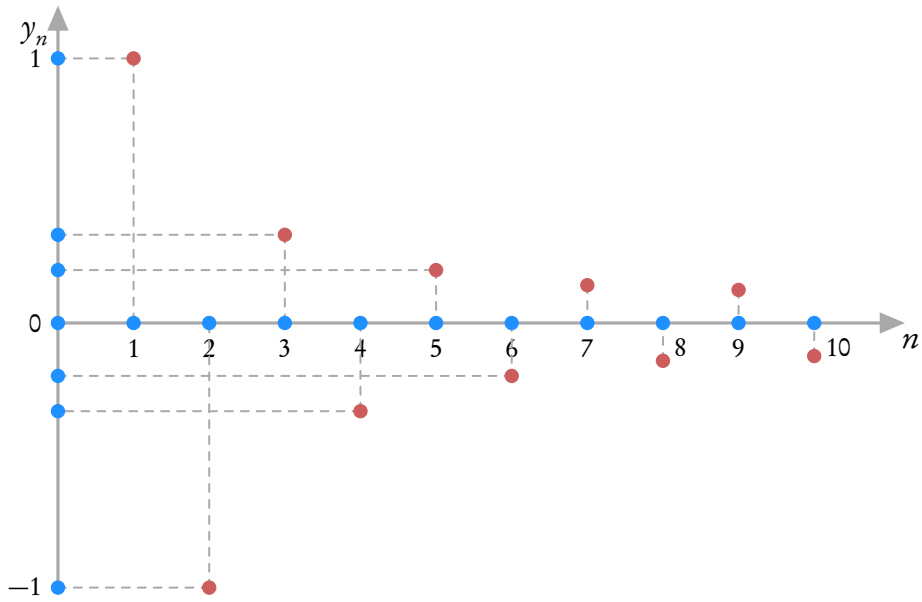


Figure 4: Series in equation (9) visualised.

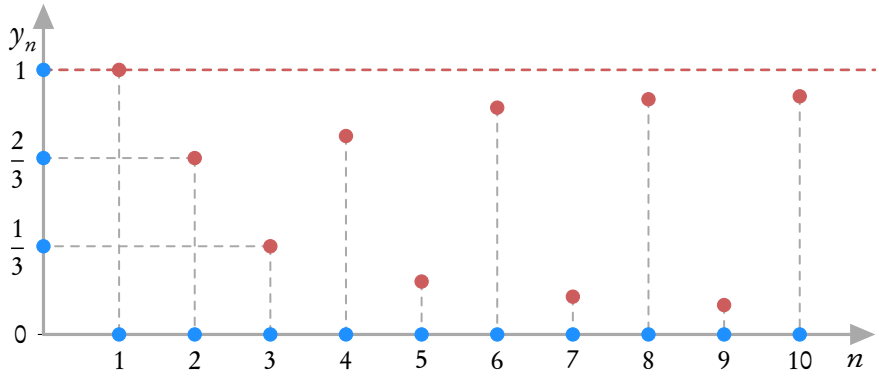


Figure 5: Series in equation (10) visualised.

AUTHOR: It can be explained by specific features of those sequences. Look at them closer.

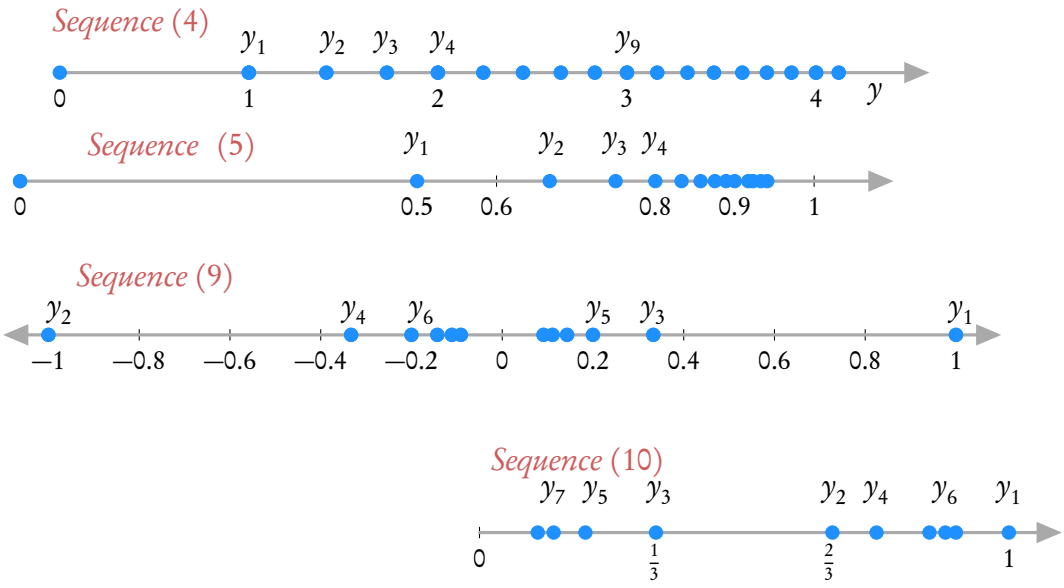


Figure 6: Mapping the sequences along a line.

Reader: The terms of sequences (4) and (5) possess the following property: each term is greater than the preceding term

$$y_1 < y_2 < y_3, \ldots < y_n < \ldots$$

It means that all the terms are arranged on the y-axis according to their serial numbers. As far as I know, such sequence, are called *increasing*.

Author: A more general case is that of *non-decreasing* sequences provided we add the equality sign to the above series of inequalities.

Definition

A sequence (y_n) is called nondecreasing if

$$y_1 \leqslant y_2 \leqslant y_3, \ldots \leqslant y_n \leqslant \ldots$$

A sequence (y_n) is called non-increasing if

$$y_1 \geqslant y_2 \geqslant y_3, \ldots \geqslant y_n \geqslant \ldots$$

Non-decreasing and non-increasing sequences come under the name of monotonic sequences.

Please, identify monotonic sequences among examples (1) -(13).

Reader: Sequences (1), (2), (3), (4), (5), (11), (12), and (13) are nondecreasing, while (6) and (7) are non-increasing. Sequences (8), (9), and (10) are not monotonic.

Author: Let us formulate one more definition.

Definition

A sequence (y_n) is bounded if there are two numbers A and B labelling the range which encloses all the terms of a sequence

$$A \leqslant y_n \leqslant B \ (n = 1, 2, 3, \ldots)$$

If it is impossible to identify such two numbers (or, in particular, one can find only one of the two such numbers, either the least or the greatest), such a sequence is *unbounded*. Do you find bounded sequences among our examples?

Reader: Apparently, (5) is bounded.

Author: Find the numbers A and B for it.

Reader: $A = 1/2, B = 1$.

Author: Of course, but if there exists even one pair of A and B, one

may find any number of such pairs. You could say, for example, that $A = 0$, $B = 2$, or $A = -100$, $B = 100$, etc., and be equally right.

READER: Yes, but my numbers are more accurate.

AUTHOR: From the viewpoint of the bounded sequence definition, my numbers A and B are not Letter and not worse than yours. However, your last sentence is peculiar. What do you mean by saying "more accurate?"

READER: My A is apparently the greatest of all possible lower bounds, while my B is the least of all possible upper hounds.

AUTHOR: The first part of your statement is doubtlessly correct, while the second part of it, concerning B, is not so self-explanatory. It needs proof.

READER: But it seemed rather obvious. Because all the terms of (5) increase gradually, and evidently tend to unity, always remaining less than unity.

AUTHOR: Well, it is right. But it is not yet evident that $B = 1$ is the least number for which $y_n \leqslant B$ is valid for all n: I stress the point again: your statement is not self-evident, it needs proof.

I shall note also that "self-evidence" of your statement about $B = 1$ is nothing but your subjective impression; it is not a mathematically substantiated corollary.

READER: But how to prove that $B = 1$ is, in this particular case, the least of all possible upper bounds?

AUTHOR: Yes, it can be proved. But let us not move too fast and by all means beware of excessive reliance on so-called self-evident impressions. The warning becomes even more important in the light of the fact that

the bounded- ness of a sequence does not imply at all that the greatest A or the least B must be known explicitly.

Now, let us get back to our sequences and find other examples of bounded sequences. .

Reader: Sequence (7) is also bounded (one can easily find $A = 0, B = 1$). Finally, bounded sequences are (9) (e.g. $A = -1, B = 1$), (10) (e.g. $A = 0, B = 1$), and (13) (e.g. $A = 3, B = 4$). The remaining' sequences are un-bounded.

Author: You are quite right. Sequences (5), (7), (9), (10), and (13) are bounded. Note that (5), (7), and (13) are bounded and at the same time monotonic. Don't you feel that this fact is somewhat puzzling?

Reader: What's puzzling about it?

Author: Consider, for example, sequence (5). Note that each subsequent term is greater than the preceding one. I repeat, each term! But the sequence contains an infinite number of terms. Hence, if we follow the sequence far enough, we shall see as many terms with increased magnitude (compared to the preceding term) as we wish. Nevertheless, these values will never go beyond a certain "boundary", which in this case is unity. Doesn't it puzzle you?

Reader: Well, generally speaking, it does. But I notice that we add to each preceding term an increment which gradually becomes less and less.

Author: Yes, it is true. But this condition is obviously insufficient to make such a sequence bounded. Take, for example, sequence (4). Here again the "increments" added to each term of the sequence gradually decrease; nevertheless, the sequence is not bounded.

Reader: We must conclude, therefore, that in (5) these "increments" diminish faster than in (4).

Author: All the same, you have to agree that it is not immediately clear that these "increments" may decrease at a rate resulting in the boundedness of a sequence.

Reader: Of course, I agree with that.

Author: The possibility of infinite but bounded sets was not known, for example, to ancient Greeks. Suffice it to recall the famous paradox about Achilles chasing a turtle.

Let us assume that Achilles and the turtle are initially separated by a distance of 1 km, Achilles moves to times faster than the turtle. Ancient Greeks reasoned like this: during the time Achilles covers 1 km the turtle covers 100 m. By the time Achilles has covered these 100 m, the turtle will have made another 10 m, and before Achilles has covered these 10 m, the turtle will have made 1 m more, and so on. Out of these considerations a paradoxical conclusion was derived that Achilles could never catch up with the turtle.

This "paradox" shows that ancient Greeks failed to grasp the fact that a monotonic sequence may be bounded.

Reader: One has to agree that the presence of both the monotonicity and boundedness is something not so simple to understand.

Author: Indeed, this is not so simple. It brings us close to a discussion on the limit of sequence. The point is that if a sequence is both monotonic and bounded, it should necessarily have a limit.

Actually, this point can be considered as the "beginning" of calculus.

Dialogue 2

LIMIT OF SEQUENCE

AUTHOR: What mathematical operations do you know?

READER: Addition, subtraction, multiplication, division, involution (raising to a power), evolution (extracting n root), and taking a logarithm or a modulus.

AUTHOR: In order to pass from elementary mathematics to higher mathematics, this "list" should be supplemented with one more mathematical operation, namely, that of finding the limit of sequence; this operation is called sometimes the limit transition (or passage to the limit). By the way, we shall clarify below the meaning of the last phrase of the previous dialogue, stating that calculus "begins" where the limit of sequence is introduced.

READER: I heard that higher mathematics uses the operations of *differentiation* and *integration*.

AUTHOR: These operations, as we shall see, are in essence nothing but the variations of the limit transition.

Now, let us get down to the concept of the *limit of sequence*. Do you know what it is?

READER: I learned the definition of the limit of sequence. However, I doubt that I can reproduce it from memory.

AUTHOR: But you seem to "feel" this notion somehow? Probably, you can indicate which of the sequences discussed above have limits and what the value of the limit is in each case.

READER: I think I can do this, The limit is 1 for sequence (5), zero for (7) and (9), and π for (13).

AUTHOR: That's right. The remaining sequences have no limits.

READER: By the way, sequence (9) is not monotonic ...

AUTHOR: Apparently, you have just remembered the end of our previous dialogue where it was stated that if a sequence is both monotonic and bounded, it has a limit.

READER: That's correct. But isn't this a contradiction?

AUTHOR: Where do you find the contradiction? Do you think that from the statement "If a sequence is both monotonic and bounded, it has a limit" one should necessarily draw a reverse statement like "If a sequence has a limit, it must be monotonic and bounded?" Later we shall see that a necessary condition for a limit is only the boundedness of a sequence. The monotonicity is not mandatory at all; consider, for example, sequence (9).

Let us get back to the concept of the limit of sequence. Since you have correctly indicated the sequences that have limits, you obviously have some understanding of this concept. Could you formulate it?

READER: A limit is a number to which a given sequence tends (converges).

AUTHOR: What do you mean by saying "converges to a number"?

READER: I mean that with an increase of the serial number, the terms of a sequence converge very closely to a certain value.

AUTHOR: What do you mean by saying "very closely"?

READER: Well, the "difference" between the values of the terms and the given number will become infinitely small. Do you think any additional explanation is needed?

AUTHOR: The definition of the limit of sequence which you have suggested can at best be classified as a subjective impression. We have already discussed a similar situation in the previous dialogue.

Let us see what is hidden behind the statement made above. For this purpose, let us look at a rigorous definition of the limit of sequence which we are going to examine in detail.

Definition

The number a is said to be the limit of sequence (y_n) if for any positive number ε there is a real number N such that for all $n > N$ the following inequality holds:

$$|y_n - a| < \varepsilon \tag{14}$$

READER: I am afraid, it is beyond me to remember such a definition.

AUTHOR: Don't hasten to remember. Try to comprehend this definition, to realize its structure and its inner logic. You will see that every word in this phrase carries a definite and necessary content and that no other

definition of the limit of sequence could be more succinct (more delicate, even).

First of all, let us note the logic of the sentence. A certain number is the limit provided that for any $\varepsilon > 0$ *there is* a number N such that for all $n > N$ inequality (14) holds. In short, it is necessary that for any ε a certain number N should exist.

Further, note two "delicate" aspects in this sentence. First, the number N should exist for any positive number ε. Obviously, *there is* an infinite set of such ε. Second, inequality (14) should hold always (i.e. for each ε) for *all* $n > N$. But there is an equally infinite set of numbers n!

Reader: Now, the definition of the limit has become more obscure.

Author: Well, it is natural. So far we have been examining the definition "piece by piece". It is very important that the "delicate" features, the "cream", so to say, are spotted from the very outset. Once you understand them, everything will fall into place.

In Figure 7(a) there is a graphic image of a sequence. Strictly speaking, the first 40 terms have been plotted on the graph. Let us assume that if any regularity is noted in these 40 terms, we shall conclude that the regularity does exist for $n > 40$.

Can we say that this sequence converges to the number a (in other words, the number a is the limit of the sequence)?

Reader: It seems plausible.

Author: Let us however, act not on the basis of our impressions but on the basis of the definition of the limit of sequence. So, we want to verify whether the number a is the limit of the given sequence. What does our

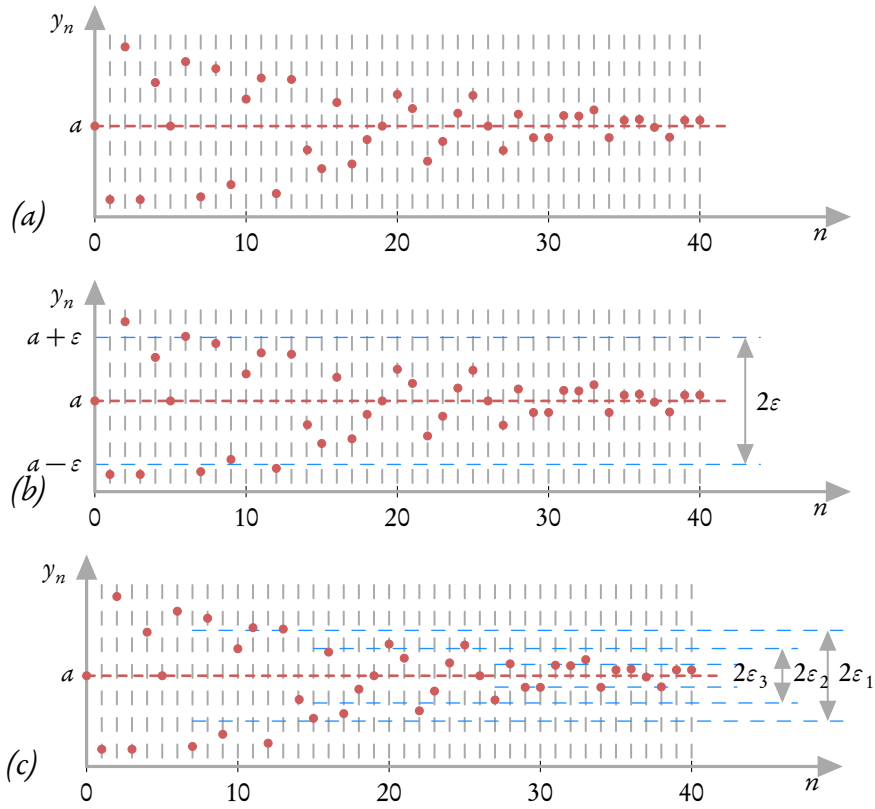


Figure 7: Finding limit of a sequence.

definition of the limit prescribe us to do?

READER: We should take a positive number ε.

AUTHOR: Which number?

READER: Probably. it must be small enough,

AUTHOR: The words "small enough" are neither here nor there. The number ε must be arbitrary.

Thus, we take an arbitrary positive ε. Let us have a look at Figure 7and

layoff on the y-axis an interval of length ε, both upward and downward from the same point a. Now, let us draw through the points $y = a + \varepsilon$ and $y = a - \varepsilon$ the horizontal straight lines that mark an "allowed" band for our sequence. If for any term of the sequence inequality (14) holds, the points on the graph corresponding to these terms fall inside the "allowed" band. We see in Figure 7(b) that starting from number ε, all the terms of the sequence stay within the limits of the "allowed" band, proving the validity of (14) for these terms. We, of course, assume that this situation will realize for all $n > 40$, i.e. for the whole infinite "tail" of the sequence not shown in the diagram.

Thus, for the selected ε the number N does exist. In this particular case we found it to be 7.

READER: Hence, we can regard a as the limit of the sequence.

AUTHOR: Don't you hurry. The definition clearly emphasizes: "for *any* positive ε". So far we have analyzed only one value of ε. We should take another value of ε and find N not for a larger but for a smaller ε. If for the second ε the search of N is a success, we should take a third, even smaller ε, and then a fourth, still smaller ε, etc., repeating each time the operation of finding N.

In Figure 7(c) three situations are drawn up for $\varepsilon_1, \varepsilon_2$and ε_3 (in this case $\varepsilon_1 > \varepsilon_2 > \varepsilon_3$). Correspondingly, three "allowed" bands are plotted on the graph. For a greater clarity, each of these bands has its own starting N. We have chosen $N_1 = 7$, $N_2 = 15$, and $N_3 = 27$.

Note that for each selected ε we observe the same situation in Figure 7(c): up to a certain n, the sequence, speaking figuratively, may be "indisciplined" (in other words, some terms may fallout of the limits of the cor-

responding "allowed" band). However, after a certain n is reached, a very rigid law sets in, namely, *all the remaining terms of the sequence* (their number is infinite) *do stay within the band*.

READER: Do we really have to check it for an infinite number of ε values?

AUTHOR: Certainly not. Besides, it is impossible. We must be sure that *whichever* value of $\varepsilon > 0$ we take, there is such N after which the whole infinite "tail" of the sequence will get "locked up" within the limits of the corresponding "allowed" band.

READER: And what if we are not so sure?

AUTHOR: If we are not and if one can find a value of ε_1 such that it is impossible to "lock up" the infinite "tail" of the sequence within the limits of its "allowed" hand, then a is not the limit of our sequence.

READER: And when do we reach the certainty?

AUTHOR: We shall talk this matter over at a later stage because it has nothing to do with the essence of the definition of the limit of sequence. I suggest that you formulate this definition anew. Don't try to reconstruct the wording given earlier, just try to put it in your own words.

READER: I'll try. The number a is the limit of a given sequence if for any positive ε there is (one can find) a serial number n such that for all subsequent numbers (i.e. for the whole infinite "tail" of the sequence) the following inequality holds: $|y_n - a| < \varepsilon$.

AUTHOR: Excellent. You have almost repeated word by word the definition that seemed to you impossible to remember.

READER: Yes, in reality it all has turned out to be quite logical and rather easy.

AUTHOR: It is worthwhile to note that the dialectics of thinking was clearly at work in this case: a concept becomes "not difficult" because the "complexities" built into it were clarified. First, we break up the concept into fragments, expose the "complexities", then examine the "delicate" points, thus trying to reach the "core" of the problem. Then we recompose the concept to make it integral, and, as a result, this reintegrated concept becomes sufficiently simple and comprehensible. In the future we shall try first to find the internal structure and internal logic of the concepts and theorems.

I believe we can consider the concept of the limit of sequence as thoroughly analyzed. I should like to add that, as a result, the meaning of the sentence "the sequence converges to a" has been explained. I remind you that initially this sentence seemed to you as requiring no additional explanations.

READER: At the moment it does not seem so self-evident any more. True, I see now quite clearly the idea behind it.

AUTHOR: Let us get back to examples (5), (7), and (9). Those are the sequences that we discussed at the beginning of our talk. To begin with, we note that the fact that a sequence (y_n) converges to a certain number a is conventionally written as

$$\lim_{n\to\infty} y_n = a$$

(it reads like this: "The limit of y_n for n tending to infinity is a.")

Using the definition of the limit, let us prove that

$$\lim_{n\to\infty} \frac{n}{n+1} = 1; \quad \lim_{n\to\infty} \frac{1}{n} = 0;$$

$$\lim_{n\to\infty} \left\{ \frac{1}{2n}[1-(-1)^n] - \frac{1}{2(n-1)}[1+(-1)^n] \right\} = 0$$

You will begin with the first of the above problems.

READER: I have to prove that

$$\lim_{n\to\infty} \frac{n}{n+1} = 1$$

I choose an arbitrary value of ε, for example, $\varepsilon = 0.1$.

AUTHOR: I advise you to begin with finding the modulus of $|y_n - a|$.

READER: In this case, the-modulus is

$$\left|\frac{n}{n+1} - 1\right| = \frac{1}{n+1}$$

AUTHOR: Apparently ε needn't be specified, at least at the beginning.

READER: O.K. Therefore. for an arbitrary positive value of ε, I have to find N such that for all $n > N$ the following inequality holds

$$\frac{1}{n+1} < \varepsilon$$

AUTHOR: Quite correct. Go on.

READER: The inequality can be rewritten in the form t

$$n > \frac{1}{\varepsilon} - 1$$

It follows that the unknown N may be identified as an integral part of $\frac{1}{\varepsilon} - 1$. Apparently, for all $n > N$ the inequality in question will hold.

AUTHOR: That's right. Let, for example, $\varepsilon = 0.01$.

READER: Then $N = \frac{1}{\varepsilon} - 1 = 100 - 1 = 99$.

AUTHOR: Let $\varepsilon = 0.001$.

READER: Then $N = \frac{1}{\varepsilon} - 1 = 1000 - 1 = 999$.

AUTHOR: Let $\varepsilon = 0.00015$.

READER: Then $N = \frac{1}{\varepsilon} - 1 = 6665$.

AUTHOR: It is quite evident that for any ε (no matter how small) we can find a corresponding N.

As to proving that the limits of sequences (7) and (9) are zero, we shall leave it to the reader as an exercise.

READER: But couldn't the proof of the equality $\lim\limits_{n\to\infty} \frac{n}{n+1} = 1$ be simplified?

AUTHOR: Have a try.

READER: Well, first I rewrite the expression in the following way:

$$\lim_{n\to\infty} \frac{n}{n+1} = \lim_{n\to\infty} \frac{1}{\frac{1}{n+1}}$$

Then I take into consideration that with an increase in n, fraction $\frac{1}{n}$ will tend to zero, and, consequently, can be neglected against unity. Hence, we may reject $\frac{1}{n}$ and have: $\lim\limits_{n\to\infty} \frac{1}{1} = 1$.

AUTHOR: In practice this is the method generally used. However one should note that in this case we have assumed, first, that $\lim\limits_{n\to\infty} \frac{1}{n} = 0$, and, second, the validity of the following rules

$$\lim_{n\to\infty} \frac{x_n}{y_n} = \frac{\lim\limits_{n\to\infty} x_n}{\lim\limits_{n\to\infty} y_n} \tag{15}$$

$$\lim_{n\to\infty} (x_n + z_n) = \lim_{n\to\infty} x_n + \lim_{n\to\infty} z_n \tag{16}$$

where $x_n = 1$, $y_n = 1 + \frac{1}{n}$ and $z_n = \frac{1}{n}$. Later on we shall discuss these rules, but at this juncture I suggest that we simply use them to compute several limits. Let us discuss two examples.

Example 1 Find $\lim\limits_{n\to\infty} \frac{3n-1}{5n-6}$.

READER: It will be convenient to present the computation in the form:

$$\lim_{n\to\infty} \frac{3n-1}{5n-6} = \lim_{n\to\infty} \frac{3-\frac{1}{n}}{5-\frac{6}{n}} = \frac{\lim\limits_{n\to\infty}\left(3-\frac{1}{n}\right)}{\lim\limits_{n\to\infty}\left(5-\frac{6}{n}\right)} = \frac{3}{5}$$

AUTHOR: Ok. **Example 2** Compute

$$\lim_{n\to\infty} \frac{6n^2-1}{5n^2+2n-1}$$

READER: We write

$$\lim_{n\to\infty} \frac{6n^2-1}{5n^2+2n-1} = \lim_{n\to\infty} \frac{6n-\frac{1}{n}}{5n+2-\frac{1}{n}}.$$

AUTHOR: Wait a moment! Did you think about the reason for dividing both the numerator and denominator of the fraction in the previous example by n? We did this because sequences $(3n-1)$ and $(5n-6)$ obviously have no limits, and therefore rule (15) fails. However, each of sequences $\left(3-\frac{1}{n}\right)$ and $\left(5-\frac{6}{n}\right)$ has a limit.

READER: I have got your point. It means that in **Example 2** I have to divide both the numerator and denominator by n^2 to obtain the sequences

with limits in both. Accordingly we obtain

$$\lim_{n\to\infty}\frac{6n^2-1}{5n^2+2n-1}=\lim_{n\to\infty}\frac{6-\dfrac{1}{n^2}}{5+\dfrac{2}{n}-\dfrac{1}{n^2}}$$
$$=\frac{\lim\limits_{n\to\infty}\left(6-\dfrac{1}{n^2}\right)}{\lim\limits_{n\to\infty}\left(5+\dfrac{2}{n}-\dfrac{1}{n^2}\right)}$$

AUTHOR: Well, we have examined the concept of the limit of sequence. Moreover, we have learned a little how to calculate limits. Now it is time to discuss some properties of sequences with limits. Such sequences are called *convergent*.

Dialogue 3

CONVERGENT SEQUENCE

AUTHOR: Let us prove the following theorem:

Theorem

If a sequence has a limit, it is bounded.

We assume that a is the limit of a sequence (y_n). Now take an arbitrary value of ε greater than 0. According to the definition of the limit, the selected ε can always be related to N such that for all $n > N$, $Iy_n - a| < \varepsilon$ starting with $n = N + 1$, all the subsequent terms of the sequence satisfy the following inequalities

$$a - \varepsilon < y_n < a + \varepsilon$$

As to the terms with serial numbers from 1 to N, it is alway possible to select both the greatest (denoted by B_1) and the least (denoted by A_1) terms since the number of these term is *finite*.

Now we have to select the least value from $a - \varepsilon$ and A_1 (denoted by A) and the greatest value from $a + \varepsilon$ and B_1 (denoted by B). It is obvious

that $A \leqslant y_n \leqslant B$ for all the terms of our sequence, which proves that the sequence y_n is bounded.

READER: I see.

AUTHOR: Not too well, it seems. Let us have a look at the logical structure of the proof. We must verify that if the sequence has a limit, there exist two numbers A and B such that $A \leqslant y_n \leqslant B$ for each term of the sequence. Should the sequence contain a *finite* number of terms, the existence of such two numbers would be evident. However, the sequence contains an *infinite* number of terms, the fact that complicates the situation.

READER: Now it is clear! The point is that if a sequence has a limit a, one concludes that in the interval from $a-\varepsilon$ to $a+\varepsilon$ we have an infinite set of y_n starting from $n=N+1$ so that *outside of this interval* we shall find only a finite number of terms (not larger than N).

AUTHOR: Quite correct. As you see, the limit "takes care of" all the complications associated with the behaviour of the infinite "tail" of a sequence. Indeed, $|y_n-a|<\varepsilon$ for *all* $n>N$, and this is the main "delicate" point of this theorem. As to the first N terms of a sequence, it is essential that their set is finite.

READER: Now it is all quite lucid. But what about ε Its value is not preset, we have to select it.

AUTHOR: A selection of a value for ε affects only N. If you take a smaller ε, you will get, generally speaking, a larger N. However, the number of the terms of a sequence which do not satisfy $|y_n-a|<\varepsilon$ will remain finite.

And now try to answer the question about the validity of t.he converse theorem: If a sequence is bounded, does it imply it is convergent as well?

Reader: The converse theorem is not true. For example, sequence (10) which was discussed in the first dialogue is bounded. However, it has no limit.

Author: Right you are. We thus come to a corollary:

Corollary

The boundedness of a sequence is a necessary condition for its convergence; however, it is not a sufficient condition. If a sequence is convergent, it is bounded. If a sequence is unbounded, it is definitely non-convergent,

Reader: I wonder whether there is a sufficient condition for the convergence of a sequence?

Author: We have already mentioned this condition in the previous dialogue, namely, simultaneous validity of both the boundedness and monotonicity of a sequence. The **Weierstrass Theorem** states:

Weierstrass Theorem

If a sequence is both bounded and monotonic, it has a limit.

Unfortunately, the proof of the theorem is beyond the scope of this book; we shall not give it. I shall simply ask you to look again at sequences (5), (7), and (13) (see Dialogue One), which satisfy the conditions of the Weierstrass theorem.

Reader: As far as I understand, again the converse theorem is not true. Indeed, sequence (9) (from Dialogue One) has a limit but is not monotonic.

Author: That is correct. We thus come to the following conclusion.

Conclusion

If a sequence is both monotonic and bounded, it is a sufficient (but not necessary) condition for its convergence.

READER: Well, one can easily get confused.

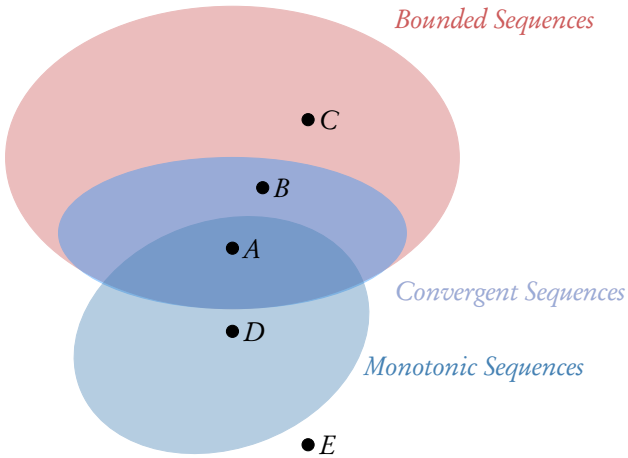


Figure 8: Bounded, monotonic and convergent sequences.

AUTHOR: In order to avoid confusion, let us have a look at another illustration (Figure 8). Let us assume that all bounded sequences are "collected" (as if we were picking marbles scattered on the floor) in an area shaded by horizontal lines, all monotonic sequences are collected in an area shaded by tilted lines, and, finally, all convergent sequences are collected in an area shaded by vertical lines. Figure 8 shows how all these areas overlap, in accordance with the theorems discussed above (the actual shape of all the areas is, of course, absolutely arbitrary). As follows from the figure, the area shaded vertically is completely included into the area shaded horizontally. It means that *any convergent sequence must be also bounded*. The overlapping of the areas shaded horizontally and by

tilted lines occurs inside the area shaded vertically. It means that *any sequence that is both bounded and monotonic must be convergent as well*. It is easy to deduce that only five types of sequences are possible. In the figure the points designated by A, B, C, D, and E identify five sequences of different types. Try to name these sequences and find the corresponding examples among the sequences discussed in Dialogue One.

READER: Point A falls within the intersection of all the three areas. It represents a sequence which is at the same time bounded, monotonic, and convergent. Sequences (5), (7), and (13) are examples of such sequences.

AUTHOR: Continue, please.

READER: Point B represents a bounded, convergent hut non-monotonic sequence. One example is sequence (9).

Point C represents a bounded but neither convergent nor monotonic sequence. One example of such a sequence is sequence (10).

Point D represents a monotonic but neither convergent nor bounded sequence. Examples of such sequences are (1), (2), (3), (4), (6), (11), and (12).

Point E is outside of the shaded areas and thus represents it sequence neither monotonic nor convergent nor bounded. One example is sequence (8).

AUTHOR: What type of sequence is impossible then?

READER: There can be no bounded, monotonic, and non-convergent sequence. Moreover, it is impossible to have both unboundedness and convergence in one sequence.

AUTHOR: As you see, Figure 8 helps much to understand t.he relationship

between such properties of sequences as *boundedness, monotonicity, and convergence*.

In what follows, we shall discuss only convergent sequences. We shall prove the following theorem:

Theorem

A convergent sequence has only one limit.

This is the theorem of the *uniqueness of the limit*. It means that a convergent sequence cannot have two or more limits. Suppose the situation is contrary to the above statement.

Consider a convergent sequence with two limits a_1 and a_2 and select a value for $\varepsilon < \frac{|a_1 - a_2|}{2}$. Now assume, for example, that $\varepsilon = \frac{|a_1 - a_2|}{2}$. Since a_1 is a limit, then for the selected value of ε there is N_1 such that for all $n > N_1$ the terms of the sequence (its infinite "tail") must fall inside the interval **1** (Figure 9). It means that we must have $|y_n - a_1| < \varepsilon$. On the other hand, since at is a_2 limit there is N_2 such that for all $n > N_2$ the terms of the sequence (again its infinite "tail") must fall inside the interval **2**. It means that we must have$|y_n - a_2| < \varepsilon$. Hence, we obtain that for all N greater than the largest among N_1 and N_2 the impossible must hold, namely, the terms of the sequence must simultaneously belong to the intervals **1** and **2**. This contradiction proves the theorem. This proof contains at least two rather "delicate" points. Can you identify them?

READER: I certainly notice one of them. If a_1 and a_2 are limits, no matter how the sequence behaves *at the beginning*, its terms in the long run have to concentrate *simultaneously* around a_1 and a_2 which is, of course, impossible

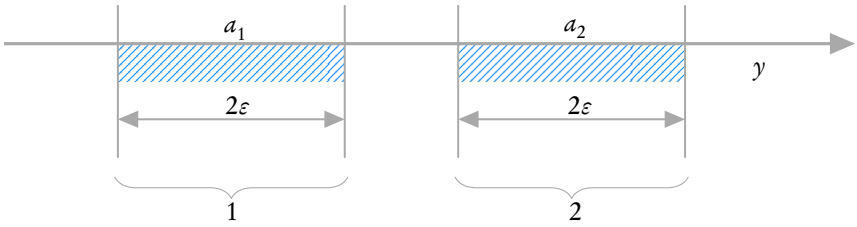


Figure 9: Proving the uniqueness of the limit.

Author: Correct. But there is one more "delicate" point, namely, no matter how close a_1 and a_2 are, the; should inevitably be spaced by a segment (a gap) of a small but definitely nonzero length.

Reader: But it is self-evident.

Author: I agree. However, this "self-evidence" is connected to one more very fine aspect without which the very calculus could not be developed. As you probably noted, one cannot identify on the real line two neighbouring points. If one point is chosen, it is impossible, in principle, to point out its "neighbouring" point. In other words, no matter how carefully you select a pair of points on the real line, it is always possible to find any number of points between the two.

Take, for example, the interval $[0, 1]$. Now, exclude the point 1. You will have a half-open interval $[0, 1|$. Can you identify the largest number over this interval?

Reader: No, it is impossible.

AUTHOR: That's right. However, if there were a point neighbouring 1, after the removal of the latter this "neighbour" would have become the largest number. I would like to note here that many "delicate" points and many "secrets" in the calculus theorems are ultimately associated with the impossibility of identifying two neighbouring points on the real line, or of specifying the greatest or least number on an open interval of the real line.

But let us get back to the properties of convergent sequences and prove the following theorem:

Theorem

1f sequences (y_n) and (z_n) are convergent (we denote their limits by a and b, respectively), a sequence $(y_n + z_n)$ is convergent too, its limit being $a + b$.

READER: This theorem is none other than rule (16) discussed in the previous dialogue.

AUTHOR: Thats right. Nevertheless, I suggest you try to prove it.

READER: If we select an arbitrary $\varepsilon > 0$, then there is a number N_1 such that for all the terms of the first sequence with $n > N_1$ we shall have $|y_n - a| < \varepsilon$. In addition, for the same ε there is N_2 such that for all the terms of the second sequence with $n > N_2$ we shall have $|z_n - a| < \varepsilon$. Now we select the greatest among N_1 and N_2 (we denote it by N), then for all $n > N$ both $|y_n - a| < \varepsilon$ and $|z_n - a| < \varepsilon$. Well, this is as far as I can go.

AUTHOR: Thus, you have established that for an arbitrary ε there is N such that for all $n > N$ both $|y_n - a| < \varepsilon$ and $|z_n - a| < \varepsilon$ simultaneously.

And what can you say about the modulus $|(y_n+z_n)-(a+b)|$ (for all n)? I remind you that $|A+B| \leqslant |A|+|B|$.

READER: Let us look at

$$\begin{aligned} |(y_n+z_n)-(a+b)| &= |(y_n-a)+(z_n-b)| \\ &\leqslant [|y_n-a|+|z_n-b|] \\ &< (\varepsilon+\varepsilon) = 2\varepsilon \end{aligned}$$

AUTHOR: You have proved the theorem, haven't you?

READER: But we have only established that there is N such that for all $n > N$ we have $|(y_n+z_n)-(a+b)| < 2\varepsilon$. But we need to prove that

$$|(y_n+z_n)-(a+b)| < \varepsilon$$

AUTHOR: Ah, that's peanuts, if you forgive the expression. In the case of the sequence (y_n+z_n) you select a value of ε, but for the sequences (y_n) and (z_n) you must select a e value of $\dfrac{\varepsilon}{2}$ and namely for this value find N_1 and N_2.

Thus, we have proved that if the sequences (y_n) and (z_n) are convergent, the sequence (y_n+z_n) is convergent too. We have even found a limit of the sum. And do you think that the converse is equally valid?

READER: I believe it should be.

AUTHOR: You are wrong. Here is a simple illustration:

$$\begin{aligned} (y_n) &= \frac{1}{2}, \frac{2}{3}, \frac{1}{4}, \frac{4}{5}, \frac{1}{6}, \frac{6}{7}, \frac{1}{8}, \cdots \\ (z_n) &= \frac{1}{2}, \frac{1}{3}, \frac{3}{4}, \frac{1}{5}, \frac{5}{6}, \frac{1}{7}, \frac{7}{8}, \cdots \\ (y_n+z_n) &= 1, 1, 1, 1, 1, 1, 1, \ldots \end{aligned}$$

As you see, the sequences (y_n) and (z_n) are not convergent, while the sequence $(y_n + z_n)$ is convergent, its limit being equal to unity.

Thus, if a sequence $(y_n + z_n)$ is convergent, two alternatives are possible:

- sequences (y_n) and (z_n) are convergent as well, or
- sequences (y_n) and (z_n) are divergent.

READER: But can it be that the sequence (y_n) is convergent, while the sequence (z_n) is divergent?

AUTHOR: It may be easily shown that this is Impossible.

To begin with, let us note that if the sequence (y_n) has a limit a, the sequence $-(y_n)$ is also convergent and its limit is $-a$. This follows from an easily proved equality

$$\lim_{n\to\infty}(c\,y_n) = c\lim_{n\to\infty} y_n$$

where c is a constant. Assume now that a sequence (y_n+z_n) is convergent to A, and that (y_n) is also convergent and its limit is a. Let us apply the theorem on the sum of convergent sequences to the sequences $(y_n + z_n)$ and $-(y_n)$. As a result, we obtain that the sequence $(y_n + z_n) - (y_n)$, i.e. (z_n), is also convergent, with the limit $A - a$.

READER: Indeed (z_n) cannot be divergent in this case.

AUTHOR: Very well. Let us discuss now one important particular case of convergent sequences, namely, the so-called *infinitesimal sequence*, or simply, *infinitesimal*. This is the name which is given to a convergent sequence with a limit equal to zero. Sequences (7) and (9) from Dialogue One are examples of infinitesimals.

Note that to any convergent sequence (y_n) with a limit a there corresponds an infinitesimal sequence (α_n), where $\alpha_n = y_n - a$. That is why mathematical analysis is also called calculus of infinitesimals.

Now I invite you to prove the following theorem:

Theorem

If (y_n) is a bounded sequence and (α_n) is infinitesimal, then $(y_n \alpha_n)$ is infinitesimal as well.

Reader: Let us select an arbitrary $\varepsilon > 0$. We must prove that there is N such that for all $n > N$ the terms of the sequence $(y_n \alpha_n)$ satisfy the inequality $|y_n \alpha_n| < \varepsilon$).

Author: Do you mind a hint? As the sequence (y_n) is bounded, one can find M such that $|y_n| \leqslant M$ for any n.

Reader: Now all becomes very simple. We know that the sequence (α_n) is infinitesimal. It means that for any $\varepsilon' > 0$ we can find N such that for all $n > N |\alpha_n| < \varepsilon'$. For ε', I select $\dfrac{\varepsilon}{M}$ Then, for $n > N$ we have

$$|y_n \alpha_n| = |y_n||\alpha_n| \leqslant M|\alpha_n| < M\frac{\varepsilon}{M} = \varepsilon$$

This completes the proof.

Author: Excellent. Now, making use of this theorem, it is very easy to prove another theorem:

Theorem

A sequence $(y_n z_n)$ is convergent to ab if sequences (y_n) and (z_n) are convergent to a and b, respectively.

Suppose $y_n = a + \alpha_n$ and $z_n = b + \beta_n$. Suppose also that the sequences

(α_n) and (β_n) are infinitesimal. Then we can write:

$$y_n z_n = ab + \gamma_n \text{ where } \gamma_n = b\alpha_n + a\beta_n + \alpha_n\beta_n$$

Making use of the theorem we have just proved, we conclude that the sequences $(b\alpha_n)$, $(a\beta_n)$ and $(\alpha_n\beta_n)$ are infinitesimal.

READER: But what justifies your conclusion about the sequence $(\alpha_n\beta_n)$?

AUTHOR: Because any convergent sequence (regardless of whether it is infinitesimal or not) is bounded. From the theorem on the sum of convergent sequences we infer that the sequence (y_n) is infinitesimal, which immediately yields

$$\lim_{n\to\infty}(y_n z_n) = ab$$

This completes the proof.

READER: Perhaps we should also analyze inverse variants in which the sequence $(y_n z_n)$ is convergent. What can be said in this case about the sequences (y_n) and (z_n)?

AUTHOR: Nothing definite. in the general case. Obviously, one possibility is that (y_n) and (z_n) are convergent. However, it is also possible, for example, for the sequence (y_n) to be convergent, while the sequence (z_n) is divergent. Here is a simple illustration:

$$(y_n) = 1, \frac{1}{4}, \frac{1}{9}, \frac{1}{16}, \frac{1}{25}, \cdots \frac{1}{n^2}, \cdots$$
$$(z_n) = 1, 2, 3, 4, 5, \ldots, n, \ldots$$
$$(y_n z_n) = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \cdots \frac{1}{n}, \cdots$$

By the way, note that here we obtain an infinitesimal sequence by multiplying an infinitesimal sequence by an unbounded sequence. In the general case, however, such multiplication needn't produce an infinitesimal.

Finally, there is a possibility when the sequence $(y_n z_n)$ is convergent, and the sequences (y_n) and (z_n) are divergent. Here is one example:

$$(y_n) = 1, \frac{1}{4}, 3, \frac{1}{16}, 5, \ldots \frac{1}{36}, 7, \ldots$$
$$(z_n) = 1, 2, \frac{1}{9}, 4, \frac{1}{25}, 6, \frac{1}{49}, \ldots$$
$$(y_n z_n) = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \ldots$$

Now, let us formulate one more theorem:

Theorem

If (y_n) and (z_n) are sequences convergent to a and b when $b \neq 0$, then a sequence $\left(\frac{y_n}{z_n}\right)$ is also convergent, its its limit being $\frac{a}{b}$.

We shall omit the proof of this theorem,

Reader: And what if the sequence (z_n) contains zero terms?

Author: Such terms are possible. Nevertheless, the number of such terms can be only finite. Do you know why?

Reader: I think, I can guess. The sequence (z_n) has a non-zero limit b.

Author: Let us specify $b > 0$.

Reader: Well, I select $\varepsilon = \frac{b}{2}$. There must be an integer N such that $|z_n - b| < \frac{b}{2}$ for all $n > N$. Obviously all z_n (the whole infinite "tail" of the sequence) will be positive. Consequently, the zero terms of the sequence (z_n) may only be encountered among a *finite* number of the first N terms.

Author: Excellent. Thus, the number of zeros among the terms of (z_n)

can only be finite. If such is the case, one can surely drop these terms. Indeed, *an elimination of any finite number of terms of a sequence does not affect its properties*. For example, a convergent sequence still remains convergent, with its limit unaltered. An elimination of a finite number of terms may only change N (for a given ε), which is certainly unimportant.

READER: It is quite evident to me that by *eliminating* a finite number of terms one does not affect the convergence of a sequence. But could an *addition* of a finite number of terms affect the convergence of a sequence?

AUTHOR: A finite number of new terms does not affect the convergence of a sequence either. No matter how many new terms are added and what their new serial numbers are, one can always find the greatest number N after which the whole infinite "tail" of the sequence is unchanged. No matter how large the number of new terms may be and where you insert them, the finite set of new terms cannot change the infinite "tail" of the sequence. And it is the "tail" that determines the convergence (divergence) of a sequence.

Thus, we have arrived at the following:

Conclusion

Elimination, addition, and any other change of a finite number of terms of a sequence do not affect either its convergence or its limit (if the sequence is convergent).

READER: I guess that an elimination of an infinite number of terms (for example, every other term) must not affect the convergence of a sequence either.

AUTHOR: Here you must be very careful. If an initial sequence is convergent, an elimination of an infinite number of its terms (provided that the number of the remaining terms is also infinite) does not affect either convergence or the limit of the sequence. If, however, an initial sequence is divergent, an elimination of an infinite number of its terms may, in certain cases, convert the sequence into a convergent one. For example, if you eliminate from divergent sequence (10) (see Dialogue One) all the terms with even serial numbers, you will get the convergent sequence.

$$1, \frac{1}{3}, \frac{1}{5}, \frac{1}{7}, \frac{1}{9}, \frac{1}{11}, \frac{1}{13}, \dots$$

Suppose we form from a given convergent sequence two new convergent sequences. The first new sequence will consist of the terms of the initial sequence with *odd* serial numbers, while the second will consists of the terms with *even* serial numbers. What do you think are the limits of these new sequences?

READER: It is easy to prove that the new sequences will have the same limit as the initial sequence.

AUTHOR: You are right.

Note that from a given convergent sequence we can form not only two but a *finite number m* of new sequences converging to the same limit. One way to do it is as follows. The first new sequence will consist of the 1st, $(m+1)$st, $(2m+1)$st, $(3m+1)$st, etc., terms of the initial sequence. The second sequence will consist of the 2nd, $(m+2)$nd, $(2m+2)$nd, $(3m+2)$nd, etc., terms of the initial sequence.

Similarly we can form the third, the fourth, and other sequences.

In conclusion, let us see how one can "spoil" a convergent sequence by turning it into divergent. Clearly, different "spoiling" approaches are pos-

sible. Try to suggest something simple.

Reader: For example, we can replace all the terms with even serial numbers by a constant that is not equal to the limit of the initial sequence. For example, convergent sequence (5) (see Dialogue One) can be "spoilt" in the following manner:

$$\frac{1}{2}, 2, \frac{3}{4}, 2, \frac{5}{6}, 2, \frac{7}{8}, 2, \ldots$$

Author: I see that you have mastered very well the essence of the concept of a convergent sequence. Now we are ready for another substantial step, namely, consider one of the most important concepts in calculus: the definition of a function.

Dialogue 4

FUNCTION

Reader: Functions are widely used in elementary mathematics.

Author: Yes, of course. You are familiar with *numerical functions*. Moreover, you have worked already with different numerical functions. Nevertheless, it will be worthwhile to dwell on the concept of the function. To begin with, what is your idea of a function?

Reader: As I understand it, a function is a certain correspondence between *two variables*, for example, between x and y. Or rather, it is a dependence of a variable y on a variable x.

Author: What do you mean by a "variable"?

Reader: It is a quantity which may assume different values.

Author: Can you explain what your understanding of the expression "a quantity assumes a value" is? What does it mean? And what are the reasons, in particular, that make a quantity to assume this or that value? Don't you feel that the very concept of a variable quantity (if you are

going to use this concept) needs a definition?

READER: O.K., what if I say: a function $y = f(x)$ symbolizes a dependence of y on x, where x and y are numbers.

AUTHOR: I see that you decided to avoid referring to the concept of a *variable quantity*. Assume that x is a number and y is also a number. But then explain, please, the meaning of the phrase "a dependence between two numbers".

READER: But look, the words "an independent variable" and "a dependent variable" can be found in any textbook on mathematics.

AUTHOR: The concept of a variable is given in textbooks on mathematics after the definition of a function has been introduced.

READER: It seems I have lost my way.

AUTHOR: Actually it is not all that difficult "to construct" an image of a numerical function. I mean *image*, not *mathematical definition* which we shall discuss later.

In fact, a numerical function may be pictured as a "black box" that *generates a number at the output in response to a number at the input*. You put into this "black box" a number (shown by x in Figure 10) and the "black box" outputs a new number (y in Figure 10). Consider, for example, the following function:

$$y = 4x^2 - 1$$

If the input is $x = 2$, the output is $y = 15$; if the input is $x = 3$, the output is $y = 35$; if the input is $x = 10$, the output is $y = 399$, etc.

READER: What does this "black box" look like? You have stressed that Figure 10 is only symbolic.

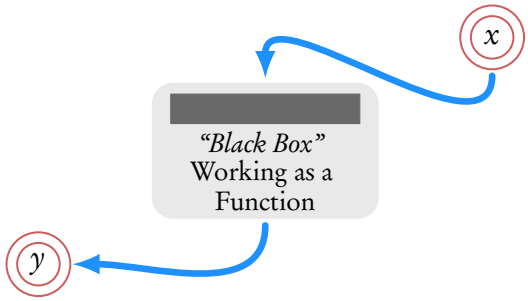


Figure 10: A numerical function as a black box.

AUTHOR: In this particular case it makes no difference. It does not influence the essence of the concept of a function. But a function can also be "pictured" like this:

$$4\square^2 - 1$$

The square in this picture is a "window" where you input the numbers. Note that there may be more than one "window". For example,

$$\frac{4\square^2 - 1}{|\square| + 1}$$

READER: Obviously, the function you have in mind is

$$\frac{4x^2 - 1}{|x| + 1}$$

AUTHOR: Sure. In this case each specific value should be input into both "windows" simultaneously. "Black box" working as a function Figure 10.

By the way, it is always important to see such a "window" (or "windows") in a formula describing the function. Assume, for example, that one needs to pass from a function $y = f(x)$ to a function $y = f(x-1)$ (on a graph of a function this transition corresponds to a displacement of the curve

in the positive direction of the x-axis by 1). If you clearly understand the role of such a "window" ("windows"), you will simply replace in this "window" (these "windows") x by $x-1$. Such an operation is Illustrated by Figure 11 which represents the following function

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

Obviously, as a result of substitution of $x-1$ for x we arrive at a new

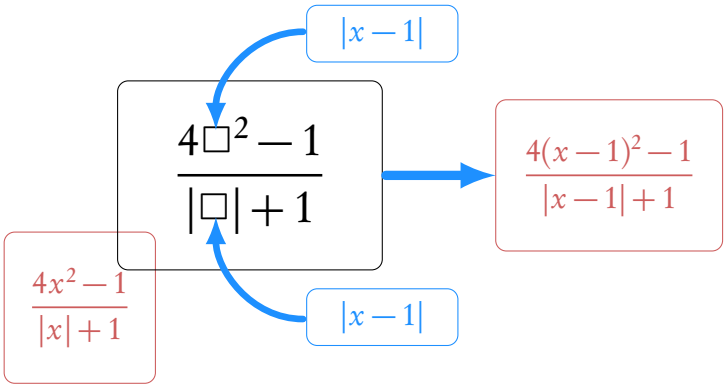


Figure 11: Change in a function from $f(x) \to f(x-1)$.

function (new "black box")

$$\frac{4(\square-1)^2-1}{|\square-1|+1}, \quad y = \frac{4(x-1)^2-1}{|x-1|+1}$$

READER: I see. If, for example, we wanted to pass from $y = f(x)$ to $y = f\left(\frac{1}{x}\right)$, the function pictured in Figure 11 would be transformed as follows:

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

AUTHOR: Correct. Now. try to find $y = f(x)$ if

$$2f\left(\frac{1}{x}\right) - f(x) = 3x$$

READER: I am at a loss.

AUTHOR: As a hint, I suggest replacing x by $\frac{1}{x}$.

READER: This yields

$$2f(x) - f\left(\frac{1}{x}\right) = \frac{3}{x}$$

Now it is clear. Together with the initial equation, the new equation forms a system of two equations for $f(x)$ and $f\left(\frac{1}{x}\right)$:

$$\left.\begin{aligned} 2f\left(\frac{1}{x}\right) - f(x) &= 3x \\ 2f(x) - f\left(\frac{1}{x}\right) &= \frac{3}{x} \end{aligned}\right\}$$

By multiplying all the terms of the second equation by 2 and then adding them to the first equation, we obtain

$$f(x) = x + \frac{2}{x}$$

AUTHOR: Perfectly true.

READER: In connection with your comment about the numerical function as a "black box" generating a numerical output in response to a numerical input, I would like to ask whether other types of "black boxes" are possible in calculus.

AUTHOR: Yes, they are. In addition to the numerical function, we shall discuss the concepts of an operator and a functional.

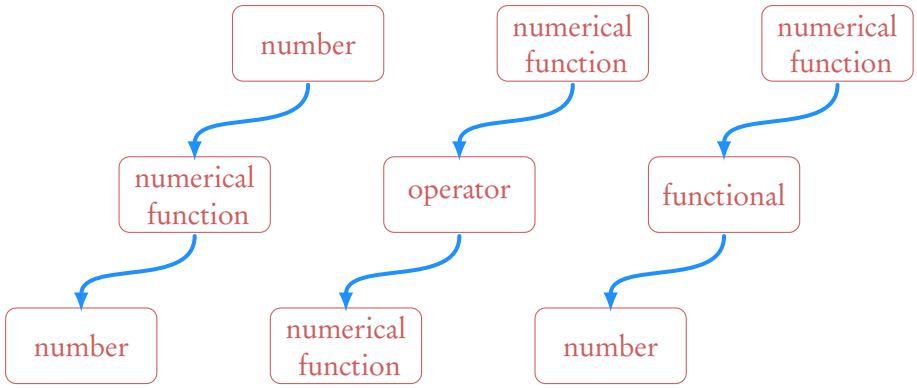


Figure 12: Understanding place and role of the numerical function as a mathematical tool.

READER: I must confess I have never heard of such concepts.

AUTHOR: I can imagine. I think, however, that Figure 12 will be helpful. Besides, it will elucidate the place and role of the numerical function as a mathematical tool. Figure 12 shows that:

- *a numerical function is a* "black box" *that generates a number at the output in response to a number at the input;*
- *an operator is a* "black box" *that generates a numerical function at the output in response to a numerical function at the input*; it is said that all operator applied to a function generates a new function;
- *a functional is a* "black box" *that generates a number at the output in response to a numerical junction at the input*, i.e. a concrete number is obtained "in response" to a concrete function.

READER: Could you give examples of operators and functionals?

AUTHOR: Wait a minute. In the next dialogues we shall analyze both

the concepts of an operator and a functional. So far, we shall confine ourselves to a general analysis of both concepts. Now we get back to our main object, the numerical function.

The question is: How to construct a "black box" that generates a numerical function.

READER: Well, obviously, we should find a relationship, or a law, according to which the number at the "output" of the "black box" could be forecast for each specific number introduced at the "input".

AUTHOR: You have put it quite clearly. Note that such a law could be naturally referred to as the *law of numerical correspondence*. However, the law of numerical correspondence would not be a sufficient definition of a numerical function.

READER: What else do we need?

AUTHOR: Do you think that any number could be fed into a specific "black box" (function)?

READER: I see. I have to define a set of numbers acceptable as inputs of the given function.

AUTHOR: That's right. This set is said to be the *domain of a function*.

Thus, the definition of a numerical function is based on two "cornerstones";

the domain of a function (a certain set of numbers), and the law of numerical correspondence.

According to this law, *every number from the domain of a function is placed in correspondence with a certain number, which is called the value of the*

function; the values form the range of the function.

READER: Thus, we actually have to deal with two numerical sets. On the one hand, we have a set called the domain of a function and, on the other, we have a set called the range of a function.

AUTHOR: At this juncture we have come closest to a mathematical definition of a function which will enable us to avoid the somewhat mysterious word "black box".

Look at Figure 13. It shows the function $y = \sqrt{1-x^2}$. Figure 13 pictures two numerical sets, namely, D (represented by the interval $[-1, 1]$) and E (the interval $[0, 1]$). For your convenience these sets are shown on two different real lines.

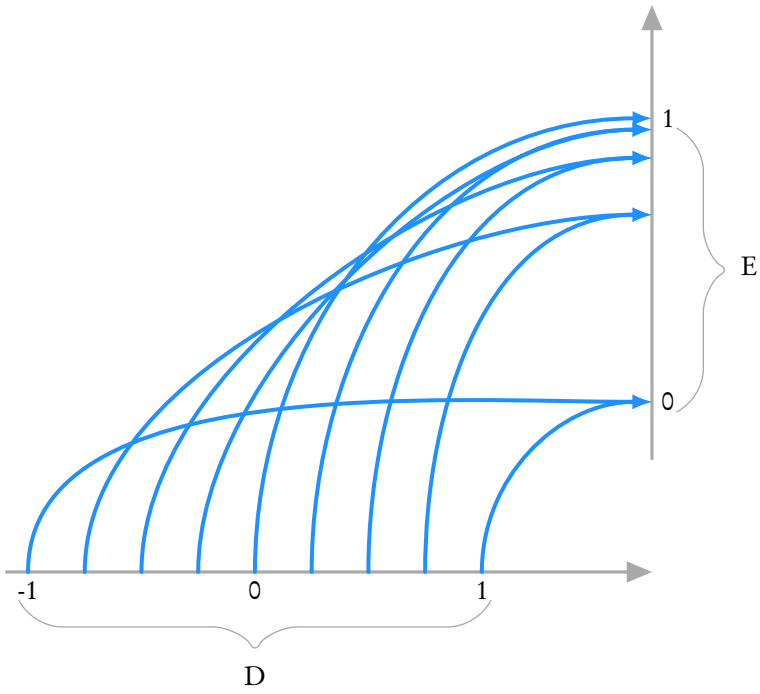


Figure 13: Range and domain of a function.

The set D is the domain of the function, and E is its range. Each number in D corresponds to one number in E (every input value is placed in correspondence with one output value). This correspondence is shown in Figure 13 by arrows pointing from D to E.

READER: But Figure 13 shows that two *different* numbers in D correspond to one number in E.

AUTHOR: It does not contradict the statement "each number in D corresponds to one number in E." I never said that *different* numbers in D must correspond to different numbers in E. Your remark (which actually stems from specific characteristics of the chosen function) is of no principal significance. Several numbers in D may correspond to one number in E. An inverse situation, however, is forbidden. It is not allowed for one number in D to correspond to more than one number in E. I emphasize that each number in D must correspond to *only one* (not more!) number in E.

Now we can formulate a mathematical definition of the numerical function.

Definition

Take two numerical sets D and E in which each element x of D (this is denoted by $x \in D$) is placed in one-to-one correspondence with one element y of E. Then we say that a function $y = f(x)$ is set in the domain D, the range of the function being E. It is said that the argument x of the function y passes through D and the values of y belong to E.

Sometimes it is mentioned (but more often omitted altogether) that both D and E are subsets of the set of real numbers R (by definition, R is the

real line).

On the other hand, the definition of the function can be reformulated using the term "mapping". Let us return again to Figure 13. Assume that the number of arrows from the points of D to the points of E is infinite (just imagine that such arrows have been drawn from each point of D). Would you agree that such a picture brings about an idea that D is *mapped* onto E?

READER: Really, it looks like mapping.

AUTHOR: Indeed, this *mapping* can be used to define the *function*.

Definition

A numerical function is a mapping of a numerical set D (which is the domain of the function) onto another numerical set E (the range of this function).

Thus, the numerical function is a *mapping of one numerical set onto another numerical set.* The term "mapping" should be understood as a kind of numerical correspondence discussed above. In the notation $y = f(x)$, symbol f means the function itself (i.e. the mapping), with $x \in D$ and $y \in E$.

READER: If the *numerical function* is a mapping of one numerical set onto another numerical set, then the operator can be considered as a mapping of a set of numerical function onto another set of functions, and the *functional* as a mapping of a set of functions onto a numerical set.

AUTHOR: You are quite right.

READER: I have noticed that you persistently use the term "numerical

function" (and I follow suit), but usually one simply says "function". Just how necessary is the word "numerical"?

AUTHOR: You have touched upon a very important aspect. The point is that in modern mathematics the concept of a function is substantially broader than the concept of a numerical function. As a matter of fact, the concept of a function includes, as particular cases, a numerical function as well as an operator and a functional, because the essence in all the three is a mapping of one set onto another independently of the nature of the sets. You have noticed that both operators and functionals are mappings of certain sets onto certain sets. In a particular case of mapping of a numerical set onto a numerical set we come to a *numerical* function. In a more general case, however, sets to be mapped can be *arbitrary*. Consider a few examples.

Example 1 Let D be a set of working days in an academic year, and E a set of students in a class. Using these sets, we can define a function realizing a schedule for the students on duty in the classroom. In compiling the schedule, each element of D (every working day in the year) is placed in one-to-one correspondence with a certain element of E (a certain student). This function is a *mapping of the set of working days onto the set of students*. We may add that the domain of the function consists of the working days and the range is defined by the set of the students.

READER: It sounds a bit strange. Moreover, these sets have finite numbers of elements.

AUTHOR: This last feature is not principal.

READER: The phrase "the values assumed on the set of students" sounds somewhat awkward.

AUTHOR: Because you are used to interpret "value" as "numerical value". Let us consider some other examples.

Example 2 Let D be a set of all triangles, and E a set of positive real numbers. Using these sets, we can define two functions, namely, the area of a triangle and the perimeter of a triangle. Both functions are mappings (certainly, of different nature) *of the set of the triangles onto the set of the positive real numbers*. It is said that the set of all the triangles is the domain of these functions and the set of the positive real numbers is the range of these functions.

Example 3 Let D be a set of all triangles, and E a set of all circles. The mapping of D onto E can be either a circle inscribed in a triangle, or a circle circumscribed around a triangle. Both have the set of all the triangles as the domain of the function and the set of all the circles as the range of the function.

By the way, do you think that it is possible to "construct" an *inverse* function in a similar way, namely, to define a function with all the circles as its domain and all the triangles as its range?

READER: I see no objections.

AUTHOR: No, it is impossible. Because any number of different triangles can be inscribed in or circumscribed around a circle. In other words, each element of E (each circle) corresponds to an Infinite number of different elements of D (i.e, an infinite number of triangles). It means that there is no function since no mapping can be realized.

However, the situation can be improved if we restrict the set of triangles.

READER: I guess I know how to do it. We must choose the set of all the *equilateral* triangles as the set D. Then it becomes possible to realize

both a mapping of D onto E (onto the set of all the circles) and an inverse mapping, i.e. the mapping of E onto D, since only one equilateral triangle could be inscribed in or circumscribed around a given circle.

AUTHOR: Very good. I see that, you have grasped the essence of the concept of functional relationship. I should emphasize that from the broadest point of view this concept is based on the idea of mapping one set of objects onto another set of objects. *It means that a function can be realized as a numerical function, an operator, or a functional*. As we have established above, a *function may be represented by an area or perimeter of a geometrical figure, such, as a circle inscribed in a triangle or circumscribed around it, or it may take the form of a schedule of students on duty in a classroom, etc*. It is obvious that a list of different functions may be unlimited.

READER: I must admit that such a broad interpretation of the concept of a function is very new to me.

AUTHOR: As a matter of fact, in a very diverse set of possible functions (mappings), we shall use only *numerical functions*, *operators*, and *functionals*. Consequently, we shall refer to numerical functions as simply *functions*, while *operators* and *functionals* will be pointed out specifically.

And now we shall examine the already familiar concept of a numerical sequence as an example of mapping.

READER: A numerical sequence is, apparently, a mapping of a set of natural numbers onto a different numerical set. The elements of the second set are the terms of the sequence. Hence, a numerical sequence is a particular case of a numerical function. The domain of a function is represented by a set of natural numbers.

AUTHOR: This is correct. But you should bear in mind that later on we

shall deal with numerical functions whose domain is represented by the *real line*, or by its *interval* (or *intervals*), and whenever we mention a function, we shall imply a numerical function.

In this connection it is worthwhile to remind you of the classification of intervals. In the previous dialogue we have already used this classification, if only partially.

First of all we should distinguish between the intervals of finite length:

- a *closed interval* that begins at a and ends at b is denoted by $[a, b]$; the numbers x composing this interval meet the inequalities $a \leqslant x \leqslant b$;
- an *open interval* that begins at a and ends at b is denoted by $]a, b[$; the numbers x composing this interval meet the inequalities $a < x < b$;
- a *half-open interval* is denoted either by $]a, b]$ or $[a, b[$, the former implies that $a < x \leqslant b$, and the latter that $a \leqslant x < b$.

The intervals may also be *infinite*:

$$
\begin{array}{lll}
]-\infty, \infty[& (-\infty < x < \infty) & \text{the real line} \\
]a, \infty] & (a < x < \infty) & \\
[a, \infty[& (a \leqslant x < \infty) & \\
]-\infty, b[& (-\infty < x < b) & \\
]-\infty, b] & (-\infty < x \leqslant b) &
\end{array}
$$

Let us consider several specific examples of numerical functions. Judging by the appearance of the formulas given below, point out the intervals

constituting the domains of the following functions:

$$y = \sqrt{1-x^2} \tag{17}$$

$$y = \sqrt{x-1} \tag{18}$$

$$y = \sqrt{2-x} \tag{19}$$

$$y = \frac{1}{\sqrt{x-1}} \tag{20}$$

$$y = \frac{1}{\sqrt{2-x}} \tag{21}$$

$$y = \sqrt{x-1} + \sqrt{2-x} \tag{22}$$

$$y = \frac{1}{\sqrt{x-1}} + \frac{1}{\sqrt{2-x}} \tag{23}$$

$$y = \sqrt{2-x} + \frac{1}{\sqrt{x-1}} \tag{24}$$

$$y = \sqrt{x-1} + \frac{1}{\sqrt{2-x}} \tag{25}$$

READER: It is not difficult. The domain of function (17) is the interval $(-1, 1]$; that of (18) is $[1, \infty[$; that of (19) is $]-\infty, 2]$; that of (20) is $]1, \infty[$; that of (21) is $]\infty, 2[$; that of (22) is $[1, 2]$, etc.

AUTHOR: Yes, quite right, but may I interrupt you to emphasize that if a function is a sum (a difference, or a product) of two functions, its domain is represented by the *intersection* of the sets which are the domains of the constituent functions. It is well illustrated by function (22). As a matter of fact, the same rule must be applied to functions (23)-(25). Please, continue.

READER: The domains of the remaining functions are (23) $]1, 2[$; (24) $]1, 2]$; (25) $[1, 2[$.

AUTHOR: And what can you say about the domain of the function $y = \sqrt{x-2} + \sqrt{1-x}$?

Reader: The domain of $y = \sqrt{x-2}$ is $[2, \infty[$, while that of $y = \sqrt{1-x}$ is $]-\infty, 1]$. These intervals do not intersect.

Author: It means that the formula $y = \sqrt{x-2} + \sqrt{1-x}$ does not define any function.

Dialogue 5

MORE ON FUNCTION

AUTHOR: Let us discuss the methods of defining functions. One of them has already been employed quite extensively. I mean the *analytical description* of a function by some *formula*, that is, an *analytical expression* (for example, expressions (17) through (25) examined at the end of the preceding dialogue).

READER: As a matter of fact, my concept of a function was practically reduced to its representation by a' formula. It was a formula that I had in mind whenever I spoke about a dependence of a variable y on a variable x.

AUTHOR: Unfortunately, the concept of a function as a formula relating x and y has long been rooted in the minds of students. This is, of course, quite wrong. A function and its formula are very different entities. It is one thing to define a function as a mapping of one set (in our case it is a numerical set) onto another, in other words, as a "black box" that generates a number at the output in response to a number at the input.

It is quite another thing to have just a formula, which represents only one of the ways of defining a function. It is wrong to *identify* a function with a formula giving its analytical description (unfortunately, it happens sometimes).

READER: It seems that after the discussion in the previous dialogue about the function, such identification in a general case is automatically invalidated. However, if we confine ourselves only to numerical functions and if we bear in mind that working with a function we always use a formula to describe it, a question arises: Why is it erroneous to identify these two notions? Why should we always emphasize the difference between the function and its formula?

AUTHOR: I'll tell you why. First, not every formula defines a function. Actually, at the end of the previous dialogue we already had such an example. I shall give you some more:

$$y = \frac{1}{\sqrt{x}} + \frac{1}{\sqrt{-x}}$$
$$y = \log x + \log(-x),$$
$$y = \sqrt{\sin x - 2},$$
$$y = \log(\sin x - 2), \text{ etc.}$$

These formulas do not represent any functions.

Second (and this is more important), not all functions can be written as formulas. One example is the so-called *Dirichlet function* which is defined on the real line:

$$y = \begin{cases} 1 & \text{if } x \text{ is a rational number} \\ 0 & \text{if } x \text{ is an irrational number} \end{cases}$$

READER: You call *this* a function?

AUTHOR: It is certainly an unusual function, but still a function. It is a mapping of a set of rational numbers to unity and a set of irrational numbers to zero. The fact that you cannot suggest any analytical expression for this function is of no consequence (unless you invent a special symbol for the purpose and look at it as a formula).

However, there is one more, third and probably the most important, reason why functions should not be identified with their formulas. Let us look at the following expression:

$$y = \begin{cases} \cos x & x < 0 \\ 1 + x^2 & 0 \leqslant x \leqslant 2 \\ \log(x-1) & x > 2 \end{cases}$$

How many functions have I defined here?

READER: Three functions: a cosine, a quadratic function, and a logarithmic function.

AUTHOR: You are wrong. The *three formulas* ($y = \cos x$, $y = 1 + x^2$, and $y = \log(x-1)$) define in this case a *single function*. It is defined on the real line, with the law of numerical correspondence given as $y = \cos x$ over the interval $]-\infty, 0[$, as $y = 1 + x^2$ over the interval $[0, 2]$, and as $y = \log(x-1)$ over the interval $]2, \infty[$.

READER: I've made a mistake because I did not think enough about the question.

AUTHOR: No, you have made the mistake because subconsciously you identified a function with its analytical expression, i.e. its formula. Later on, operating with functions, we shall use formulas rather extensively. However, you should never forget that a formula is not all a function is, It is only one way of defining it.

The example above illustrates, by the way, that one should not identify such notions as the *domain of a function* and the *range of* x on which an analytical expression is defined (i.e. the domain of an analytical expression). For example, the expression $1+x^2$ is defined on the real line. However, in the example above this expression was used to define the function only over the interval $[0,2]$.

It should be emphasized that the question about the domain of a function is of principal significance. It goes without saying that the domain of a function cannot be wider than the domain of an analytical expression used to define this function. But it can be narrower.

READER: Does it mean that a cosine defined, for example, over the interval $[0,\pi]$ and a cosine defined over the interval $[\pi,3\pi]$ are two different functions?

AUTHOR: Strictly speaking, it does. A cosine defined, for example, on the real line is yet another function. In other words, using cosine we may, if we wish, define any number of different functions by varying the domain of these functions.

In the most frequent case, when the domain of a function coincides with the domain of an analytical expression for the function, we speak about a *natural* domain of the function. Note that in the examples in the previous dialogue we dealt with the natural domains of the functions. A natural domain is always meant if the domain of a function in question is not specified (strictly speaking, the domain of a function should be specified in every case).

READER: It turns out that one and the same function can be described by different formulas and, vice versa, one and the same formula can be used

to “construct” different functions.

Author: In the history of mathematics the realization of this fact marked the final break between the concept of a function and that of its analytical expression. This actually happened early in the 19th century when Fourier, t.he French mathematician, very convincingly showed t.hat it is quite irrelevant whether one or many analytical expressions are used to describe a function. Thereby an end was put to the very long discussion among mathematicians about. Identifying a function with its analytical expression.

It should be noted that similarly to other basic mathematical concepts, the concept of a function went through a long history of evolution. The term “function” was introduced by the German mathematician Leibnitz late in the 17th century. At that time this term had a rather narrow meaning and expressed a relationship between geometrical objects. The definition of a functional relationship, freed from geometrical objects, was first formulated early in the 18th century by Bernoulli. The evolution of the concept of a function can be conventionally broken up into three main stages. During the first stage (the 18th century) a function was practically identified with its analytical expression. During the second stage (the 19th century) the modern concept of a function started to develop as a mapping of one numerical set onto another. With the development of the general theory of sets, the third stage began (the 20th century) when the concept of a function formerly defined only for numerical sets was generalized over the sets of an arbitrary nature.

Reader: It appears that by overestimating the role of a formula we inevitably slip back to the concepts of the 18th century.

Author: Let us discuss now one more way of defining a function, namely,

the *graphical method*. The *graph* of a function $y = f(x)$ is a set of points on the plane (x, y) whose abscissas are equal to the values of the independent variable (x), and whose ordinates are the corresponding values of the dependent variable (y). The idea of the graphical method of defining a function is easily visualized. Figure 14(*a*) plots the graph of the function

$$y = \begin{cases} \cos x & x < 0 \\ 1 + x^2 & 0 \leqslant x \leqslant 2 \\ \log(x-1) & x > 2 \end{cases}$$

discussed earlier. For a comparison, the graphs of the functions $y = \cos x$, $y = 1 + x^2$, and $y = \log(x-1)$ within their natural domains of definition in tho same figure (cases (b), (c), and (d)).

READER: In Figure 14(*a*) I notice an open circle. What does it mean?

AUTHOR: This circle graphically represents a point excluded from the graph. In this particular case the point $(2, 0)$ does not belong to the graph of the function. Figure 15 plots the graphs of the functions that were discussed at the end of the previous dialogue. Let us have a close look at them.

READER: Obviously, in all the cases shown in Figure 15 the domain of the function is supposed coinciding with the domain of the corresponding analytical expression.

AUTHOR: Yes, you are right. In cases (b), (c), (d), and (e) these domains are infinite intervals. Consequently, only a part of each graph could be shown.

READER: In other cases, however, such as (g), (h), and (i), the domains of the functions are intervals of finite length. But here as well the figure has space for only a part of each graph.

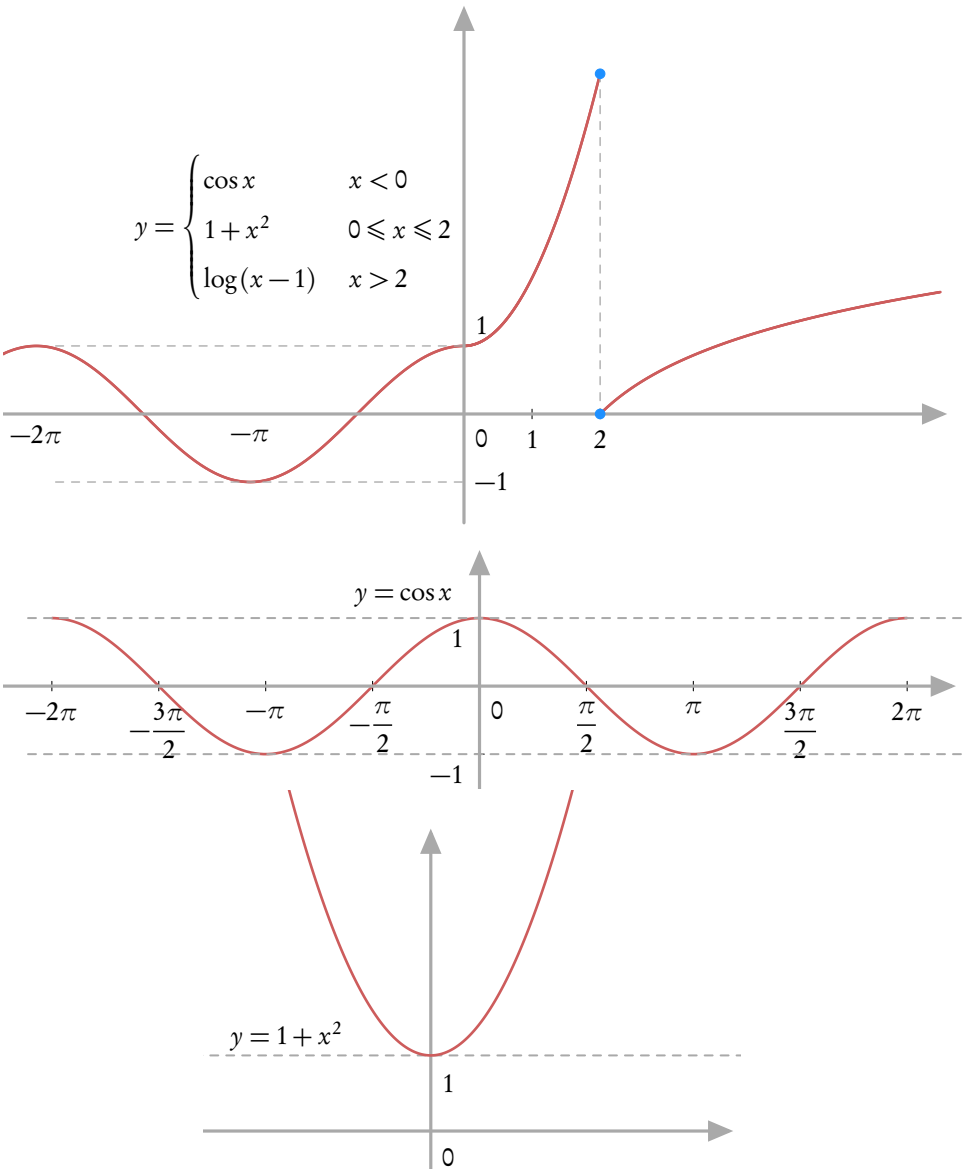


Figure 14: The graphs of the functions $y = \cos x$, $y = 1 + x^2$, and $y = \log(x-1)$.

Author: That is right. The graph is presented in its complete form only in cases (a) and (f). Nevertheless, the behaviour of the graphs is quite

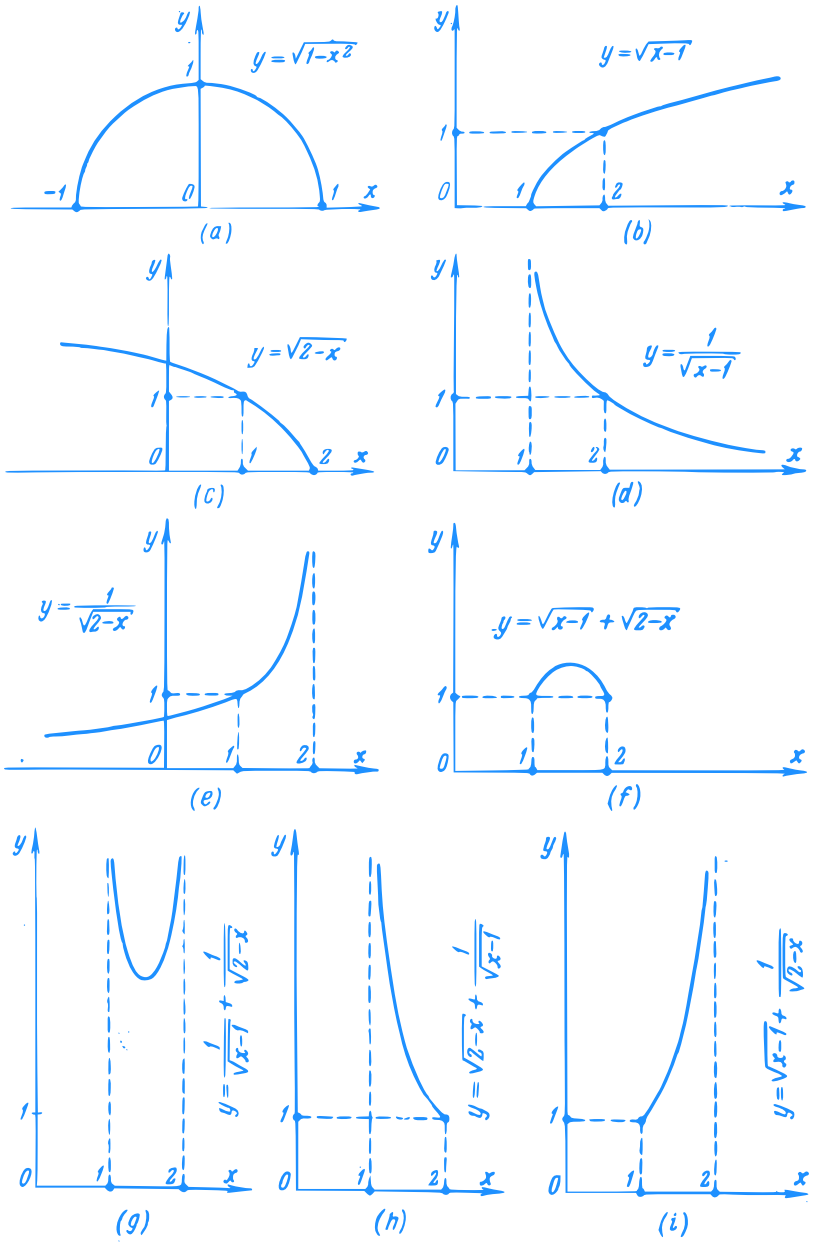


Figure 15: A variety of functions and their domains.

clear for all the functions in Figure 15.

The cases which you noted, i.e. (g), (h), and (i), are very interesting. Here we deal with the unbounded function defined over the finite interval. The notion of boundedness (unboundedness) has already been discussed with respect to numerical sequences (see Dialogue One). Now we have to extrapolate this notion to functions defined over intervals.

Definition

A function $y = f(x)$ is called *bounded* over an interval D if one can indicate two numbers A and B such that

$$A \leqslant f(x) \leqslant B \quad \text{for all } x \in D.$$

If not, the function is called *unbounded*.

Note that within infinite intervals you may define both bounded and unbounded functions. You are familiar with examples of bounded functions: $y = \sin x$ and $y = \cos x$. Examples of unbounded functions are in Figure 15 (cases (b), (c), (d), and (e)).

READER: Over the intervals of finite length both bounded and unbounded functions may also be defined. Several illustrations of such functions are also shown in Figure 15: the functions in cases (a) and (f) are bounded; the functions in cases (g), (h), and (i) are unbounded.

AUTHOR: You are right.

READER: I note that in the cases that I have indicated the bounded functions are defined over the closed intervals ($[-1,1]$ for (a) and $[1,2]$ for (f)), while the unbounded functions are defined both over the open and half-open intervals ($]1,2[$ for (g), $]1,2]$ for (h), and $[1,2[$ for (i)).

AUTHOR: This is very much to the point. However, you should bear in mind that it is possible to construct bounded functions defined over open (half-open) intervals, and unbounded functions defined over closed intervals. Here are two simple illustrations:

Example 1: $y = x^2 \quad 0 \leqslant x < 2$

Example 2:

$$y = \begin{cases} \dfrac{1}{x} & 0 \leqslant x < 2 \\ 1 & x = 0 \end{cases}$$

The graphs of these functions are shown in Figure 16.

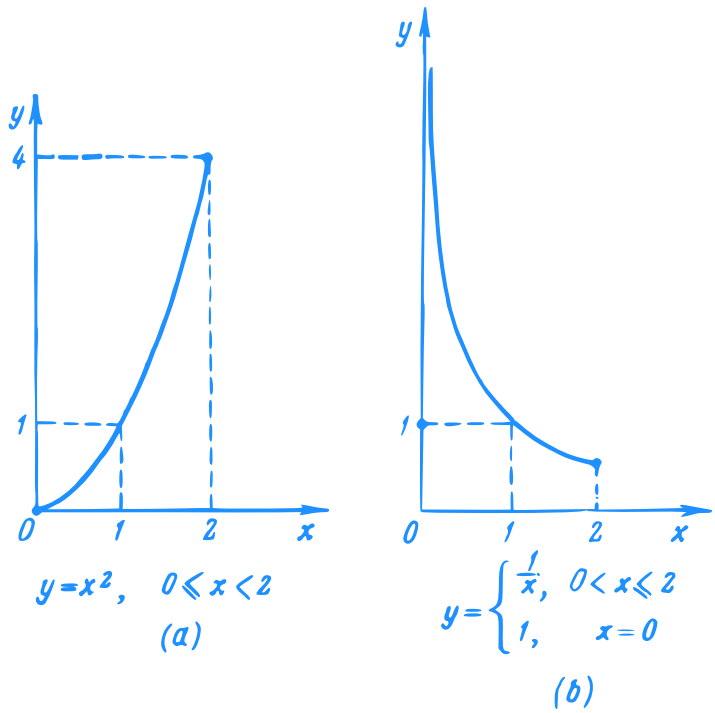


Figure 16: Examples of bounded functions defined over open (half-open) intervals, and unbounded functions defined over closed intervals.

Reader: It seems that the boundedness (unboundedness) of a function and the finiteness of the interval over which it is defined are not interrelated, Am I right?

Author: Not completely. There is, for example, the following theorem.

Theorem

If a function is defined over a closed interval and if it is monotonic, the function is bounded.

Reader: Obviously, the monotonicity of a function is determined similarly to the monotonicity of a numerical sequence.

Author: Yes, it is. Monotonic functions can be classified, as sequences, into non-decreasing and non-increasing:

Definition

A function $y = f(x)$ is said to be *non-decreasing* over an interval D it for any x_1 and x_2 from this interval $f(x_1) \leqslant f(x_2)$ if $(x_1 \leqslant x_2$. If, however, $f(x_1) \geqslant f(x_2)$, the function is said to be *non-increasing*.

Can you prove the theorem formulated above?

Reader: Let the function $y = f(x)$ be defined over the closed interval $[a, b]$. We denote $f(a) = y_a$ and $f(b) = y_b$. To make the case more specific, let us assume that the function is non-decreasing. It means that $y_a \leqslant y_b$. I don't know how to proceed.

Author: Select an arbitrary point x over the interval $[a, b]$.

Reader: Since $a \leqslant x$ and $x \leqslant b$, then, according to the condition of the above theorem, $y_a \leqslant f(x)$ and $f(x) \leqslant y_b$. Thus, we get that $y_a \leqslant f(x) \leqslant$

y_b for all x in the domain of the function. This completes the proof.

AUTHOR: Correct. So, if a *monotonic* function is defined over a closed interval, it is bounded. As to a *non-monotonic* function defined over a closed interval, it may be either bounded (Figure 15(a) and (f)) or unbounded (Figure 15(b)).

And now answer the following question: Is the function $y = \sin x$ monotonic?

READER: No, it isn't.

AUTHOR: Well, your answer is as vague as my question. First we should determine the domain of the function. If we consider the function $y = \sin x$ as defined on the natural domain (on the real line), then you are quite right. If, however, the domain of the function is limited to the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ the function becomes monotonic (non-decreasing).

READER: I see that the question of the boundedness or monotonicity of any function should be settled by taking into account both the type of the analytical expression for the function and the interval over which the function is defined.

AUTHOR: This observation is valid not only for the boundedness or monotonicity but also for other properties of functions. For example, is the function $y = 1 - x^2$ an *even* function?

READER: Evidently the answer depends on the domain of the function.

AUTHOR: Yes, of course. If the function is defined over an interval symmetric about the origin of coordinates (for example, on the real line or over the interval $[-1, 1]$), the graph of the function will be symmetric about the straight line $x = 0$. In this case $y = 1 - x^2$ is an even func-

tion. If, however, we assume that the domain of the function is $[-1,2]$, the symmetry we have discussed above is lost (Figure 17) and, as a result, $y = 1 - x^2$ is not even.

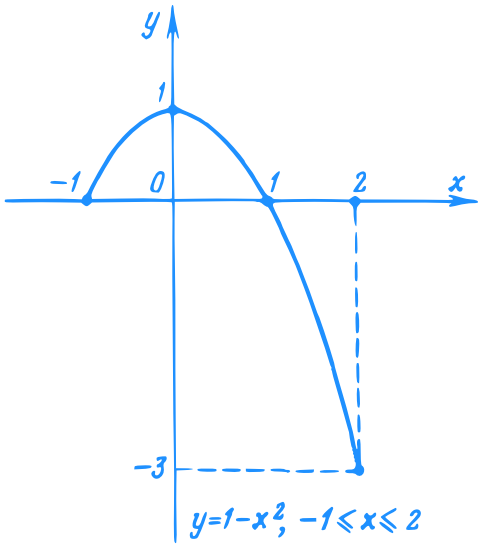


Figure 17: Is the function $y = 1 - x^2$ *even*?

READER: It is obvious that your remark covers the case of odd functions as well.

AUTHOR: Yes, it does. Here is a rigorous definition of an even function.

Definition

A function $y = f(x)$ is said to be *even* if it is defined on a set D symmetric about the origin and if $f(-x) = f(x)$ for all $x \in D$.

By substituting $f(-x) = -f(x)$ for $f(-x) = f(x)$, we obtain the definition of an odd function.

But let us return to monotonic functions. If we drop the equality sign

in the definition of a monotonic function (see p. 71) (in $f(x_1) \leqslant f(x_2)$ or $f(x_1) \leqslant f(x_2)$), we obtain a so-called *strictly monotonic function*. In this case a non-decreasing function becomes an *increasing function* (i.e. $f(x_1) < f(x_2)$) . Similarly, a non-increasing function becomes a *decreasing function* (i.e. $f(x_1) > f(x_2)$). In all the previous illustrations of monotonic functions we actually dealt with strictly monotonic functions (either increasing or decreasing). Strictly monotonic functions possess an

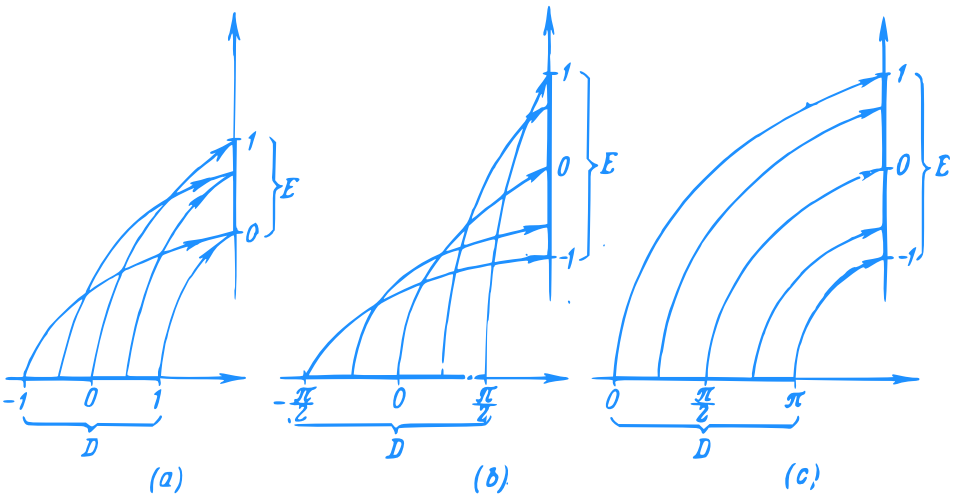


Figure 18: Concept of an *inverse* function.

interesting property: *each has an inverse function*.

READER: The concept of an *inverse function* has already been used in the previous dialogue in conjunction with the possibility of mapping a set of equilateral triangles onto a set of circles. We saw that the *inverse mapping*, i.e. the mapping of the set of circles onto the set of equilateral triangles, was possible.

AUTHOR: That's right. Here we shall examine the concept of an inverse function in greater detail (but for numerical functions). Consider Fig-

ure 18. Similarly to the graphs presented in Figure 13, it shows three functions:

(a) $y = \sqrt{1-x^2}$ $\quad -1 \leqslant x \leqslant 1$

(b) $y = \sin x$ $\quad -\dfrac{\pi}{2} \leqslant x \leqslant \dfrac{\pi}{2}$

(c) $y = \cos x$ $\quad 0 \leqslant x \leqslant \pi$

Here we have three mappings of one numerical set onto another. In other words, we have three mappings of an interval onto another interval. In case (a) the interval $[-1, 1]$ is mapped onto the interval $[0, 1]$; in (b) the interval $\left[-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right]$ is mapped onto the interval $[-1, 1]$; and in (c) the interval $[0, \pi]$ is mapped onto the interval $[-1, 1]$. What is the difference between mappings (b) and (c), on the one hand, and mapping (a), on the other?

READER: In cases (b) and (c) we have a one-to-one correspondence, i.e. each point of the set D corresponds to a single point of the set E and vice versa, i.e. each point of E corresponds to only one point of D. In case (a), however, there is no one-to-one correspondence.

AUTHOR: Yes, you are right. Assume now that the directions of all the arrows in the figure are reversed. Now, will the mappings define a function in all the three cases?

READER: Obviously, in case (a) we will not have a function since then the reversal of the directions of the arrows produces a forbidden situation, namely, one number corresponds to two numbers. In cases (b) and (c) no forbidden situation occurs so that in these cases we shall have some new functions.

AUTHOR: That is correct. In case (b) we shall arrive at the function $y = \arcsin x$, which is the inverse function with respect to $y = \sin x$ defined

over the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

In case (c) we arrive at the function $y = \arccos x$, which is the inverse function with respect to $y = \cos x$ defined over $[0, \pi]$.

I would like to place more emphasis on the fact that in order to obtain an inverse function from an initial function, it is necessary to have a one-to-one correspondence between the elements of the sets D and E. That is why the functions $y = \sin x$ and $y = \cos x$ were defined not on their natural domains but over such intervals where these functions are either increasing or decreasing. In other words, the initial functions in cases (b) and (c) in Figure 18 were defined as strictly monotonic. A strict monotonicity is a *sufficient condition* for the above-mentioned one-to-one correspondence between the elements of D and E. No doubt you can prove without my help the following theorem.

Theorem

If a function $y = f(x)$ is strictly monotonic, different x are mapped onto different y.

READER: Thus, a sufficient condition for the existence of the inverse function- is the strict monotonicity of the initial function. Is this right?

AUTHOR: Yes, it is.

READER: But isn't the strict monotonicity of the initial function also a *necessary condition* for the existence of the inverse function?

AUTHOR: No, it is not. A one-to-one correspondence may also take place

in the case of a non-monotonic function. For example,

$$y = \begin{cases} 1 - x & 0 < x < 1 \\ x & 1 \leqslant x \leqslant 2 \end{cases}$$

Have a look at the graph of this function shown in Figure 19. If a function is strictly monotonic, it has the inverse function. However, the converse is not true.

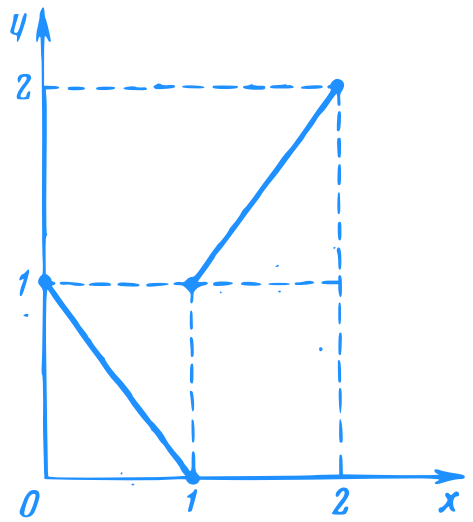


Figure 19: A non-monotonic function.

Reader: As I understand it, in order to obtain an inverse function (when it exists), one should simply reverse the roles of x and y in the equation $y = f(x)$ defining the initial function. The inverse function will then be 'given by the equation $x = F(y)$. As a result the range of the initial function becomes the domain of the inverse function.

Author: That is correct. In practice a conversion of the initial function to the inverse function can be easily performed on a graph. The graph of the inverse function is always *symmetric* to the graph of the initial

function about a straight line $y = x$. It is illustrated in Figure 20, which shows several pairs of graphs of the initial and inverse functions. A list of some pairs of functions with their domains is given below:

	Initial	Domain	Inverse	Domain
(a)	x^3	$-\infty < x < \infty$	$\sqrt[3]{x}$	$-\infty < x < \infty$
(b)	x^2	$0 \leqslant x < \infty$	$\sqrt{x}$	$0 \leqslant x < \infty$
(c)	10^x	$-\infty < x < \infty$	$\log x$	$0 < x < \infty$
(d)	$\sin x$	$-\frac{\pi}{2} \leqslant x \leqslant \frac{\pi}{2}$	$\arcsin x$	$-1 \leqslant x \leqslant 1$
(e)	$\cos x$	$0 \leqslant x \leqslant \pi$	$\arccos x$	$-1 \leqslant x \leqslant 1$
(f)	$\tan x$	$-\frac{\pi}{2} < x < \frac{\pi}{2}$	$\arctan x$	$-\infty < x < \infty$
(g)	$\cot x$	$0 < x < \infty$	$\operatorname{arccot} x$	$-\infty < x < \infty$

All the domains of the inverse functions shown in the list case of are the natural domains of the functions (however, in the case of $y = \sqrt[3]{x}$ the natural domain is sometimes assumed to be restricted to the interval $[0, \infty[$ instead of the whole real line). As to the initial functions, only two of them ($y = x^3$ and $y = 10^x$) are considered in this case as defined on their natural domains. The remaining functions are defined over shorter intervals to ensure the strict monotonicity of the functions.

Now we shall discuss the concept of a *composite function*.

Let us take as an example the function $h(x) = \sqrt{1 + \cos^2 x}$. Consider also the functions $f(x) = \cos x$ and $g(y) = \sqrt{1 + y^2}$.

READER: This $f(x)$ notation is something new. So far we used to write $y = f(x)$,

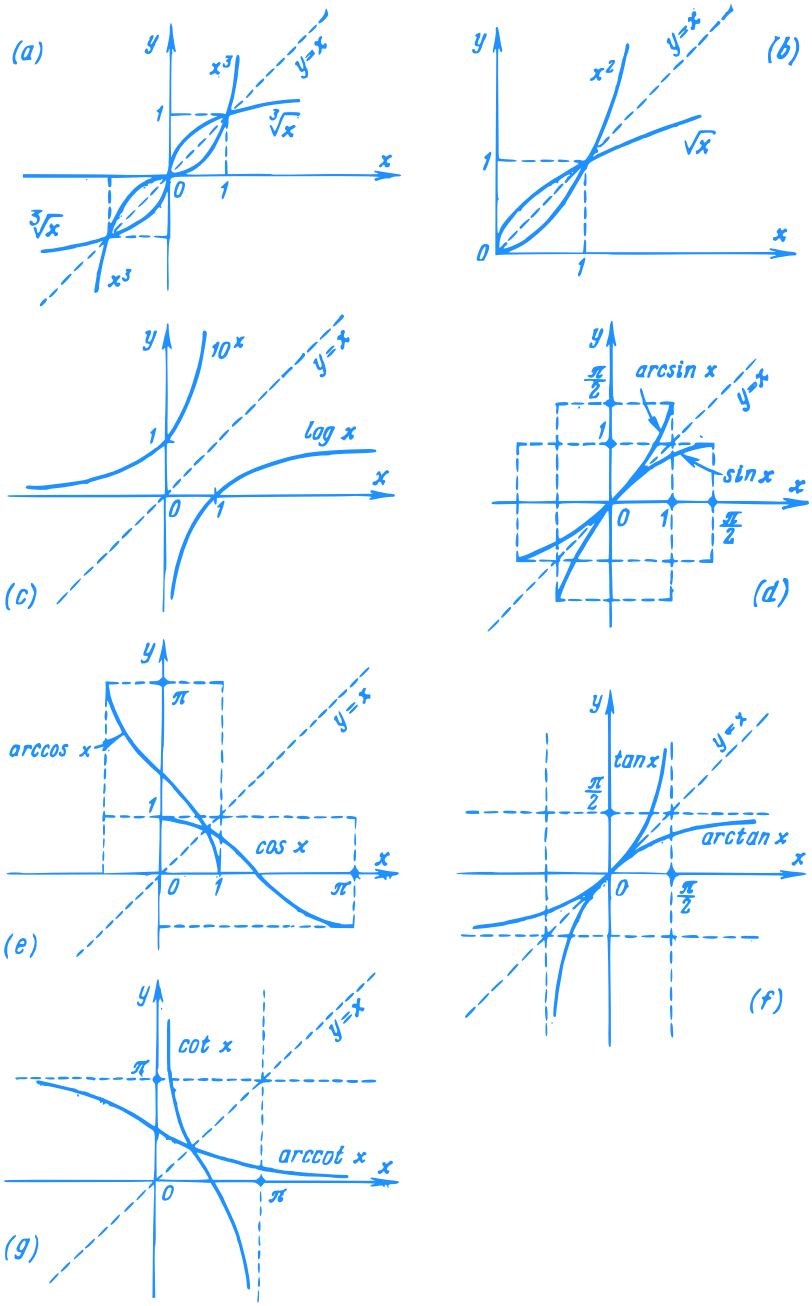


Figure 20: Functions and their inverses showing symmetry along the line $y = x$.

AUTHOR: You are right. However, it is expedient to simplify the notation.

Consider the three functions: $h(x)$, $f(x)$, and $g(y)$. The function $h(x)$ is a composite function composed of $f(x)$ and $g(y)$:

$$h(x) = g[f(x)]$$

READER: I understand. Here, the values of $f(x)$ are used as the values of the independent variable (argument) for $g(x)$.

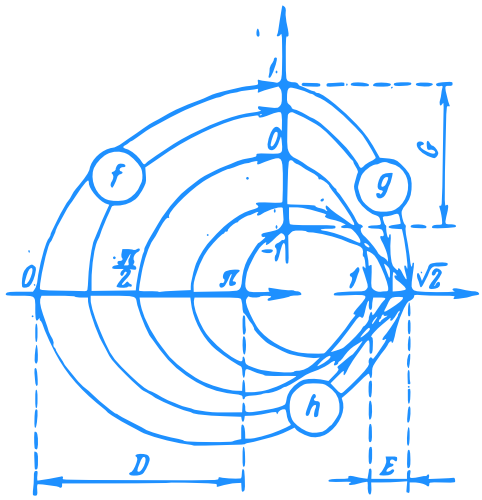


Figure 21: Understanding the composite function.

AUTHOR: Let us have a look at Figure 21, which pictures the mappings of sets in the case of our composite function, $h(x) = \sqrt{1+\cos^2 x}$, with $f(x) = \cos x$ defined over the interval $[0, \pi]$.

We see that the function f is a mapping of D (the interval $[0, \pi]$) onto G (the interval $[-1, 1]$), that is, the mapping f. The function g (the function $\sqrt{1+y^2}$) is a mapping of G onto E (the interval $[1, \sqrt{2}]$), that is, the mapping g. Finally, the function h (the function $\sqrt{1+\cos^2 x}$ defined over the interval $[0, \pi]$) is a mapping of D onto E, that is, the mapping h.

The mapping h is a result of the consecutive mappings f and g, and is said to be the composition of mappings; the following notation is used

$$h = g \circ f$$

(the right-hand side of the equation should be read from right to left: the mapping f is used first and then the mapping g).

READER: Obviously, for a composite function one can also draw a diagram shown in Figure 22.

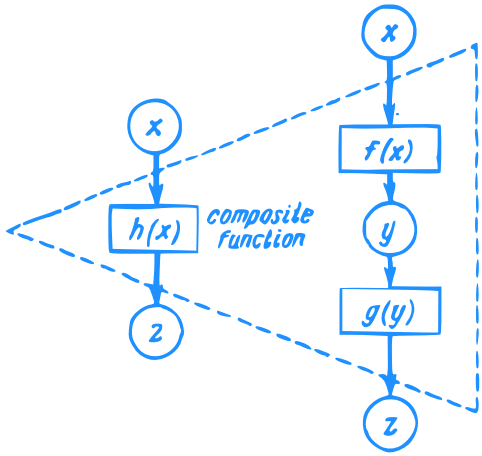


Figure 22: Understanding the composite function.

AUTHOR: I have no objections. Although I feel that. we better proceed from the concept of a mapping of one set onto another, as in Figure 21.

READER: Probably, certain "difficulties" may arise because the range of f is at the. same time the domain of g?

AUTHOR: In any case, this observation must always be kept in mind. One should not forget that the natural domain of a composite function $g[f(x)]$ is a portion (subset) of the natural domain of $f(x)$ for which the

values of f belong to the natural domain of g. This aspect was unimportant in the example concerning $g[f(x)] = \sqrt{1+\cos^2 x}$ because all the values of f (even if $\cos x$ is defined on the whole real line) fall into the natural domain of $g(y) = \sqrt{1+y^2}$. I can give you, however, a different example:

$$h(x) = \sqrt{\sqrt{x-1}-2}, \quad f(x) = \sqrt{x-1}, \quad g(y) = \sqrt{y-2}$$

The natural domain of $f(x)$ is $[1, \infty[$ Not any point in this interval, however, belongs to the domain of the composite function $h(x)$, Since the expression $\sqrt{y-2}$ is meaningful only if $y \geqslant 2$, and for $y = 2$ we have $x = 5$, the natural domain of this composite function is represented by $[5, \infty]$, i.e. a subset smaller than the natural domain of $f(x)$.

Let us examine one more example of a composite function. Consider the function $y = \sin(\arcsin x)$, You know that $\arcsin x$ can be regarded as an angle the sine of which is equal to x. In other words, $\sin(\arcsin x) = x$, Can you point out the difference between the composite function $y = \sin(\arcsin x)$ and the function $y = x$?

READER: Yes, I can. The natural domain of the function $y = x$ is represented by the whole real line. As to the composite function $y = \sin(\arcsin x)$, its natural domain coincides with the natural domain of the function $\arcsin x$, i.e. with $[-1, 1]$. The graph of the function $y = \sin(\arcsin x)$ is shown in Figure 23.

AUTHOR: Very good. In conclusion, let us get back to the problem of the graphical definition of a function. Note that there are functions whose graphs cannot be plotted in principle, the whole curve or a part of it. For example, it is impossible to plot the graph of the function $y = \sin\dfrac{1}{x}$ in the vicinity of $x = 0$ (Figure 24). It is also impossible to have the graph of the Dirichlet function mentioned above.

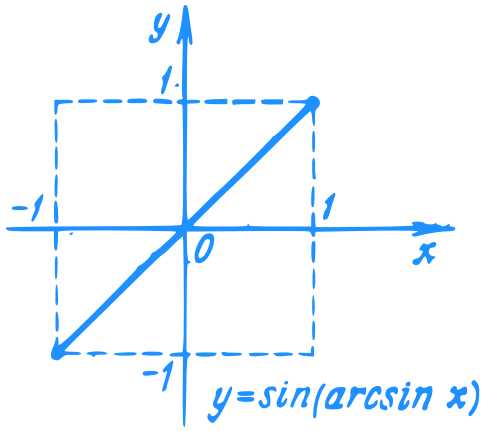


Figure 23: The graph of the function $y = \sin(\arcsin x)$.

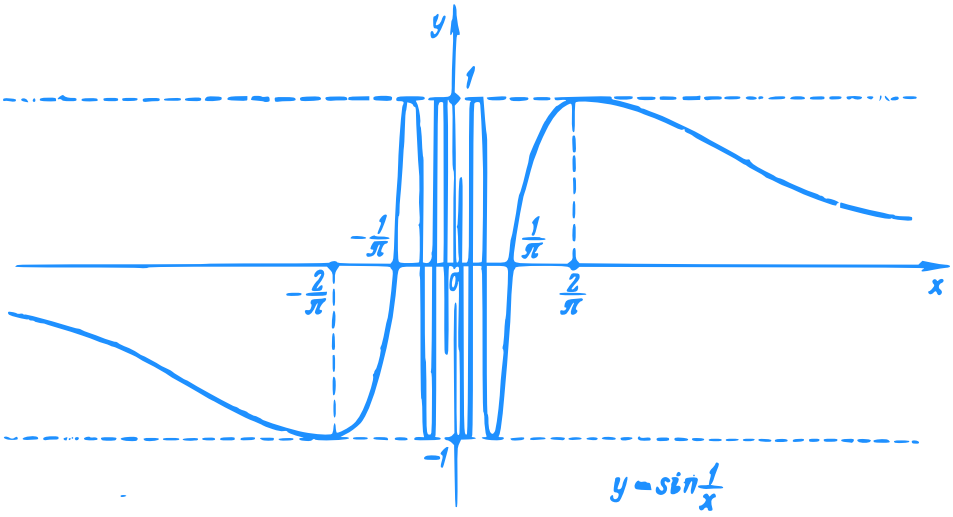


Figure 24: The graph of the function $y = \sin\dfrac{1}{x}$.

Reader: It seemed to me that the Dirichlet function had no graph at all.

Author: No, this is not the case. Apparently, your idea of a graph of a function is always a curve.

Reader: But all the graphs that we have analyzed so far were curves, and

rather smooth curves, at that.

AUTHOR: In the general case, such an image is not obligatory. But it should be stressed that *every function has its graph, this graph being unique*.

READER: Does this statement hold for functions that are not numerical

AUTHOR: Yes, it does. In the most general case we can give the following definition.

Definition

The graph of a function f defined on a set D with a range on a set E is a set of all pairs (x, y) such that the first element of the pair x belongs to D, while the second element of the pair y belongs to E, y being a function of x ($y = f(x)$).

READER: So it turns out that the graph of a function such as the area of a circle is actually a set of pairs each consisting of a circle (an element x) and a positive number (an element y) representing the area of a given circle.

AUTHOR: Precisely so. Similarly, the graph of a function representing a schedule of students on duty in a classroom is a set of pairs each containing a date (an element x) and the name of a student (an element y) who is on duty on this date. Note also that in practice this function indeed takes a graphic form.

If in a particular case both elements of the pair (both x and y) are numbers, we arrive at the graph of the function represented by a set of points on the coordinate plane. This is the familiar graph of a numerical function.

Dialogue 6

LIMIT OF FUNCTION

AUTHOR: Consider now the concept of the limit of function.

READER: But we have already covered rather extensively the concept of the limit of a numerical sequence. But a sequence is nothing else but a function defined on a set of natural numbers. Thus, having discussed the limit of sequence, we become acquainted with the limit of function as well. I wonder whether there is any point in a special discussion of the concept of the *limit of function*.

AUTHOR: Undoubtedly, a further discussion will be very much to the point. The functions we are concerned with substantially differ from sequences (I have already emphasized this fact) because they are defined over *intervals* and not on sets of natural numbers. This fact makes the concept of the *limit of function specific*. Note, for example, that every specific convergent sequence has only one limit. It means that the words "the limit of a given sequence" are self-explanatory. As for a function defined over an interval, one can speak of an infinite number of "limits"

because the limit of function is found for each specific point $x = a$ (or, as we say, for x tending to a). Thus the phrase "the limit of a *given* function" is meaningless because "the limit of a *given* function must be considered only at each given point a". Besides, this point a should either belong to the domain of the function or coincide with one of the ends of the domain.

READER: In this case the definition of the limit of function should be very different from that of the limit of sequence.

AUTHOR: Certainly, there is a difference.

Note, first of all, that we analyze a function $y = f(x)$, which is defined over a segment, and a point a in this segment (which may coincide with one of its ends when the function is defined over an open or half-open interval).

READER: Do you mean to say that at the point $x = a$ the function $f(x)$ may not be defined at all?

AUTHOR: That is quite correct. Now let us formulate the definition of the limit of function.

Definition

A number b is said to be the limit of a function $f(x)$ at x tending to a (the limit at point a) if for any positive value of ε there is a positive value of δ such that for all x satisfying the conditions x belongs to the domain of the function; $x \neq a$ and

$$|x - a| < \delta \tag{26}$$

we have

$$|f(x) - b| < \varepsilon \tag{27}$$

The standard notation is

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

READER: The definition of the limit of function is noticeably longer and more complicated than that of the limit of sequence.

AUTHOR: Note, first of all, that according to (26), point x should belong to the interval $]a-\delta, a+\delta[$. Point $x = a$ should be *eliminated* from this interval. The interval $]a-\delta, a+\delta[$ without point $x = a$ is called a *punctured δ-neighbourhood* of point a.

We select an arbitrary positive number ε. For ε we want to find *another* positive number δ such that the value of the function at *any* point x from the punctured δ-neighbourhood of point a must be inside the interval $]b-\varepsilon, b+\varepsilon[$ (speaking about *any* point x we imply only the points x in the domain of the function). If there is such b for any $\varepsilon > 0$, b is said to be the limit of the function at point a. Otherwise, b is not the limit of the function at point a.

READER: And what does your "otherwise" mean in practice?

AUTHOR: Assume that the search for δ has been successful for n diminishing numbers $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. But then you notice that for a certain number ε' it is impossible to find the required number δ, i.e. for any value of δ no matter how small) there is *always at least one* point x from the punctured δ-neighbourhood of point a at which the value of the function *lies outside* the interval $]b-\varepsilon\ b+\varepsilon'[$.

READER: But can it happen that we reduce the δ-neighbourhood of point a so much that not a single point x, belonging to the domain of the function, remains in the δ-neighbourhood?

AUTHOR: Obviously this is impossible. Because the function is defined over an interval, and point a is taken either from this interval or coincides with its end point.

READER: Everything seems clear. Apparently, in order to root all this firmly in my mind we should discuss the graph of a function.

AUTHOR: It is a good idea. Let us analyze, for the sake of convenience, the graph of the function $y = \sqrt{x}$ (Figure 25). This figure illustrates only two situations. One of them represents the selection of ε_1 (see the figure). It is easy to infer that δ_1 is the value that we look for: the values of the function at all points x from the δ_1-neighbourhood of point a are inside the interval $]b-\varepsilon_1\, b+\varepsilon_1[$ These values are represented by the portion of the graph between points A and B. The second situation represents the selection of ε_2.

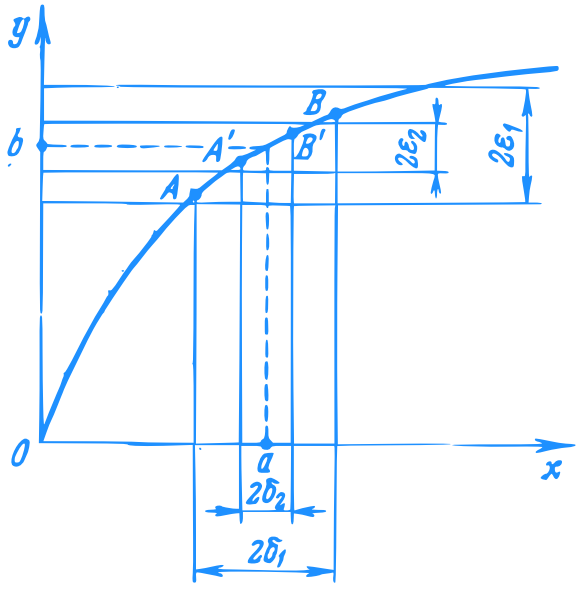


Figure 25: Understanding limit of the function $y = \sqrt{x}$.

In this case the number that we seek for is δ_1: the values of the function at points x from the δ_2-neighbourhood of point a are represented by the portion of the graph between points A' and B'.

READER: Everything you have just described looks so obvious that I see no "cream", to use your own words.

AUTHOR: "The cream" consists in the following. No matter how small $]b-\varepsilon\ b+\varepsilon[$ is, one may always select a δ-neighbourhood for point a such that for all points x in this δ-neighbourhood (all points, with the exception of point a itself and those at which the function is not defined) the values of the function should by all means lie within the indicated interval.

READER: Could you give an example of a function violating this rule?

AUTHOR: For instance, the function $y=\sin\dfrac{1}{x}$ in the vicinity of point $x=0$. The graph of the function is plotted in Figure 24. Obviously, the smaller is $|x|$ the greater is the frequency with which the graph of the function oscillates about the x-axis. For an infinitely small $|x|$ the frequency of the oscillations tends to infinity. It is easy to prove that the function $y=\sin\dfrac{1}{x}$ has no limit at $x=0$.

READER: But this function is not defined at zero.

AUTHOR: You are right. However, this fact is irrelevant from the viewpoint of the existence (or absence) of the limit of the function at $x=0$. This function is defined over $]-\infty,0[$ and $]0,\infty[$. Point $x=0$ is a common boundary between the intervals over which the function $\sin\dfrac{1}{x}$ is defined.

But let us return to the concept of the limit. Can we, for example, state

that $b = 0$ is the limit of the function $\sin\frac{1}{x}$ at point $x = 0$?

READER: It seems that I get the point. As long as we select $\varepsilon > 1$, everything is O.K. But for any $\varepsilon < 1$ it becomes impossible to find a δ-neighbourhood of point $x = 0$ such that at all points $x \neq 0$ in this δ-neighbourhood the values of the function $\sin\frac{1}{x}$ are inside the interval]-8, e]. No matter how *small* the δ-neighbourhood of point $x = 0$ is, it is the segment of *finite* length, so that the graph of our function will oscillate *infinitely many times* and thus will *infinitely many times* go *beyond* $]-\varepsilon, \varepsilon[$.

AUTHOR: That's right. Note also that in order to be convinced that a function has no limit, it is sufficient to find a violation even more "modest". Namely, it is sufficient that the graph of the function leave the interval $]-\varepsilon, \varepsilon[$ *at least once* for any δ-neighbourhood.

READER: Apparently, not only $b = 0$ but no other $b \neq 0$ can be the limit of the function $y = \sin\frac{1}{x}$ at $x = 0$. Because for any $b \neq 0$ we can use the same arguments as for $b = 0$.

AUTHOR: Hence, we have proved that the function $y = \sin\frac{1}{x}$ has no limit at point $x = 0$.

READER: The reason for the absence of the limit at $x = 0$ lies in oscillations of the graph of the function. These oscillations become more and more frequent while approaching $x = 0$.

AUTHOR: But the reason is not confined only to the infinitely increasing frequency of oscillations of the graph. Another reason is the constancy of the amplitude of oscillations. Let us "slightly correct" our function by multiplying $\sin\frac{1}{x}$ by x, The graph of the function $y = x\sin\frac{1}{x}$ is shown in

Figure 26. Do you think that $b = 0$ is the limit of this function at $x = 0$?

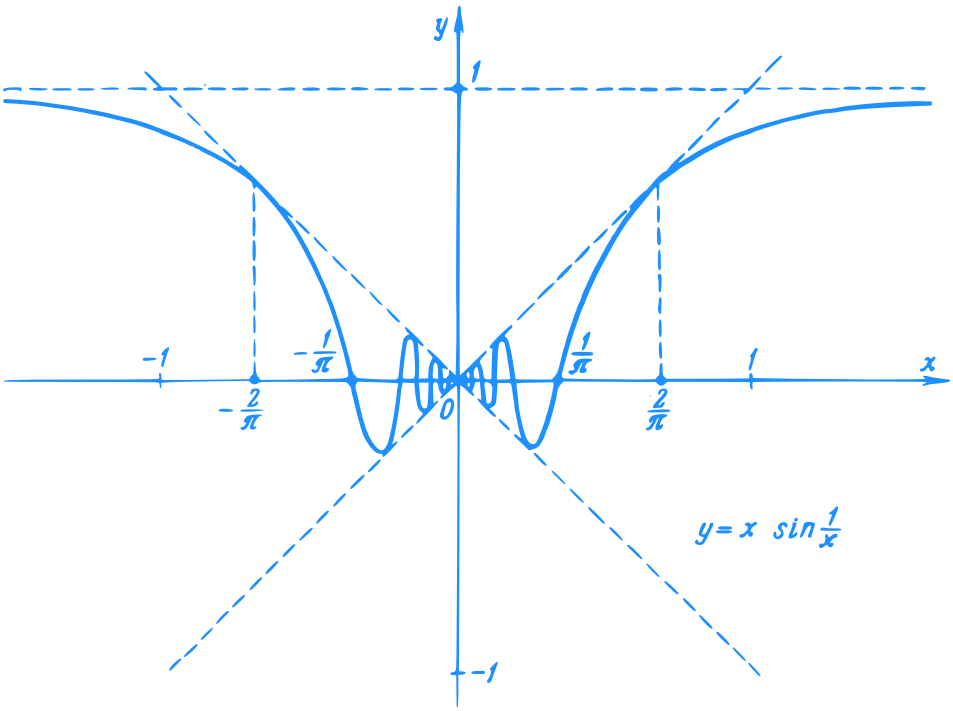


Figure 26: Finding the limit of the function $y = x \sin \dfrac{1}{x}$.

READER: I am at a loss.

AUTHOR: I'll answer this question myself. Yes, it is. The proof is within your reach if you use the definition of the limit of function. You are welcome.

READER: We select an arbitrary $\varepsilon > 0$. We should find $\delta > 0$such that $\left| x \sin \dfrac{1}{x} - 0 \right| < \varepsilon$ for all x (excluding $x = 0$) satisfying the condition $|x - 0| < \delta$. It seems to me that δ we look for is $\delta = \varepsilon$.

AUTHOR: You are quite right. Because if $|x| < \delta = \varepsilon$, it becomes evident that $\left| x \sin \dfrac{1}{x} \right| = |x| \left| \sin \dfrac{1}{x} \right| < \varepsilon$. (since $\left| \sin \dfrac{1}{x} \right| \leqslant 1$).

Reader: Really, the existence of the limit is proved without considerable difficulties.

Author: But, certainly, not always. Consider, for example, a well-known function $y = \sqrt{x}$ and prove (using the definition of the limit of function) that $b = 1$ is the limit of the function at point $x = 1$. To begin with, consider the following inequality:

$$\left|\sqrt{x} - 1\right| < \varepsilon$$

Try to find a function $g(\varepsilon)$ such that $|x-1| < g(\varepsilon)$ for any x satisfying the condition $\left|\sqrt{x} - 1\right| < \varepsilon$

Reader: I understand that $g(\varepsilon)$ is actually the desired δ 6 corresponding to an arbitrary ε.

Author: Yes, of course. We begin with some transformations. We shall proceed from the inequality:

$$\left|\sqrt{x} - 1\right| < \varepsilon \tag{28}$$

which can be rewritten in the form:

$$(1-\varepsilon) < \sqrt{x} < (1+\varepsilon)$$

Since $\sqrt{x} \geqslant 0$, the selection of $\varepsilon < 1$ *a fortiori* (which, of course, does not impair the generality of our proof) allows us to square the last inequalities

$$(1-\varepsilon)^2 < x < (1+\varepsilon)^2$$

On removing the parentheses, we obtain

$$(-2\varepsilon - \varepsilon^2) < (x-1) < (2\varepsilon + \varepsilon^2) \tag{29}$$

Note that inequalities (29) are equivalent to (28) (provided that $0 < \varepsilon < 1$). Now let us proceed from (29) to a more exacting inequality:

$$|x-1| < (2\varepsilon - \varepsilon^2) \tag{30}$$

(since $0 < \varepsilon < 1$, we have $(2\varepsilon - \varepsilon^2) > 0$). It is easy to conclude that if (30) holds, inequalities (29) and, consequently (28) will hold all the more. Thus, for an arbitrary ε within $0 < \varepsilon < 1$, it is sufficient to take $\delta = 2\varepsilon - \varepsilon^2$

READER: What happens if $\varepsilon \gg 1$?

AUTHOR: Then δ determined for any $\varepsilon < 1$ will be adequate *a fortiori*.

READER: Apparently, we may state that

$$\lim_{x\to 2}\sqrt{x} = \sqrt{2}, \quad \lim_{x\to 3}\sqrt{x} = \sqrt{3}$$

and, in general, $\lim\limits_{x\to a}\sqrt{x} = \sqrt{a}$.

AUTHOR: Yes, that's right.

READER: But could we generalize it to

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

AUTHOR: Yes, it is often the case. But not always. Because the function $f(x)$ may be undefined at point a. Remember that the limit of the function $y = x\sin\dfrac{1}{x}$ at point $x = 0$ is zero, but the function itself is not defined at point $x = 0$.

READER: But perhaps the equality lim f (x) = f (a) x a can be considered as valid in all the cases when f (x) is defined at point a?

AUTHOR: This may not be correct either. Consider, for example, a function which is called the "fractional part of x". The standard notation for

this function is $\{x\}$. The function is defined on the whole real line. We shall divide the real line into half-intervals $[n, n+1[$. For x in $[n, n+1[$ we have $\{x\} = x - n$. The graph of the function $y = \{x\}$ is shown in Figure 27.

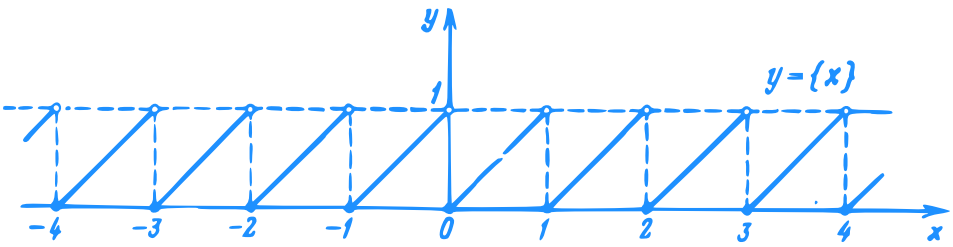


Figure 27: Graph of the function $y = \{x\}$.

Take, for example, $x = 1$. It is obvious that $\{x\}$ is defined at point $x = 1(\{x\} = 0)$. But does the function have the limit at $x = 1$?

READER: It clearly has no limit. In any δ-neighbourhood of point $x = 1$ there may exist concurrently both the points at which $\{x\}$ assumes values greater than, for example, $\frac{2}{3}$, and the points at which $\{x\}$ assumes values less than $\frac{1}{3}$.

It means that neither $b = 1$ nor $b = 0$ can be the limit of the function at point $x = 1$, if only because it is impossible to find an adequate δ for $\varepsilon = \frac{1}{3}$.

AUTHOR: I see that you have come to be rather fluent in operating with limits of functions. My compliments.

By the way, you have just proved the theorem on the *uniqueness of the limit of function* at a given point.

Theorem

A function cannot have two (or more) limits at a given point.

Now let us return to the equality

$$\lim_{x\to a} f(x) = f(a) \tag{31}$$

You already know that there are situations when $\lim\limits_{x\to a} f(x)$ exists but $f(a)$ does not exist and, vice versa, when $f(a)$ exists but $\lim\limits_{x\to a} f(x)$ does not exist. Finally, a situation is possible when both $\lim\limits_{x\to a} f(x)$ and $f(a)$ exist, but their values are not equal. I'll give you an example:

$$f(x) = \begin{cases} x^2 & \text{if } x \neq 0 \\ 1 & \text{if } x = 0 \end{cases}$$

The graph of this function is shown in Figure 28. It is easy to see that $f(0) = 1$, while $\lim\limits_{x\to a} f(x) = 0$.

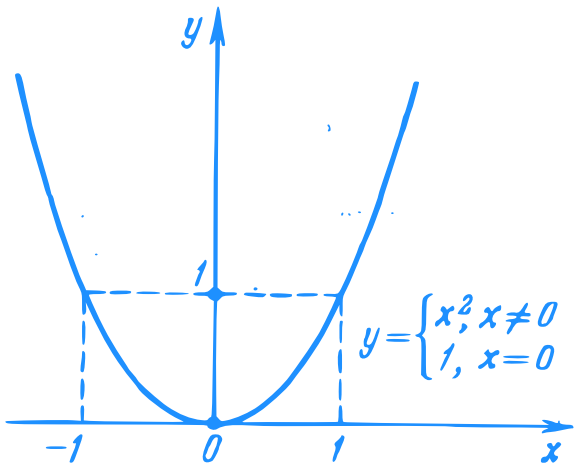


Figure 28: Graph of the function $y = f(x)$, such that $f(x) = x^2$ for $x \neq 0$ and $f(x) = 1$ for $x = 0$.

You must be convinced by now that equality (31) is not always valid.

READER: But presumably, it is often true, isn't it?

AUTHOR: Yes, and if it is, the function $f(x)$ is said to be *continuous* at $x = a$.

Thus, we have arrived at a new important concept, namely, that of the *continuity of a function at a point*. Let us give the following definition.

Definition

A function $f(x)$ is said to be *continuous at a point* $x = a$ if

(1) it is defined at $x = a$,

(2) there is the limit of the function at $x = a$,

(3) this limit equals the value of the function at $x = a$;

or, in other words, the function $f(x)$ is called continuous at a point a if

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

I believe that the preceding discussion has brought us so closely to this definition that it needs no additional explanation. I would only like to emphasize that the concept of the continuity of a function is essentially *local*. Similarly to the concept of the limit of function, it is related to a particular point x. A function may be either continuous at all points of an interval over which it is defined, or discontinuous at some of its points.

Taking the examples given above, can you single out those functions that are discontinuous at particular points?

READER: To begin with, I may refer to the function whose graph is plotted in Figure 28. This function is discontinuous at $x = 0$.

AUTHOR: Why?

READER: Because at this point the function assumes the value $y = 1$, though the limit of the function at this point is apparently zero.

AUTHOR: Very good. Can you give other examples?

READER: The function $y = \{x\}$ (see Figure 27) is discontinuous at points $x = 0, \pm 1, \pm 2, \pm 3, \ldots$ The function $y = \sin \dfrac{1}{x}$ (see Figure 24) is discontinuous at $x = 0$ where it is undefined and, moreover, has no limit. The function whose graph is shown in Figure 14(*a*) (see the previous, dialogue) is discontinuous at $x = 2$. The function $y = \tan x$ is discontinuous at points

$$x = \pm \frac{\pi}{2}, \pm \frac{3\pi}{2}, \pm \frac{5\pi}{2}, \pm \frac{7\pi}{2}$$

AUTHOR: That will do. Note that the points at which the continuity of a function is violated are called *discontinuity points*. We say that at these points a function has a discontinuity. In passing through a discontinuity point a graph of a function manifests a singularity. This fact is well illustrated by the examples you have just indicated,

READER: The discontinuity points in all these examples result in an *interruption of the curve* plotting the function. One exception is the function $y = \sin \dfrac{1}{x}$ since it is simply impossible to trace a graph of the function at $x = 0$.

AUTHOR: I may add that neither could you plot the function $y = \tan x$ at its discontinuity points (since you cannot draw a line which "goes into infinity").

READER: In any case, if a function is continuous everywhere in the domain (has no discontinuity points), its graph is a continuous line: it can be drawn without lifting the pencil from the paper.

AUTHOR: I agree. I would like to emphasize that the *continuity of a function at a point x guarantees that a very small displacement from this point will result in a very small change in the value of the function*.

Let us turn to Figure 27 which is the graph of the function $y = \{x\}$. Consider, for instance, $x = 0.5$. The function is continuous at this point. It is quite evident that at a very small displacement from the point (either to the left or to the right) the value of the function will also change only a little. Quite a different situation is observed if $x = 1$ (at one of the discontinuity points). At $x = 1$ the function assumes the value $y = 0$. But an infinitesimal shift *to the left* from the point $x = 1$ (take, for example, $x = 0.999$, or $x = 0.9999$, or any other point no matter how close to $x = 1$) will bring a sharp change in the value of the function, from $y = 0$ to $y \approx 1$.

READER: Quite clear. I must admit, however, that the local nature of the concept of a continuous function (i.e. the fact that the continuity of a function is always related to a specific point x) does not quite conform to the conventional idea of continuity. Because continuity typically implies a process and, consequently, a sort of an interval. It seems that continuity should be related not to a specific moment of time, but to an interval of time.

AUTHOR: It is an interesting observation. This local character is a manifestation of one of the specific features of calculus. When analyzing a function at a given point x, you used to speak about its value only at this specific point; but calculus operates not only with the value of a function

at a point but also with the limit of the function (or its absence) at this point, with the continuity of the function at the point. It means that on the basis of the information about a function at a given point we may construct an image of the *behaviour of the function in the vicinity of this point*. Thus we can predict the behaviour of the function if the point is slightly shifted from x.

So far we have made only the first step in this direction. The next step will be the introduction of the concept of a derivative. This will be the subject of discussion in Dialogues Eight and Nine.

READER: Nevertheless, I would like to note that in the above examples a function was found to be either continuous everywhere over any interval of finite length or discontinuous at a *finite* number of points. In this sense the local nature of the concept of a discontinuity point is evident. The continuity of the function, however, is always observed over a certain interval.

AUTHOR: First, the continuity of a function within an interval does not interfere with the local nature of continuity. *A function is continuous over an interval if it is continuous at all points of this interval.*

Second, it is not difficult to construct an example in which the number of discontinuity points over an interval of finite length is *infinitely large*. Let us look, for example, at the following function:

$$y = \begin{cases} 2 & \text{for } x = \pm 1, \pm\frac{1}{2}, \pm\frac{1}{4}, \pm\frac{1}{8}, \pm\frac{1}{16} \dots \\ x^2 & \text{for all the remaining points of the real line, including } x = 0 \end{cases}$$

The graph of this function is illustrated in Figure 29. It is easy to conclude that in any δ-neighbourhood of point $x = 0$ the function has an infinite number of discontinuity points.

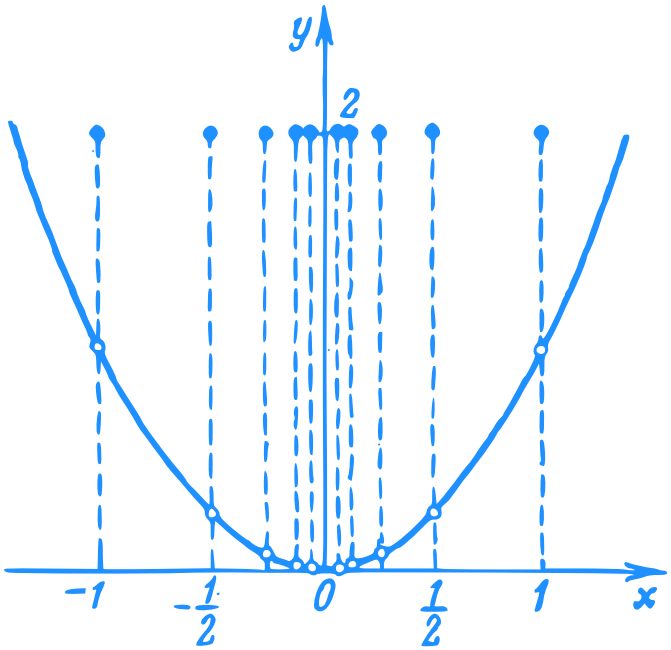


Figure 29: Graph of the function $y = f(x)$, such that $f(x) = 2$ for $x = \pm 1, \pm\frac{1}{2}, \pm\frac{1}{4}, \pm\frac{1}{8}, \pm\frac{1}{16} \ldots$ and $f(x) = x^2$ for all the remaining points of the real line, including $x = 0$.

Finally, I can give an example of a function which is discontinuous *at all points of an infinite interval*. This is a function you already know, the Dirichlet function (see the previous dialogue). Being defined on the whole real line, the function has no limit at any point of the real line; consequently, it is discontinuous at each point.

READER: This is the reason why we in principle cannot plot the Dirichlet function by a graph.

AUTHOR: As to the most frequent functions, such as *power, exponential, logarithmic, trigonometric*, and *inverse trigonometric*, they are continuous at all points of the natural domains of the corresponding analytical expres-

sions. The same can be said about composite functions obtained from the above elementary functions. The continuity of all these functions is proved in the more advanced courses of calculus. We limit ourselves to a mere stating of the fact.

Dialogue 7

MORE ON THE LIMIT OF FUNCTION

Reader: Comparing the definition of the limit of a function at a point with the definition of the limit of a numerical sequence, I come to the conclusion that these two limits are of different nature.

Author: And I understand why. In fact, I did emphasize the difference myself in the previous dialogue, pointing out, as you probably remember, that a sequence is a function defined on a set of integers, while the functions we are discussing at the moment are defined over intervals. I doubt, however, that you are justified in speaking about the difference in the nature of the limit of function and that of sequence. In the final analysis (and this is essential) *the limit of a function at a point may be defined on the basis of the limit of a numerical sequence*.

Reader: This is very interesting.

Author: Let us forget, for the time being, about the definition of the

limit of function given in the previous dialogue. Consider a new definition. We shall consider, as before, a function, $f(x)$ defined over an interval, and a point $x = a$ either taken within the interval or coinciding with its end.

AUTHOR: The two are *equivalent*.

READER: But in form they are quite different.

AUTHOR: We can prove their equivalence. To begin with, let the definition using a δ-neighbourhood of point a be called "definition *1*", and the definition using numerical sequences, "definition *2*".

Now, what two theorems must be proved to demonstrate the equivalence of definitions *1* and *2*? Can you formulate these theorems?

READER: We have to prove two theorems, one direct and the other converse. We want to prove that definition *2* follows from definition *1* and vice versa (i.e. definition *1* follows from definition *2*).

AUTHOR: Correct. First, I shall prove the following theorem.

Theorem

If a number b is the limit of a function $f(x)$ at a point a, in terms of definition *1*, it is the limit of the function $f(x)$ at a in terms of definition *2* as well.

Since b is the limit of the function $f(x)$ at point a in terms of definition *1* (this is given), consequently, for any $\varepsilon > 0$ there is $\delta > 0$ such that $|f(x) - b| < \varepsilon$ for all $x \neq a$ from a δ-neighbourhood of point a. Then we "construct" an arbitrary sequence (x_n) , requiring that it be convergent to point a (any x_n belong to the domain of the function and $x_n \neq a$ for any n). As a result we obtain a sequence of the corresponding values of

the function (the sequence $[f(x_n)]$) . We want to prove that the sequence $[f(x_n)]$ is convergent to b.

First, I select an arbitrary $\varepsilon > 0$. I should find a number N such that $|f(x_n) - b| < \varepsilon$ for all $n > N$.

I cannot immediately find such N for an arbitrary ε. However, I can indicate for an arbitrary ε such δ that $|f(x) - b| < \varepsilon$ if $|x - a| < \varepsilon$. Then I take this δ and find a sequence (x_n) convergent to a. Evidently, since (x_n) is convergent to a, ε(as any other positive real number) can be placed in correspondence with a number N such that $|x_n - a| < \delta$ for all $n > N$. And, consequently, we also have that $|f(x_n) - b| < \varepsilon$ for all $n > N$. Hence we find that the thus found number N is actually the desired number. It proves the convergence of the sequence $[f(x_n)]$ to b. Since the sequence (x_n), which is convergent to a was chosen ("constructed") arbitrarily, we conclude that the theorem's proof is completed.

If the line of reasoning is clear to you, try briefly to recapitulate the logical structure of the proof.

READER: I shall try to present the structure of the proof as a diagram (Figure 30).

AUTHOR: Your diagram is correct. Will you expand on it.

READER: The *first step* of the proof: we find for an arbitrary $\varepsilon > 0$ a number $\delta > 0$ such that $|f(x) - b| < \varepsilon$ if $|x - a| < \delta$.

The *second step* of the proof: we take δ selected at the first step; choose a sequence (x_n) convergent to a, and find a number N such that $|x - a| < \delta$ for all $n > N$. Having in mind the arguments used at the first step, we conclude that $|f(x) - b| < \varepsilon$.

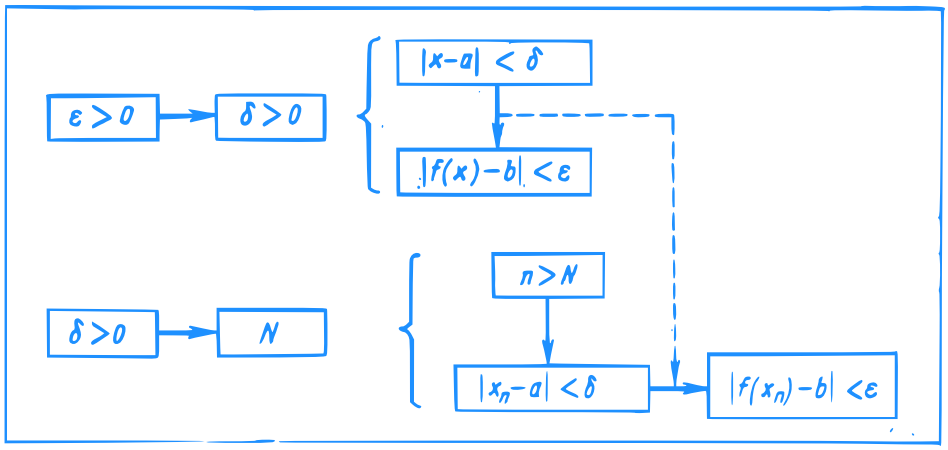


Figure 30: Schematic diagram of the proof that two definitions of the limit of a function are equal.

We have thus found for an arbitrary $\varepsilon > 0$ a number N such that $|f(x_n) - b| < \varepsilon$ for all $n > N$. This completes the proof.

AUTHOR: Correct. In conclusion I want to emphasize several essential points on which the proof hinges. We know that $|f(x) - b| < \varepsilon$ for any x from the δ-neighbourood of a. Since a sequence (x_n) is convergent to a, all x_n (the whole infinite "tail" of the sequence (x_n) starting from a certain number $N + 1$) are contained inside the δ-neighbourhood of point a. It then follows that all $f(x_n)$ (the whole infinite "tail" of the sequence $[f(x_n)]$ starting from the same number N + 1.) are contained inside the interval $]b - \varepsilon, b + \varepsilon[$, This proves that the sequence $[f(x_n)]$ converges to b.

READER: I understand.

AUTHOR: Now I am going to prove the following converse theorem.

Theorem

If a number b is the limit of a function $f(x)$ at a point a in terms of definition *2*, it is also the limit of the function $f(x)$ at a in terms of definition *1*.

In this case I shall use the proof by contradiction. Assume the contrary to what is to be proved, namely, assume that b, the limit of $f(x)$ in terms of definition *2*, is not, however, the limit of $f(x)$ in terms of definition *1*. Can you formulate the last proposition (more exactly, the assumption)?

READER: As far as I remember, a similar formulation has already been discussed in the previous dialogue. If b is not the limit of the function $f(x)$ at point a (in terms of definition *1*), it means that there is $\varepsilon' > 0$ such that it is impossible to find a necessary $\delta > 0$. Namely, no matter what δ we select, each time the function $f(x)$assumes a value outside of $]b - \varepsilon', b + \varepsilon'[$ for at least one point x from the δ-neighbourhood of point a, i.e. the inequality $|f(x) - b| < \varepsilon'$ is violated.

AUTHOR: Correct. Assume that we have selected precisely this $\varepsilon' > 0$. Next take an arbitrary $\delta > 0$, for instance, $\delta_1 = 1$. As you have said, in any δ-neighbourhood of point a and, hence, in the δ-neighbourhood of this point there is at least one point x (denoted by x_1) such that $|f(x_1 - b| \geqslant \varepsilon'$.

READER: What happens if the δ-neighbourhood contains many such points x?

AUTHOR: It makes no difference. The important fact is that there is *at least one such point*. If there are several such points, take anyone of them and denote it by x_1.

Now we take a new δ, for instance, $\delta_2 = \frac{1}{2}$. According to our assumption, the δ_2-neighbourhood of point a will contain at least one point x (denoted by x_2) such that $|f(x_2 - b| \geqslant \varepsilon'$.

Further we take $\delta_3 = \frac{1}{3}$.The δ_3-neighbourhood of point a will also contain at least one point x (point x_3) such that $|f(x_3 - b| \geqslant \varepsilon'$.

We can continue this process for a sequence of the δ-neighbourhoods of point a

$$\delta_1 = 1,\ \delta_2 = \frac{1}{2},\ \delta_3 = \frac{1}{3},\ \ldots \delta_n = \frac{1}{n},\ \ldots$$

Note that the δ-neighbourhoods are selected in such a way that the sequence (δ_n) converges to zero (is infinitesimal).

If each time we select from each δ-neighbourhood one point x in which $f(x)$ assumes a value outside of the interval $]b - \varepsilon', b + \varepsilon'[$, we obtain a sequence composed of points

$$x_1, x_3, x_3, \ldots x_n, \ldots$$

Since the sequence (δ_n) converges to zero, the sequence (x_n) inevitably converges to a. A sequence composed of the corresponding values of the function (the sequence $[f(x_n)]$ is not convergent to b because for all n we have $|f(x_n - b| \geqslant \varepsilon'$. It means that we obtained a sequence (x_n) convergent to a for which the sequence $[f(x_n)]$ is divergent.

This contradicts the condition of the theorem which states that b is the limit of the function at a in terms of definition 2. It means that for any sequence (x_n) convergent to a the corresponding sequence $[f(x_n)$ must be convergent to b. And the sequence (x_n) that we have found contradicts this condition.

Hence, the assumption that b, being the limit of the function in terms of

definition *2*, is not at the same time the limit of the function in terms of definition *1*, is invalidated. This completes the proof of the theorem.

READER: I must admit of being wrong when I spoke about different natures, of the limit of numerical sequence and the limit of function at a point.

AUTHOR: These limits differ but *their nature is the same*. The concept of the *limit of function at a point* is based, as we have seen, on the concept of the *limit of numerical sequence*.

That is why basic theorems about the limits of functions are analogous to those about the limits of sequences.

READER: We have already noted one of such theorems: the *theorem on the uniqueness of the limit of function at a point*.

AUTHOR: This theorem is analogous to that about the uniqueness of the limit of numerical sequence.

I shall also give (without proof) the *theorems on the limit of the sum; the product, and the ratio of functions*.

Theorems

If functions $f(x)$ and $g(x)$ have limits at a point a, then functions

$$[f(x)+g(x)],\ [f(x)g(x)],\ \left(\frac{f(x)}{g(x)}\right)$$

also have limits at this point. These limits equal the sum, product, and ratio, respectively, of the limits of the constituent functions (in the last case it is necessary that the limit of the function $g(x)$ at a be different from zero).

Thus,

$$\lim_{x\to a}[f(x)+g(x)]=\lim_{x\to a}f(x)+\lim_{x\to a}g(x)$$

$$\lim_{x\to a}[f(x)\,g(x)]=\lim_{x\to a}f(x)\lim_{x\to a}g(x)$$

$$\lim_{x\to a}\left(\frac{f(x)}{g(x)}\right)=\frac{\lim\limits_{x\to a}f(x)}{\lim\limits_{x\to a}g(x)}$$

under an additional condition $\lim\limits_{x\to a}g(x)\neq 0$

READER: We have already discussed the similar theorems for numerical sequences.

AUTHOR: Next I wish to make two remarks, using for the purpose specially selected examples.

Note 1. It is illustrated by the following example. Obviously

$$\lim_{x\to a}\sqrt{1-x^2}=0 \text{ and } \lim_{x\to a}\sqrt{x-1}=0.$$

Does it mean that $\lim\limits_{x\to a}(\sqrt{1-x^2}+\sqrt{x-1})=0$?

READER: The limit of the function $\sqrt{1-x^2}$ at $x=1$ exists and is equal to zero. The limit of the function $\sqrt{x-1}$ at $x=1$ also exists and is also equal to zero. According to the theorem on the limit of the sum, the limit of $f(x)=\sqrt{1-x^2}+\sqrt{x-1}$ must exist and be equal to the sum of the two preceding limits, i.e. to zero.

AUTHOR: Nevertheless, $f(x)=\sqrt{1-x^2}+\sqrt{x-1}$ no limit at $x=1$ for a simple reason that the expression $\sqrt{1-x^2}+\sqrt{x-1}$ has meaning only at a single point (point $x=1$). Applying the theorem on the limit of the sum, you have not taken into account the domains of the functions $\sqrt{1-x^2}$ and $\sqrt{x-1}$. The former has the natural domain over $[-1, 1]$, while the latter over $[1, \infty[$.

READER: Apparently your note also covers the cases when the theorems on the limit of the product and the limit of the ratio of functions are used.

AUTHOR: It goes without saying. Working with functions, you must always consider their domains. The natural domains of functions may intersect (or even coincide), but sometimes they may not. This aspect must never be over-looked. Why do you think we never have such complications when working with sequences?

READER: Obviously because all numerical sequences have one and the same domain, i.e. a set of natural numbers.

AUTHOR: Correct. Now we come to

Note 2. Do you think the limit

$$\lim_{x\to 0}\frac{\sin x}{x} \text{ exists?}$$

READER: In any case the theorem on the limit of the ratio is not valid here because $\lim\limits_{x\to 0} x = 0$.

AUTHOR: In fact, if $\lim\limits_{x\to 0} f(x) = 0$ and $\lim\limits_{x\to 0} g(x) = 0$, the limit of their ratio i.e., the limit of the function $\dfrac{f(x)}{g(x)}$ may exist.

READER: What is this limit?

AUTHOR: It depends on the functions $f(x)$ and $g(x)$. Let us show, for example, that

$$\boxed{\lim_{x\to 0}\frac{\sin x}{x} = 1}$$

Note that the function $\dfrac{\sin x}{x}$ is not defined at $x = 0$. This fact, however, does not interfere with searching for the limit of the function at $x = 0$.

We shall start with well-known inequalities:

$$\sin x < x < \tan x \left(0 < x < \frac{\pi}{2}\right)$$

An assumption that $\left(0 < x < \frac{\pi}{2}\right)$ will not affect the generality of our results. Dividing $\sin x$ by each term of these inequalities, we obtain

$$1 > \frac{\sin x}{x} > \cos x$$

hence

$$0 < \left(1 - \frac{\sin x}{x}\right) < (1 - \cos x)$$

Next we take into account that

$$1 - \cos x = 2\sin^2 \frac{x}{2} < 2\sin \frac{x}{2} < 2\frac{x}{2} = x$$

Thus we have

$$0 < \left(1 - \frac{\sin x}{x}\right) < x$$

or

$$-x < -\left(1 - \frac{\sin x}{x}\right) < 0$$

whence

$$\left|1 - \frac{\sin x}{x}\right| < |x|$$

We thus arrive at the following inequality valid for $|x| < \frac{\pi}{2}$

$$\left|\frac{\sin x}{x} - 1\right| < |x| \tag{32}$$

By using this inequality, we can easily prove that the function $\frac{\sin x}{x}$ has the limit at $x = 0$, and this limit is unity. It will be convenient to use definition *1* for the limit of function at a point.

Select an arbitrary $\varepsilon > 0$, demanding for the sake of simplicity that $\varepsilon < \frac{\pi}{2}$. For δ, it is sufficient to take $\delta = \varepsilon$ since, according to (32), the condition $|x - 0| < \delta$ immediately leads to

$$\left| \frac{\sin x}{x} - 1 \right| < \delta = \varepsilon$$

Thus, unity is indeed the limit of the function $\frac{\sin x}{x}$ at $x = 0$.

READER: Do we really have to resort to a similar line of reasoning, based on the definition of the limit of function at a point, each time we have to find the limit of $\frac{f(x)}{g(x)}$ when both $\lim_{x \to 0} f(x) = 0$ and $\lim_{x \to 0} g(x) = 0$.

AUTHOR: No, of course not. The situation we are speaking about is known as an *indeterminate form* of the type $\frac{0}{0}$. There are rules which enable one to analyze such a situation in a relatively straightforward manner and, so to say, "resolve the indeterminacy". In practice it is usually not difficult to solve the problem of existence of the limit of a function $\frac{f(x)}{g(x)}$ at a specific point and find its value (if it exists). A few rules of evaluation of indeterminate forms of the type $\frac{0}{0}$ (and other types as well) will be discussed later. A systematic analysis of such rules, however, goes beyond the scope of our dialogues.

It is important to stress here the following principle (which is significant for further considerations): although the theorem on the limit of the ratio is not valid in the cases when $\lim_{x \to 0} g(x) = 0$, the limit of a function $\frac{f(x)}{g(x)}$ at a point a may exist if $\lim_{x \to 0} f(x) = 0$. The example of the limit of the function $\frac{\sin x}{x}$ at $x = 0$ is a convincing illustration of this principle.

READER: Presumably, a similar situation may take place for numerical sequences as well?

AUTHOR: It certainly may. Here is a simple example:

$$(x_n) = 1, \frac{1}{8}, \frac{1}{27}, \frac{1}{64}, \ldots \frac{1}{n^3}, \ldots \quad (\lim_{n\to\infty} x_n = 0)$$
$$(y_n) = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots \frac{1}{n}, \ldots \quad (\lim_{n\to\infty} y_n = 0)$$

It is readily apparent that the limit of the sequence $\left(\frac{x_n}{y_n}\right)$ is the limit of the sequence $\left(\frac{1}{n^2}\right)$. This limit does exist and is equal to zero.

READER: You mentioned that the existence of the limit of a function $\frac{f(x)}{g(x)}$ at a, when both $\lim_{x\to 0} f(x) = 0$ and $\lim_{x\to 0} g(x) = 0$ the existence of the limit of the type $\frac{0}{0}$, is very important for further considerations. Why?

AUTHOR: The point is that one of the most important concepts in calculus, namely, that of *derivative*, is based on the limit of the type $\frac{0}{0}$. This will be clear in the subsequent dialogues.

Dialogue 8

VELOCITY

AUTHOR: We are practically ready to tackle the concept of a *derivative*. This concept, alongside with the concepts of the *limit of numerical sequence* and the *limit of function*, is one of the most important special concepts in calculus.

We may approach the concept of a derivative by considering, for instance, a quantity widely used in physics: the *instantaneous velocity* of nonuniform motion of a body.

READER: We have been familiarized with this notion when studying kinematics in the course of physics, or, to be precise, the kinematics of nonuniform motion in a straight line.

AUTHOR: Exactly. What is your idea of the instantaneous velocity?

READER: The instantaneous velocity of a body is defined as the velocity of a body at a given moment of time (at a given point of its trajectory).

AUTHOR: And what is your idea of the *velocity at a given moment of time*?

READER: Well, I see it as ...If a body moves uniformly, at different moments of time its velocity remains the same. If a body moves non-uniformly (accelerating or decelerating), its velocity will, in the general case, vary from moment to moment.

AUTHOR: Don't you feel that the phrase "velocity at a given moment of time" is merely a *paraphraze* of the "instantaneous velocity"? Six of one and half a dozen of the other, eh? The term "velocity at a given moment of time" calls for an explanation as much as the term "instantaneous velocity".

To measure the velocity of a body, one should obviously measure a certain distance (path) covered by the body, and the time interval during which the distance is covered. But, what path and period of time are meant when we refer to the velocity *at a given moment of time*?

READER: Yes, in order to *measure* velocity, one must actually know a certain path and time interval during which the path is covered. But our subject is not the measurement, it is a *definition* of the *instantaneous velocity*.

AUTHOR: For the time being we shall not bother about a formal definition. It is more important to realize its essential meaning. In order to do this, we cannot avoid the aspect of measurements. Now, how would you find away to measure the velocity of a body at a given moment of time?

READER: I can take a short time interval Δt, that is, the period from the given moment of time t to the moment $t+\Delta t$. During this time interval the body covers a distance Δs. If Δt is sufficiently small, the ratio $\dfrac{\Delta s}{\Delta t}$ will give the velocity of the body at the moment t.

AUTHOR: What do you mean by a *sufficiently short* time interval? What

do you compare it with? Is this interval sufficiently small in comparison with a year, a month, an hour, a minute, a second, or a millisecond?

READER: Perhaps, neither a year, a month, an hour nor a minute will do in this case. I see now that the instantaneous velocity can only be measured with a certain degree of accuracy. The smaller is the smaller is the error with which, the instantaneous velocity is measured.

AUTHOR: In principle, the concept of the *instantaneous velocity* (or, in other words, "velocity at a given moment of time") must be independent of the measurement accuracy.

The velocity you are talking about, that is, the ratio $\frac{\Delta s}{\Delta t}$ is nothing more than the *average velocity during* Δt. It is not the instantaneous velocity at all. Of course, you are right when you say that the smaller is Δt the closer is the value of the average. velocity to the value of the instantaneous velocity. However, no matter how small is Δt; the ratio $\frac{\Delta s}{\Delta t}$ is always only the average velocity during lit.

READER: Then a better definition of the instantaneous velocity is beyond me.

AUTHOR: Consider a graph of distance covered by a body plotted as a function of time, that is, the graph of the function $s = s(t)$. This graph is shown in Figure 31 by a solid line. Note that in physics one typically uses the same symbol to denote both a function and its values (in this case we use the symbol s).

READER: The figure also shows several thin lines.

AUTHOR: The thin lines (parabolas: and straight lines) are shown only to indicate how the graph of $s = s(t)$ was plotted. This graph is thus

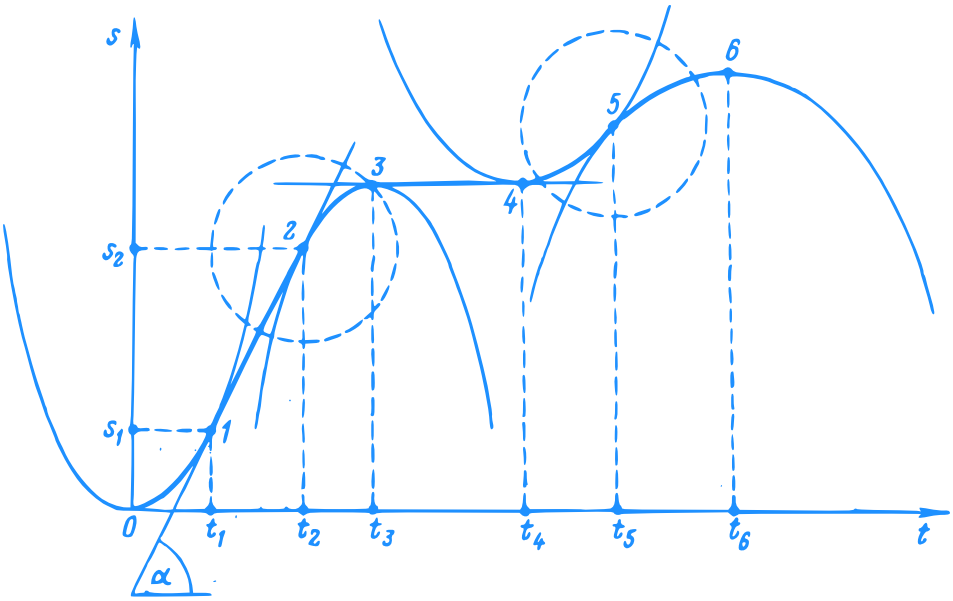


Figure 31: Analysing the concept of instantaneous velocity.

composed of "pieces" of parabolas and straight lines. For instance, for the time interval from 0 to t_1 the graph is represented by a "piece" of the extreme left-hand parabola (portion 0-1 of the graph). Please recall the formula for the distance covered in a uniformly accelerated motion with zero initial velocity.

READER: This formula is

$$s(t) = \frac{at^2}{2} \tag{33}$$

where a is acceleration.

AUTHOR: And the extreme left-hand parabola is the graph of the function represented by your formula.

READER: So for the time interval from 0 to t_1 the body moves at a constant acceleration.

AUTHOR: Exactly.

READER: I see. For the time interval from t_1 to t_2 the body moves uniformly (portion 1-2 of the graph is a straight line); from t_2 to t_3 the body moves at a constant deceleration (the graph is an inverse parabola); from t_3 to t_4 the body is not moving at all; from t_4 to t_5 it moves at a constant acceleration, and from t_5 to t_6 it moves at a constant deceleration.

AUTHOR: Precisely so. Now let us consider the graph of the function $s(t)$ shown in Figure 31 from a purely mathematical standpoint. Let us pose the following question: How strongly do the values of the function change in response to the value of its argument t in different portions of the graph?

READER: In portion 3-4 the values of the function $s(t)$ do not change at all, while in other portions they do. A slower rate of change of the function is observed in the vicinity of points 0, 3, 4, and 6; a faster rate of change is observed in the vicinity of points 1, 2, and 5. As a matter of fact, the rate of change is equally fast throughout portion 1-2.

AUTHOR: You are a keen observer. And where do you think the rate of change is faster, at point 2 or at point 5?

READER: Of course, at point 2. Here the graph of the function has a much steeper slope than at point 5.

AUTHOR: Let us turn to Figure 32. Here in column *A* two portions of the graph of the function $s(t)$ are shown separately, namely, those in the vicinity of points 2 and 5 (in Figure 31 these portions are identified by dash circles). In column *B* the portions of the graph close to points 2 and 5 are shown again, but this time with a two-fold increase in scale. Column *C* shows the result of another two-fold scale increase. Obviously, as the

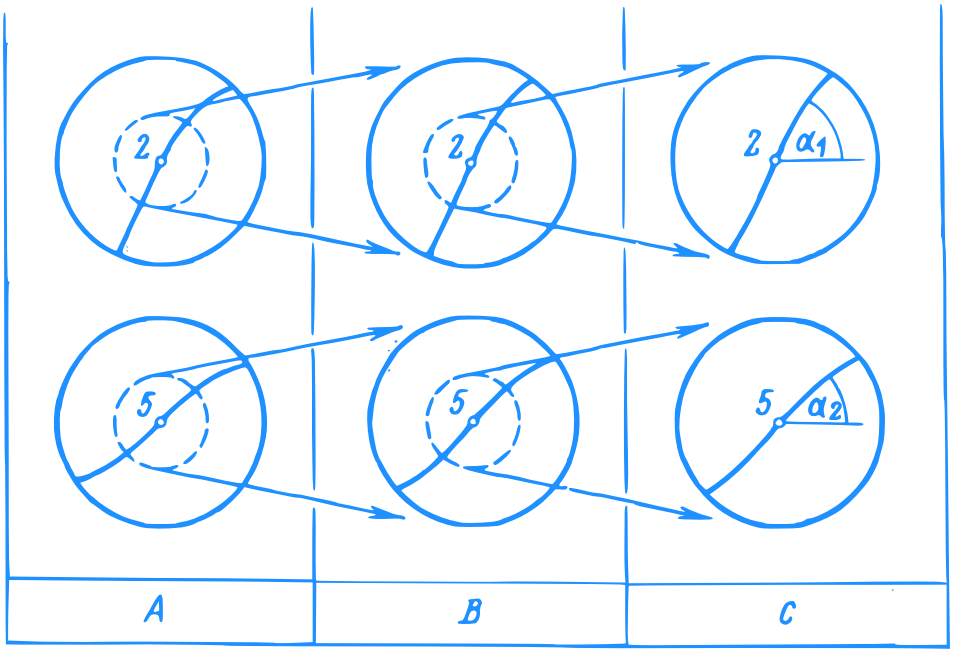


Figure 32: Analysing the concept of instantaneous velocity mathematically.

scale increases, the curvature of the graph $s(t)$ becomes less noticeable. We may say that the graph has a property of "linearity on a small scale", which enables us to consider the slope of the graph *at a specific point*. In Figure 32 (in column C) it is shown that the slope of the graph at point 2 is (α_1 (the slope is measured relative to the t-axis), while the slope at point 5 is α_2, and clearly $\alpha_2 < \alpha_1$.

Denote the slope of the curve $s(t)$ at the moment t by $\alpha(t)$. Then $\tan\alpha(t)$ is said to be the rate of change of the function $s(t)$ at the moment t, or simply the instantaneous velocity.

READER: But why tangent?

AUTHOR: You immediately come to it by considering portion 1-2 of the graph in Figure 31. This portion represents a uniform motion of the

body, the rate of change of $s(t)$ being identical at all points. Obviously, it equals the average velocity during the time interval $t_2 - t_1$, which is

$$\frac{s_2 - s_1}{t_2 - t_1} = \tan\alpha$$

READER: In Figure 32 you have demonstrated a "straightening" of the graph by increasing its scale. But this straightening is only *approximate*. Why have you stopped at a mere four-fold scale increase?

AUTHOR: We can get rid of this approximation and formulate a more rigorous definition of a slope at a point. To be more specific, we consider a segment of the graph $s(t)$ close to point 5. In this segment we select an arbitrary point B and draw a secant through points 5 and B (Figure 33). Next, on the same graph between points 5 and B we select an arbitrary point C and draw a new secant $5C$. Further, we select an arbitrary point D in the segment between 5 and C and draw a new secant $5D$. We may continue this process infinitely long and, as a result, we obtain a sequence of secants which converges to a certain straight line (line $5A$ in Figure 33). This straight line is said to be tangent to the curve at point 5. *The slope of the tangent is said to be the slope of the graph at a given point.*

READER: If I understand you correctly we are now in a position to formulate strictly the answer to the question about the instantaneous velocity.

AUTHOR: Try to do it, then.

READER: The instantaneous velocity of a body at a moment of time t is the rate of change of $s(t)$ at the moment t. *Numerically it is equal to the tangent of the slope of the tangent line to the graph of the function $s(t)$ at the moment t.*

AUTHOR: Very good. But you should have mentioned that $s(t)$ expresses

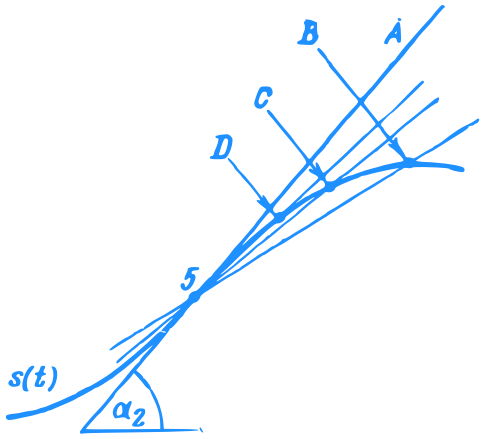


Figure 33: Analysing the concept of instantaneous velocity mathematically.

the distance covered by the body as a function of time.

READER: This is true, my definition of the instantaneous velocity is tied to the *graph* of $s(t)$. What if the function $s(t)$ is not defined graphically?

AUTHOR: Anyway, a graph for $s(t)$ always exists. The only "inconvenience" in your definition is that it is necessary to take into account the scale of units on the coordinate axes. If the unit of time (on the t-axis) and the unit of length (on the s-axis) are represented by segments of identical length, the instantaneous velocity at time t is

$$\tan\alpha(t)\,\frac{\text{unit of length}}{\text{unit of time}}$$

If, however, the segment representing one unit of length is n times greater than the segment representing one unit of time, the instantaneous velocity is

$$\frac{1}{n}\tan\alpha(t)\,\frac{\text{unit of length}}{\text{unit of time}}$$

This "inconvenience", however, has no principal significance. But it is also possible to formulate a definition of the instantaneous velocity in a

form free of graphic images.

Look at Figure 34 which carries Figure 33 one step further.

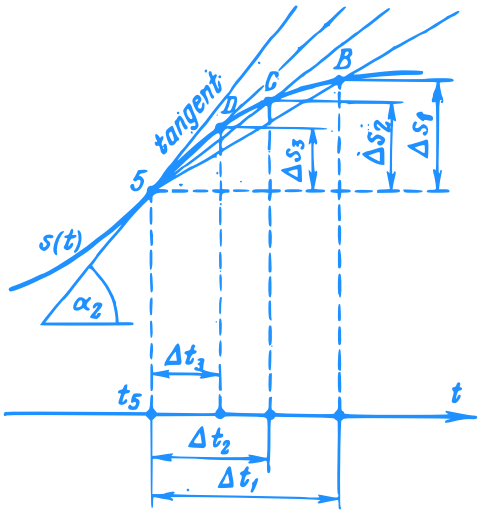


Figure 34: Analysing the concept of instantaneous velocity mathematically.

Figure 34 shows that the slope of the secant $5B$ is a ratio $\frac{\Delta s_1}{\Delta t_1}$. In other words, this is the average velocity for the time interval from t_5 to $t_5+\Delta t_1$. The slope of the secant $5C$ is $\frac{\Delta s_2}{\Delta t_2}$ that is, the average velocity for the time interval from t_5 to $t_5+\Delta t_2$ $(\Delta t_2 < \Delta t_1)$. The slope of the secant $5D$ is $\frac{\Delta s_3}{\Delta t_3}$ that is, the average velocity for the time interval from t_5 to $t_5+\Delta t_3$ $(\Delta t_3 < \Delta t_2)$, etc. Thus, a sequence of the secants converging to the tangent line (drawn at point 5 of the graph $s(t)$) corresponds to a sequence of the average velocities converging to the slope α_2 of the tangent line, that is, to the value of the instantaneous velocity at the time moment t_5.

READER: It comes out that the *instantaneous velocity is the limit of a sequence of average velocities*.

AUTHOR: Precisely. The instantaneous velocity is in fact the limit of a sequence of average velocities, provided that the time interval over which the averaging is made tends to zero converging to the moment of time t (viz., t_5 in Figure 34).

Now let us formulate the definition in a more rigorous manner. What we want to define is the instantaneous velocity of a body at a moment of time t. Consider an arbitrary time interval from t to $t+\Delta t_1$. The distance covered by the body during this interval is Δs_1. The average velocity of the body during this time interval is

$$v_{\text{av}}(t, \Delta t_1) = \frac{\Delta s_1}{\Delta t_1}$$

Next we select a shorter time interval δt_2, from t to $t+\Delta t_2$, ($\Delta t_2 < \Delta t_1$), during which a distance Δs_2 is covered. Consequently, the average velocity over Δt_2 is

$$v_{\text{av}}(t, \Delta t_2) = \frac{\Delta s_2}{\Delta t_2}$$

We continue this process of selecting shorter and shorter time intervals starting at the moment of time t. As a result, we obtain a sequence of the average velocities

$$v_{\text{av}}(t, \Delta t_1),\ v_{\text{av}}(t, \Delta t_2),\ v_{\text{av}}(t, \Delta t_3), \ldots$$

The limit of this sequence for $\Delta t \to 0$ is the instantaneous velocity at the moment of time t:

$$v(t) = \lim_{\Delta t \to 0} v_{\text{av}}(t, \Delta t) \tag{34}$$

Taking into account that

$$v_{\text{av}}(t, \Delta t) = \frac{s(t+\Delta t) - s(t)}{\Delta t}$$

we rewrite expression (35) in the following form

$$v(t) = \lim_{\Delta t \to 0} \frac{s(t+\Delta t) - s(t)}{\Delta t} \tag{35}$$

As a result, we can formulate the following definition of the instantaneous velocity.

Definition

The instantaneous velocity at a moment of time t is the limit of a sequence of average velocities over time intervals from t to $t + \Delta t$ for $\Delta t \to 0$.

READER: Now I realize that instead of talking about a sufficiently small time interval Δt (I am referring to our talk about the ratio $\frac{\Delta s}{\Delta t}$ at the beginning of the dialogue), the argument should have been based on the limit transition for $\Delta t \to 0$. In other words, the instantaneous velocity is not $\frac{\Delta s}{\Delta t}$ at a sufficiently small Δt but $\lim_{\Delta t \to 0} \frac{\Delta s}{\Delta t}$

AUTHOR: Exactly. The definition formulated above for the instantaneous velocity not only exposes the gist of the concept but gives a rule for its calculation, provided that an analytical expression for $s(t)$ is known. Let us make such a calculation assuming that $s(t)$ is given by expression (33).

READER: We should substitute (33) for (35). This gives

$$v(t) = \lim_{\Delta t \to 0} \frac{\frac{a(t+\Delta t)^2}{2} - \frac{at^2}{2}}{\Delta t}$$

AUTHOR: Go ahead. Remove the parentheses.

READER: This will give

$$\begin{aligned} v(t) &= \lim_{\Delta t \to 0} \frac{a(t^2 + 2t\Delta t + \Delta t^2 - t^2)}{2\Delta t} \\ &= \lim_{\Delta t \to 0} \left(at + \frac{\Delta t}{2} \right) \\ &= at \end{aligned}$$

We have arrived at a familiar formula for the velocity of uniformly accelerated motion with zero initial velocity:

$$\boxed{v(t) = at} \tag{36}$$

AUTHOR: You are absolutely right. I must congratulate you: for the first time in your life you have carried the so-called operation of *differentiation*. In other words, you have determined for a given function $s(t)$ its *derivative*, that is, the function $v(t)$.

READER: Does it mean that the instantaneous velocity is a derivative?

AUTHOR: Note that a derivative exists only *with respect to* a known *initial* function. If the initial function is $s(t)$ (path as a function of time), the derivative is the instantaneous velocity.

Let us return now to the graph $s(t)$ shown in Figure 31. Our previous arguments and, in particular, relation (36), allow us to transform the graph $s(t)$ into a graph of the derivative, that is, the function $v(t)$. A comparison of the two graphs is given in Figure 35. I recommend that you carefully analyze Figure 35, interpreting it as a comparison of the graph of a function $s(t)$ and the graph of its rate of change.

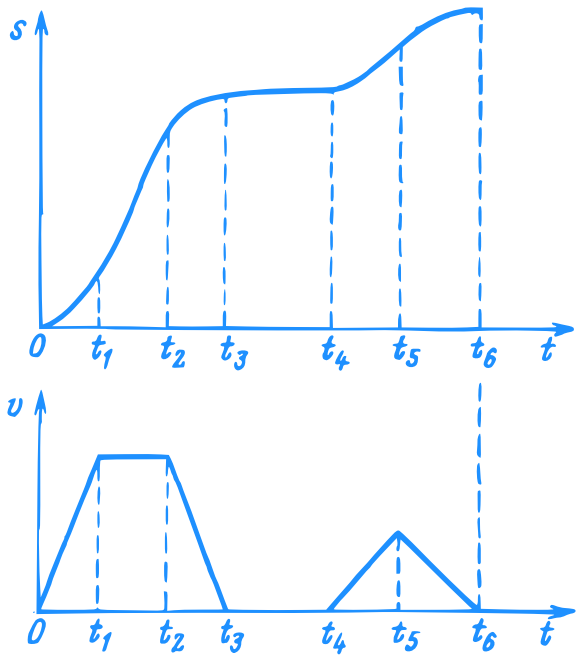


Figure 35: Comparing the graph of a function $s(t)$ and the graph of its rate of change.

Dialogue 9

DERIVATIVE

AUTHOR: The previous dialogue gave us an opportunity to introduce the concept of a *derivative* for a specific example from physics (the instantaneous velocity of a body moving non-uniformly along a straight line). Now let us examine this concept from a purely mathematical viewpoint without assigning any physical meaning to the mathematical symbols used.

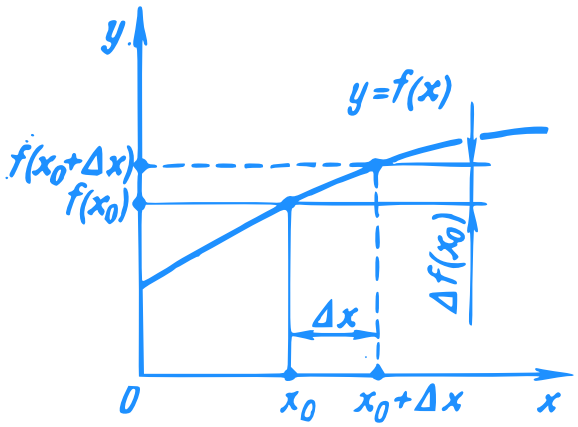


Figure 36: Examining concept of a derivative on an arbitrary function.

Figure 36 shows a graph of an arbitrary function $y = f(x)$. Let us select an arbitrary point $x = x_0$ from the domain of the function. In the subsequent argument this point is considered as *fixed*. Now consider another point x from the domain of the function and introduce a notation $\Delta x = x - x_0$. The value Δx is called the *increment of the independent variable*. The increment is considered with respect to the fixed point x_0. Depending on the point x, the value of Δx may. be larger or smaller, positive or negative.

Now let us examine a difference between the values of the function at points $x = x_0 + \Delta x$ and $x = x_0$:

$$\Delta f(x_0) = f(x_0 + \Delta x) - f(x_0)$$

The difference $\Delta f(x_0)$ is said to be the *increment of a function f at a point x_0*. Since x_0 is fixed, $\Delta f(x_0)$ should be considered as a function of a variable increment Δx of the independent variable.

READER: Then it is probably more logical to denote this function by $\Delta f(\Delta x)$, and not by $\Delta f(x_0)$, isn't it?

AUTHOR: Probably, you are right. However, the accepted notation is $\Delta f(x_0)$. Such a notation emphasized the fact that the increment of f (in other words, the given function of Δx is referred to point x_0.

With the concepts of the increment introduced, it is not difficult to evaluate the rate of change of f close to x_0.

READER: This rate should be described by the ratio $\dfrac{\Delta f(x_0)}{\Delta x}$. For instance, if we compare $\Delta f(x_0)$ with an increment of t at another point from the domain of the function (say, point $x = x_1$), we may obtain an inequality

$$\frac{\Delta f(x_1)}{\Delta x} > \frac{\Delta f(x_0)}{\Delta x}$$

and therefore conclude that the rate of change of f close to point x_1 is greater than that close to x_0.

AUTHOR: Please, be careful. You have not said anything about the vague increment Δx. If Δx is too large, the inequality you have just mentioned may lead to a wrong conclusion. I shall make myself clearer by referring to Figure 37. As you see,

$$\frac{\Delta f(x_1)}{\Delta x} > \frac{\Delta f(x_0)}{\Delta x}$$

You must agree, however, that close to point x_0 the function changes much faster (the graph of the function has a steeper slope) than in the vicinity of x_1.

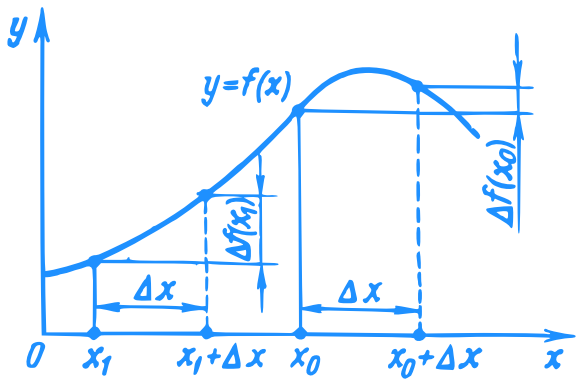


Figure 37: Examining concept of a derivative on an arbitrary function.

READER: It is necessary that the value of the increment Δx be sufficiently small. The smaller is Δx the more accurate is the information about the rate of change of the function close to the point under consideration.

AUTHOR: Well, we can do even better than this. We may, for example, consider the limit of the ratio $\dfrac{\Delta f(x_0)}{\Delta x}$ for $\Delta x \to 0$ (remember the previous dialogue).

READER: This limit will characterize the *rate of change* of the function f directly at $x = x_0$.

AUTHOR: Exactly. Let us calculate the limit in detail:

$$\lim_{\Delta x \to 0} \frac{\Delta f(x_0)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta f(x + x_0) - f(x_0)}{\Delta x} \tag{37}$$

and examine first of all the mathematical nature of this limit.

READER: Since point x_0 is fixed, it is evidently the limit of the ratio of two functions of Δx for $\Delta x \to 0$.

AUTHOR: Let us denote these functions by F and G:

$$F(\Delta x) = f(x_0 + \Delta x) - f(x_0), \qquad G(\Delta x) = \Delta x$$

READER: Limit (37) is then

$$\lim_{\Delta x \to 0} \frac{F(\Delta x)}{G(\Delta x)}$$

where $\lim_{\Delta x \to 0} F(\Delta x) = 0$ and $\lim_{\Delta x \to 0} G(\Delta x) = 0$. Hence, we have here a limit similar to that discussed at the end of Dialogue Seven, namely, a limit of the type $\frac{\infty}{\infty}$.

AUTHOR: Right. This limit, that is, the limit of the type $\frac{\infty}{\infty}$ the main subject of this dialogue.

The primary requirement in this case is the *existence* of the limit. It means that the function f should be such that

$$\lim_{\Delta x \to 0} F(\Delta x) = 0$$

The necessary condition for satisfying this equality is the *continuity* of f at $x = x_0$ But we shall discuss this problem later.

If the limit of the type $\frac{\infty}{\infty}$ (in other words, limit (37)) does exist, it is called "the derivative of the function f at point $x = x_0$ and usually denoted by $f'(x_0)$."

Definition

The derivative of a function f at a point x_0 (denoted by $f'(x_0)$) is the limit of the ratio of an increment of the function f at the point x_0 (denoted by $\Delta f(x_0)$) to an increment Δx the independent variable for $\Delta x \to 0$:

$$f'(x_0) = \lim_{\Delta x \to 0} \frac{\Delta f(x_0)}{\Delta x}$$

or, in a more detailed notation,

$$f'(x_0) = \lim_{\Delta x \to 0} \frac{\Delta f(x_0)}{\Delta x} \tag{38}$$

Note that you are already familiar with the right-hand side of equation (38) (cf. expression (35) from the previous dialogue).

READER: Actually the derivative of the function f at point 0 is the limit of the function

$$\frac{F}{G} = \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

at $\Delta x = 0$. The independent variable of the function $\frac{F}{G}$ is the increment Δx.

AUTHOR: You are quite right. However, in what follows you must use the definition of the derivative as formulated above. This definition does not involve the function $\frac{F}{G}$ of Δx since this function plays, as you understand, only an auxiliary role. We should simply bear in mind that the phrase "the limit of the ratio of an increment $\Delta f(x_0)$ to an increment Δx

for $\Delta x \to 0$" describes the limit of a function of Δx, i.e. the function $\dfrac{F}{G}$, which is considered at $\Delta x = 0$.

The derivative can be also interpreted *in terms of geometry*.

READER: Shall we do it by using again the tangent to the graph of a function?

AUTHOR: Yes, of course. Let us take the graph $y = f(x)$ (Figure 38), fixing a point $x = x_0$. Consider an increment Δx_1 of the argument; the corresponding increment of the function at point x_0 is $\Delta f_1(x_0)$. Denote the slope of the chord AB_1 by α_1 it is readily apparent that

$$\frac{\Delta f_1(x_0)}{\Delta x_1} = \tan \alpha_1$$

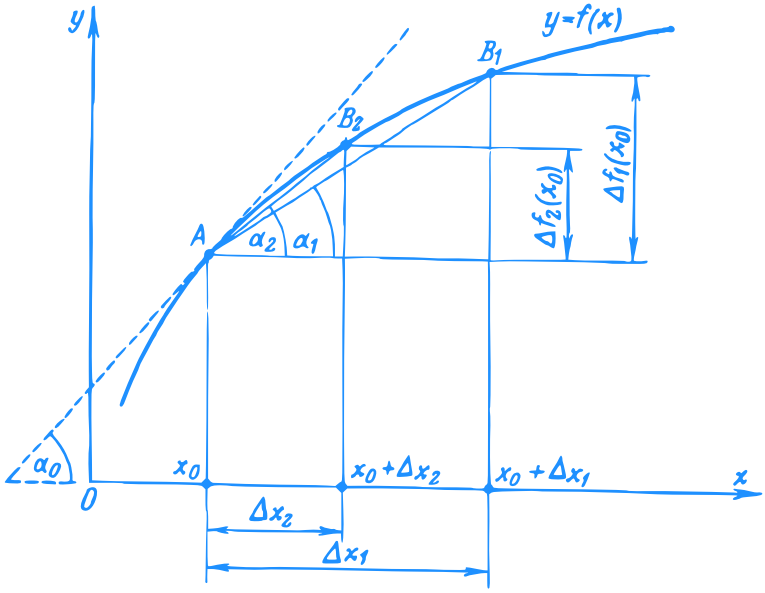


Figure 38: Examining concept of a derivative geometrically.

Next take an increment Δx_1 (so that $\Delta x_2 < \Delta x_1$). This increment corresponds to the increment $f_2(x_0)$ of the function f at point x_0. Denote the

slope of the chord AB_2 by α_2; it is similarly quite apparent that

$$\frac{\Delta f_2(x_0)}{\Delta x_2} = \tan \alpha_2$$

Further, take an increment Δx_3 (so that $\Delta x_3 < \Delta x_2$), and so on. As a result, we obtain an infinitesimal sequence of increments of the independent variable:

$$\Delta x_1, \Delta x_2, \Delta x_3, \Delta x_4, \ldots \Delta x_n, \ldots$$

and the corresponding infinitesimal sequence of increments of the function at point x_0:

$$\Delta f_1(x_0), \Delta f_2(x_0), \Delta f_3(x_0), \Delta f_4(x_0), \ldots \Delta f_n(x_0), \ldots$$

This leads to a new sequence of the values of the tangent of the slopes of the chords $AB_1, AB_2, AB_3, \ldots AB_n, \ldots$. obtained as a sequence of the ratios of the two sequences given above

$$\tan \alpha_1, \tan \alpha_2, \tan \alpha_3, \tan \alpha_4, \ldots \tan \alpha_n, \ldots \tag{39}$$

Both sequences Δx_n and $\Delta f_n(x_0)$ converge to zero. And what can be said about the convergence of the sequence $(\tan \alpha_n)$ or, in other words, the sequence $\dfrac{\Delta f_n(x_0)}{\Delta x_n}$?

Reader: Obviously, the sequence $\dfrac{\Delta f_n(x_0)}{\Delta x_n}$ converges to $f'(x_0)$. In other words, the limit of $\dfrac{\Delta f_n(x_0)}{\Delta x_n}$ is the derivative of f at x_0.

Author: What are the grounds for this conclusion?

Reader: Why, isn't it self-evident?

Author: Let me help you. Your conclusion is based on *Definition 2* of the limit of function at a point. Don't you think so?

READER: Yes, I agree. Indeed, a certain number (in this case $f'(x_0$ is the limit of a function $\Phi(\Delta x)$ (in this case $\Phi = \frac{F}{G}$) at $\Delta x = 0$ if for any sequence Δx_n convergent to zero the corresponding sequence $(\Phi(\Delta x_n))$ converges to this number. Sequence (39) is precisely the sequence $(\Phi(\Delta x_n))$ in our case.

AUTHOR: Correct. We have thus found that $\lim\limits_{n\to\infty} \tan\alpha_n = f'(x_0)$. Now look at Figure 38 and tell me which direction is the limit for the sequence composed of the chords $AB_1, AB_2, AB_3, \ldots AB_n, \ldots$?

READER: It is the direction of the *tangent* to the graph $f(x)$ at point $x = x_0$.

AUTHOR: Correct. Denote the slope of the tangent line by α_0. Thus

$$\lim_{n\to\infty} \tan\alpha_n = \tan\alpha_0$$

Consequently,

$$f'(x_0) = \tan\alpha_0$$

We thus obtain the following geometrical interpretation of the derivative:

> The derivative of a function f at a point x_0 is defined by the slope of the tangent to the graph of the function f at the point $x = x_0$.

Note that the slope of the tangent is measured relative to the positive direction of the abscissa axis, so that the derivative of f at point x_0 in Figure 39 is positive (at this point $\tan\alpha_0 > 0$), while at point x_0 the derivative of f is negative $\tan\alpha_0 < 0$).

But the geometrical interpretation of the derivative must not upstage the basic idea that

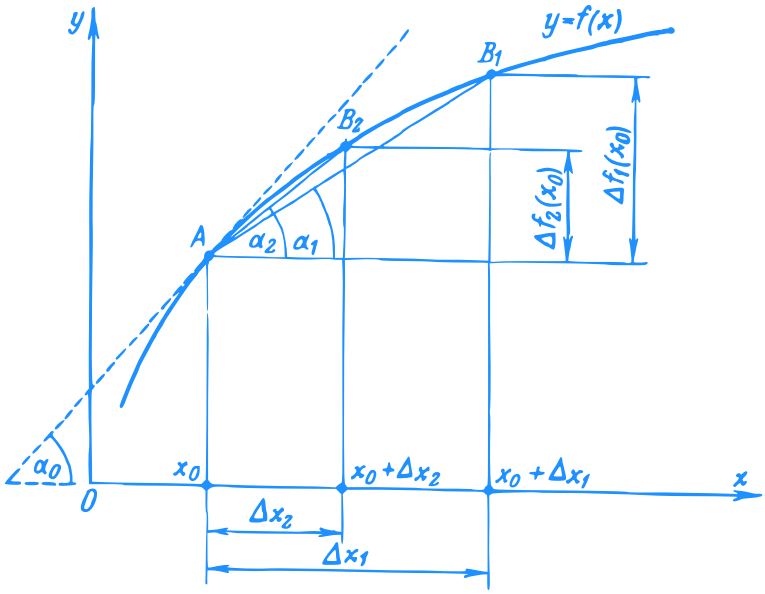


Figure 39: Examining concept of a derivative geometrically.

> The derivative of a function f at a point x_0 is the rate of change of f at this point.

In the previous dialogue we analyzed the function $s(t)$ describing the dependence of the distance covered by a body during the time t. In this case the derivative of $s(t)$ at a point $t = t_0$ is the *velocity* of the body at the moment of time $t = t_0$. If, however, we take $v(t)$ as the initial function (the instantaneous velocity of a body as a function of time), the derivative at $t = t_0$ will have the meaning of the acceleration of the body at $t = t_0$. Indeed, *acceleration* is the *rate of change of the velocity* of a body.

READER: Relation (38) seems to allow a very descriptive (if somewhat simplified) interpretation of the derivative. We may say that

The derivative of a function $y = f(x)$ at a point $x = x_0$ shows how much steeper the change in y is in comparison with the change in x in the neighbourhood of $x = x_0$.

AUTHOR: This interpretation of the derivative is quite justified, and it may be useful at times.

Getting back to the geometrical interpretation of the derivative, we should note that it immediately leads to the following rather important

Conclusions:

§ The derivative of a function $f = \text{const}$ (the derivative of a constant) is zero at all the points.

§ The derivative of a junction $f = ax + b$ (where a and b are constants) is constant at all the points and equals a.

§ The derivative of a function $f = \sin x$ is zero at the points $x = +\pi n$ (at these points the tangent to the graph of the function is horizontal).

This "list" could, of course, be expanded.

Next I would like to attract your attention to the following: from the viewpoint of mathematics a derivative of a function must also be considered as a certain function.

READER: But the derivative is a limit and, consequently, a *number*!

AUTHOR: Let us clarify this. We have fixed a point $x = x_0$ and obtained

for a function $f(x)$ at this point the number

$$\lim_{\Delta x \to 0} \frac{\Delta f(x_0)}{\Delta x}$$

For each point x (from the domain of f) we have, in the general case, its own number

$$\lim_{\Delta x \to 0} \frac{\Delta f(x)}{\Delta x}$$

This gives a mapping of a certain set of numbers x onto a different set of numbers $\lim\limits_{\Delta x \to 0} \dfrac{\Delta f(x)}{\Delta x}$. The function which represents this mapping of one numerical set onto another is said to be the derivative and is denoted by $f'(x)$.

READER: I see. So far we have considered only one value of the function $f(x)$, namely, its value at the point $x = x_0$.

AUTHOR: I would like to remind you that in the previous dialogue we analyzed $v(t)$ which was the derivative of $s(t)$. The graphs of the two functions (i.e. the initial function $s(t)$ and its derivative $v(t)$) were even compared in Figure 35.

READER: Now it is clear.

AUTHOR: I would like to make two remarks with regard to $f'(x)$.

Note 1. A function $f'(x)$ is obtained only by using a function $f(x)$. Indeed,

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} \tag{40}$$

It is as if there is a certain operator (recall Dialogue Four) which generates $f'(x)$ at the output in response to $f(x)$ at the input. In other words, this operator, applied to the function $f(x)$, "generates" $f'(x)$. This operator

is usually denoted by $\frac{d}{dx}$. This notation should be interpreted as a single entity and not as a ratio (it reads: "d over dx").

Consider an "image"

$$\frac{d}{dx}\boxed{1}=\boxed{2}.$$

The squares in this expression symbolize the familiar "windows". "Window" 1 is to input $f(x)$, while "window" 2 outputs $f'(x)$. Thus,

$$\boxed{\frac{d}{dx}f(x)=f'(x)} \tag{41}$$

Definition:

The operation of obtaining $f'(x)$ from $f(x)$ is said to be the differentiation of $f(x)$.

The operator $\frac{d}{dx}$ performs this operation over $f(x)$ and is said to be the operator of differentiation.

READER: But what exactly is $\frac{d}{dx}$ doing with $f(x)$?

AUTHOR: It is exactly the operation prescribed by (40). We may say that $\frac{d}{dx}$ "constructs" the ratio

$$\frac{(x+\Delta x)-f(x)}{\Delta x}$$

from $f(x)$ and determines the limit of this ratio (regarded as a function of Δx) at $dx=0$.

READER: In other words, the operator dx performs a certain limit transition operation, doesn't it?

AUTHOR: Certainly. The whole differential calculus (and with it, integral calculus) can be formulated in terms of certain limit transitions.

READER: Why should we introduce an operator $\frac{d}{dx}$ if it represents nothing else but the limit transition operation described by (40)?

AUTHOR: You have posed a very important question. The problem is that if we had formulated differential calculus in terms of limits, using the relations of type (40), all books on calculus should have been increased in their volume several-fold and become hardly readable. The use of the relations of type (41), instead of (40), makes it possible to avoid this.

READER: But how can we use the relations of type (41) *without implicitly applying* the relations of type (40)?

AUTHOR: What is done is this.

First, using (40), we find the result of applying the operator $\frac{d}{dx}$ to a sum, product, and ratio of functions, and to composite or inverse functions *provided* that the result of applying the operator to the initial function (or functions) is *known*. In other words, the first step is to establish the *rules for the differentiation of functions*.

Second, using (40), we find out the result of applying $\frac{d}{dx}$ to some basic elementary functions (for instance, $y = x^n$, $y = \sin x$, and $y = \log x$).

After these two steps are completed you can practically forget about the relations of type (40). *In order to differentiate a function, it is sufficient to express the function via basic elementary functions (the derivatives of which were obtained earlier) and apply the rules for differentiation.*

READER: Does it mean that the relations of type (40) could be put aside after they have been used, first, for compiling a *set of differentiation rules*

and, second, for making a *table of derivatives for basic elementary functions*?

AUTHOR: Yes, this is the procedure. Using the differentiation rules and the table of derivatives for some basic elementary functions you are in a position to forget about the relations of type (40) and are free to proceed further by using the "language" of the relations of type (41). A formal course of differential calculus could skip the analysis of limit transition operations, that is, the relations of type (40). It is quite sufficient for a student to learn a set of differentiation rules and a table of derivatives of some functions.

READER: I certainly prefer to be given the foundation.

AUTHOR: Our next dialogue will be devoted to a discussion of the programme of actions as outlined above. At the first step of the programme, the main rules for differentiation will be established on the basis of the relations of (40) and, in addition, the derivatives of three functions $y = x^2$, $y = \sin x$, and $y = \log_a x$ will be obtained. At the second step, we shall obtain (without reference to the relations of type (40)) the derivatives of the following functions:

$y = x^n$, $y = x^{-n}$, $y = \sqrt{x}$, $y = \cos x$, $y = \tan x$, $y = \cot x$, $y = \arcsin x$, $y = \arccos x$, $y = \arctan x$, $y = \operatorname{arccot} x$, and $y = a^x$.

READER: I'll be looking forward to the next dialogue. By the way, you wanted to make one more remark about the derivative $f'(x)$.

AUTHOR: Note 2 concerns the natural domain of a derivative. Let a set D be the domain of $f(x)$. The question is whether D is also the domain of $f'(x)$.

READER: In any case, the domain of $f'(x)$ cannot be wider than the do-

main of $f(x)$ because in order to find $f'(x)$ we use $f(x)$.

AUTHOR: A carefully balanced answer, to be sure. The domain of $f'(x)$ is in the general case a subset of D. It is obtained from D as a result of elimination of those points x for which $\lim\limits_{\Delta x \to 0} \dfrac{\Delta f(x)}{\Delta x}$ does not exist. By the way, this subset is called the *domain of differentiability* of $f(x)$.

READER: What are the *conditions of differentiability* of $f(x)$ at any specific point x?

AUTHOR: Obviously, these conditions are identical to those of the existence of $\lim\limits_{\Delta x \to 0} \dfrac{\Delta f(x)}{\Delta x}$ at point x. We have already observed that it is the limit of the type $\dfrac{\infty}{\infty}$, which necessitates that both the numerator and denominator tend to zero. It means that $f(x)$ must be continuous at x. The following theorem could be proved rigorously.

Theorem:

The continuity of a function $f(x)$ at a point x is a necessary condition for the existence of $f'(x)$ at x.

However, we shall not give the proof of this theorem here. The simple qualitative arguments given above will suffice.

READER: I wonder whether the continuity of a function is also a *sufficient* condition for its differentiability.

AUTHOR: No, it is not. Consider, for example, the function $y = |\log x|$. It is sufficient only to look at its graph (Figure 40) to conclude that at $x = 1$ the tangent to the graph of the function is, strictly speaking, nonexistent (on approaching $x = 1$ from the left we have one tangent, viz., the straight line AA, while on approaching $x = 1$ from the right we have another

tangent, viz., the straight line BB). It means that $y = |\log x|$ does not have a derivative at $x = 1$, although the function is continuous at this point.

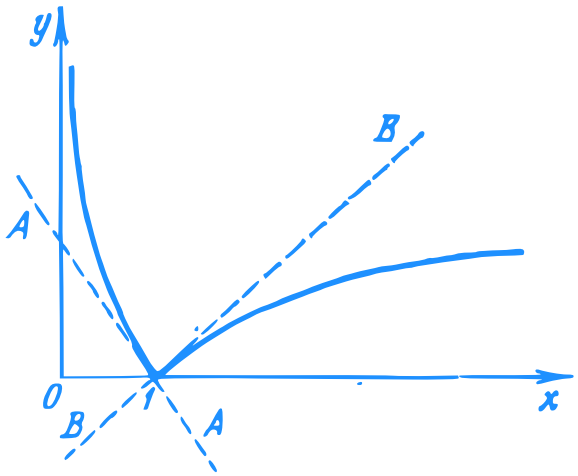


Figure 40: Examining the derivative of the function $y = |\log x|$.

In conclusion, let us turn to one interesting property of differentiable functions.

Let $f(x)$ be a differentiable function, and its increment Δf at x be related to the increment Δx of the argument as follows:

$$\Delta f = f'(x)dx + \eta(\Delta x)\Delta x \tag{42}$$

where $\eta(\Delta x)$is a function of Δx.

By dividing both parts of (42) by Δx, we obtain

$$\frac{\Delta f}{\Delta x} = f'(x) + \eta(\Delta x)$$

Passing to a limit in both sides of the last equation for Δx gives $\lim\limits_{\Delta x \to 0} \eta(\Delta x) = 0$.

Consequently, $\eta(\Delta x)$ is an *infinitesimal* (we use the same terminology as for numerical sequences, see Dialogue Three).

Conclusion:

An increment Δf of at a point x of a function $f(x)$ differentiable at this point can be represented by two summands, namely, a summand proportional to the increment Δx of the argument (this summand is $f'(x)\Delta x$) and a summand negligible in comparison with the first for sufficiently small Δx (this summand is $\eta(\Delta x)\Delta x$, where $\eta(\Delta x)$ is infinitesimal).

READER: It seems that this is a formulation of the property of "linearity on a small scale" that you mentioned in the previous dialogue (see Figure 32).

AUTHOR: Quite true. The main part of the increment of a differentiable function (a summand linear with respect to Δx) is called the *differential* of the function.

Dialogue 10

DIFFERENTIATION

AUTHOR: Now our aim is a practical realization of the programme outlined in the previous dialogue. This dialogue could be considered as a drill on the calculation of derivatives. We shall divide the talk into three parts.

1. Differentiation rules.
2. Differentiation of elementary functions $y = x^2$, $y = \sin x$, and $y = \log_a x$.
3. Application of differentiation rules to different functions.

Before the start I would like to remind you that the differentiation of a function $f(x)$ is defined as the operation of obtaining $f'(x)$ from $f(x)$. This operation is performed by using the operator of differentiation $\dfrac{d}{dx}$

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

The Differentiation Rules

AUTHOR: **Rule One.** We shall prove the following theorem:

Theorem

The derivative of the sum of two functions equals the sum of their derivatives provided that they exist, i.e.

$$\frac{d}{dx}[f(x)+g(x)]=\frac{d}{dx}f(x)+\frac{d}{dx}g(x) \tag{43}$$

Denote $f(x)+g(x)=u(x)$. Then the theorem can be written as follows: $u'(x)=f'(x)+g'(x)$. Try to prove this theorem.

READER: First I write

$$\begin{aligned}
f'(x)&=\lim_{\Delta x\to 0}\frac{f(x+\Delta x)-f(x)}{\Delta x}\\
g'(x)&=\lim_{\Delta x\to 0}\frac{g(x+\Delta x)-g(x)}{\Delta x}\\
u'(x)&=\lim_{\Delta x\to 0}\frac{u(x+\Delta x)-u(x)}{\Delta x}\\
&=\lim_{\Delta x\to 0}\frac{f(x+\Delta x)+g(x+\Delta x)-f(x)-g(x)}{\Delta x}
\end{aligned}$$

But I don't know what to do next.

AUTHOR: We shall repeat your writing but drop the limit signs:

$$\begin{aligned}
\frac{\Delta u(x)}{\Delta x}&=\frac{f(x+\Delta x)+g(x+\Delta x)-f(x)-g(x)}{\Delta x}\\
&=\frac{f(x+\Delta x)-f(x)}{\Delta x}+\frac{g(x+\Delta x)-g(x)}{\Delta x}\\
&=\frac{\Delta f(x)}{\Delta x}+\frac{\Delta g(x)}{\Delta x}
\end{aligned}$$

This gives

$$\frac{\Delta u(x)}{\Delta x} = \frac{\Delta f(x)}{\Delta x} + \frac{\Delta g(x)}{\Delta x}$$

READER: I see. Next we use the well-known theorem on the limit of the sum of functions, and the proof is complete.

AUTHOR: Quite correct. **Rule Two.** Let us prove the next theorem:

Theorem

A constant multiplier is factored out of the derivative, that is

$$\frac{d}{dx}[a f(x)] = a \frac{d}{dx} f(x) \tag{44}$$

The theorem is immediately proved if we use the following obvious equality

$$\frac{\Delta [a f(x)]}{\Delta x} = a \frac{\Delta f(x)}{\Delta x}$$

Rule Three. Now we shall consider the theorem on the derivative of the product of two functions.

Theorem

The derivative of a function $u(x) = f(x)\, g(x)$ is calculated by using the following formula:

$$u'(x) = f'(x)\, g(x) + f(x)\, g'(x) \tag{45}$$

provided that the derivatives $f'(x)$ and $g'(x)$ exist.

Formula (45) is called the *Leibnitz formula*. Another expression for the same formula is:

$$\frac{d}{dx}(f\, g) = g \frac{d}{dx} f + f \frac{d}{dx} g$$

READER: Apparently, as in the proof of the first theorem, we must express $\frac{\Delta u(x)}{\Delta x}$ through $\frac{\Delta f(x)}{\Delta x}$ and $\frac{\Delta g(x)}{\Delta x}$. But how to do it?

AUTHOR: The simplest way is

$$\begin{aligned} u+\Delta u &= (f+\Delta f)(g+\Delta g) \\ &= fg+g\Delta f+f\Delta g+\Delta f\Delta g \end{aligned}$$

Hence,

$$\Delta u = g\Delta f+f\Delta g+\Delta f\Delta g$$

consequently,

$$\frac{\Delta u(x)}{\Delta x} = g(x)\frac{\Delta f(x)}{\Delta x}+f(x)\frac{\Delta g(x)}{\Delta x}+\frac{\Delta f(x)}{\Delta x}\Delta g(x)$$

Now we find the limit for $\Delta x \to 0$. Notice that neither $g(x)$ nor $f(x)$ depends on Δx, and $\Delta g(x)$ tends to zero. As a result,

$$\lim_{\Delta x\to 0}\frac{\Delta u(x)}{\Delta x} = \lim_{\Delta x\to 0} g(x)\frac{\Delta f(x)}{\Delta x}+\lim_{\Delta x\to 0} f(x)\frac{\Delta g(x)}{\Delta x}$$

The theorem is thus proved.

Rule Four. The next theorem is related to the derivative of the ratio of two functions.

Theorem

The derivative of a function $u(x)=\dfrac{f(x}{g(x)}$ is:

$$u'(x) = \frac{f'(x)g(x)-f(x)g'(x)}{g^2(x)} \tag{46}$$

provided that the derivatives $f'(x)$ and $g'(x)$ exist, and that $g(x)\neq 0$.

It can be written in a different form:

$$\frac{d}{dx}\left(\frac{f}{g}\right)=\frac{g\dfrac{d}{dx}f-f\dfrac{d}{dx}g}{g^2}$$

Try to prove this theorem.

READER: I shall proceed by analogy with the preceding proof. I can write Hence,

$$\begin{aligned}\Delta u&=\frac{f+\Delta f}{g+\Delta g}-\frac{f}{g}\\&=\frac{g\dfrac{\Delta f}{\Delta x}-f\dfrac{\Delta g}{\Delta x}}{g^2+g\Delta g}\end{aligned}$$

This yields,

$$\frac{\Delta u}{\Delta x}=\frac{g\dfrac{\Delta f}{\Delta x}-f\dfrac{\Delta g}{\Delta x}}{g^2+g\Delta g}$$

Passing then to the limit for $\Delta x\to 0$, I take into account that neither g nor f depend on Δx, and that Δg also tends to zero. Using the known theorems on the limit of the product and the sum of functions, we obtain

$$\begin{aligned}\lim_{\Delta x\to 0}\frac{\Delta u}{\Delta x}&=\lim_{\Delta x\to 0}\left(\frac{1}{g^2+g\Delta g}\right)\lim_{\Delta x\to 0}\left(g\frac{\Delta f}{\Delta x}-f\frac{\Delta g}{\Delta x}\right)\\&=\frac{1}{g^2}\left(g\lim_{\Delta x\to 0}\frac{\Delta f}{\Delta x}-f\lim_{\Delta x\to 0}\frac{\Delta g}{\Delta x}\right)\end{aligned}$$

This completes the proof.

AUTHOR: Very good. Now we shall discuss the, problem of the differentiation of a *composite function* (for composite functions, see Dialogue Five. Let $w=h(x)$ be a composite function, and $h(x)=g[f(x)]$. This

composite function is the composition of two functions $w = g(y)$ and $y = f(x)$.

I remind you that the derivative $f'(x)$ indicates how faster y charges compared to x and the derivative $g'(y)$ indicates how faster w changes compared to y. Consequently, the product $g'(y)f'(x)$ must indicate how faster w changes compared to x, i.e. it equals the derivative $h'(x)$.

Rule Five. Thus we arrive at the differentiation rule for composite functions.

Theorem

The derivative of a composite function $h(x) = g[f(x)]$ is:

$$h'(x) = g'(y)f'(x) \tag{47}$$

READER: We have arrived at this rule using very simple arguments. I wonder whether they can be regarded as a proof of the rule.

AUTHOR: No, of course not. Therefore I am going to give the proof of the differentiation rule for composite functions.

Let the independent variable x have an increment Δx such that $x + \Delta x$ belongs to the domain of $f(x)$. Then the variable y will have an increment $\Delta y = f(x+\Delta x) - f(x)$, while the variable w will have an increment $\Delta w = g(y + \Delta y) - g(y)$. Since the derivative $g'(y)$ exists. the increment Δw can be expressed as follows

$$\Delta w = g'(y)\Delta y + \eta \Delta y$$

where $\eta \to 0$ for $\Delta y \to 0$ (see expression (6) from the previous dialogue).

Dividing both sides of the equation by Δx, we obtain

$$\frac{\Delta w}{\Delta x} = g'(y)\frac{\Delta y}{\Delta x} + \eta\frac{\Delta y}{\Delta x}$$

Next we pass to the limit for $\Delta x \to 0$

$$\lim_{\Delta x\to 0}\frac{\Delta w}{\Delta x} = g'(y)\lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x} + \lim_{\Delta x\to 0}\left(\eta\frac{\Delta y}{\Delta x}\right)$$

Since $\lim\limits_{\Delta x\to 0}\frac{\Delta w}{\Delta x} = h'(x)$ and $\lim\limits_{\Delta x\to 0}\frac{\Delta y}{\Delta x} = f'(x)$, we have

$$h'(x) = g'(y)\,f'(x) + f'(x)\lim_{\Delta x\to 0}\eta$$

And since $\Delta y \to 0$ for $\Delta x \to 0$,

$$\lim_{\Delta x\to 0}\eta = \lim_{\Delta y\to 0}\eta = 0$$

Hence we arrive at (48), namely, at the rule for the differentiation of composite functions.

Rule Six. Finally, I shall give (without proof) the rule for the differentiation of inverse functions.

Theorem

If a derivative $y'(x)$ of an initial monotonic function $y(x)$ exists and is not equal to zero, the derivative of the inverse junction $x(y)$ is calculated by the formula:

$$x'(y) = \frac{1}{y'(x)} \qquad (48)$$

READER: It seems that this formula can be easily obtained if we make use of the geometrical interpretation of the derivative. Really, consider the graph of a monotonic function $y(x)$ (Figure 41); its derivative at point

x_0 is $\tan\alpha$. The same curve can, obviously, be regarded as the graph of the inverse function $x(y)$, with y considered as the independent variable instead of x, and x considered as the dependent variable instead of y. But the derivative of the inverse function at point y_0 is $\tan\beta$ (see the figure). Since $\alpha+\beta=\dfrac{\pi}{2}$ we have

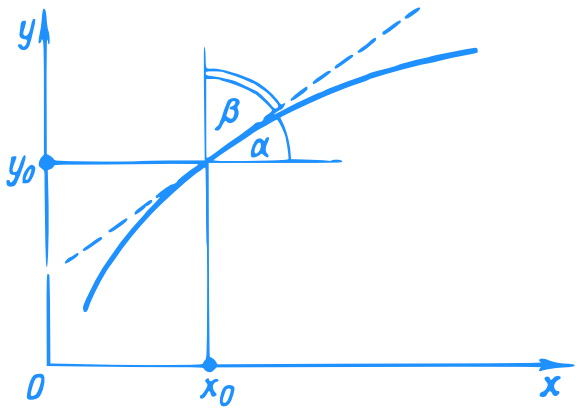


Figure 41: Derivative of a function and its inverse.

$$\tan\beta=\frac{1}{\tan\alpha}$$

This gives the above-cited differentiation rule for inverse functions.

AUTHOR: I must admit that although your line of reasoning is not a rigorous mathematical proof, it is an example of an effective application of geometrical concepts.

The Differentiation of Functions

$y = x^2$, $y = \sin x$, and $y = \log_a x$

AUTHOR: Using (40) from the previous dialogue, calculate the derivatives of the three indicated functions. Start with $y = x^2$. Go ahead.

READER: I write

$$y + \Delta y = (x + \Delta x)^2 = x^2 + 2x\Delta x + \Delta x^2$$

Hence,

$$\Delta y = 2x\Delta x + \Delta x^2$$

Consequently,

$$\frac{\Delta y}{\Delta x} = 2x + \Delta x$$

Further we pass to the limit for $\Delta x \to 0$

$$\lim_{\Delta x \to 0} \frac{\Delta y(x)}{\Delta x} = 2x$$

Therefore,

$$y'(x) = 2x$$

AUTHOR: You have thus obtained the result of applying the operator $\frac{d}{dx}$ to the function $y = x^2$:

$$\boxed{\frac{d}{dx} x^2 = 2x} \tag{49}$$

We observe that for a quadratic function $y = x^2$ at the input of the operator $\frac{d}{dx}$ we obtain a linear function $y = 2x$ at the output.

Now try to differentiate the function $y = \sin x$.

READER: I shall write

$$y + \Delta y = \sin(x + \Delta x)^2 = \sin x \cos \Delta x + \cos x \sin \Delta x$$

Hence,

$$\Delta y = \sin x \cos \Delta x + \cos x \sin \Delta x - \sin x$$

AUTHOR: You had better use here the formula for the difference between two sines, not the formula for the sine of the sum. Represent Δy in the form

$$\Delta y = \sin(x + \Delta x) - \sin x = 2 \sin \frac{\Delta x}{2} \cos\left(x + \frac{\Delta x}{2}\right)$$

Next we obtain

$$\frac{\Delta x}{\Delta y} = \frac{\sin \frac{\Delta x}{2}}{\frac{\Delta x}{2}} \cos\left(x + \frac{\Delta x}{2}\right)$$

In taking the limit for $\Delta x \to 0$, recall a result obtained in Dialogue Seven:

$$\lim_{\Delta x \to 0} \frac{\sin \Delta x}{\Delta x} = 1$$

READER: Yes, I see. Therefore,

$$\begin{aligned} \lim_{\Delta x \to 0} \frac{\Delta x}{\Delta y} &= \lim_{\Delta x \to 0} \frac{\sin \frac{\Delta x}{2}}{\frac{\Delta x}{2}} \lim_{\Delta x \to 0} \cos\left(x + \frac{\Delta x}{2}\right) \\ &= \lim_{\Delta x \to 0} \cos\left(x + \frac{\Delta x}{2}\right) \\ &= \cos x \end{aligned}$$

AUTHOR: The operator $\dfrac{d}{dx}$ applied to the function $y = \sin x$ thus generates the function $y = \cos x$:

$$\boxed{\frac{d}{dx} \sin x = \cos x} \tag{50}$$

Now we have to differentiate the function $y = \log_a x$. This time, however, we should start with a discussion of the transcendental number e (which is usually called the "base of natural or Napierian logarithms"). The number e may be defined as the limit of a numerical sequence

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n \tag{51}$$

The approximate value of e is: $e = 2.718\,281\,828\,459\,0\ldots$ Using (51), we can show that e is also the limit of $y = (1+x)^{\frac{1}{x}}$ for x tending to zero

$$\boxed{e = \lim_{x\to 0}(1+x)^{\frac{1}{x}}} \tag{52}$$

We shall omit the proof of (52).

READER: It seems that (52) follows logically from (51).

AUTHOR: Far from it. Don't forget that in (51) we deal with the limit of a numerical sequence, while in (52) with the limit of a function at a point. While n are integers, x belongs to the real line (with the exception of $x = 0$). Therefore, the transition from (51) to (52) requires a good deal of time and space.

Now turn to the differentiation of $y = \log_a x$. Follow the line adopted above.

Start.

READER: Obviously,

$$y + \Delta y = o\,g_a(x + \Delta x)$$

Hence,

$$\Delta y = \log_a(x + \Delta x) - \log_a x$$

$$= \log_a \frac{x+\Delta x}{\Delta x}$$

and, consequently,

$$\frac{\Delta x}{\Delta y} = \frac{1}{\Delta x} \log_a \frac{x+\Delta x}{x}$$

At this point I would have to find the limit for $\Delta x \to 0$.

AUTHOR: I shall give you a hand here. We can rewrite

$$\begin{aligned}
\frac{\Delta x}{\Delta y} &= \log_a \left(\frac{x+\Delta x}{x}\right)^{\frac{1}{\Delta x}} \\
&= \log_a \left(1+\frac{\Delta x}{x}\right)^{\frac{x}{\Delta x}\cdot\frac{1}{x}} \\
&= \frac{1}{x} \log_a \left(1+\frac{\Delta x}{x}\right)^{\frac{x}{\Delta x}}
\end{aligned}$$

READER: I see. This gives

$$\frac{\Delta x}{\Delta y} = \frac{1}{x} \log_a \left(1+\frac{\Delta x}{x}\right)^{\frac{x}{\Delta x}}$$

To find the limit for $\Delta x \to 0$, we use (52). As a result

$$\begin{aligned}
\lim_{\Delta x \to 0} \frac{\Delta x}{\Delta y} &= \frac{1}{x} \lim_{\Delta x \to 0} \log_a \left(\frac{x+\Delta x}{x}\right)^{\frac{x}{\Delta x}} \\
&= \frac{1}{x} \log_a e \\
&= \frac{1}{x} \cdot \frac{1}{\ln a}
\end{aligned}$$

(symbol ln is the standard notation for the *natural logarithm*).

AUTHOR: We have thus found that the operator $\frac{d}{dx}$ applied to the function $y = \log_a x$ gives $y = \frac{1}{x} \cdot \frac{1}{\ln a}$:

$$\boxed{\frac{d}{dx} \log_a x = \frac{1}{x} \cdot \frac{1}{\ln a}} \tag{53}$$

Notice that the natural domain of the function $y = \frac{1}{x} \cdot \frac{1}{\ln a}$ in (53) $]0, \infty[$.

We can sum up our conclusions now.

Using relation (40) from Dialogue Nine, first, we have established the six differentiation rules and, second, we have differentiated three functions. The results are summarized in Table on the next page, and Figure 42 graphically represents the result of the action of the operator $\frac{d}{dx}$ on the three selected functions. The left-hand column in the figure lists the graphs of the three functions $f(x)$, and the right-hand column shows the graphs of the corresponding derivatives $f'(x)$.

In what follows we shall not use formulas of type (40) from Dialogue Nine, that is, we shall not operate in terms of limit transitions. Using the results obtained above, we shall find the derivatives for a number of elementary functions without calculating the relevant limits.

The Differentiation Rules

Rule 1: the differentiation of the sum of functions	$\frac{d}{dx}(f+g)=\frac{d}{dx}f+\frac{d}{dx}g$
Rule 2: the differentiation of the function with constant term	$\frac{d}{dx}(af)=a\frac{d}{dx}f\ (a=\text{constant})$
Rule 3: the differentiation of the product of functions	$\frac{d}{dx}(fg)=f\frac{d}{dx}g+g\frac{d}{dx}f$
Rule 4: the differentiation of the ratio of functions	$\frac{d}{dx}\frac{f}{g}=\frac{g\frac{d}{dx}f-f\frac{d}{dx}g}{g^2}$
Rule 5: the differentiation of composite functions	$\frac{d}{dx}g[f(x)]=\left(\frac{d}{df}g(f)\right)\frac{d}{dx}f(x)$
Rule 6: the differentiation of inverse functions	$\frac{d}{dy}x(y)=\frac{1}{\frac{d}{dx}y(x)}$

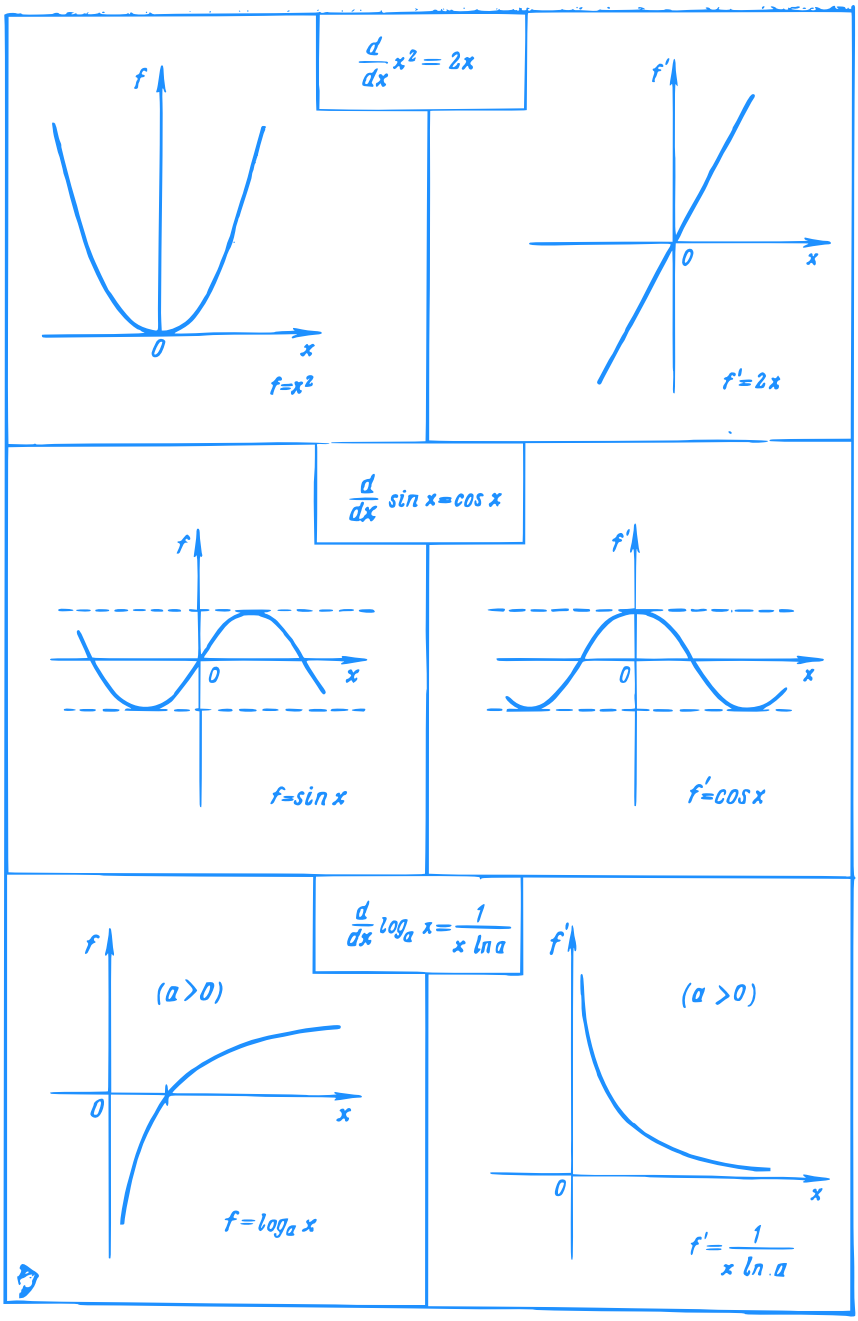


Figure 42: Graphs of some basic functions and their derivatives.

The Application of the Differentiation Rules to Different Functions

AUTHOR: As a first example, consider the function $y = x^n$. Prove that differentiation gives $y = nx^{n-1}$, that is,

$$\boxed{\frac{d}{dx}x^n = nx^{n-1}} \tag{54}$$

Prove this proposition by using the *method of mathematical induction*.

READER: For $n = 2$ formula (54) holds and yields (49). Assume now that (54) holds for $n = m$. We have to prove that it is also true for $n = m+1$. We write $x^{m+1} = x^m x$ and use the *Leibnitz formula* (**Rule Three**):

$$\frac{d}{dx}(x^m x) = x\frac{d}{dx}x^m + x^m\frac{d}{dx}x$$

Since $\dfrac{d}{dx}x = 1$ and according to the assumption $\dfrac{d}{dx}x^m = mx^{m-1}$, we obtain

$$\frac{d}{dx}x^{m+1} = mxx^{m-1} + x^m = (m+1)x^m$$

The proof is completed.

AUTHOR: The next example is the function $y = x^{-n}$. Differentiate this function using **Rule Four** and (54).

READER: This is simple. Applying **Rule Four**, we obtain

$$\frac{d}{dx}\left(\frac{1}{x^n}\right) = \frac{-\dfrac{d}{dx}x^n}{x^{2n}}$$

By virtue of (54),

$$\boxed{\frac{d}{dx}x^{-n} = -nx^{-(n+1)}} \tag{55}$$

AUTHOR: One particular result that follows from (57) is

$$\boxed{\frac{d}{dx}\left(\frac{1}{x}\right) = -\frac{1}{x^2}} \tag{56}$$

The next example is the function $y = \sqrt{x}$.

READER: Here I shall use **Rule Six** (the differentiation rule for inverse functions). The inverse function involved is $x = y^2$. Its derivative is given by (49). As a result,

$$\frac{d}{dx}\sqrt{x} = \frac{1}{\frac{d}{dx}y^2} = \frac{1}{2y} = \frac{1}{2\sqrt{x}}$$

Thus

$$\boxed{\frac{d}{dx}\sqrt{x} = \frac{1}{2\sqrt{x}}} \tag{57}$$

AUTHOR: Now we can pass to the *trigonometric* functions. Consider the function $y = \cos x$.

READER: I propose to use (58) and the identity $\sin^2 x + \cos^2 = 1$. By differentiating both sides of the identity and using **Rule One**, we obtain

$$\frac{d}{dx}\sin^2 x + \frac{d}{dx}\cos^2 x = 0$$

Next, by applying **Rule Five** (the differentiation rule for composite functions) in conjunction with (49), we find

$$2\sin x \frac{d}{dx}\sin x + 2\cos x \frac{d}{dx}\cos x = 0$$

From (58), $\frac{d}{dx}\sin x = \cos x$ so that

$$\frac{d}{dx}\cos x = -\sin x$$

AUTHOR: That is correct, although the result can be obtained in a simpler way. Better use the identity $\cos x = \sin\left(\frac{\pi}{2} - x\right)$. Further, applying **Rule Five**, we obtain

$$\frac{d}{dx}\sin\left(\frac{\pi}{2} - x\right) = \left(\frac{d}{dy}\sin y\right)\frac{d}{dx}\left(\frac{\pi}{2} - x\right)$$

(here $y = \frac{\pi}{2} - x$)

Making use of (58), we find

$$\frac{d}{dx}\sin\left(\frac{\pi}{2} - x\right) = -\frac{d}{dy}\sin y = -\cos y = -\sin x$$

Using now the suggested identity, we arrive at the final result:

$$\boxed{\frac{d}{dx}\sin x = -\cos x} \tag{58}$$

READER: The operation of differentiation thus "turns" the sine into the cosine and, vice versa, that is, the cosine in to the sine.

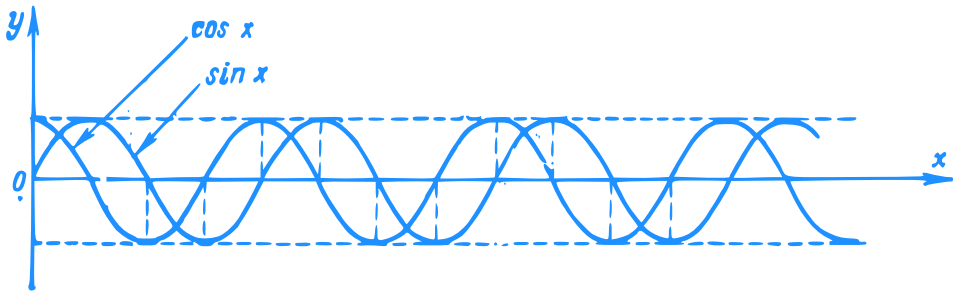


Figure 43: Graphs of the sine and cosine functions.

AUTHOR: . Yes, it does. But in the last case the sign changes too, that is, the cosine is transformed into the sine with a negative sign. If you plot the graphs of $\sin x$ and $\cos x$ in the same system of coordinates (Figure 43),

you will find that at points x where one of the functions reaches maximum or minimum (takes the value 1 or -1) the other function vanishes. It is readily apparent that this fact has direct relation to your remark. If, for example, at a certain point x the function $\sin x$ assumes its maximum value, the tangent to its graph at the same point will, obviously, be horizontal.

Consequently, the derivative of the function (i.e, $\cos x$) must vanish at this point. I recommend that you carefully analyze Figure 43. In particular, follow the correspondence between the slope of the tangent to the graph of the function drawn at different points and the sign of the derivative at the same points.

Now turn to the next example, the function $y = \tan x$. Difierentiate this function using the results of the differentiation of $\sin x$ and $\cos x$ and applying **Rule Four**.

READER: This will be easy:

$$\begin{aligned}\frac{d}{dx}\left(\frac{\sin x}{\cos x}\right) &= \frac{\cos x \dfrac{d}{dx}\sin x - \sin x \dfrac{d}{dx}\cos x}{\cos^2 x} \\ &= \frac{\cos^2 x + \sin^2 x}{\cos^2 x} \\ &= \frac{1}{\cos^2 x}\end{aligned}$$

Finally,

$$\boxed{\frac{d}{dx}\tan x = \frac{1}{\cos^2 x}} \tag{59}$$

AUTHOR: The result for $y = \cot x$ can be obtained similarly:

$$\boxed{\frac{d}{dx}\cot x = -\frac{1}{\sin^2 x}} \tag{60}$$

In order to differentiate $y = \arcsin x$, we use **Rule Six**

$$\begin{aligned}\frac{d}{dx}\arcsin x &= \frac{1}{\dfrac{d}{dy}\sin y}\\ &= \frac{1}{\cos y}\\ &= \frac{1}{\cos(\arcsin x)}\end{aligned}$$

Since

$$\cos(\arcsin x) = \sqrt{1-x^2}$$

we obtain

$$\boxed{\frac{d}{dx}\arcsin x = \frac{1}{\sqrt{1-x^2}}} \tag{61}$$

In order to differentiate $y = \arccos x$, it is sufficient to use (60) and the identity

$$\arcsin x + \arccos x = \frac{\pi}{2}$$

Therefore,

$$\boxed{\frac{d}{dx}\arccos x = -\frac{1}{\sqrt{1-x^2}}} \tag{62}$$

Using **Rule Six**, we differentiate the function $y = \arctan x$

$$\begin{aligned}\frac{d}{dx}\arctan x &= \frac{1}{\dfrac{d}{dy}\tan y}\\ &= \cos^2 y\\ &= [\cos(\arctan x)]^2\end{aligned}$$

Since

$$\cos(\arccos x) = \frac{1}{\sqrt{1+x^2}}$$

we obtain

$$\boxed{\frac{d}{dx}\arctan x = \frac{1}{\sqrt{1+x^2}}} \tag{63}$$

And, finally, the differentiation of $y = \operatorname{arccot} x$ is carried out by using the identity

$$\arctan x + \operatorname{arccot} x = \frac{\pi}{2}$$

and yields

$$\boxed{\frac{d}{dx}\operatorname{arccot} x = -\frac{1}{\sqrt{1+x^2}}} \tag{64}$$

We have thus performed the differentiation of all elementary trigonometric and inverse trigonometric functions,

In conclusion, let us examine the exponential function $y = a^x$. Using (53) and **Rule Six**, we obtain

This gives

$$\boxed{\frac{d}{dx}a^x = a^x \ln a} \tag{65}$$

Result (65) is very interesting. We see that the differentiation of the exponential function $y = a^x$ again yields the exponential function a^x multiplied by the constant term $\ln a$. In a particular case of $a = e$, we have $\ln e = 1$,and therefore

$$\boxed{\frac{d}{dx}e^x = e^x} \tag{66}$$

The exponential function $y = e^x$ is simply called the *exponential curve*. From (66) it follows that differentiation transforms this function into itself.

Dialogue 11

ANTIDERIVATIVE

READER: Differentiation is an operation of finding a function $f'(x)$ for a given function $f(x)$. Presumably, an inverse operation is possible as well, isn't it?

AUTHOR: An inverse operation indeed exists. It is called integration. Integration of a function $f(x)$ is an operation by which the so-called antiderivative is found for the given function $f(x)$.

Definition

An antiderivative is defined as a function $F(x)$ whose derivative equals an initial function $f(x)$:

$$f(x) = \frac{d}{dx}F(x) \tag{67}$$

READER: Quite clear. In the preceding dialogue we were seeking a derivative $f'(x)$ for a given function $f(x)$, and now we deal with a situation in which the given function $f(x)$ is the derivative of a yet unknown function

$F(x)$.

AUTHOR: Absolutely right. Take, for example, a function $f(x) = 2x^2 - 3x$. The differentiation of this function gives its *derivative*

$$f'(x) = 4x - 3$$

and its integration gives the antiderivative

$$F(x) = \frac{2}{3}x^3 - \frac{3}{2}x^2$$

READER: But how did you find this antiderivative?

AUTHOR: This was simple. I resorted to the well-known rules of differentiation but in a *reverse* order. In other words, I mentally searched for a function that would yield our function $f(x) = 2x^2 - 3x$ after differentiation. You can easily verify that

$$F'(x) = \frac{2}{3}3x^2 - \frac{3}{2}2x = 2x^2 - 3x$$

READER: But then why not take as this antiderivative, for example, a function $F(x) = \frac{2}{3}x^3 - \frac{3}{2}x^2 + 2$. It will again yield $F'(x) = 2x^2 - 3x$.

AUTHOR: You noticed a very important feature. Indeed, an antiderivative found for a given function is not unique. If $F(x)$ is an antiderivative (for a function f), then any function $F(x) + C$, where C is an arbitrary constant, is also an antiderivative for the initial function because

$$\frac{d}{dx}[F(x) + C] = \frac{d}{dx}F(x) + \frac{d}{dx}C = \frac{d}{dx}F(x)$$

READER: This means, therefore, that each given function $f(x)$ corresponds to a *family* of antiderivatives, $F(x) + C$ doesn't it?

AUTHOR: Precisely. Take a graph of one of the anti-derivatives. By translating it along the y-axis, you will obtain a family of the curves of antiderivatives for a given function f. For example, let $f(x) = \sin x$, The curves of antiderivatives for this function are plotted in Figure 44. These curves plot functions

$$F(x) = -\cos x + C$$

(the dash curve is the graph of the function $f(x) = \sin x$) . The constants C were taken with an increment of 0.5. By reducing this increment, one can obviously obtain a pattern of arbitrarily high density of $F(x)$ curves.

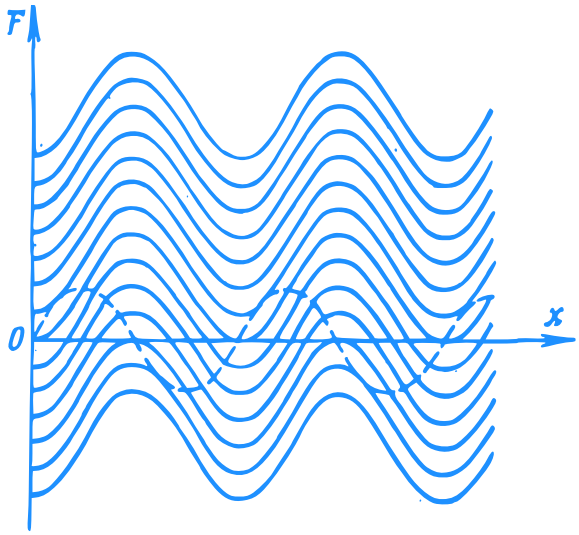


Figure 44: Graphs of the anti-derivatives of the function $f(x) = \sin x$.

The figure clearly shows that all the antiderivatives belong to one family (in other words, correspond to the same initial function f). This may not always be as clear if the function is represented in an analytical form. Take, for example, functions $F_1 = -\cos x$ and $F_2 = 3 - 2\cos^2 \frac{x}{2}$. It would be difficult to say at the first glance that these two functions are the antiderivatives of *one and the same function* (namely, $f = \sin x$). However,

since $2\cos^2\frac{x}{2} = 1 + cos\, x$, we find

$$F_2(x) = 3 - 1 - \cos x = -\cos x + 2$$

READER: I guess it would be possible to find directly that $F_1'(x) = F_2'(x)$, wouldn't it?

AUTHOR: Of course, it would:

$$\frac{d}{dx}F_2(x) = -2\frac{d}{dx}\cos^2\frac{x}{2} = 4\cos\frac{x}{2}\sin\frac{x}{2}\cdot\frac{1}{2} = \sin x$$

$$\frac{d}{dx}F_1(x) = \frac{d}{dx}(-\cos x) = \sin x$$

But the easiest way is to notice that $F_2 - F_1 = C$.

We could find numerous such examples. For instance, it is not difficult to check that the following pairs of functions belong to the same family of antiderivatives (each pair to its own family):

(a) $F_1 = x^2 - 2x + 3$, $F_2 = (x-1)^2$

(b) $F_1 = \arcsin x$, $F_2 = 1 - \arccos x$

(c) $F_1 = \tan x\, \sin x + \cos x$, $F_2 = (2\cos x + 1)\dfrac{1}{\cos x}$

Thus, in case (a) we find $F_2 - F_1 = -2$; both functions are the antiderivatives of the function $f = 2x - 2$.

Please, check cases (b) and (c) yourself.

READER: In case (b) $F_2 - F_1 = (1 - \arccos x + + \arcsin x) = 1 - \dfrac{\pi}{2}$ both functions are the antiderivatives of the function $f = \dfrac{1}{1-x^2}$

Case (c) is more intricate. Some preliminary manipulations are necessary:

$$F_1(x) = \tan x \sin x + \cos x = \frac{\sin^2 x + \cos^2 x}{\cos x} = \frac{1}{\cos x}$$

$$F_2(x) = \frac{2\cos x + 1}{\cos x} = 2 + \frac{1}{\cos x}$$

Therefore, $F_2 - F_1 = 2$. Both functions (F_1 and F_2) are the antiderivatives of $f = \dfrac{\sin x}{\cos^2 x}$

AUTHOR: Correct. Now, taking into account the results obtained in the previous dialogue, we can compile a table (see Table 2) which gives various functions $f(x)$ in the first column the corresponding derivatives $f'(x)$ in the second column, and the antiderivatives $F(x) + C$, corresponding to the functions $f(x)$, in the third column. I want to stress once more: the transformation $f(x) \to f'(x)$ is the operation of differentiation of the function $f(x)$, and the transformation $f(x) \to [F(x)+C]$ is the operation of integration of the function $f(x)$.

A List of Derivatives and Antiderivatives for Selected Functions

	Function $f(x)$	Derivative $f'(x)$	Antiderivative F(x)
(1)	a	0	$ax+C$
(2)	x^n	nx^{n-1}	$\frac{1}{n+1}x^{n+1}+C$
(3)	e^x	e^x	e^x+C
(4)	$\frac{1}{x}$	$-\frac{1}{x^2}$	$\ln x+C$
(5)	$\sqrt{x}$	$\frac{1}{2\sqrt{x}}$	$\frac{2}{3}x\sqrt{x}+C$
(6)	$\sin x$	$\cos x$	$-\cos x+C$
(7)	$\cos x$	$-\sin x$	$\sin x+C$
(8)	$\frac{1}{\cos^2}x$	$2\frac{\sin x}{\cos^3 x}$	$-\cot x+C$
(9)	$\frac{1}{\sin^2 x}$	$-2\frac{\cos x}{\sin^3 x}$	$\tan x+C$
(10)	$\frac{1}{1+x^2}$	$-\frac{2x}{(1+x^2)^2}$	$\arctan x+C$

READER: Examples (8), (9), and (10) in Table 2 give an impression that the transformation $f(x) \to f'(x)$ is more complicated than the transformation $f(x) \to [F(x) + C]$.

AUTHOR: This impression stems from a special selection of the functions $f(x)$. Thus, it is easier to differentiate the function $\tan x$ than the function $\dfrac{1}{\cos^2 x}$. Indeed, in the latter case we have to use the rules for differentiation of a ratio of two functions or of a composite function.

In general, it should be noted that the operation of integration is substantially more complicated than that of differentiation. The differentiation of elementary functions invariably gives elementary functions. By employing the rules for differentiation discussed in the previous dialogue, you will be able (and with no difficulties, as a rule) to differentiate practically any elementary function. But integration is quite a different proposition. The rules for the integration of elementary functions comprise numerous techniques, and we would need several special dialogues to scan them. But the main point is that not every elementary function has an elementary function for its antiderivative. As one example, I shall mention the antiderivatives of such elementary functions as $\dfrac{1}{\log x}$ or $\dfrac{1}{\sqrt{1+x^2}}$. As a rule, in such cases one is forced to resort to the methods of the so-called *numerical integration*.

READER: I was very attentive and want to pose two questions. First: What is meant by the term *elementary function*?

AUTHOR: In Dialogue Nine I gave examples of the so-called *fundamental elementary functions* (x^n, x^{-n}, $x^{1/n}$, $\sin x$, $\cos x$, $\tan x$, $\cot x$, $\arcsin x$, $\arccos x$, $\arctan x$, $\operatorname{arccot} x$, a^x, $\log_a x$]. An *elementary function* is any function which can be formed of *fundamental elementary functions* by a *finite*

number of the operations of addition, subtraction, multiplication, division, involution, evolution, and taking a modulus, as well as by using the rules for obtaining inverse and composite functions. All the functions used in the previous dialogues are elementary (with an exception of the Dirichlet function mentioned in Dialogue Five), and many of them are fundamental elementary functions.

READER: My second question concerns the *rules for integration* you refer to. Could you give at least some examples?

AUTHOR: I shall quote three simplest rules.

Three Simple Rules for Integration

1. If F is an antiderivative for f. and G is an anti-derivative for g, an antiderivative for the sum of the functions $f+g$ is a function $F+G$.
2. If F is an antiderivative for f an antiderivative for a function af, where a is a constant, is a function aF.
3. If $F(x)$ is an antiderivative for $f(x)$, and a and b are constants, an antiderivative for a function $f(ax+b)$ is a function $\frac{1}{a}F(ax+b)$.

All the three rules are proved readily by using the rules for differentiation (in the third rule one has to apply the rule for the differentiation of composite functions). Indeed,

1. $\frac{d}{dx}(F+G)=\frac{d}{dx}F+\frac{d}{dx}G=f+g$

$$\textcircled{2}\quad \frac{d}{dx}(aF)=a\frac{d}{dx}F=af$$

$$\textcircled{3}\quad \frac{d}{dx}\left(\frac{1}{a}F(ax+b)\right)=\frac{1}{a}\frac{d}{dx}F(ax+b)=\frac{1}{a}\frac{d}{dy}F(y)\frac{d}{dx}(ax+b)$$
$$=\frac{1}{a}f(y)a=f(y)=f(ax+b)$$
$$\text{here: } y=ax+b$$

Of course, the three rules cited above do not exhaust a rich collection of integration rules available in calculus.

But here these three rules will be sufficient since our goal is quite modest: to give the fundamental idea of an anti-derivative.

READER: Our discussion of a derivative covered its geometrical interpretation as well. Is there a geometrical interpretation of an antiderivative?

AUTHOR: Yes, there is. Let us find it (besides, we shall need it later).

Consider a function $f(x)$. For the sake of simplicity, assume that this function is *monotonic* (and even *increasing*). Later we shall drop the monotonicity of a function. The most important is that the function be *continuous* over the chosen interval (i.e. over the interval on which it is defined). Figure 45 shows a shaded area (the so-called *curvilinear trapezoid*) bounded by the graph of the function $f(x)$, the interval $[a, x]$ of the x-axis, and two perpendiculars erected from points a and x on the x-axis. Let point a be fixed; as for point x (the right-hand end of the interval $[a, x]$), it is not fixed and can assume values from a upward (within the domain of definition of the function). Obviously, the *area* of the curvilinear trapezoid shaded in the figure is a *function* of x. We shall denote it by $S(x)$,

Now turn to Figure 46. Let us give an increment Δx to the independent

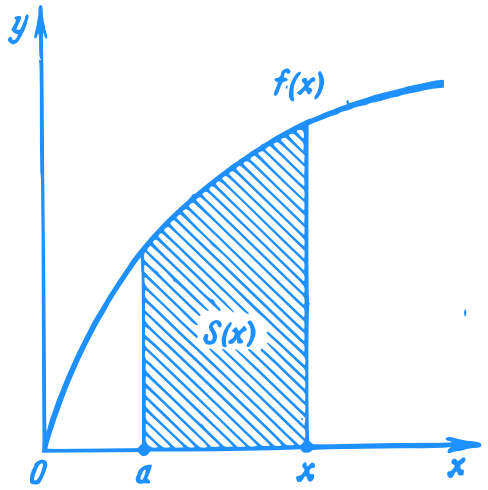


Figure 45: Examining the geometrical interpretation of the antiderivative.

variable x. The interval $[a, x+\Delta x]$ corresponds to the area $S(x+\Delta x)$. Denote $\Delta S(x) = S(x+\Delta x) - S(x)$. The increment $\Delta S(x)$ is, obviously, the area of the shaded curvilinear trapezoid. The figure shows that

$$\text{area } ADEF < \Delta S(x) < \text{area } ABCF$$

But the area $ADEF$ is equal to $f(x)\Delta x$, and the area $ABCF$ is equal to $f(x+\Delta x)\Delta x$. Therefore,

$$\text{area } ADEF < \Delta S(x) < \text{area } ABCF$$

or

$$f(x)\Delta x < \frac{\Delta S(x)}{\Delta x} < f(x+\Delta x)$$

or

$$0 < \left(\frac{\Delta S(x)}{\Delta x} - f(x)\right) < [f(x)+\Delta x) - f(x)] = \Delta f(x)$$

Now we find the limiting values of these inequalities for Δx tending to zero. By virtue of the continuity of the function $f(x)$ we conclude that

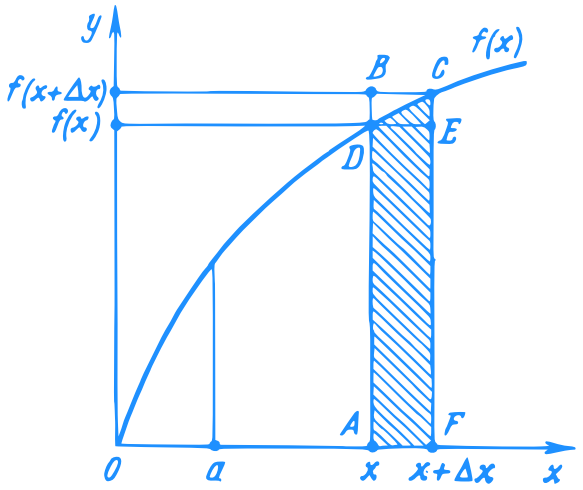


Figure 46: Examining the geometrical interpretation of the antiderivative.

$\lim\limits_{\Delta x \to 0} \Delta f(x) = 0$. Consequently,

$$\lim_{\Delta x \to 0} \left(\frac{\Delta S(x)}{\Delta x} - f(x) \right) = 0$$

As the function $f(x)$ is independent of Δx, the last relation yields

$$\lim_{\Delta x \to 0} \frac{\Delta S(x)}{\Delta x} = f(x) \tag{68}$$

By the definition of derivative,

$$\lim_{\Delta x \to 0} \frac{\Delta S(x)}{\Delta x} = S'(x)$$

Consequently, relation (68) signifies that

$$f(x) = S'(x) \tag{69}$$

Thus, in terms of geometry, *the antiderivative of the function f, taken at point x, is the area of curvilinear trapezoid bounded by the graph of the function $f(x)$ over the interval $[a, x]$ of the x-axis.*

READER: Presumably, it is one of the possible anti-derivatives, isn't it?

AUTHOR: Definitely.

READER: But it is evident that the area $S(x)$ also depends on the choice of point a.

AUTHOR: Absolutely correct. By choosing different points a, we shall have different areas of curvilinear trapezoids and, correspondingly, different antiderivatives. But all of them will be the antiderivatives of the function f taken at point x. It is only important that in all cases $a < x$.

READER: Then why is it that point a vanishes from the final results?

AUTHOR: Your bewilderment is understandable. Let us reformulate the results obtained above. Let $F(x)$ be an antiderivative of a function $f(x)$ taken at point x. According to (69), we can write

$$S(x) = F(x) + C$$

(here we have used the following theorem: if two functions have equal derivatives, the functions will differ by a constant term). The constant C is found readily since $S(a) = 0$. Therefore,

$$S(a) = F(a) + C = 0$$

Hence, $C = -F(a)$. This gives

$$S(x) = F(x) - F(a) \tag{70}$$

Conclusion

If $F(x)$ is an antiderivative of a function $f(x)$, then the area $S(x)$ of a curvilinear trapezoid bounded by the graph of the junction $f(x)$

over the interval $[a, x]$ is given by the difference $F(x) - F(a)$.

You see now that point a is introduced explicitly.

READER: Now everything is clear,

AUTHOR: Relation (69) (and from it, (70)) can be obtained for every continuous function; the *monotonicity* of a function is *not* a necessary condition. Consider a function $f(x)$ whose graph is plotted in Figure 47. We choose a point x and wish to prove that for any $\varepsilon > 0$ there is $\delta > 0$ such that

$$\left| \frac{\Delta S(x)}{\Delta x} - f(x) \right| < \varepsilon \tag{71}$$

for all Δx satisfying the condition $|\Delta x| < \delta$.

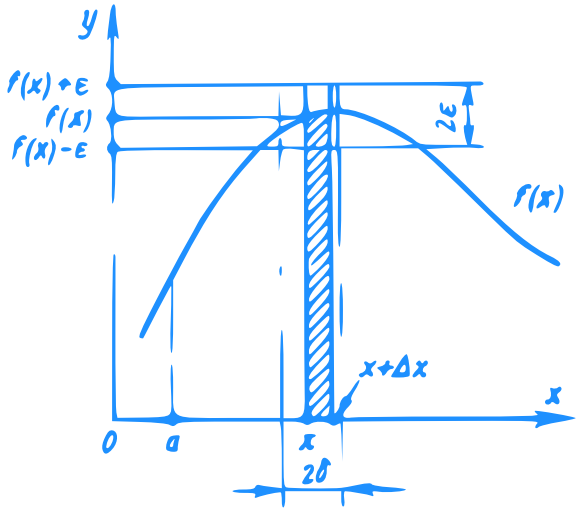


Figure 47: Examining the geometrical interpretation of the antiderivative.

READER: Shall we consider point x as fixed?

AUTHOR: Yes. Increments Δx and, correspondingly, $\Delta S(x)$, are always

considered for a definite point x. So we take an arbitrary number $\varepsilon > 0$ (shown in the figure). As f (x) is a continuous function, there is a number $\delta > 0$ such that

$$|f(x)+\Delta x-f(x)|<\varepsilon \tag{72}$$

for all Δx satisfying the condition $|\Delta x| < 0$. This number δ is the one we were to find.

Indeed, let us choose, for definiteness, that $\Delta x > 0$ but specify that $\Delta x < \delta$. The area of the curvilinear trapezoid shaded in Figure 47 will be denoted by $\Delta S(x)$ (this trapezoid is bounded by the graph of the function $f(x)$ over the interval $[x, x+\Delta x]$. Inequality (72) yields (see the figure):

$$[f(x)-\varepsilon]\Delta x<\Delta S(x)<[f(x)-\varepsilon]\Delta x$$

or

$$[f(x)-\varepsilon]<\frac{\Delta S(x)}{\Delta x}<[f(x)-\varepsilon]$$

or,

$$-\varepsilon<\left(\frac{\Delta S(x)}{\Delta x}-f(x)\right)<\varepsilon$$

or, finally.

$$\left|\frac{\Delta S(x)}{\Delta x}-f(x)\right|<\varepsilon$$

which is what we wanted to prove.

You see that a function f needn't be monotonic: relation (69) (and with it, (70) is easily generalized to the case of an arbitrary continuous function f.

Now let us turn again to Figure 44 that gives a family of graphs of the antiderivative $F(x) = -\cos x + C$ for the function $f(x) = \sin x$. Indicate which of these graphs (which antiderivative) stands for $S(x)$ in each of the following three cases: (a) $a = 0$, (b) $a = \frac{\pi}{2}$, and (c) $a = \pi$.

READER: The question is clear. I denote the sought functions by $S_1(x)$, $S_2(x)$, and $S_3(x)$, respectively. These functions are plotted in Figure 48.

Obviously, we can write

$$S_1(x) = F(x) - F(0)$$
$$S_2(x) = F(x) - F(\frac{\pi}{2})$$
$$S_3(x) = F(x) - F(\pi)$$

AUTHOR: Correct. It is important to underline that in each of the above three equalities the function $F(x)$ is a function chosen arbitrarily from the family of antiderivatives of f, shown in Figure 44.

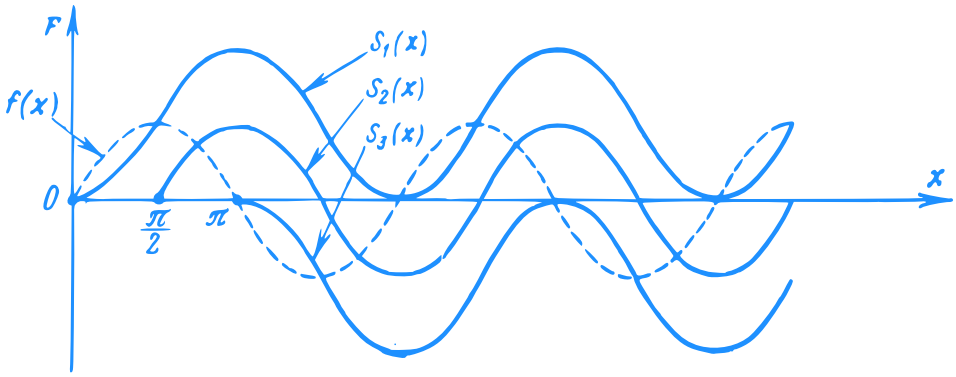


Figure 48: Particular graphs of the antiderivative for specific value of a.

READER: It looks as if whatever the selected antiderivative of the function f is, the difference between its values at two points depends only on the choice of these points but not on the choice of a specific antiderivative.

AUTHOR: You have pointed out a property of principal significance. It is so important that deserves a special dialogue.

Dialogue 12

INTEGRAL

AUTHOR: We know already that the difference between the values of an antiderivative at two arbitrary points depends only on the choice of these points (and, evidently, on the type of the initial function $f(x)$). As these two points we choose points a and b, that is, consider an increment of an antiderivative, $F(b)-F(a)$, This increment plays a very important role among the tools of calculus; it is called the *integral*.

Definition

The increment of an anti-derivative F of a function f, i,e. $F(b)-F(a)$, is said to be the integral of f from a to b.

The notation of the integral is:

$$\int_a^b f(x)\,dx$$

(it reads: "integral of f of x, dx, from a to b"). The numbers a and b are the *lower* and *upper limits of integration*. The function f is said to be *integrand*, and x the *integration variable*.

Consequently, if F is one of the antiderivatives of the function f, then the definition of an integral states that

$$\int_a^b f(x)\,dx = F(b) - F(a) \tag{73}$$

Formula (73) is known in the literature on mathematics as the *Newton-Leibnitz formula*. Remember that F here is an arbitrary antiderivative of the function f.

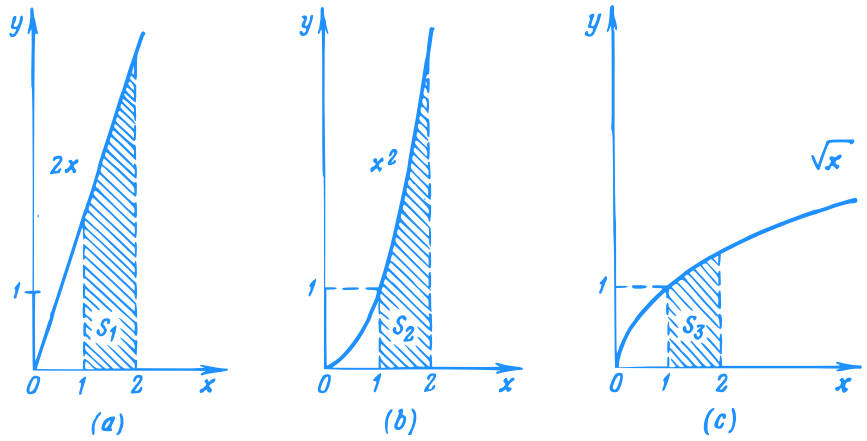


Figure 49: Examining the geometrical interpretation of the integral as area.

Reader: As far as I understand, the integral of the function f from a to b is precisely the area of the curvilinear trapezoid bounded by the graph of the function $f(x)$ over the interval $[a, b]$. Is that right?

Author: Absolutely. The expression

$$\int_a^b f(x)\,dx$$

is nothing less than the area of this geometrical figure. Figure 49 shows three cases plotting different integrands: (a) $f(x) = 2x$, (b) $f(x) = x^2$, (c) $f(x) = \sqrt{x}$.

The limits of integration are chosen identical in all the three cases: $a = 1$, $b = 2$. The corresponding areas of the curvilinear trapezoids are shaded in the figure:

$$S_1 = \int_1^2 2x\,dx$$
$$S_2 = \int_1^2 x^2\,dx$$
$$S_3 = \int_1^2 \sqrt{x}\,dx$$

The numbers S_1, S_2, and S_3 are different because the integrands $f(x)$ are different.

We thus find that the expression

$$\int_a^b f(x)\,dx$$

works as a *functional* (recall Dialogue Four). You "input" in it a function f, and it "outputs" a number S.

By the way, you can easily find how this functional works. To achieve this, use formula (73) and take into account that the antiderivative of the function $f(x) = 2x$ is $F(x) = x^2 + C$, that of $f(x) = x^2$ is $F(x) = \frac{1}{3}x^3 + C$, and that of $f(x) = \sqrt{x}$ is $F(x) = \frac{2}{3}x\sqrt{x} + C$.

The standard notation is: $F(b) - F(a) = F(x)\big|_a^b$. Therefore,

$$\int_1^2 2x\,dx = x^2\big|_1^2 = 4 - 1 = 3$$
$$\int_1^2 x^2\,dx = \frac{1}{3}x^3\big|_1^2 = \frac{1}{3}(8 - 1) = \frac{7}{3}$$
$$\int_1^2 \sqrt{x}\,dx = \frac{2}{3}x\sqrt{x}\big|_1^2 = \frac{2}{3}(2\sqrt{2} - 1)$$

With the function $2x$ at the "input" of the functional $\int_1^2 f(x)\,dx$, we obtain at the "output" the number 3; with x^2 at the "input", we obtain at the "output" the number $\frac{7}{3}$; and with $\sqrt{x}$ at the "input", we obtain at the "output" the number $\frac{2}{3}(2\sqrt{2}-1)$.

READER: I see that we can rather easily find the areas of various curvilinear trapezoids!

AUTHOR: More than only curvilinear trapezoids. For instance, try to find the area of the figure shaded in Figure 50.

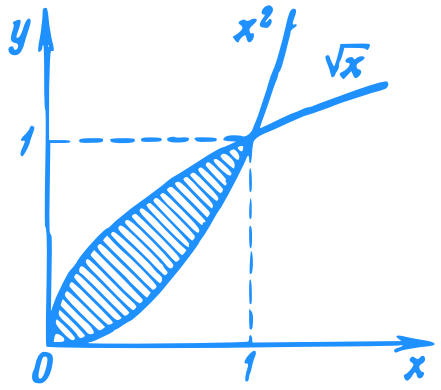


Figure 50: Curvilinear trapezoids as integral of a function.

READER: This area is the difference between the areas of two curvilinear trapezoids:

$$S = \int_0^1 \sqrt{x}\,dx - \int_0^1 x^2\,dx$$

Therefore,

$$S = \frac{2}{3}x\sqrt{x}\Big|_0^1 - \frac{1}{3}x^3\Big|_0^1 = \frac{2}{3} - \frac{1}{3} = \frac{1}{3}$$

AUTHOR: Correct. Consider another example. Find the area of the shaded figure in Figure 51.

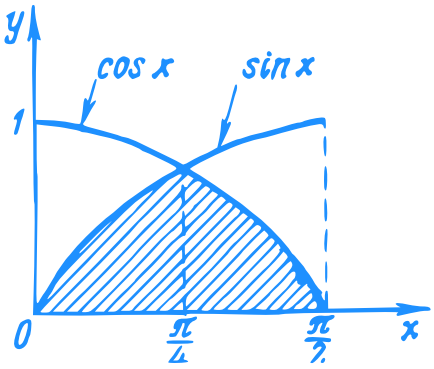


Figure 51: Curvilinear trapezoid as the integral of a function.

READER: The graphs of the functions $\sin x$ and $\cos x$ intersect at the point $x = \frac{\pi}{4}$. Consequently, one has to use the antiderivative of the function $\sin x$ over the interval $\left[0, \frac{\pi}{4}\right]$ and that of the function $\cos x$ over the interval $\left[\frac{\pi}{4}, \frac{\pi}{2}\right]$. Hence,

$$\begin{aligned} S &= \int_0^{\frac{\pi}{4}} \sin x\, dx + \int_{\frac{\pi}{4}}^{\frac{\pi}{2}} \cos x\, dx = -\cos x\Big|_0^{\frac{\pi}{4}} + \sin x\Big|_{\frac{\pi}{4}}^{\frac{\pi}{2}} \\ &= -\left(\cos\frac{\pi}{4} - \cos 0\right) + \left(\sin\frac{\pi}{2} - \sin\frac{\pi}{4}\right) \\ &= -\left(\frac{\sqrt{2}}{2} - 1\right) + \left(1 - \frac{\sqrt{2}}{2}\right) \\ &= 2 - \sqrt{2} \end{aligned}$$

AUTHOR: Perfectly right. Now we shall discuss one "fine point", returning to formula (73) and rewriting it in the form

$$\boxed{\int_a^x f(t)\,dt = F(x) - F(a)} \tag{74}$$

What has been changed by this rewriting?

READER: First, we have replaced the *constant* upper limit of integration (the number b) by the *variable* limit of integration (the variable x). Second, we have substituted the integration variable t for the integration variable x.

AUTHOR: Only the first of these changes is significant. The second (the substitution of the integration variable) is of no consequence. It is easy to see that the formulas

$$\int_a^b f(x)\,dx \quad \int_a^b f(t)\,dt \quad \int_a^b f(y)\,dy \quad \int_a^b f(z)\,dz$$

are *equivalent* since all the four give $F(b)-F(a)$. So it does not matter what *symbol* is used for the integration variable in each particular case.

READER: Why, then, did you have to substitute the variable t for the integration variable x in (74)?

AUTHOR: Only not to confuse the integration variable with the variable upper limit. These are different variables and, of course, must be denoted by different symbols. The expression

$$\int_a^x f(t)\,dt$$

is called the *integral with a variable upper limit*. It is important that in contrast to the expression

$$\int_a^b f(t)\,dt$$

this expression yields not a number but a function. According to (74), this function is $F(x)-F(a)$.

READER: But if the $\int_a^b \square\,dt$ "black box" is a *functional*, then the $\int_a^x \square\,dt$ "black box" is an operator? (I have used here our symbolic notation of "windows" into which the function f must be input).

AUTHOR: Correct. This is immediately clear in the following unusual table.

An Unusual Table

$f(x)$	$\int_1^2 f(t)\,dt$	$\int_1^x f(t)\,dt$
$2x$	3	x^2-1
$3x^2$	7	x^3-1
$4x^3$	15	x^4-1
$5x^4$	31	x^5-1
$6x^5$	63	x^6-1

The second and third columns of this table show *what* the "output" of the two "black boxes", $\int_1^2 f(t)\,dt$ and $\int_1^x f(t)\,dt$, is when the "input" is a function f of the first column.

The integral $\int_a^x (\ldots)\,dt$ is thus indeed an operator. Note that its effect on a function is *opposite* to that of the operator $\frac{d}{dx}$ (we discussed this operator in Dialogue Nine).

Indeed, take a function f and first apply to it the operator $\int_a^x (\ldots)\,dt$ and then the operator $\frac{d}{dx}$: $\frac{d}{dx}\left(\int_a^x (\ldots)\,dt\right)$. This gives

$$\frac{d}{dx}\left(\int_a^x (\ldots)\,dt\right) = \frac{d}{dx}[F(x)-F(a)] = \frac{d}{dx}F(x) = f(x)$$

i.e. we obtain the initial function f.

READER: We could apply these operators to the function in the *reverse* order, couldn't we?

AUTHOR: Yes, we could. This means that the expression

$$\int_a^x \left(\frac{d}{dt} f(t)\right) dt$$

also gives the initial function f. At least, to within a constant term.

READER: Can it be verified?

AUTHOR: Yes, and very easily. What function is the antiderivative for $f'(x)$?

READER: Obviously, the function $f(x)+c$.

AUTHOR: Therefore,

$$\int_a^x \left(\frac{d}{dt} f(t)\right) dt = \int_a^x f'(t)\, dt = f(x) - f(a)$$

READER: Will it be correct to say that while the operator $\dfrac{d}{dx}$ performs the *operation of differentiation*, the operator $\displaystyle\int_a^x (\ldots)\, dt$ performs the *operation of integration*?

AUTHOR: Precisely. It might seem that the topic is exhausted, but the discussion would be incomplete without a clarification of one essential "subtlety". Throughout this dialogue we operated with something we called "the area of a curvilinear trapezoid" and found that this is the meaning of the integral. But what is the "area of a curvilinear trapezoid?"

READER: But surely this is self-evident. One glance at the figures is enough.

AUTHOR: Look, for instance, at Figure 45. It shows a shaded geometrical

figure called a curvilinear trapezoid. But it says nothing about the *area* of the trapezoid.

READER: The area is a standard concept in geometry.

AUTHOR: No objections. But do not forget that in geometry you normally apply this concept to a well-defined set of figures: triangles, trapezoids, etc. And you remember that difficulties arise when you try to determine the *area of a circle*. By definition, the area of a circle is the limit of the sequence of the areas of regular polygons inscribed in, or circumscribed around, the circle, for an infinitely increasing number of the sides of the polygon.

READER: Presumably, the area of a curvilinear trapezoid can also be defined as the *limit of a specific sequence of areas*?

AUTHOR: Yes, this is the normal approach. Consider a curvilinear trapezoid bounded by the graph of a function $f(x)$ over the interval $[a, b]$ (Figure 52). Let us subdivide the interval $[a, b]$ into n subintervals of identical length $\Delta x = \dfrac{b-a}{n}$ (in Figure 52 $n = 10$). The end points of these n subintervals are denoted from left to right:

$$x_0 = a, x_1, x_2, x_3, \ldots x_n = b$$

On each Δx-long interval, used as a base, we construct a *rectangle* of altitude $f(x_{k-1})$, where k is the subscript of the right-hand end of this subinterval (this choice is arbitrary; the left-hand end would do equally well). The area of this rectangle is

$$f(x_{k-1})\Delta x$$

Consider now a sum of the areas of all such rectangles (this total area is

shaded in Figure 52):

$$S_n(a,b) = f(x_0)\Delta x + f(x_1)\Delta x + \ldots + f(x_{n-1})\Delta x$$
$$= [f(x_0) + f(x_1) + \ldots + f(x_{n-1})]\,\frac{b-a}{n}$$

As the function $f(x)$ is continuous, the ensemble of all these rectangles for sufficiently large n (sufficiently small Δx) will be very close to the curvilinear trapezoid in question, and, at any rate, the closer the larger n is (the smaller Δx). It is, therefore, logical to assume the following definition:

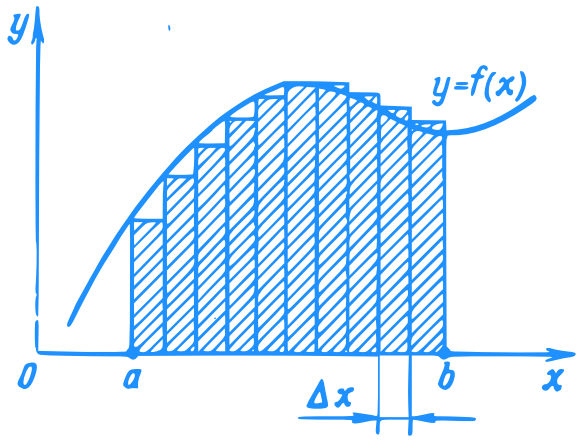


Figure 52: Curvilinear trapezoids as integral of a function.

Definition:

The sequence of sums $(S_n(a,b))$ with n tending to infinity has the limit said to be the area of the given curvilinear trapezoid $S(a,b)$:

$$S(a,b) = \lim_{n\to\infty} S_n(a,b) \qquad (75)$$

READER: The area $S(a,b)$ of a curvilinear trapezoid was shown earlier to be the integral $\int_a^b f(x)\,dx$; consequently, definition (75) is a new defini-

tion of the integral:

$$\boxed{\int_a^b f(x)\,dx = \lim_{n\to\infty} S_n(a,b)} \tag{76}$$

Do you agree?

AUTHOR: Yes, certainly. And note that definition (76) is *independent*, that is, it is not based on the concept of the antiderivative.

Historically, by the way, the integral appeared as (76), the fact that explains the origin of the standard notation. Indeed, if definition (76) is rewritten in a slightly different form

$$\int_a^b f(x)\,dx = \lim_{n\to\infty}\left(\sum_{k=1}^{n} f(x_{k-1})\Delta x\right) \tag{77}$$

you may notice a certain similarity in the form of the left- and right-hand sides of this equality. The very symbol $\int$ (the integral sign) originated from the letter S which was often used to denote sums. The product $f(x_{k-1})dx$ evolved to $f(x)dx$. In the 17th century mathematicians did not use the concept of the limit. They treated integrals as "sums of an infinitely large number of infinitely small addends", with $f(x)dx$ being these infinitesimal addends. In this sense, the area of a curvilinear trapezoid S was defined as the "sum of an infinitely large number of infinitely small areas $f(x)dx$."

You realize, I hope, that such concepts were obviously lacking mathematical rigorousness.

READER: This illustrates what you termed on many occasions "subjective impressions".

AUTHOR: It must be clear to you by now that a strict mathematical interpretation of the concept of the *integral* is possible only if the *limit*

transition is used. I have already emphasized that the limit transition is the foundation of calculus. If the concept of the limit is avoided ("limit of sequence" or "limit of function"), neither the derivative nor integral can he treated rigorously.

READER: But the integral can be defined without resorting to (76). It is quite sufficient to use the Newton-Leibnitz formula (73). And this formula does not involve any limit transitions.

AUTHOR: But this formula involves the antiderivative. And the antiderivative involves, in the long run, the concept of the derivative, that is, the unavoidable limit transition.

By the way, your last remark makes me touch the aspects of introducing the integral in the literature. Two methodically distinct approaches are possible.

The *first approach* (the one used in those dialogues) assumes that the operation of integration is directly introduced as an operation inverse to differentiation. The Newton-Leibnitz formula (73) then serves, in fact, as the definition of the integral: it is defined as an *increment of the antiderivative*.

The *second approach* assumes that the operation of integration is introduced as an *independent* operation, the integral being defined as the *limit* of a sequence formed of the appropriate sums (see formula (76)). This approach corresponds to the historical progress in mathematics; indeed, originally integral calculus was evolving independently of differential calculus. The profound relationship between the two branches of mathematics had been discovered only by the end of the 17th century when the main problems of the two were understood as *mutually inverse*. The

Newton-Leibnitz formula (73) was precisely a reflection of this relationship: it was demonstrated that the integral is none other than an increment of the antiderivative.

Dialogue 13

DIFFERENTIAL EQUATIONS

AUTHOR: You are, certainly, familiar with various types of equations: algebraic, logarithmic, exponential, trigonometric. They have a common feature: by solving these equations one arrives at *numbers* (these are the so-called "roots" of equations). Now we are going to deal with a very different type of equations, namely, *equations whose solutions are functions*. Among the equations subsumed into this class are the so-called *differential equations*.

Consider a function $f(x)$. We denote its first derivative (the first-order derivative) by $f'(x)$, its second derivative by $f''(x)$, its third derivative by $f'''(x)$, and so on.

Definition:

A differential equation is an equality relating x, $f(x)$, $f'(x)$, $f''(x)$, etc. A solution of a differential equation is a junction $f(x)$.

READER: So far you have never mentioned the concepts of second deriva-

tive or third derivative.

AUTHOR: True, and this is what we are going to do right now.

READER: It is readily apparent that since a derivative $f(x)$ is a *function*, it can be differentiated, thus yielding a *derivative of the derivative*; I guess, this must be the second derivative of the original function $f(x)$.

AUTHOR: By differentiating the function $f(x)$ n times o(of course, if this can be done with the given function), we obtain a derivative of the nth order (in other words, "the nth derivative"). Thus, the third derivative of $f(x)$ is obviously,

$$f'''(x) = \frac{d}{dx}\left[\frac{d}{dx}\left(\frac{d}{dx}f(x)\right)\right]$$

Note that we are, in fact, familiar with the second derivative. As the function $f(x)$ is the first derivative of an anti-derivative $F(x)[f(x) = F'(x)]$, the function $f''(x)$ can be considered as the second derivative of the antiderivative $F(x)$:

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

READER: We know that the derivative of $f(x)$ (to be exact, its *first* derivative) is the rate of change of this function. Its magnitude is reflected in the slope of the graph of the function $f(x)$ at each point and is measured as the tangent of the angle between the tangent line to the graph and the abscissa axis. Could anything of this type be said about the *second* derivative of $f(x)$?

AUTHOR: Evidently, the second derivative of $f(x)$ characterizes the rate at which *the rate of change of the function changes with* x, so that it is a finer characteristic of the behaviour of the initial function. Look at Figure 53. What is the difference between functions f_1 and f_2 at point $x = x_0$?

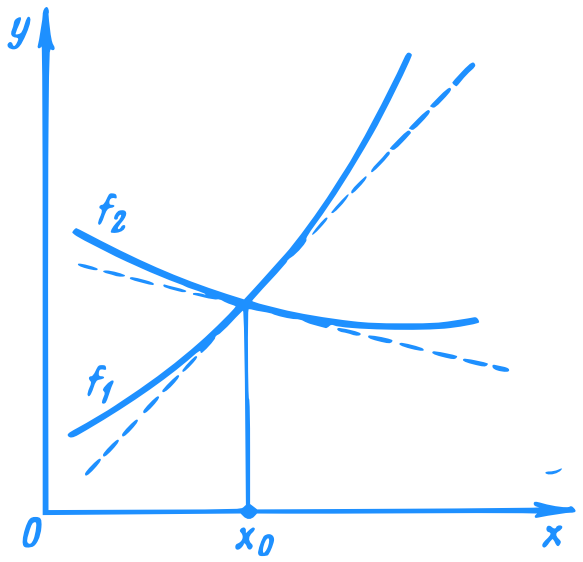


Figure 53: Examining behavior of functions and their derivatives at a point.

READER: They have different first derivatives. I can write:

$$f_1(x_0) = f_2(x_0), \quad f_1'(x_0) \neq f_2'(x_0)$$

AUTHOR: To complete the picture, note that at the point in question the derivatives differ both in magnitude (the figure clearly shows that $|f_2'(x_0| < |f_1'(x_0)|$ and in sign: $f_1'(x_0) > 0$, $f_2'(x_0) < 0$. We say, therefore, that the function f_1 *increases* (and rather rapidly) at point $x = x_0$, while the function f_2 *decreases* (and comparatively slowly).

Now turn to Figure 54. We observe that not only the values of the functions f_1 and f_2 but also the values of their first derivatives coincide at point $x = x_0$:

$$f_1(x_0) = f_2(x_0), \quad f_1'(x_0) = f_2'(x_0)$$

However, the graph shows a difference in the behaviour of the functions f_1 and f_2 in the vicinity of x_0. Try to describe this difference.

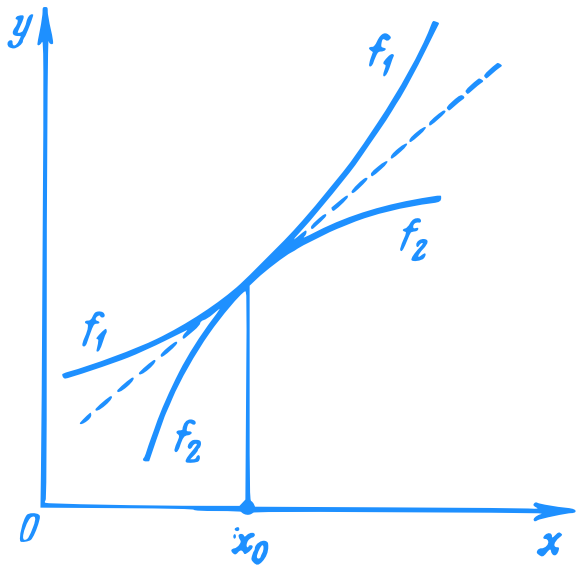


Figure 54: Examining behavior of functions and their derivatives at a point.

READER: In the vicinity of x_0 the graph of the function f_1 is convex downward, while that of the function f_2 is convex upward. Besides, the curvature is greater for the function f_2 than for f_1.

AUTHOR: These are precisely the finer features of the behaviour of f_1 close to $x = x_0$, and they can be identified by finding the value of the second derivative at x_0 (by calculating the value of $f''(x_0)$). In the case shown in Figure 54 we have

$$f_1''(x_0) \neq f_2''(x_0)$$

You will immediately see that $f_1''(x_0) > 0$ and $f_1''(x_0) < 0$. Indeed, the slope of f_1 at x_0 steadily increases; hence, the slope of f_1' is *positive*. On the contrary, the slope of f_2 at x_0 steadily decreases; hence, the slope of f_2' is *negative*. It is quite obvious (see the figure) that

$$|f_1''(x_0)| < |f_2''(x_0)|$$

READER: In all likelihood, the third derivative of $f(x)$, i.e, $f'''(x_0)$, is a still finer characteristic of the behaviour of $f(x)$ at $x = x_0$. Am I right?

AUTHOR: Precisely. Unfortunately, it is virtually impossible to illustrate this simply enough on a graph of the function $f(x)$.

I think that it is enough for a discussion of derivatives of different orders; let us move on to *differential equations*. Note, first of all, that an equation of the type

$$f'(x) = \varphi(x) \tag{78}$$

where $\varphi(x)$ is a given function, can be considered as the simplest particular case in the theory of differential equations; its solution is obtained by a straightforward integration.

Two simple (and, incidentally, very frequently encountered) types of differential equations are

$$\boxed{f'(x) = p f(x)} \tag{79}$$

$$\boxed{''(x) = -q f(x) \quad (q > 0)} \tag{80}$$

where p and q are constants.

Equation (79) is called the *differential equation of exponential growth (decay)*, and equation (80) is the differential equation of harmonic oscillations.

Let us look at these equations more closely. We begin with the differential equation of exponential growth (decay). What conclusions can be drawn from the form of this equation?

READER: The form of equation (79) shows that the rate of change of the function $f(x)$ coincides with the value of the function, to within a con-

stant factor p; at each point x, In other words, the function $f(x)$ and its first derivative $f'(x)$ coincide, to within the mentioned factor, at each point x.

AUTHOR: Please, recall Dialogue Ten and tell me what functions could serve as solutions of this equation. What are the functions for which the derivative coincides with the function itself? In other words, what functions are transformed by differentiation into themselves?

READER: This property is typical of the exponential function a^x for $a = e$. It is called the exponential curve and is often denoted by $\exp(x)$. We have found in Dialogue Ten that

$$\frac{d}{dx}\exp(x) = \exp(x)$$

AUTHOR: Correct. This means that the function $f(x) = \exp(px)$ must be taken as a solution of the equation $f'(x) = p\,f(x)$. Indeed,

$$\begin{aligned}\frac{d}{dx}\exp(px) &= \left(\frac{d}{dy}\exp(y)\right)\frac{d}{dx}(px)\\ &= p\exp(y) = p\exp(px)\end{aligned}$$

For this reason equation (79) is called the differential equation of *exponential* growth (decay). Obviously, we have growth if $p > 0$, and decay if $p < 0$.

READER: Apparently, any function

$$f(x) = C\exp(px)$$

where C is an arbitrary constant factor, is a solution of this equation, because the constant C is factored out of the derivative.

AUTHOR: You are absolutely right.

A solution of a differential equation $f'(x) = p\,f(x)$ is a family of functions

$$f(x) = C\exp(px)$$

with an arbitrary constant factor C (usually referred to as the integration constant).

Some of functions $C\exp(px)$ are plotted in Figure 55 (we have specified $p > 0$).

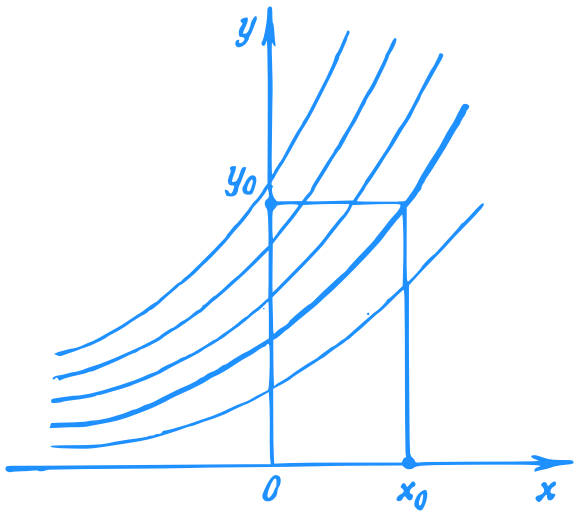


Figure 55: Graphs of some of functions $C\exp(px)$.

The formula $f(x) = C\exp(px)$, describing the whole family of functions, is called the *general solution* of a given differential equation. By fixing (i.e. specifying) a value of C, one selects (singles out) a *particular solution* from the general solution.

READER: How can it be done?

AUTHOR: Oh, this is elementary. It is sufficient to prescribe a specific

value to the function $f(x)$ at a certain point. For example, let us prescribe

$$f(x_0) = y_0$$

In this case we are interested in a single curve among the curves of the whole family (see Figure 55; the selected curve is shown by a thicker solid line). This curve is a graph of the function $C\exp(px)$ for which $C\exp(px_0) = y_0$, and, therefore, $C = y_0\exp(-px_0)$. Consequently, the particular solution we are seeking for has the form

$$f(x) = y_0\exp[p(x-x_0)] \tag{81}$$

READER: We thus obtain that in order to find a specific (particular) solution of the differential equation $f'(x) = pf(x)$, it is necessary to supplement the equation with an additional condition: $f(x_0) = y_0$.

AUTHOR: Precisely. This condition is called the *initial condition.*

Let us turn now to differential equation (80):

$$f''(x) = -qf(x) \quad (q > 0)$$

READER: In this case the value of the function $f(x)$ coincides at each point not with the rate of change of the function but with the rate of change of its rate of change, with the sign reversed.

AUTHOR: In other words, the function $f(x)$ is equal, to within a constant factor, to its second derivative $f''(x)$. Recall what functions have this property.

READER: I guess that the solutions of equation (80) are functions $\sin x$ or $\cos x$.

AUTHOR: To be precise: $\sin(\sqrt{q}x)$ or $\cos(\sqrt{q}x)$. Indeed,

$$\frac{d}{dx}\left(\frac{d}{dx}\sin(\sqrt{q}x)\right) = \sqrt{q}\frac{d}{dx}\cos(\sqrt{q}x) = -q\sin(\sqrt{q}x)$$

or

$$\frac{d}{dx}\left(\frac{d}{dx}\cos(\sqrt{q}x)\right)=-\sqrt{q}\frac{d}{dx}\sin(\sqrt{q}x)=-q\cos(\sqrt{q}x)$$

This is why the equation in question is called the differential equation of *harmonic oscillations*.

It is easily seen that the general solution of equation (80) can be written in the form

$$f(x)=C_1\sin(\sqrt{q}x)+C_2\cos(\sqrt{q}x) \tag{82}$$

where C_1 and C_2 are arbitrary constants (integration constants). Indeed,

$$\begin{aligned} f'(x)&=\sqrt{q}\,C_1\cos(\sqrt{q}x)+-\sqrt{q}\,C_2\sin(\sqrt{q}x)\\ f''(x)&=-[q\,C_1\sin(\sqrt{q}x)+\,C_2\sin(\sqrt{q}x)]\\ &=-q\,f(x)\end{aligned}$$

Reader: But this gives us *two* integration constants instead of one, as in the preceding case.

Author: Yes, and the reason is that differential equation (80) contains the *second* derivative. Hence, it is necessary to integrate *twice* in order to obtain the function $f(x)$. And we know that each integration leads to a family of anti-derivatives, that is, generates an integration constant. In the general case, the number of integration constants in the general solution of a specific differential equation equals the *maximum* order of derivative in this equation. The general solution of equation (79) has a single integration constant because it contains only the first derivative of the sought function and does not involve derivatives of higher order. The general solution of equation (80) has two integration constants because the equation contains the second-order derivative of the sought function and no derivatives of higher order.

READER: And how do we write the initial condition for equation (80)?

AUTHOR: One has to prescribe at a point $x = x_0$ a value not only to the sought function but also to its first derivative. In this case the initial conditions are written as follows:

$$f(x_0) = f_0 \quad f'(x_0) = f_0' \tag{83}$$

READER: And if a differential equation involved the third derivative, and the general solution contained, as a result, not two but three integration constants?

AUTHOR: In this case the initial conditions would prescribe values to the required function, its first derivative, and its second derivative at a point $x = x_0$:

$$f(x_0) = f_0 \quad f'(x_0) = f_0' \quad f''(x_0) = f_0''$$

But let us return to the general solution of equation (80). It is usually written not in form (82) but in a somewhat different form. Namely, either

$$f(x) = A \sin(\sqrt{q}x + \alpha) \tag{84}$$

or

$$f(x) = A \cos(\sqrt{q}x + \beta) \tag{84a}$$

Formula (86) is obtained from (84) if we set $\alpha = \beta + \dfrac{\pi}{2}$.

In what follows we shall use notation (84). In this form the role of the integration constants C_1 and C_2 in general solution (82) is played by constants A and α. Formula (82) is easily transformed by trigonometry to (84), by using the formula for the sine of a sum. Indeed,

$$A \sin(\sqrt{q}x + \alpha) = A \sin(\sqrt{q}x) \cos\alpha + A \cos(\sqrt{q}x) \sin\alpha$$

so that

$$C_1 = A\cos\alpha, \quad C_2 = A\sin\alpha$$

Now try to obtain from general solution (84) a particular solution satisfying initial conditions (83).

READER: We shall obtain it by expressing the constants A and a via f_0 and f_0' the first derivative of $f(x)$:

Equality (84) yields an expression for

$$f'(x) = A\sqrt{q}\cos(\sqrt{q}x + \alpha)$$

In this case initial conditions (83) take the form

$$\begin{aligned} \sin(\sqrt{q}x_0 + \alpha) &= \frac{f_0}{A} \\ \cos(\sqrt{q}x_0 + \alpha) &= \frac{f_0'}{A\sqrt{q}} \end{aligned} \tag{85}$$

System (85) must be solved for the unknown constants A and α. Squaring both equations of the system and summing them up, we obtain (taking into account that $\sin^2\nu + \cos^2\nu = 1$)

$$\left(\frac{f_0'}{A}\right)^2 + \left(\frac{f_0'}{A\sqrt{q}}\right)^2 = 1$$

This yields

$$A = \sqrt{f_0^2 + \left(\frac{f_0^{2\prime}}{\sqrt{q}}\right)} \tag{86}$$

Dividing the first equation of system (85) by the second, we obtain

$$\tan(\sqrt{q}x_0 + \alpha) = \frac{f_0}{f_0'}\sqrt{q} \tag{87}$$

From (87) we can find constant α.

The constants A and α, expressed in terms of f_0 and $f_0\prime$ must be substituted into (84); the result is the particular solution satisfying initial conditions (83).

AUTHOR: Assume that initial conditions (83) are

$$f(0)=0,\ f\prime(0)=f_0' \tag{88}$$

READER: In this case formulas (86) and (87) yield

$$A=\frac{f_0\prime}{\sqrt{q}} \tag{89}$$

If $\tan\alpha=0$, then $\alpha=\pi n$, where $n=0,\pm1,\pm2$, And since, first,

$$\sin(\sqrt{q}x+\alpha)=\sin(\sqrt{q}x)\cos\alpha+\cos(\sqrt{q}x)\sin\alpha$$

and, second, in this particular case $\sin\alpha=0$ and $\cos\alpha=\pm1$, we conclude that either

$$f(x)=\frac{f_0\prime}{\sqrt{q}}\sin(\sqrt{q}x)$$

or

$$f(x)=-\frac{f_0\prime}{\sqrt{q}}\sin(\sqrt{q}x)$$

AUTHOR: The second variant is unacceptable because it violates the condition $f\prime(0)=f_0'$.

READER: Hence, the required particular solution is

$$f(x)=\frac{f_0\prime}{\sqrt{q}}\sin(\sqrt{q}x) \tag{90}$$

AUTHOR: Very good. Now consider the initial conditions in the form

$$f(0)=f_0,\ f\prime(0)=0 \tag{91}$$

READER: Formula (86) yields $A = f_0$. However, formula (87) is no help in this case since $f_0' = 0$.

AUTHOR: I advise you to use the relation derived earlier, namely, the second equation in system (85). In this case it takes the form $\cos\alpha = 0$.

READER: We obtain then

$$A = f_0, \ \cos\alpha = 0 \tag{92}$$

This yields $\alpha = \dfrac{\pi}{2} + \pi n$, and therefore

$$f(x) = f_0 \sin\left(\sqrt{q}x + \frac{\pi}{2} + \pi n\right) = f_0 \cos(\sqrt{q}x + \pi n)$$

AUTHOR: It can he readily found that the particular solution satisfying initial conditions (92) is of the form

$$f(x) = f_0 \cos(\sqrt{q}x + \pi n) \tag{93}$$

Pay attention to the *periodicity* of the functions representing solutions (general or particular) of differential equation (80).

READER: Relation (90) or (94) clearly shows that the *period* of these functions is

$$x_1 = \frac{2\pi}{\sqrt{q}} \tag{94}$$

AUTHOR: Right. Now I want to dwell on one feature of principal significance. The point is that the differential equations discussed above describe quite *definite processes*, and this is especially clear if we use *time* as the independent variable. Denoting this variable by t, we can rewrite equations (79) and (80) in the form

$$f'(t) - p f(t) = 0 \tag{79a}$$

$$f''(t)+qf(t)=0 \quad (q>0) \tag{80a}$$

Equation (79a) describes a process of exponential growth ($p>0$) or exponential decay ($p<0$). Equation (80a) describes a process of harmonic oscillations with the period $T\frac{2\pi}{\sqrt{q}}$.

READER: Would it be correct to say that any differential equation describes a process? I assume that f is a function of time.

AUTHOR: Quite true. This is a point worthy of maximum attention. In a sense, it reflects the principal essence of differential equations. Note: a differential equation relates the values assumed by a function and some of its derivatives at an *arbitrary moment of time* (at an *arbitrary point in space*), so that a solution of the equation gives us a picture of the *process evolving in time (in space)*. In other words, a differential equation embodies a *local relation* (a relation at a point x, at a moment t) between $f, f', f'', \ldots$, thus yielding a certain *picture as a whole*, a certain *process*, an *evolution*. This is the principal idea behind the differential equations.

READER: And what is the role played by initial conditions?

AUTHOR: The role of initial (and boundary) conditions is obvious. A differentia I equation *per se* can only describe the *character* of evolution, of a given process. But a specific pattern of evolution in a process is determined by concrete *initial conditions* (for example, the coordinates and velocity of a body at the initial moment of time).

READER: Can the character of the process "hidden" in a differential equation be deduced simply from the form of this equation?

AUTHOR: An experienced mathematician is normally able to do it. One glance at equation (79a) is sufficient to conclude that the process is an

exponential growth (decay).

Equation (80a) is a clear message that the process involves oscillations (to be precise, harmonic oscillations). Assume, for example, that differential equation has the following form

$$f''(t) - p f'(t) + q f(t) = 0 \quad (p < 0,\ q > 0) \tag{95}$$

(compare it to equations (79a) and (80a)). We shall not analyze this equation in detail. We only note that what it "hides" is not a harmonic oscillatory process but a process of *damped* oscillations. It can be shown (although we shall not do it) that in this process the amplitude of oscillations will steadily diminish with time by the exponential law $\exp(pt)$.

READER: Does it mean that equation (95) describes a process which combines an oscillatory process and a process of exponential decay?

AUTHOR: Precisely. It describes an *oscillatory process*, but the amplitude of these oscillations *decays* with time.

Dialogue 14

MORE ON DIFFERENTIAL EQUATIONS

AUTHOR: All the preceding dialogues (with an exception of Dialogue Eight) left out, or very nearly so, any possible *physical content* of the mathematical concepts and symbols we were discussing. I wish to use this dialogue, which concludes the book, to "build a bridge" between higher mathematics and physics, with differential equations as a "building material". We shall analyze differential equations of exponential decay and those of harmonic oscillations filling them with a specific physical content.

READER: In other words, you suggest discussing specific *physical processes*?

AUTHOR: Yes, I do. I emphasize that differential equations play an outstanding role in physics. First, any more or less real physical process cannot, as a rule, be described without resorting to differential equations. Second, a typical situation is that in which *different* physical processes are

described by *one and the same* differential equation. It is said then that the physical processes are *similar*. Similar physical processes lead to identical mathematical problems. Once we know a solution of a specific differential equation, we actually have the result for all similar physical processes described by this particular differential equation.

Let us turn to the following specific problem in physics. Imagine an ensemble of decaying radioactive atomic nuclei. Denote by $N(t)$ a function describing the number of atomic nuclei per unit volume which have not decayed by the moment of time t. We know that at the moment $t = t_0$ the number of non-decayed nuclei (per unit volume) is N_0, and that the rate of decrease in the number of non-decayed nuclei at the moment t is proportional to the number of non-decayed nuclei at the given moment:

$$\boxed{-N'(t) = \frac{1}{\tau} N(t)} \tag{96}$$

Here $\dfrac{1}{\tau}$ is a proportionality factor; evidently, τ has the dimension of time; its physical meaning will be clarified later.

We are to find the function $N(t)$.

This is our specific physical problem. Let us look at it from the mathematical viewpoint.

READER: Equation (96) is a differential equation of type (80a) from the preceding dialogue, in which $p = -\dfrac{1}{\tau}$. The initial condition in this case is $N(t_0) = N_0$. By using result (81) of the preceding dialogue, we immediately obtain

$$\boxed{N(t) = N_0 \exp\left(-\frac{1}{\tau}(t - t_0)\right)} \tag{97}$$

AUTHOR: Correct. The formula that you have written, i.e, (97), describes

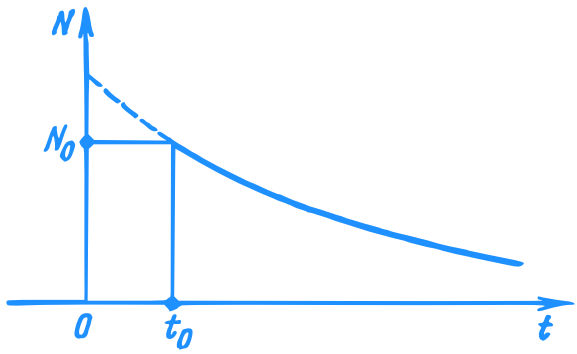


Figure 56: Graphically depicting radioactive decay of nuclei.

the *law of radioactive decay*; we find that this decay is exponential. The number of non-decayed nuclei decreases with time exponentially (Figure 56).

By taking the logarithm of equality (97) (using natural logarithms), we obtain

$$\ln N(t) = \ln N_0 - \frac{t - t_0}{\tau}$$

This yields

$$\tau = \frac{t - t_0}{\ln \dfrac{N_0}{N(t)}}$$

The constant τ is, therefore, such a time interval during which the number of non-decayed nuclei diminishes by a factor of e (i.e. approximately by a factor of 2.7); indeed, in this case $\ln \dfrac{N_0}{N(t)} = \ln e = 1$.

Let us turn now to a different physical problem. Let a light wave with intensity I_0 be incident perpendicularly at a boundary (the so-called interface) of a medium; the wave propagates through the medium with gradually attenuating intensity. We choose the x-axis as the wave propagation direction and place the origin (point $x = 0$) on the interface (Figure 57).

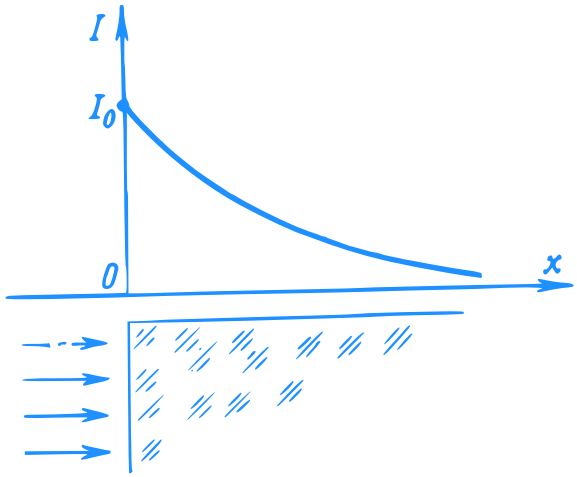


Figure 57: A light wave propagating through a interface medium.

We want to find $I(x)$, that is, the light intensity as a function of the depth of penetration into the medium (in other words, on the path traversed within this medium). We also know that the rate of attenuation at a given point x (i.e. the quantity $-I'(x)$ is proportional to the intensity at this point:

$$\boxed{-I'(x) = \eta\, I(x)} \tag{98}$$

Here η is the proportionality factor whose dimension is, obviously, that of inverse length; its physical meaning will be clear somewhat later.

This, therefore, is the formulation of the physical problem.

READER: It is readily apparent that, as in the preceding case, we deal here with a differential equation of exponential decay. The initial condition is $I(0) = I_0$. By using result (81) of the preceding dialogue, we obtain

$$I(x) = I_0 \exp(-\eta x) \tag{99}$$

AUTHOR: Formula (99) describes *Bouguer's law*, well known in optics:

as light penetrates the matter, its intensity decays exponentially (see Figure 57). We readily see that the constant η is a quantity inverse to the length along which the light intensity diminishes by a factor of e. The constant η is called the *linear absorption coefficient*.

Note that results (97) and (99) describe two *different* physical problems from different fields of physics. We describe here two *different physical processes*. Nevertheless, the mathematical nature of these physical processes is the same: both are described by the *same differential equation*.

Let us consider a different physical problem. Assume that a ball with mass m, attached to fixed walls by elastic springs, vibrates along the x-axis (Figure 58). The origin $x = 0$ is chosen in the position in which the ball is at equilibrium, that is, half-way between the walls. The motion of the ball is governed by *Newton's second law*:

$$ma = F \tag{100}$$

where a is acceleration, and F is the restoring force. We assume that

$$F = -kx \tag{101}$$

where k is the elasticity factor characterizing the elasticity of the spring. We shall consider the displacement of the ball from the equilibrium position (i.e. the quantity x) as a function of time, $x(t)$. This is the function we want to find.

We remind the reader that acceleration is the second derivative of a function which describes path as a function of time: $a = x''(t)$. Consequently, we can rewrite (100), taking into account (101) in the form

$$mx''(t) + k\,x(t) = 0$$

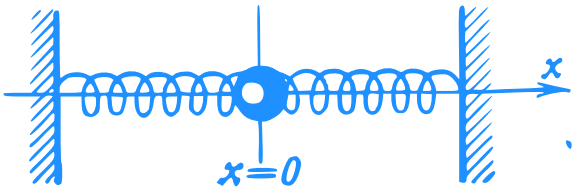


Figure 58: A mass attached between two elastic springs.

or

$$\boxed{x''(t)+\frac{k}{m}x(t)=0} \tag{102}$$

READER: This is a differential equation of type (80a) of the preceding dialogue provided that $q=\dfrac{k}{m}$.

AUTHOR: This means that the general solution must be of the form

$$x(t)=A\sin\left(\sqrt{\frac{k}{m}}\,t+\alpha\right) \tag{103}$$

We thus find that the ball in the problem vibrates harmonically around its equilibrium position $x=0$. The parameter A is, obviously, the *amplitude of vibrations*. The parameter α is called the *initial phase of vibrations*. Recalling relation (94) of the previous dialogue, we conclude that the *period of vibrations* is

$$T=2\pi\sqrt{\frac{m}{k}} \tag{104}$$

Instead of the period T, the so-called *angular frequency* ω is often used: $\omega=\dfrac{2\pi}{T}$. Formula (104) yields

$$\omega=\sqrt{\frac{k}{m}} \tag{105}$$

By using (106), we rewrite general solution (103) in the form

$$x(t) = A\sin(\omega t + \alpha) \tag{106}$$

READER: And what about the initial conditions in this case?

AUTHOR: Assume that the ball is at rest at $t < 0$. By setting specific initial conditions at $t = 0$, we choose a method by which vibrations are initiated at the moment $t = 0$. For example, let the initial conditions be given by relations (88) of the previous dialogue:

$$x(0) = 0, \quad x'(0) = v_0 \tag{107}$$

This means that at the moment $t = 0$ the ball which is at the equilibrium position ($x = 0$) starts moving at a velocity v_0. According to relation (90) of the previous dialogue, we obtain the following particular solution:

$$x(t) = \frac{v_0}{\omega}\sin(\omega t) \tag{108}$$

Now try to discern the physical meaning of the initial conditions of type (92) of the previous dialogue.

READER: These conditions have the form:

$$x(0) = x_0, \quad x'(0) = 0 \tag{109}$$

This means that at the initial moment $t = 0$ the ball was displaced from the equilibrium position by $x = x_0$ and let go. The corresponding particular solution, following from relation (94) of the previous dialogue, takes the form

$$x(t) = x_0\cos(\omega t) \tag{110}$$

AUTHOR: AUTHOR: In the first case we thus initiate vibrations by imparting the initial velocity v_0 to the ball at the equilibrium position (in this

case the amplitude A of vibrations is $\frac{v_0}{\omega}$, and the initial phase α, can be set equal to zero, in 0) accordance with (110)). In the second case the vibrations are initiated by displacing the ball from the equilibrium position by x_0 and then letting it go (in this case $A = x_0$, and the initial phase a can be set equal to $\frac{\pi}{2}$, in accordance with (110).

READER: Could we consider a case in which at $t = 0$ the ball is displaced from the equilibrium position by x_1 and simultaneously given an initial velocity v_1?

AUTHOR: Of course, this is one of the possible situations. Figure 59 shows four vibration modes (four particular solutions) corresponding to four different initial conditions

(four different methods of starting the vibrations of the ball):

(1) $x(0) = 0,\ x'(0) = v_0$; in this case $A = \frac{v_0}{\omega},\ \alpha = 0$.

(2) $x(0) = x_0,\ x'(0) = 0$; in this case $A = x_0,\ \alpha = \frac{\pi}{2}$.

(3) $x(0) = x_1,\ x'(0) = v_1$; (the initial velocity imparted to the ball has the same direction as the initial displacement); in this case $A = A_1,\ \alpha = \alpha_1$ (see the figure).

(4) $x(0) = x_1,\ x'(0) = -v_1$; (the initial velocity imparted to the ball has the direction opposite to that of the initial displacement); in this case $A = A_1,\ \alpha = \pi - \alpha_1$ (see the figure).

As follows from relation (86) of the preceding dialogue,

$$A_1 = \sqrt{x_1^2 + \left(\frac{v_1}{\omega}\right)^2} \tag{111}$$

and according to (106),

$$\alpha_1 = \arctan\left(\frac{x_1\omega}{v_1}\right) \tag{112}$$

Reader: I notice that by fixing specific initial conditions (in other words, by initiating the vibrations of the ball by a specific method), we predetermine the amplitude and initial phase of the vibrations.

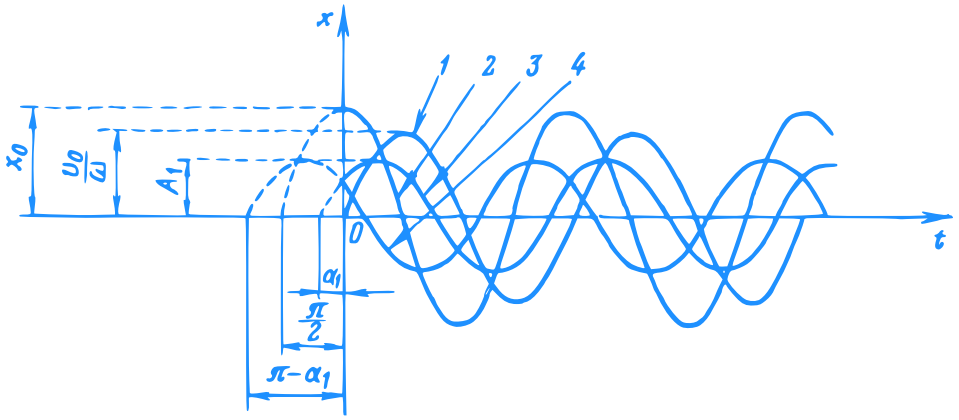


Figure 59: Four vibration modes (four particular solutions) corresponding to four different initial conditions.

Author: Precisely. This is clearly shown in Figure 59. By the way, the same figure shows that the period of vibrations (their frequency) remains constant regardless of the initial conditions.

To summarize, we note that a harmonic oscillation is characterized by three parameters (see (106)) the amplitude A, initial phase α, and frequency ω, The first two parameters are determined by the choice of initial conditions, and the last parameter is independent of them.

The above-described process of vibrations is one of the *mechanical* processes. Let us turn now to a process of an essentially different physical

nature. We shall analyze the motion of electric charges in a circuit consisting of a capacitor with capacitance C and a coil with inductance L (Figure 60).

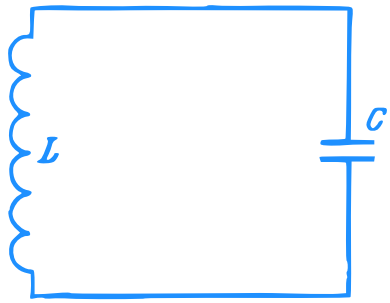


Figure 60: Analysing electrical oscillations in a circuit with a capacitor and an inductor.

Let the capacitor plates have a charge $Q(t)$ at a moment t; correspondingly, the potential difference between the capacitor plates will be $\frac{Q(t)}{C}$. If the current in the circuit at the moment t is $i(t)$, then the potential difference generated in the coil is $-Li'(t)$. We know that it must be balanced out by the potential difference across the capacitor plates:

$$-Li'(t) = \frac{Q(t)}{C} \tag{113}$$

Let us differentiate relation (114). This gives

$$-Li''(t) = \frac{Q'(t)}{C} \tag{114}$$

Now we shall take into account that

$$Q'(t) = i(t)$$

(current intensity, or simply current, is the rate of change of charge). As a result, equation (114) can be rewritten in the form:

$$-Li''(t) = \frac{1}{C} i(t)$$

or

$$\boxed{i''(t)+\frac{1}{LC}i(t)=0} \tag{115}$$

The resultant differential equation is quite familiar, isn't it?

READER: This is a differential equation of type (80a) of the preceding dialogue provided that $q=\dfrac{1}{LC}$. We con- clude, therefore, that the process in the circuit is harmonic.

AUTHOR: Note, however, that the process is not that of mechanical vibrations of a ball attached to springs but the process of *electromagnetic oscillations* in an electric circuit.

READER: As $q=\dfrac{1}{LC}$ and using relation (94) of the previous dialogue, we obtain a relation for the period of electromagnetic oscillations in the circuit:

$$T=2\pi\sqrt{LC} \tag{116}$$

The general solution of equation (115) is then

$$i(t)=A\sin\left(\frac{1}{\sqrt{LC}}t+\alpha\right) \tag{117}$$

AUTHOR: Absolutely correct. The two physical processes, namely, the *mechanical vibrations* of a ball attached to springs and the *electromagnetic oscillations* in a circuit, are *mathematically similar*. They are described by the same differential equation. Otherwise you couldn't write, nearly automatically as you did, the period of oscillations (formula (116) and the general solution (formula 117).

In our dialogues we have discussed only two (and rather simple) types of differential equations: those of exponential growth (decay) and of harmonic oscillations. And we have illustrated them with a number of physical processes of very different kind.

Reader: I guess that the list of different differential equations, and certainly the list of physical processes described by these equations, could be substantially enlarged.

Author: No doubt. This concludes our discussion of differential equations. I want to note in conclusion that differential equations are widely applied not only in physics but in chemistry, biology, cybernetics, sociology, and other fields of science as well.

PROBLEMS

Problem 1 Find a formula for the nth term from the first several terms of the sequence:

(a) $\frac{1}{11}, \frac{1}{21}, \frac{1}{31}, \frac{1}{41}, \frac{1}{51}, \cdots$

(b) $1, \frac{1}{4}, \sqrt{3}\,\frac{1}{16}, \sqrt{5}, \frac{1}{36}, \sqrt{7}\ \ldots$

(c) $1, -\left(\frac{1}{2}\right)^2, \left(\frac{1}{2\cdot 3}\right)^3, -\left(\frac{1}{2\cdot 3\cdot 4}\right)^4, \cdots$

(d) $\frac{3}{4}, -\left(\frac{6}{7}\right)^2, \left(\frac{9}{10}\right)^3, -\left(\frac{12}{13}\right)^4, \left(\frac{15}{16}\right)^5, \cdots$

Answer 1

(a) $y_n = \dfrac{1}{10n+1}$

(b) $y_n = \dfrac{\sqrt{n}}{2}[1-(-1)^n] + \dfrac{1}{2n^2}[1+(-1)^n]$

(c) $y_n = \dfrac{(-1)^{n+1}}{(n!)^n}$

(d) $y_n = (-1)^{n+1}\left(\dfrac{3n}{3n+1}\right)^n$

Problem 2 Find the least term of each sequence:

(a) $y_n = n^2 - 5n + 1$

(b) $y_n = n + \dfrac{100}{n}$

(c) $y_n = n + 5\sin\dfrac{\pi n}{2}$

Answer 2

(a) $y_2 = y_3 = -5$

(b) $y_{10} = 20$

(c) $y_3 = -2$

Problem 3 Find the largest term of each sequence:

(a) $y_n = \dfrac{90n}{n^2+9}$

(b) $y_n = \dfrac{10^n}{n!}$

Answer 3

(a) $y_3 = 15$

(b) $y_9 = y_{10} = \dfrac{10^9}{9!}$

Problem 4 Find which of the sequences given below are monotonic:

(a) $y_n = 3n^2 - n$

(b) $y_n = n^2 - 3n$

(c) $y_n = 7n - n^2$

(d) $y_n = \log\left(\dfrac{3}{4}\right)^n$

Answer 4 (a) Increasing, (b) nondecreasing; (c) non-monotonic; (d) decreasing.

Problem 5 There are two sequences (y_n) and (z_n) such that $0 \leqslant y_n \leqslant z_n$ for all n. The sequence (z_n) is convergent and its limit is zero. Prove that the sequence (y_n) is convergent to zero.

Problem 6 Prove that

(a) $\lim\limits_{n\to\infty} \dfrac{n}{2^n} = 0$;

(b) $\lim\limits_{n\to\infty} (\sqrt{n+1} - \sqrt{n-1}) = 0$

Hint. In problem (a) transform

$$2^n = (1+1)^n = \left[1 + n + \ + \ldots\right] > \left[n + \frac{n(n-1)}{2}\right] > \frac{n^2}{2}$$

and use the theorem proved in problem 5.

In problem (b) transform

$$\sqrt{n+1} - \sqrt{n-1} = \frac{2}{\sqrt{n+1} + \sqrt{n-1}} < \frac{2}{\sqrt{n-1}}$$

and use the theorem proved above.

Problem 7 Find the limits of the following sequencers:

(a) $y_n = \dfrac{2n + \dfrac{1}{n} + 3}{\left(\sqrt{n} + \sqrt{3}\right)^2}$

(b) $y_n = \dfrac{5n^2\left(1 + \dfrac{1}{n}\right)^n}{\dfrac{1}{n} - 3n^2}$

(c) $y_n = \dfrac{\left(1 + \dfrac{1}{n}\right)^n + \left(1 + \dfrac{1}{2n}\right)^{2n}}{2 + \dfrac{1}{n} + \dfrac{1}{\sqrt{n}}}$

(d) $y_n = \dfrac{2^n + n}{2^n + \sqrt{n}}(\sqrt{n+1} - \sqrt{n-1})$

Answer 7 (a) 2; (b) $-\dfrac{5}{3}e$; (c) e; (d) 0.

Problem 8 Find the function $f(x)$ if $3f(x-1) - f\left(\frac{1-x}{x}\right) = 2x$.

Answer 8 $f(x) = \dfrac{3}{4}(x+1) + \dfrac{1}{4(x+1)}$.

Problem 9 Find analytical relations and the natural domains of the following functions:

(a) $f(1-x)$;

(b) $f\left(\dfrac{1}{x}\right)$

for $f(x) = \log(x^2 - 1)$,

Answer 9 (a) $\log(x^2 - 2x)$; $x < 0$, $x > 2$; (b) $\log \dfrac{1-x^2}{x^2}$; $0 < |x| < 1$.

Problem 10 Analyze the continuity and differentiability of the function $f(x) = \arcsin(sinx)$ within the limits of the natural domain of the function.

Answer 10 The natural domain of the function $f(x)$ is $]-\infty, \infty[$; the function is continuous everywhere; it is differentiable at all points with the exception of points

$$x = \pm\frac{\pi}{2}, \pm\frac{3\pi}{2}, \pm\frac{5\pi}{2}, \ldots$$

Problem 11 Prove that the function $f(x) = \sqrt[3]{x^2}$ has no derivative at point $x = 0$.

Problem 12 Prove that $3x^6 - 5x^3 - 30x < 40$ if $|x| \leqslant 2$.

Hint. Find first that the maximum value of the polynomial $f(x) = 3x^6 - 5x^3 - 30x$: over the interval $[-2, 2]$ is below 40. To do this, find the values of $f(x)$ at the end points of the interval $[-2, 2]$ and at the points at which the derivative of $f(x)$ is zero (if these points belong to the indicated interval).

Problem 13 Find the maximum and minimum values of the function $f(x) = x - 2\ln x$ over the interval $[1, e]$.

Answer 13 The minimum value is $f(2) = 2 - 2\ln 2$, the maximum value is $f(1) = 1$.

Problem 14 Find a point x_0 at which the tangent to the graph of the function $f(x) = x^2 + 1$ is parallel to the straight line $y = 3x$.

Answer 14 $x_0 = \dfrac{3}{2}$.

Problem 15 Write the equation of the tangent to the graph of the function $f(x) = x^2 - 4x + 5$ at point $x_0 = 1$.

Answer 15 $y = -2x + 4$.

Note. The equation of the tangent to the graph of the function $f(x)$ at $x = x_0$ is: $y = f(x_0) + f'(x_0)(x - x_0)$, where $f'(x_0)$ is the value of the derivative of the function at x_0.

Problem 16 Find the derivatives of the following functions:

(a) $f(x) = \sqrt[n]{x}$;

(b) $f(x) = \dfrac{1}{x^2 + 3}$;

(c) $f(x) = \sqrt{x^2 + 5}$;

(d) $f(x)=\tan^2 x$;

(e) $f(x)=\sin^4 5x$;

(f) $f(x)=\arcsin\sqrt{x}$;

(g) $f(x)=\ln\dfrac{x^2-1}{10}$;

(h) $f(x)=\ln\dfrac{1+x}{\sqrt{1+x^2}}$.

Answer 16

(a) $\dfrac{1}{n\sqrt[n]{x^{n-1}}}$;

(b) $-\dfrac{2x}{(x^2+3)^2}$;

(c) $\dfrac{x}{\sqrt{x^2+5}}$;

(d) $\dfrac{2\sin x}{\cos^2 x}$;

(e) $20\sin^2 5x\cos 5x$;

(f) $\dfrac{1}{2\sqrt{x(1-x)}}$;

(g) $\dfrac{2x}{x^2-1}$;

(h) $\dfrac{1-x}{(1+x)(1+x^2)}$.

Problem 17 Verify that the functions

$$F_1=\cos^2 x+\cos^4 x$$

$$F_2=\cos 2x-\frac{1}{4}\sin^2 2x$$

$$F_3=\cos^4 x+3\cos^2 x+\sin^2 x$$

are the antiderivatives of one and the same function. Find F_2-F_1 and F_3-F_1.

Answer 17 $F_2 - F_1 = 1$; $F_3 - F_1 = 2$.

Problem 18 Find the area of the curvilinear trapezoid described by the graph of the function $f(x) = x^2 + 1$ over the inter- val $[-3, 3]$.

Answer 18 24.

Problem 19 Find the difference in areas of the curvilinear trapezoids defined by the graphs of the functions $f_1 = e^x$ and $f_2 = e^{-x}$ over the interval $[0, 1]$.

Answer 19 $e + \frac{1}{e}$.

Problem 20 Find the value of a minimizing the area of the curvilinear trapezoid defined by the graph of the function $f(x) = (x - a)^2 + a^2$ over the interval $[0, 1]$.

Answer 20 $a = \frac{1}{4}$.

Problem 21 Find the area of a figure bounded by the graph of the function $f(x) = x^2 - 2x + 2$ and two tangents to this graph, drawn at points $x_1 = 0$ and $x_2 = 2$.

Answer 21 $\frac{2}{3}$.

Problem 22 Find the numbers obtained by evaluating the integral $\int_1^2 f(x)\,dx$ of the following functions:

(a) $f(x) = \frac{1}{x}$;

(b) $f(x) = \frac{1}{x^2}$;

(c) $f(x) = \frac{1}{x^3}$;

(d) $f(x)=\dfrac{1}{x^4}$;

Answer 22 (a) $\ln 2$; (b) $\dfrac{1}{2}$; (c) $\dfrac{3}{8}$; (d) $\dfrac{17}{24}$.

Problem 23 Find the functions obtained by evaluating the integral $\int_1^x f(t)\,dt$ of the following functions:

(a) $f(t)=\dfrac{1}{t}$;

(b) $f(t)=\dfrac{1}{t^2}$;

(c) $f(t)=\dfrac{1}{t^3}$;

(d) $f(t)=\dfrac{1}{t^4}$;

Answer 23 (a) $\ln x$; (b) $-\dfrac{1}{x}+1$; (c) $-\dfrac{2}{x^2}+\dfrac{1}{2}$; (d) $-\dfrac{1}{3x^3}+\dfrac{1}{3}$.

Problem 24 Verify that $f(x)=(x+1)e^x$ satisfies the equation $f'(x)-f(x)=e^x$.

Problem 25 Find a particular solution of the equation $f'(x)=f(x)$ such that $f(x)=2$ for $x=2$.

Answer 25 $f(x)=2\exp(x-2)$.

Problem 26 Consider the equation $f'(x)=f(x)$. Find a particular solution for which the tangent to the graph at point x_0 intersects the ordinate axis at point y_1.

Answer 26 $f(x)=\dfrac{y_1}{1-x_0}\exp(x-x_0)$.

Hint. Make use of the equation of tangent (see Note to problem 15).

Problem 27 Consider the graphs of different particular solutions of the equation $f'(x) = f(x)$. Verify that at the same point x_0 the tangents to all these graphs intersect the abscissa axis at a common point $x = x_0 - 1$.

Problem 28 Find the nth derivative of the following functions:

(a) $f(x) = \sin x$;

(b) $f(x) = \cos x$.

Answer 28

(a) $f(x) = \sin\left(x + \dfrac{n\pi}{2}\right)$;

(b) $f(x) = \cos\left(x + \dfrac{n\pi}{2}\right)$.

Problem 29 Find the value of the fourth derivative of the function $f(x) = \ln x$ for $x = 2$.

Answer 29 $-\dfrac{3}{8}$.

Problem 30 Find the area of the curvilinear trapezoid defined by the graph of the third derivative of the function $f(x) = x^5 - 2x^2 + x - 1$ over the interval $[0, 1]$.

Answer 30 20.

The book introduces basic concepts of calculus for high school students. The book is presented as a relatively free-flowing dialogue between the AUTHOR and the READER. From one discussion to another the AUTHOR will lead the inquisitive and receptive READER to different notions, ideas, and theorems of calculus, emphasizing especially complicated or delicate aspects, stressing the inner logic of proofs, and attracting the reader's attention to special points. I hope that this form of presentation will help a reader of the book in learning new definitions such as those of derivative, antiderivative, definite. integral, differential equation, etc. The author also expects that it will lead the reader to better understanding of such concepts as numerical sequence, limit of sequence, and function.

MIR PUBLISHERS MOSCOW