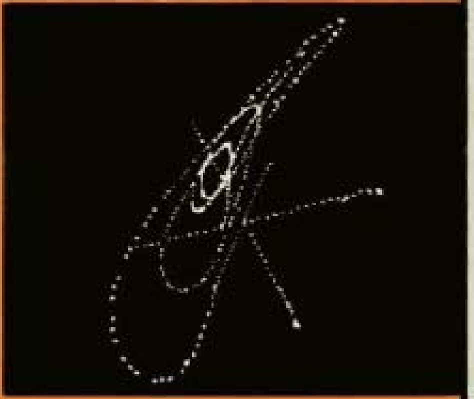
Elements of Applied Mathematics

Ya. B. Zeldovich
A. D. Myškis



Mir
Publishers
Moscow

Academician Yakov Borisovich ZELDOVICH is an outstanding Soviet theoretical physicist, one of the founders of the theory of burning, detonation, and shock waves. His main research interests cover a wide range involving physical chemistry, astrophysics, the theory of elementary particles, and nuclear physics. He is the recipient of the Lenin Prize and four State Prizes, and three times he has been named the Hero of Socialist Labour.

Ya. B. Zeldovich's previously published writings include *Higher Mathematics for Beginners* (1973), *Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena* (1967, with Yu. P. Raizer), *Elements of Gasdynamics and the Classical Theory of Shock Waves* (1969, with Yu. P. Raizer), and *Relativistic Astrophysics* (1971 - 1974, with I. D. Novikov).

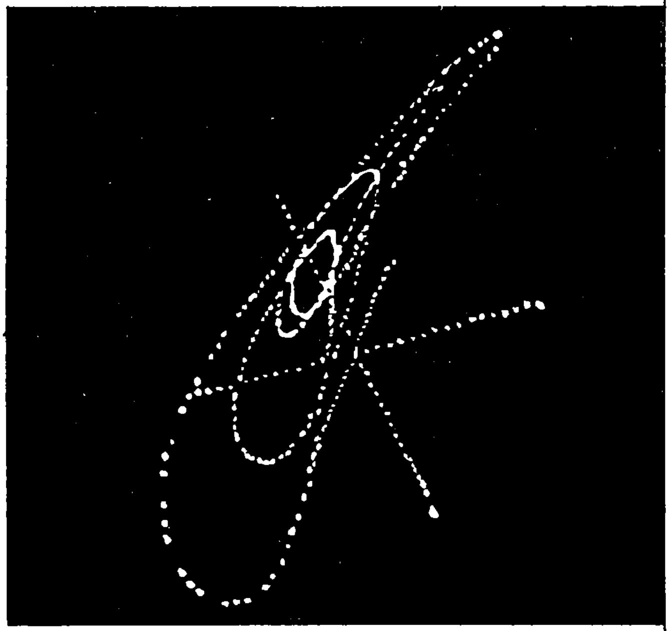
Professor Anatoly Dmitrievich MYŠKIS, D. Sc., is well known as a research mathematician and also for his original approach to the teaching of higher mathematics. He is one of the founders of the theory of delay differential equations. His publications include *Introductory Mathematics for Engineers* (1972, 1975) and *Advanced Mathematics for Engineers* (1975).

Elements of Applied Mathematics

Ya.B.Zeldovich

A.D.Myškis

ИЗДАТЕЛЬСТВО • МИР •



Ya. B. ZELDOVICH and A. D. MYŠKIS

Elements of Applied Mathematics

Translated from the Russian by
George Yankovsky

MIR PUBLISHERS MOSCOW

Я. Б. Зельдович, А. Д. Мышкис

ЭЛЕМЕНТЫ ПРИКЛАДНОЙ
МАТЕМАТИКИ

Издательство «НАУКА»

First Published 1976
Revised from the 1972 Russian edition

To the reader
Mir Publishers would be grateful
for your comments on the content,
translation and design of this book.
We would also be pleased to receive
any other suggestions you may
wish to make.

Our address is:
USSR 129820, Moscow I-110, GSP
Pervy Rizhsky Pereulok, 2
MIR PUBLISHERS

На английском языке

Printed in the Socialist Republic of Romania

CONTENTS

PREFACE

This book is not a textbook in the ordinary sense of the word but rather a reader in the mathematical sciences. Using simple examples taken from physics and a variety of mathematical problems, we have tried to introduce the reader to a broad range of ideas and methods that are found in present-day applications of mathematics to physics, engineering and other fields. Some of these ideas and methods (such as the use of the delta function, the principle of superposition, obtaining asymptotic expressions, etc.) have not been sufficiently discussed in the ordinary run of mathematics textbooks for non-mathematicians, and so this text can serve as a supplement to such textbooks. Our aim has been to elucidate the basic ideas of mathematical methods and the general laws of the phenomena at hand. Formal proofs, exceptions and complicating factors have for the most part been dropped. Instead we have strived in certain places to go deeper into the physical picture of the processes.

It is assumed that the reader has a grasp of the basic essentials of differential and integral calculus for functions of one variable, including the expansion of such functions in power series, and is able to use such knowledge in the solution of physical problems. It is sufficient (but not necessary!) to be acquainted with Zeldovich's *Higher Mathematics for Beginners* (Mir Publishers, Moscow, 1973), which we will refer to throughout this text as HM. What is more, the present text may be regarded as a continuation of *Higher Mathematics for Beginners*. In some places the material is closely related to Myškis' *Lectures in Higher Mathematics* (Mir Publishers, Moscow, 1972). Nevertheless, this text is completely independent and requires of the reader only the background indicated above. Nothing else.

The range of this book is clear from the table of contents. The reader need not study it chapter by chapter but may, in accordance with his particular interests, dip into any chapter, taking care to follow up any references to other sections of the book. For the sake of convenience, material needed in other parts of the book is usually indicated at the beginning of the chapter. Sections and formulas are numbered separately in each chapter, and, within the chapter, references to formulas are given without the number of the chapter.

This text should be useful to students majoring in engineering, physics and other fields as well, starting roughly in the latter half of the first year of study as a supplementary mathematics text. It is also suitable for the practicing engineer and specialist if he wishes to look into some aspect of mathematics that may be of use in his work.

It is undoubtedly difficult to read such diversified material in an active fashion. The reader with a specific purpose in mind will ordinarily devote his whole attention to the matter at hand and skip through any other material that does not pertain to his problem.

The question arises of whether there is any purpose in reading right through the book with the aim of storing up useful knowledge for the future. Would not such an approach make for superficial reading that would leave only hazy impressions in the end?

The authors take the following view.

Firstly, we have always attempted to give not only practical procedures, but also the underlying general ideas as well. Such ideas enrich the reader, expand his scientific horizon and enable him to take a different look of the surrounding world, all of which produces a more profound and lasting impression.

The second argument in favour of this approach has to do with the psychology of memory. The reader may not be able to reproduce all the material he has gone through, but traces are definitely left in his memory. Some of it will come to mind when brought into association with other, familiar, things; for example, when he encounters a problem that requires just such a procedure or mathematical device. Even a very hazy recollection of where such a device may be found proves to be useful.* To put it crudely, there are things, concepts and relationships which we recall only when they are needed.

So our advice is: read our book and study it. But even if there is not time enough to make a deep study, then merely read it like a novel, and it may be just the thing you will need for solving some difficult problem.

* Following Freud, we might say that forgotten theorems remain hidden in the subconscious, alongside the repressions of early childhood.

Never fear difficulties. It is only in grappling with hard problems that a person discovers latent talent and exhibits the best he has.

We will be grateful to our readers for any remarks and criticism they have with respect to the contents of this text and the exposition of the material. Quite naturally, some parts of the book reflect the different habits and interests of the authors, one of whom is a physicist and the other a mathematician. At times we pulled in different directions, and if to this one adds the efforts of the reader, then the overall effect should follow the law of addition of vector forces (see Ch. 9). One of Krylov's fables dealt with a similar situation, but we hope that the results in our case will be better.

The main purpose of this book is to provide the reader with methods and information of practical utility, while simplifying in the extreme all formal definitions and proofs.

Chapter 15 is devoted to electronic digital computers and describes the operating principles of the machines and the programming of elementary problems. This chapter is not designed to take the place of specialized texts for programmers, but it does enable one to grasp the general principles and specific nature of numerical computations.

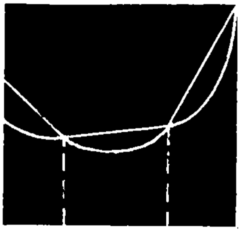
Chapters 9 and 11 give detailed definitions of vectors and tensors, their relationships with linear transformations, and examine quite thoroughly symmetric and antisymmetric tensors. Here we introduce the pseudovector, which is the equivalent of the antisymmetric tensor in three-dimensional space. In this latest edition, physical problems have been added on the motion of a particle in a central-force field and on the rotation of a rigid body. These problems have always been classical examples of the application of the theory of ordinary differential equations to physics.

In connection with the theory of spectral decomposition we have considered the problem of the phase of Fourier components; the loss of information in transitions to spectral density has become customary; it is well to recall that besides the square of the modulus, oscillations are also described by the phase.

These and other additions, it is hoped, will make the text more physical and somewhat more modern without altering its overall purpose and accessibility.

Chapter 1

CERTAIN NUMERICAL METHODS



The solution of a physical problem that has been obtained in mathematical terms, such as combinations of various functions, derivatives, integrals and the like, must be "brought to the number stage", which serves as the ultimate answer in most cases. For this purpose, the various divisions of mathematics have worked out a diversity of numerical methods. In elementary mathematics, as a rule, one considers only methods of the exact solution of problems: the solution of equations, geometrical constructions, and the like. This is a weakness since such solutions are possible only in very rare instances, which naturally reduces drastically the range of admissible problems, and frequently turn out to be extremely unwieldy. Even such a relatively simple problem as the solution of general algebraic equations of the nth degree (with arbitrary coefficients) has proved, in elementary mathematics when $n > 2$, to be unbelievably complicated and cumbersome, whereas for $n > 4$ it is simply impossible. Only the systematic application of the methods of approximate calculations based on the apparatus of higher mathematics has made it possible to bring to completion, and, what is more, in a unified manner, the solution of a broad class of mathematical problems that find important applications. Moreover, the development of numerical methods of higher mathematics and the introduction of modern computing machines have resulted in the following: if a problem has been stated in sufficiently clear-cut mathematical terms, then (with the exception of particularly intricate cases) it will definitely be solved with an accuracy that satisfies practical needs. Thus, higher mathematics not only supplies the ideas underlying an analysis of physical phenomena, but also the numerical methods that permit carrying the solution of specific problems of physics and engineering to completion.

Some of these methods are given in the first course of differential and integral calculus. For instance, the most elementary methods of computing derivatives and integrals, evaluating functions by means of series, and so forth. In this and subsequent chapters (in particular, Chs. 2, 3, 8) we examine these methods in detail. They are not directly connected in any way and so they can be read independently.

1.1 Numerical integration

When a practical problem has been reduced to evaluating a definite integral, we can say that the most difficult part of the matter is over. If the integrand $f(x)$ is such that it is possible to express the indefinite integral $F(x)$ by means of a finite number of elementary functions, then in principle it is rather simple to obtain the value of the definite integral via the formula $\int_a^b f(x)\,dx = F(b) - F(a)$. This involves carrying out a number of arithmetical operations to find the values of $F(b)$ and $F(a)$. In practice, however, this can lead to considerable complications since a very cumbersome formula may result for the indefinite integral $F(x)$.

This procedure may prove to be unsuitable altogether if, as frequently happens, it is impossible to obtain a formula for the indefinite integral.

It sometimes happens that the integral is expressible in terms of nonelementary but well-studied functions for which tables of values have been compiled (see, for example, *Tables of Higher Functions* by Jahnke, Emde, and Lösch [9]). Such, for instance, are the functions

$$\int e^{-x^2}\,dx, \quad \int \frac{\sin x}{x}\,dx, \quad \int \frac{\cos x}{x}\,dx, \quad \int \frac{e^x}{x}\,dx, \quad \int \sin x^2\,dx, \quad \int \cos x^2\,dx$$

and so forth. These integrals, which cannot be expressed in terms of elementary functions, even have special names: error function (see Ch. 13), the sine integral, etc.

In other cases the integral can be evaluated by expanding the integrand in a series of one kind or another. For example, for the integral

$$I = \int_0^1 \frac{\sin x}{x}\,dx$$

the appropriate indefinite integral cannot be expressed in terms of elementary functions. Nevertheless, by taking advantage of the Taylor series for the sine,

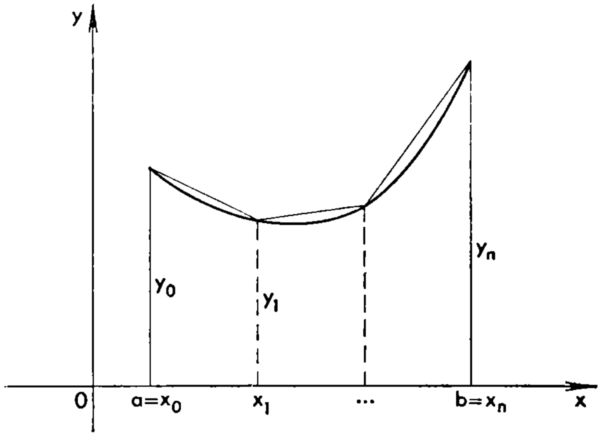
$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

we obtain

$$I = \int_0^1 \left(1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \dots\right) dx = 1 - \frac{1}{3\cdot 3!} + \frac{1}{5\cdot 5!} - \frac{1}{7\cdot 7!} + \dots$$

Computing the successive terms of this series, we stop when the desired reasonable degree of accuracy has been attained. For example, comput-

Fig. 1



ing to three decimal places we can confine ourselves to the first three terms of the series and obtain the value $I = 0.946$.

Some definite integrals can be evaluated precisely with the aid of methods of the theory of functions of a complex variable. For instance, in Sec. 5.9 we will show that

$$\int_{-\infty}^{\infty} \frac{\cos \omega x}{1 + x^2} dx = \pi e^{-\omega} (\omega > 0) \tag{1}$$

although the corresponding indefinite integral cannot be expressed in terms of elementary functions.

And if even such methods prove unsuitable, then the definite integral can be evaluated numerically with the aid of formulas of numerical integration, which we now discuss.

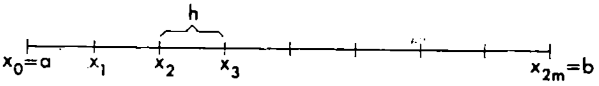
A simple and good method (discussed in HM, Sec. 2.7) consists in the following: the interval of integration is partitioned into several small equal subintervals. The integral over each subinterval is roughly equal to the product of the length of the subinterval by the arithmetic mean of the values of the integrand at the endpoints of the subinterval. This method is called the *trapezoidal rule* because it actually amounts to replacing the arc in each subinterval of the curve $y = f(x)$ by its chord, and the area under the arc (the value of the integral) is replaced by the area of the resulting trapezoid with vertical bases (Fig. 1). The appropriate formula is (but verify this)

$$\int_a^b y \, dx \approx \frac{b-a}{n} \left(\frac{y_0 + y_n}{2} + y_1 + y_2 + \dots + y_{n-1} \right) \tag{2}$$

where $f(x_i) = y_i$ for the sake of brevity.

A still more effective formula can be obtained if the curve $y = f(x)$ over a subinterval is replaced by a parabola, that is to say, the graph of a quadratic relationship. Partition the interval of integration from $x = a$ to $x = b$ into an even number, $2m$, of equal subintervals. Referring to Fig. 2, let the endpoints of the subintervals be $x_0 = a$, $x_1, x_2, \ldots, x_{2m} = b$. Denote the length of a subinterval by h so that $x_1 + h = x_0 + 2h, \ldots, x_{2m} = x_{2m-1} + h = x_0 + 2mh$.

Fig. 2



Consider

$$\int\limits_{x_0}^{x_2} f(x)\,dx = \int\limits_{x_0}^{x_0+2h} f(x)\,dx$$

or the contribution to the original integral of the first two subintervals. Replace the curve $y = f(x)$ over the interval from $x = x_0$ to $x = x_2$ by a parabola passing through the points (x_0, y_0), (x_1, y_1), (x_2, y_2), and approximately replace the area under the curve by the area under the parabola. The equation of the parabola has the form $y = mx^2 + {}$ $+ nx + l$. The coefficients m, n, l are found from the condition that the parabola passes through the three given points:

$$\begin{cases} y_0 = mx_0^2 + nx_0 + l, \\ y_1 = mx_1^2 + nx_1 + l, \\ y_2 = mx_2^2 + nx_2 + l \end{cases}$$

But this leads to rather cumbersome computations.

It is simpler to do as follows. Seek the equation of the parabola in the form

$$y = A(x - x_0)(x - x_1) + B(x - x_0)(x - x_2) + {} \\ + C(x - x_1)(x - x_2) \qquad (3)$$

It is clear that what we have on the right of (3) is a second-degree polynomial so that (3) is indeed the equation of a parabola. Put $x = x_0$ in (3) to get $y_0 = C(x_0 - x_1)(x_0 - x_2)$, or $y_0 = C2\mathrm{h}^2$. Putting $x = x_1 = x_0 + h$ in (3), we get $y_1 = -Bh^2$. Finally, putting $x = x_2 = x_0 + 2h$ in (3), we get $y_2 = 2Ah^2$, whence

$$A = \frac{y_2}{2h^2}, \quad B = -\frac{y_1}{h^2}, \quad C = \frac{y_0}{2h^2}\,^* \qquad (4)$$

[*] Note that the following problem can be solved in similar fashion: to find a polynomial of degree n that takes on specified values for $n + 1$ values of x. This *interpolation problem* is encountered in searching for empirical formulas (see Sec. 2.4), in operations involving functions given in tabular form, and elsewhere.

The area under the parabola is

$$\int_{x_0}^{x_0+2h} [A(x - x_0)(x - x_1) + B(x - x_0)(x - x_2) + C(x - x_1)(x - x_2)]\,dx = A\frac{2}{3}h^3 - B\frac{4}{3}h^3 + C\frac{2}{3}h^3$$

(When performing the integration it is convenient to make a change of variable: $x - x_1 = s$; also note that $x_0 = x_1 - h$, $x_2 = x_1 + h$). Therefore

$$\int_{x_0}^{x_2} f(x)\,dx \approx A\frac{2}{3}h^3 - B\frac{4}{3}h^3 + C\frac{2}{3}h^3$$

Using (4) we find

$$\int_{x_0}^{x_2} f(x)\,dx \approx \frac{h}{3}(y_0 + 4y_1 + y_2) \qquad (5)$$

In the same way we can evaluate $\int_{x_2}^{x_4} f(x)\,dx, \ldots, \int_{x_{2m-2}}^{x_{2m}} f(x)\,dx$. For the integral over the whole interval from $x = a$ to $x = b$ we get

$$\int_a^b f(x)\,dx \approx \frac{h}{3}[(y_0 + 4y_1 + y_2) + (y_2 + 4y_3 + y_4) + \ldots + (y_{2m-2} + 4y_{2m-1} + y_{2m})]$$

whence

$$\int_a^b f(x)\,dx \approx \frac{h}{3}[y_0 + y_{2m} + 2(y_2 + y_4 + \ldots + y_{2m-2}) + 4(y_1 + y_3 + \ldots + y_{2m-1})] \qquad (6)$$

where $h = \frac{b-a}{2m}$. This formula is known as *Simpson's formula* (*Simpson's rule*).

If we rewrite (5) as

$$\frac{1}{2h}\int\limits_{x_0}^{x_0+2h} f(x)\,dx \approx \frac{2}{3}y_1 + \frac{y_0+y_2}{6}$$

then we find that Simpson's formula is based on the approximate replacement of the mean value $\bar{y}$ of the function $y(x)$ on the interval from x_0 to $x_0 + 2h$ by

$$\bar{y} \approx \frac{2}{3}y_1 + \frac{y_0+y_2}{6} \tag{7}$$

The trapezoidal formula would have yielded

$$\bar{y} \approx \frac{1}{2}y_1 + \frac{y_0+y_2}{4} \tag{8}$$

(check this). To verify the type of formula, put $y(x) \equiv k =$ constant, whence $\bar{y} = k$. Then both formulas become exact. From the derivation of the formulas it follows that both (7) and (8) are exact for linear functions (of the type $y = px + n$), and formula (7) is exact also for quadratic functions (of the type $y = px^2 + nx + l$). It is interesting to note that (7) is actually also exact for cubic functions ($y = px^3 + nx^2 + lx + k$).

If the interval is partitioned into the same number of subintervals, Simpson's formula is as a rule much more exact than the trapezoidal rule.

Let us consider some examples.

1. Find $I = \int\limits_0^1 \frac{dx}{1+x}$.

Partition the interval of integration into two subintervals and evaluate the integral using the trapezoidal rule and then Simpson's rule.

We find the necessary values of the integrand:

x	0	0.5	1
y	1	0.6667	0.5

The trapezoidal rule yields

$$I = \frac{1}{2}\left(\frac{1+0.5}{2} + 0.6667\right) = 0.7083$$

Simpson's rule yields

$$I = \frac{1}{6}(1 + 4\cdot 0.6667 + 0.5) = 0.6944$$

Since $I = \ln(1+x)\Big|_0^1 = \ln 2 = 0.6931$ (to four decimal places), it follows that Simpson's rule yielded an error of roughly 0.2% and the trapezoidal rule 2%.

2. Find $I = \int\limits_0^1 \frac{\ln(1+x)}{1+x^2}\,dx$. (Note that $\int \frac{\ln(1+x)}{1+x^2}\,dx$ cannot be expressed in terms of elementary functions.) Partition the interval as in Example 1, then find the required values of the integrand:

x	0	0.5	1
y	0	0.324	0.346

By the trapezoidal rule we get $I = \frac{1}{2}\left(\frac{0.346}{2} + 0.324\right) = 0.248$.

By Simpson's rule we have $I = \frac{1}{6}(0.324 \cdot 4 + 0.346) = 0.274$. The exact value of the integral (to three significant places) is 0.272. Therefore, the trapezoidal rule yielded an error of about 10%, while Simpson's rule was less than 1% in error.

The advantage of Simpson's rule is particularly evident as the number n of subintervals in the partition increases. In that case we can say that the error of the trapezoidal rule decreases in inverse proportion to n^2, while the error of Simpson's rule decreases in inverse proportion to n^4 (if the third derivative of the integrand remains finite on the interval of integration).

The foregoing examples show how quickly and simply numerical methods permit finding integrals. It often happens that even when we can find an exact formula for the integral, it is easier to evaluate it by numerical integration.

The great Italian physicist Enrico Fermi, builder of the first atomic pile, preferred (as his work notes show) to evaluate an integral numerically even if it could be expressed in terms of logarithms and arc tangents.

Exercises

1. Evaluate the following integrals using Simpson's rule and the trapezoidal rule, the interval of integration being partitioned into four subintervals:

(a) $\int\limits_1^3 \frac{dx}{x}$, (b) $\int\limits_0^2 e^{-x^2}\,dx$, (c) $\int\limits_0^\pi \frac{\sin x}{x}\,dx$

Carry out the computations to within slide-rule accuracy. The values of the exponential function may be taken from any mathematics handbook, say *A Guide-Book to Mathematics* by Bronstein and Semendyayew [3].

2. Find $\int_0^3 \frac{x\,dx}{e^x+1}$ by Simpson's rule with the interval of integration partitioned into 4 subintervals. Carry the calculations to four decimal places. As a check, evaluate the same integral with a partition into 6 subintervals.
3. Verify directly that formula (8) is exact for first-degree polynomials, and formula (7) is exact for second-degree polynomials. *Hint.* First verify the exactness of these formulas for the function $y = x$, and of (7) for the function $y = x^2$.

1.2 Computing sums by means of integrals

Consider the sums of several numbers obtained by a definite law, say,

$$S_6 = 1 + 2 + 3 + 4 + 5 + 6$$

or

$$S_{96} = \sqrt{5} + \sqrt{6} + \sqrt{7} + \dots + \sqrt{99} + \sqrt{100}$$

or

$$S_{21} = 1 + \frac{1}{1.1} + \frac{1}{1.2} + \frac{1}{1.3} + \dots + \frac{1}{2.9} + \frac{1}{3.0}$$

Denote the sum by S. The subscript indicates the number of terms in the sum. Each such sum may be written more compactly if we are able to establish a relationship between the magnitude of a separate term in the sum and the number of that term. In the first case, the term a_n is simply equal to the number n, which varies from 1 to 6, and so the first sum is

$$S_6 = \sum_{n=1}^{6} a_n = \sum_{n=1}^{6} n$$

It is sometimes more convenient to start with zero; then we have to put $n - 1 = m$, that is, $n = m + 1$, which yields

$$S_6 = \sum_{m=0}^{5} (m + 1)$$

Instead of m we could write n or any other letter since the summation index is a *dummy index*, that is to say, one in which the sum does not depend on the letter used to indicate it. (In this respect, the summation index is similar to the variable of integration in a definite integral;

see HM, Sec. 2.8.) Note that in the last sum the index varies from 0 to 5 and the number of terms is accordingly equal to six.

We can write the second sum in similar fashion, thus:

$$S_{96} = \sum_{p=5}^{100} \sqrt{p}$$

However it is better if the index varies from 1 or 0; to do this, put $p - 4 = m$ or $p - 5 = n$ and we get

$$S_{96} = \sum_{m=1}^{96} \sqrt{m+4} = \sum_{n=0}^{95} \sqrt{n+5}$$

The third sum is of the form

$$S_{21} = \sum_{n=0}^{20} \frac{1}{1+0.1n}$$

We see that it may be regarded as the sum of values of the function $f(x) = \frac{1}{x}$ when x assumes values from $x = 1$ to $x = 3$ at intervals of 0.1 (this approach will be found to be useful in the sequel). In other cases as well, one often finds it convenient to view a given sum as the sum of values of a certain function $f(x)$ for x varying from $x = a$ to $x = b$ with a constant interval $\Delta x = h$.

The foregoing sums are not hard to compute, but if a sum consists of a large number of terms, computation can be a wearisome job. What is more, one often has to compare sums based on the same law but with different numbers of terms. In this case it is desirable to know not only the sum for a certain definite number of terms but also how the sum depends on the number of terms involved.

In many cases it is possible to approximate a sum with the aid of integrals. Recall that the study of definite integrals begins with an approximate representation of the integral as the sum of a finite number of terms. It is then established that exact integration of many functions by means of indefinite integrals is quite an easy affair.

Let us take advantage of this for computing sums.

Write down the trapezoidal formula that gives an approximate expression of an integral in the form of a sum. If the interval of integration between $x = a$ and $x = b$ is partitioned into n equal subintervals of length $\frac{b-a}{n} = h$ each, then

$$\int_a^b f(x)\,dx \approx h \cdot \left[\frac{1}{2} f(a) + f(a+h) + f(a+2h) + \right.$$
$$\left. \ldots + f(a+(n-1)\,h) + \frac{1}{2} f(a+nh)\right]$$

where $a + nh = b$.

From this formula we obtain the following expression for the sum:

$$f(a) + f(a + h) + f(a + 2h) + \ldots + f(b) \approx \frac{1}{h}\int_a^b f(x)\,dx + \frac{1}{2}f(a) + \frac{1}{2}f(b) \qquad (9)$$

Dropping the term $\frac{1}{2}f(a) + \frac{1}{2}f(b)$ in the right member of (9) — which is justified if

$$\frac{1}{h}\int_a^b f(x)\,dx \gg \frac{1}{2}f(a) + \frac{1}{2}f(b)$$

that is, if the number of terms of the sum is great and the function $f(x)$ does not vary too fast — we obtain a cruder formula:

$$f(a) + f(a + h) + f(a + 2h) + \ldots + f(b) \approx \frac{1}{h}\int_a^b f(x)\,dx \qquad (10)$$

If h is small, we can take advantage of the approximate formulas

$$\frac{1}{2}f(a) \approx \frac{1}{h}\int_{a-\frac{1}{2}h}^{a} f(x)\,dx, \quad \frac{1}{2}f(b) \approx \frac{1}{h}\int_{b}^{b+\frac{1}{2}h} f(x)\,dx$$

Then from (9) we get

$$f(a) + f(a + h) + f(a + 2h) + \ldots + f(b) \approx \frac{1}{h}\int_a^b f(x)\,dx + \frac{1}{h}\int_{a-\frac{1}{2}h}^{a} f(x)\,dx + \frac{1}{h}\int_{b}^{b+\frac{1}{2}h} f(x)\,dx$$

or

$$f(a) + f(a + h) + f(a + 2h) + \ldots + f(b) \approx \frac{1}{h}\int_{a-\frac{h}{2}}^{b+\frac{h}{2}} f(x)\,dx \qquad (11)$$

In deriving (10) and (11), we proceeded from formula (9), which was obtained from the trapezoidal formula, which itself is approximate. For this reason, although (9) and (11) are more exact than (10), they are still approximate formulas. The accuracy of the formulas increases as we diminish h and increase the number of terms in the sum.

Let us apply the formulas just obtained to computing the sums given at the beginning of this section.

For the sum $S_6 = 21$, $f(x) = x$, $h = 1$. By (10) we get

$$\int_1^6 x\,dx = \frac{1}{2}x^2\Big|_1^6 = 17.5$$

The error is 17%. By formula (9) we get

$$S_6 \approx \int_1^6 x\,dx + \frac{1}{2} + \frac{6}{2} = 21$$

And finally by (11) we have

$$S_6 \approx \int_{0.5}^{6.5} x\,dx = \frac{(6.5)^2}{2} - \frac{(0.5)^2}{2} = 21$$

Formulas (9) and (11) yielded the exact result. This is because in the case at hand the function $f(x)$ is linear (see Exercise 3).

For the next sum, $S_{96} = \sqrt{5} + \sqrt{6} + \dots + \sqrt{100}$, taking into account that $f(x) = \sqrt{x}$, $h = 1$, we get by (10)

$$\int_5^{100} \sqrt{x}\,dx = \frac{2}{3}x^{3/2}\Big|_5^{100} = \frac{2}{3}(1000 - 5\sqrt{5}) = 659.2$$

By formula (9) we get $S_{96} \approx 665.3$ and by (11) $S_{96} \approx 665.3$. The sum S_{96} computed to an accuracy of four significant figures is also 665.3.

For the last sum $f(x) = \frac{1}{x}$, $h = 0.1$. And so by (10) we have

$$S_{21} \approx \frac{1}{0.1}\int_1^3 \frac{dx}{x} = 10 \ln 3 = 10.99$$

By (9) we get $S_{21} \approx 11.66$ and by (11) obtain $S_{21} \approx 11.51$. Computed directly, to two decimal places, the sum $S_{21} \approx 11.66$.

In the foregoing examples, the formulas (9) and (11) give very satisfactory accuracy and so they should be used in practical situations. The crude formula (10) may be useful to formulate the law of an increasing sum when the number of terms increases without bound. However, in all numerical computations with a definite number of terms in the sum, one should use (11), which is not any more complicated than (10) but is much more accurate*.

In all the cases considered above, the separate terms of the sum had the same signs.

If the terms have different signs, then to compute the sum via an integral, one should consider the integral of a function that changes sign several times over the interval of integration. For such integrals the trapezoidal rule yields very poor results. Therefore, the formulas we have derived serve only for sums whose terms are of the same sign and cannot be applied to sums in which the terms have different signs. For example, these formulas cannot be used to find sums in which the signs of the summands alternate. Sums of this kind are called *alternating*. An example of an alternating sum is

$$S_8 = \frac{1}{4} - \frac{1}{5} + \frac{1}{6} - \frac{1}{7} + \frac{1}{8} - \frac{1}{9} + \frac{1}{10} - \frac{1}{11}$$

How do we find the sum here?

If the terms of an alternating sum decrease (or increase) in absolute value, so that the absolute value of each successive term is less than (or more than) the absolute value of the preceding term, we can use the following device.

Combine adjacent positive and negative terms in pairs. Then the problem reduces to computing a sum in which the signs of all terms are the same.

For example,

$$S_8 = \frac{1}{4} - \frac{1}{5} + \frac{1}{6} - \frac{1}{7} + \frac{1}{8} - \frac{1}{9} + \frac{1}{10} - \frac{1}{11} = \left(\frac{1}{4} - \frac{1}{5}\right) + \left(\frac{1}{6} - \frac{1}{7}\right)$$
$$+ \left(\frac{1}{8} - \frac{1}{9}\right) + \left(\frac{1}{10} - \frac{1}{11}\right) = \frac{1}{4\cdot 5} + \frac{1}{6\cdot 7} + \frac{1}{8\cdot 9} + \frac{1}{10\cdot 11}$$

We will assume that the original alternating sum consists of an even number of terms. Then if it begins with a positive term, the final term is negative. Such a sum may be written thus:

$$S_{n+1} = f(a) - f(a + h) + f(a + 2h) - f(a + 3h) + \dots - f(b) \quad (12)$$

where $b = a + nh$.

The difference between two adjacent terms of this sum can be written as

$$f(a + kh) - f(a + (k + 1)\,h) \approx -h \left.\frac{df}{dx}\right|_{x=a+\left(k+\frac{1}{2}\right)h}$$
$$= -h\cdot f'\left(a + \left(k + \frac{1}{2}\right)h\right)$$

[*] The use of (10) involves an obvious systematic error due to the dropping of two terms in the derivation of the formula. If all the terms are the same, that is, $f(x) \equiv A =$ constant and $(b - a)/h = n$, then the sum at hand is equal to $(n + 1)A$, whereas (10) only yields nA (in this case, (11) yields an exact value).

This is an approximate equation, but it is the more accurate, the smaller h is.* The sum (12) assumes the form

$$S_{n+1} \approx -h\left[f'\left(a+\frac{h}{2}\right)+f'\left(a+\frac{5}{2}h\right)+\ldots+f'\left(b-\frac{h}{2}\right)\right] \quad (13)$$

Let us apply (11) to the right-hand member. Observe that in (11) h is the difference between adjacent values of the independent variable. In formula (13) this difference is equal to $a+\left(k+\frac{1}{2}\right)h-\left[a+\left(k-\frac{3}{2}\right)h\right]=2h$. For this reason, when using (11) we have to take $2h$ instead of h; we then get

$$S_{n+1} \approx -h\cdot\frac{1}{2h}\int\limits_{a+\frac{h}{2}-h}^{b-\frac{h}{2}+h} f'(x)\,dx = -\frac{1}{2}f(x)\Big|_{a-\frac{h}{2}}^{b+\frac{h}{2}}$$

$$=\frac{f\left(a-\frac{h}{2}\right)-f\left(b+\frac{h}{2}\right)}{2} \quad (14)$$

Let us use this formula to compute the sum

$$S_8=\frac{1}{4}-\frac{1}{5}+\frac{1}{6}-\frac{1}{7}+\frac{1}{8}-\frac{1}{9}+\frac{1}{10}-\frac{1}{11}$$

Here

$$f(x)=\frac{1}{x},\quad a=4,\ b=11,\ h=1$$

and so we have

$$S_8\approx\frac{1}{2}\cdot\left(\frac{1}{3.5}-\frac{1}{11.5}\right)=0.0994$$

Direct summing gives $S_8 = 0.0968$ (with an error of less than 3%).

* If we put $a + kh = x$ and change signs, the last equation can be rewritten as $f(x+h)-f(x)\approx hf'\left(x+\frac{h}{2}\right)$. It is easy to estimate its accuracy. Indeed, by Taylor's formula,

$$f(x+h)-f(x)=f'(x)h+\frac{f''(x)}{2}h^2+\frac{f'''(x)}{6}h^3+\ldots$$

$$hf'\left(x+\frac{h}{2}\right)=h\left[f'(x)+f''(x)\frac{h}{2}+\frac{f'''(x)}{2}\left(\frac{h}{2}\right)^2+\ldots\right]$$

Thus, the expansions differ only beginning with the third order terms when compared with h.

If an alternating sum has an odd number of terms, then the first term $f(a)$ and the last term $f(b)$ are of the same sign. In this case, we first use (14) to find the sum without the last term $f(b)$, and then add that term:

$$S_{n+1}=f(a)-f(a+h)+f(a+2h)-\dots+f(b)$$
$$\approx\frac{1}{2}\left[f\left(a-\frac{h}{2}\right)-f\left(b-\frac{h}{2}\right)\right]+f(b)$$

If h is small, then to compute $f\left(b\pm\frac{h}{2}\right)$ we can confine ourselves to the first two terms of Taylor's series:

$$f(x)=f(b)+f'(b)\cdot(x-b)+\dots$$

Putting $x=b-\frac{h}{2}$ here, we find $f\left(b-\frac{h}{2}\right)\approx f(b)-\frac{h}{2}f'(b)$, and putting $x=b+\frac{h}{2}$, we get $f\left(b+\frac{h}{2}\right)\approx f(b)+\frac{h}{2}f'(b)$. And so

$$f(b)-\frac{1}{2}f\left(b-\frac{h}{2}\right)\approx f(b)-\frac{1}{2}f(b)+\frac{1}{4}hf'(b)$$
$$=\frac{1}{2}\left[f(b)+\frac{h}{2}f'(b)\right]\approx\frac{1}{2}f\left(b+\frac{h}{2}\right)$$

The final sum for the case of an odd number of terms is

$$S_{n+1}\approx\frac{f\left(a-\frac{h}{2}\right)+f\left(b+\frac{h}{2}\right)}{2} \qquad (15)$$

This formula is no less exact than (14), with the aid of which it was obtained. Consider the following example:

$$S_9=\frac{1}{4}-\frac{1}{5}+\frac{1}{6}-\frac{1}{7}+\frac{1}{8}-\frac{1}{9}+\frac{1}{10}-\frac{1}{11}+\frac{1}{12}$$

Using (15) we get $S_9\approx\frac{1}{2}\left(\frac{1}{3.5}+\frac{1}{12.5}\right)=0.1829$, whereas direct summation yields $S_9=0.1801$. The error amounts to roughly 1.5%. Observe that, as can be seen from (14) and (15), the magnitude of an alternating sum is of the same order as the separate summands. For this reason, adding another term makes an essential change in the value. Taking the foregoing example, we find that S_9 is almost twice S_8.

Note the essential difference between an alternating sum and a sum with terms of the same sign. We will increase the number of terms in the sum so that the first and last terms remain unaltered and the law of formation of the sum is the same. To do this, we reduce the difference between adjacent terms of the sum, that is, we reduce the size of h. Thus, from the sum $1 + \frac{1}{2} + \frac{1}{3}$ we can obtain the sum $1 + \frac{1}{1.1} + \frac{1}{1.2} + \frac{1}{1.3} + \dots + \frac{1}{3}$ or $1 + \frac{1}{1.01} + \frac{1}{1.02} + \dots + \frac{1}{3}$.

If all the terms of the sum have the same sign, then the sum is roughly proportional to the number of terms. This is evident, say, in the case of formula (10). Indeed, the integral in the right-hand member of (10) is preceded by a factor $1/h$, where h is inversely proportional to the number of terms in the sum. Therefore, in the process just described the magnitude of the sum increases indefinitely with growth in the number of terms.

In the case of an alternating sum with an even number of summands, the magnitude (with the number of terms increasing as described) approaches a definite number that is independent of the number of terms in the sum, namely,

$$S = \frac{f(a) - f(b)}{2} \tag{16}$$

This is readily discernible in formula (14), since for a large number of terms h will be close to zero and for this reason $f\left(a - \frac{1}{2}h\right) \approx f(a)$, $f\left(b + \frac{1}{2}h\right) \approx f(b)$. The case is similar for an odd number of terms; from (15) we obtain in the limit a different value, namely:

$$S = \frac{f(a) + f(b)}{2} \tag{17}$$

Observe that when the number of summands is small, that is, when h is great, the simplified formulas (16) and (17) are much worse than (14) and (15). Let us consider an example. Suppose $S = 1 - 2 + 3 - 4 = -2$. Using the simplified formula (16), we get $S \approx -\frac{3}{2}$ (an error of 25%), whereas formula (14) yields

$$S \approx \frac{1}{2}\left[\left(1 - \frac{1}{2}\right) - \left(4 + \frac{1}{2}\right)\right] = -2$$

which is exact.

The expressions for the sums that we have just obtained are approximate expressions, and their accuracy increases when adjacent terms are closer together, that is, when h decreases.

Exercises

1. Find the sum of $1 + \sqrt[3]{2} + \sqrt[3]{3} + \dots + \sqrt[3]{n}$ using formula (11). Compare with the results obtained by direct summation for $n = 3, 4, 5$.

2. Find the sum of

$$1+\frac{1}{2\sqrt{2}}+\frac{1}{3\sqrt{3}}+\dots+\frac{1}{20\sqrt{20}}$$

3. Prove that formulas (9) and (11) are absolutely exact if $f(x)$ is a linear function. The terms of the sum then form an arithmetic progression.

1.3 Numerical solution of equations

In practical computations one rather often has to deal with the numerical solution of equations like

$$f(x)=0 \tag{18}$$

where f is a given function. These equations may be algebraic or transcendental (that is, nonalgebraic, for instance, trigonometric and the like). They are sometimes grouped together and termed *finite*, in contrast to, say, differential equations. Today there are a large number of methods for solving the various classes of equations of type (18). We consider here only three of the more universal methods that find extensive applications in other divisions of mathematics as well.

One ordinarily begins with a rough and very approximate solution, the so-called *zero-order approximation*. If a physics problem is being tackled, this rough solution may be obtained from the physical essence of the problem itself. One way is to sketch a graph of the function $f(x)$ and obtain a rough solution by marking the point of intersection of the graph and the x-axis.

Suppose we have such a rough solution x_0. Then denote the exact solution, still unknown, by $\overline{x}=x_0+h$. Using Taylor's formula, we get

$$f(x_0+h)=f(x_0)+\frac{h}{1!}f'(x_0)+\dots \tag{19}$$

But the left-hand member must be equal to zero. Dropping the dots, that is to say, the higher-order terms, we get

$$f(x_0)+hf'(x_0)\approx 0,$$

$$h\approx -\frac{f(x_0)}{f'(x_0)}$$

and so

$$\overline{x}\approx x_0-\frac{f(x_0)}{f'(x_0)}$$

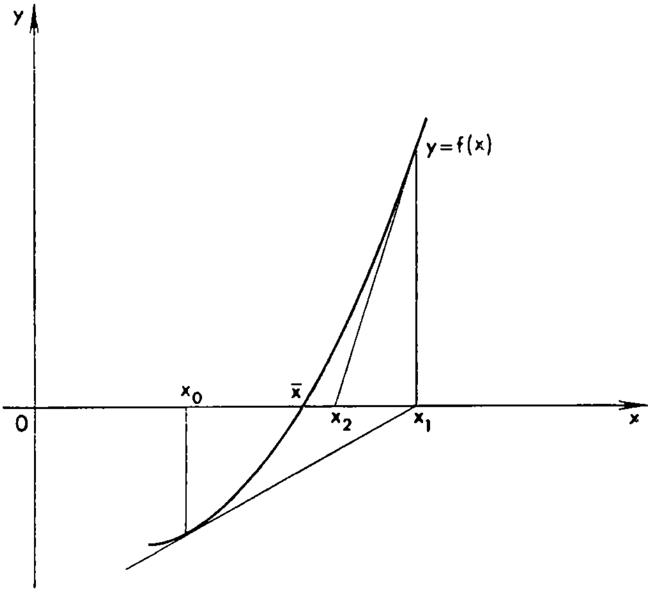


Fig. 3

Denoting the right member by x_1

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \tag{20}$$

we get the "first approximation" (the subscript indicates the number of the approximation). Doing the same thing again, we get a "second approximation":

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

and so on. Since rejecting higher terms in (19) is equivalent to replacing the graph of the function $f(x)$ by a tangent to it at $x = x_0$, the geometrical meaning of this method consists in a sequential construction of tangents to the graph and finding the points where they intersect the x-axis (Fig. 3). It is clear that the successive approximations rapidly converge to the desired solution, provided of course that the zero-order approximation was not too far away.

This is called *Newton's method* (or the *method of tangents*, in accord with its geometrical meaning). The convenience of this method is that one only needs to compute the values of the function $f(x)$ and its derivative and perform arithmetical operations, which is a simple matter in the case of a function specified by a formula.

Let us take an example. Suppose we have to solve the equation

$$x^3 - 3x - 1 = 0 \tag{21}$$

Since the left-hand side is equal to -3 for $x = 1$ and to 1 for $x = 2$, there is a root lying between $x = 1$ and $x = 2$ and it is naturally closer to 2 than to 1. So we take $x_0 = 2$. Then (20) yields

$$x_1 = 2 - \left(\frac{x^3 - 3x - 1}{3x^2 - 3}\right)_{x=2} = 1.889$$

Similarly, we get (to three decimal places)

$$x_2 = 1.879, \quad x_3 = 1.879$$

And so, to the accuracy given, the solution is $\bar{x} = 1.879$.

It is interesting to note that in this example we could have written the exact solution to the equation. Already in the 16th century the Italian mathematician Cardano published a formula for the solution of the cubic equation

$$x^3 + px + q = 0 \quad *$$

which has the form

$$x = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}$$

However, if we substitute into this formula the values of the coefficients of equation (21), then under the cube root symbol we find the imaginary numbers $0.5 \pm i\sqrt{0.75}$ and it is only the sum of these roots that is real. Which means we have to extract the roots of imaginary numbers (cf. Sec. 5.4). Thus, even in this simple example Newton's method proves to be much simpler than the "exact" formula, so what is there left to say about quartic equations where the exact formula is so unwieldy that it is not even written out in full in reference books! Or of equations higher than the fourth degree and also of most of the transcendental equations where there is no "exact" formula at all! For such equations the advantage of numerical methods is particularly obvious.

Newton's method is one of a group of *iteration methods* (or *methods of successive approximation*) in which a unified procedure is repeated in succession (iterated) to yield ever more exact approximate solutions. In general form, the iteration method as applied to equation (18) looks like this: the equation is written in the equivalent form

$$x = \varphi(x) \tag{22}$$

* The general equation of the third degree $ay^3 + by^2 + cy + d = 0$ is reduced to this form by the substitution $y = x - \frac{b}{3a}$.

Then a value $x = x_0$ is chosen for the zero-order approximation, and the subsequent approximations are computed from the formulas $x_1 = \varphi(x_0)$, $x_2 = \varphi(x_1)$, ..., and, generally,

$$x_{n+1} = \varphi(x_n) \tag{23}$$

Here two cases are possible:

(1) the process can *converge*, that is, the successive approximations tend to a finite limit $\overline{x}$; then, passing to the limit in (23) as $n \to \infty$, we see that $x = \overline{x}$ is a solution of equation (22);

(2) the process can *diverge*, that is, there is no finite limit to the constructed "approximations"; from this it does not follow that there is no solution to (22); it might simply mean that the chosen iteration process is not suitable.

We illustrate this in a simple equation that does not require any "science" to solve:

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

with the obvious solution $\overline{x} = 2$. If we put $x_0 = 0$ and compute to three decimal places, we get

$$x_1 = 1;\ x_2 = 1.5;\ x_3 = 1.75;\ x_4 = 1.875;\ x_5 = 1.938;$$
$$x_6 = 1.969;\ x_7 = 1.984;\ x_8 = 1.992;\ x_9 = 1.996;$$
$$x_{10} = 1.998;\ x_{11} = 1.999;\ x_{12} = 2.000;\ x_{13} = 2.000$$

which shows the process actually did converge. The limit can be found more quickly if one notices that in the given case the differences between successive approximations form a geometric progression with first term $a = x_1 - x_0$ and ratio $q = \frac{x_2 - x_1}{x_1 - x_0}$. Therefore the sum of the entire progression, that is, $\overline{x} - x_0$, is

$$\frac{a}{1-q} = \frac{x_1 - x_0}{1 - \frac{x_2 - x_1}{x_1 - x_0}} = \frac{(x_1 - x_0)^2}{2x_1 - x_0 - x_2}$$

whence

$$\overline{x} = x_0 + \frac{(x_1 - x_0)^2}{2x_1 - x_0 - x_2} = \frac{x_1^2 - x_0 x_2}{2x_1 - x_0 - x_2} \tag{25}$$

In more complicated examples the succesive differences are only reminiscent of a geometric progression; in such cases, formula (25) does not yield an exact solution, but does enable one to skip over a few approximations and obtain an approximation from which it is again possible to begin iteration. (This is *Aitken's method*).

If (24) is solved for the x in the right-hand member, then we can rewrite the equation in the equivalent form

$$x = 2x - 2 \tag{26}$$

and begin with $x_0 = 0$, then we successively get $x_1 = -2$, $x_2 = -6$, $x_3 = -14$. Here the process will not converge. This could have been foreseen by noting that (23) implies

$$x_{n+1} - x_n = \varphi(x_n) - \varphi(x_{n-1})$$

that is,

$$\begin{aligned} x_2 - x_1 &= \varphi(x_1) - \varphi(x_0), \\ x_3 - x_2 &= \varphi(x_2) - \varphi(x_1), \\ x_4 - x_3 &= \varphi(x_3) - \varphi(x_2), \\ &\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

Thus, if the values of the function $\varphi(x)$ vary more slowly than the values of its argument, the distances between successive approximations will become smaller and smaller, and if the values of the function $\varphi(x)$ vary faster than those of the argument, the distances between successive approximations will progressively increase and the process will diverge. Since the rate of change of the values of $\varphi(x)$ relative to x is equal to $\varphi'(x)$, we find that if

$$|\varphi'(x)| \leqslant k < 1 \tag{27}$$

then the process of iteration converges and it converges the faster, the smaller k is; but if

$$|\varphi'(x)| > 1$$

then the process diverges. These inequalities must hold for all x or at least near the desired root $\bar{x}$ of (22).

We see that the equations (24) and (26) are completely equivalent but they generate distinct iteration processes. (In order to explain this difference, there is no need even to rewrite (24) as (26); it suffices to notice that in order to obtain the next approximation, one has to substitute the subsequent approximation, via the first method, into the right member of (24) and, via the second method, into the left member of the same equation.) In other cases as well, equation (18) can be rewritten in the form (22) in many ways, each of which generates its own iteration method. Some of the methods may prove to be rapidly convergent and therefore more convenient, while others will turn out slowly convergent, and still others will even be divergent.

The above-mentioned solution of (21) can now be understood as follows: the equation is rewritten in the equivalent form

$$x = x - \frac{x^3 - 3x - 1}{3x^2 - 3}$$

after which the iteration method is applied beginning with $x_0 = 2$.

In general form, Newton's method for equation (18) reduces (see (20) and the subsequent equations) to this equation being rewritten in the equivalent form

$$x = x - \frac{f(x)}{f'(x)} \tag{28}$$

with the iteration method applied afterwards. This form might appear to be somewhat artificial, although it is easy to demonstrate the equivalence of the equations (18) and (28): from (18) follows (28) and conversely, but it is not immediately apparent how the denominator $f'(x)$ helps. However it is easy to verify that the derivative of the right member, that is,

$$\left(x - \frac{f}{f'}\right)' = 1 - \frac{f'f' - ff''}{f'^2} = \frac{ff''}{f'^2}$$

vanishes when $x = \overline{x}$, where $\overline{x}$ is the solution of (18). Hence (see the reasoning involving the estimate (27)), the closer the successive approximations approach $\overline{x}$, the faster the process converges.

What is more, since, when (27) holds, the iterations do not converge more slowly than a progression with ratio k, it follows that Newton's method converges faster than a geometric progression with any ratio! *

Let us now take up the so-called *perturbation method,* which, like the method of iterations, is one of the most universal methods in applied mathematics. A simple example will serve to illustrate this method.

Suppose it is required to find the solution of the transcendental equation

$$e^{x-1} = 2 - x + \alpha \tag{29}$$

for small $|\alpha|$. Notice that for $\alpha = 0$ a solution can be found by inspection: $x = 1$. Therefore if a solution of (29) dependent on α is sought by means of a series expansion in powers of α,

$$x = x_0 + a\alpha + b\alpha^2 + c\alpha^3 + \dots$$

* A simple typical example will make it easy to establish the rate of this convergence. Suppose we are considering the approximations by Newton's method to the zeroth root of the equation $x + x^2 = 0$. These approximations are connected by the relation

$$x_{n+1} = x_n - \frac{x_n + x_n^2}{1 + 2x_n} = \frac{x_n^2}{1 + 2x_n} < x_n^2$$

Suppose $0 < x_0 < 1$. Then we successively get

$$x_1 < x_0^2, \quad x_2 < x_1^2 < x_0^4, \quad x_3 < x_2^2 < x_0^8, \dots, x_n < x_0^{2^n}, \dots$$

It can be demonstrated that in the general case as well the rate of convergence of Newton's method is of the order of $a^{2^n} (0 < a < 1)$.

then by substituting $\alpha = 0$ we find that x_0 must be equal to 1. Now substitute the expansion

$$x = 1 + a\alpha + b\alpha^2 + c\alpha^3 + \ldots \tag{30}$$

into both members of (29) and take advantage of the familiar Taylor expansion of an exponential function to get

$$1 + \frac{a\alpha + b\alpha^2 + c\alpha^3 + \ldots}{1!} + \frac{(a\alpha + b\alpha^2 + c\alpha^3 + \ldots)^2}{2!}$$
$$+ \frac{(a\alpha + b\alpha^2 + c\alpha^3 + \ldots)^3}{3!} + \ldots = 2 - (1 + a\alpha + b\alpha^2 + c\alpha^3 + \ldots) + \alpha$$

Removing brackets and retaining terms up to α^3, we get

$$1 + a\alpha + b\alpha^2 + c\alpha^3 + \frac{a^2}{2}\alpha^2 + ab\alpha^3 + \frac{a^3}{6}\alpha^3 + \ldots$$
$$= 1 - a\alpha - b\alpha^2 - c\alpha^3 - \ldots + \alpha$$

Equating the coefficients of equal powers of α in both members, we get the relations

$$a = -a + 1,$$
$$b + \frac{a^2}{2} = -b,$$
$$c + ab + \frac{a^3}{6} = -c,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

whence we successively find

$$a = \frac{1}{2},\ b = -\frac{1}{16},\ c = \frac{1}{192},\ \ldots$$

Substituting these values into (30), we obtain the desired solution to (29) in the form of the series

$$x = 1 + \frac{\alpha}{2} - \frac{\alpha^2}{16} + \frac{\alpha^3}{192} + \ldots \tag{31}$$

that converges very well for small $|\alpha|$.

In the general case, the perturbation method is applied in the following manner. In the statement of a certain problem, let there be a certain parameter α in addition to the basic unknown quantities, and for $\alpha = 0$, let the problem be more or less easy to solve (the *unperturbed solution*). Then the solution of the problem for α close to 0 can in many cases be obtained by an expansion in powers of α to a certain degree of accuracy. Here, the first term of the expansion not containing α is obtained when $\alpha = 0$, which means it yields an unperturbed solution. Subsequent terms yield corrections for "perturbation" of the solution. These corrections are of first, second, etc. orders with respect to α. The terms themselves are often found by the *method of undetermined coefficients*, that is, the coefficients of α, α^2 and so on are designated by certain letters which are then found from the conditions given in the statement of the problem.

The perturbation method yields good results only for small $|\alpha|$. For large $|\alpha|$ the method can lead to fundamental errors because it may turn out that rejected terms (those not written down) are more essential than the ones that are left.

Thus, the perturbation method makes it possible, by proceeding from the solution of certain "central" problems, to obtain a solution of problems whose statement is close to the "central" ones, if of course the formulation does not involve any fundamental, qualitative, change in the solution.

In many problems, the very aspect of the first-order term enables one to draw some useful conclusions concerning the dependence of the solution upon the parameter in the case of a small variation of the parameter.

The perturbation method is directly connected with the iteration method; this we demonstrated in the case of equation (29). First of all, it is convenient that the unperturbed solution be the zeroth solution. This is attained by means of the substitution $x = 1 + y$, whence

$$e^y = 1 - y + \alpha \tag{32}$$

To carry out iterations, observe that if the desired root $\bar{y}$ of the equation is close to 0, the equation can be conveniently rewritten in a form $y = \varphi(y)$ such that the expansion of $\varphi(y)$ in powers of y does not contain the first power but only the constant term, and terms containing y^2, y^3 and so on. Then the expansion of $\varphi'(y)$ will not contain the constant term and so for y close to $\bar{y}$ and, hence, to 0, $|\varphi'(y)|$ will assume small values, and this yields the convergence of the iteration method. Therefore, in the original equation, transpose all terms involving the first power of y to the left-hand member, and all other terms to the right-hand member, and then perform the iterations.

In our example, the left member of equation (32), e^y, by virtue of Taylor's formula,

$$e^y = 1 + y + \frac{y^2}{2} + \frac{y^3}{6} + \dots$$

contains the term y, and so it must be written as

$$e^y = y + (e^y - y)$$

the expansion of the parenthesis no longer containing y in the first power. And so in place of (32) we write

$$y + (e^y - y) = 1 - y + \alpha$$

whence, grouping terms with y (but not removing the brackets!), we arrive at the equation

$$y=\frac{1}{2}+\frac{\alpha}{2}-\frac{1}{2}(e^{y}-y)$$

This equation, which is equivalent to (32), is already prepared for iteration. Expanding the right-hand side in powers of y, we get

$$y=\frac{\alpha}{2}-\frac{y^2}{4}-\frac{y^3}{12}-\dots \qquad (33)$$

Now we can carry out the iterations beginning with $y_0=0$. Substituting this value into the right member of (33), we get

$$y_1=\frac{\alpha}{2} \qquad (34)$$

Substituting $y=y_1$ into the right member of (33), we obtain

$$y_2=\frac{\alpha}{2}-\frac{\alpha^2}{16}-\frac{\alpha^3}{96}-\dots \qquad (35)$$

We see that the first approximation (34) coincides with the exact solution (31) up to terms of the first order, the second approximation (35), to within second-order terms. To construct the third approximation we can simply substitute into the right member of (33)

$$y=\frac{\alpha}{2}-\frac{\alpha^2}{16}$$

We then have an expansion that coincides with the exact one up to third-order terms inclusive, so that only these terms need be taken into account in computations and so on. This enables one, when computing each subsequent iteration, to deal only with finite sums.

Exercises

1. Apply Newton's method to equation (21) beginning with the value $x_0=0$, $x_0=1$.
2. Rewriting (21) as $x=\frac{x^3-1}{3}$, apply the iteration method to it beginning with the value $x_0=0,\ 1,\ 2$.
3. Solve the equation $x^3+\alpha x-1=0$ for small $|\alpha|$ by the perturbation method.

ANSWERS AND SOLUTIONS

Sec. 1.1

1. By the trapezoidal rule: (a) 1.12, (b) 0.88, (c) 1.84. By Simpson's rule: (a) 1.10, (b) 0.88, (c) 1.85.

2. Partitioning the interval of integration into four subintervals, we find $\int_0^3 \frac{x\,dx}{e^x+1} = 0.6275$. Partitioning the interval of integration into 6 intervals, we also obtain $\int_0^3 \frac{x\,dx}{e^x+1} = 0.6275$.

3. Consider the formula (8) for the function $y = x$ over the interval from x_0 to $x_0 + 2h$. Here

$$\bar{y} = \frac{1}{2h}\int_{x_0}^{x_0+2h} x\,dx = \frac{1}{2h}\,\frac{(x_0+2h)^2 - x_0^2}{2} = \frac{4x_0h + 4h^2}{4h} = x_0 + h$$

whereas

$$\frac{1}{2}y_1 + \frac{y_0 + y_2}{4} = \frac{1}{2}(x_0 + h) + \frac{x_0 + (x_0 + 2h)}{4} = x_0 + h$$

The results coincide, that is, the formula in this case is exact. From this follows the exactness of (8) for the function $y = ax$ (a = constant) as well, since a serves as a common factor in all terms. Writing down the exact equation (8) for $y = ax$ and $y = k$ (which has already been verified) and adding the left members and the right members, we find that (8) holds true for the function $y = ax + k$ as well. Equation (7) is considered similarly.

Sec. 1.2

1. In our case $f(x) = \sqrt[3]{x}$, $h = 1$, $a = 1$, $b = n$. By (11) we get

$$S_n = \int_{1/2}^{n+1/2} \sqrt[3]{x}\,dx = \int_{1/2}^{n+1/2} x^{1/3}\,dx = \frac{3}{4}\,x^{4/3}\Big|_{1/2}^{n+1/2}$$

$$= \frac{3}{4}\left[\sqrt[3]{\left(n+\frac{1}{2}\right)^4} - \sqrt[3]{\frac{1}{16}}\right]$$

or

$$S_n = \frac{3}{4}\left[\left(n+\frac{1}{2}\right)\sqrt[3]{n+\frac{1}{2}} - \frac{1}{2\sqrt[3]{2}}\right]$$

For $n = 3$ the formula yields $S_3 = 3.34$; direct computation yields $S_3 = 3.70$. The error is 11%.
For $n = 4$ we obtain $S_4 = 4.93$ by the formula and $S_4 = 5.29$ by direct calculation. The error is 7%. For $n = 5$ we have by the formula $S_5 = 6.64$, by direct calculation $S_5 = 7.00$. The error is 5%.

2. By formula (11) we get $S = 2.39$.
3. Let $f(x) = px + k$, that is, the sum is of the form

$$S_{n+1} = [pa + k] + [p(a + h) + k] + [p(a + 2h) + k] + \dots + [pb + k]$$

where $b = a + nh$. We thus have an arithmetic progression with first term $a_1 = pa + k$ and difference $d = ph$. The sum of such a progression is, as we know, equal to

$$S_{n+1} = \frac{a_1 + a_{n+1}}{2}(n + 1) = \frac{(pa + k) + (pb + k)}{2}(n + 1) = \left(p\frac{a + b}{2} + k\right)(n + 1)$$

Apply, say, (9) to get the value

$$\frac{1}{h}\int_a^b (px + k)\,dx + \frac{1}{2}(pa + k) + \frac{1}{2}(pb + k)$$

$$= \frac{1}{h}\left[p\frac{b^2 - a^2}{2} + k(b - a)\right] + \left(p\frac{a + b}{2} + k\right)$$

$$= \frac{b - a}{h}\left(p\frac{a + b}{2} + k\right) + \left(p\frac{a + b}{2} + k\right)$$

$$= n\left(p\frac{a + b}{2} + k\right) + \left(p\frac{a + b}{2} + k\right)$$

$$= \left(p\frac{a + b}{2} + k\right)(n + 1)$$

which coincides with the exact value for the sum.

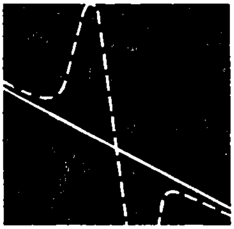
Sec. 1.3

1. For $x_0 = 0$ we obtain another root $\overline{\overline{x}} = -0.347$ of equation (21). For $x_0 = 1$ the method cannot be applied (the denominator vanishes).
2. For $x_0 = 0$ and $x_0 = 1$ we get convergence to the root $\overline{\overline{x}}$ from Exercise 1. For $x_0 = 2$ the iteration process diverges.
3. The expansion of the solution is of the form

$$x = \frac{1}{3}\alpha - \frac{1}{81}\alpha^3 + \dots$$

Chapter 2

MATHEMATICAL TREATMENT OF EXPERIMENTAL RESULTS



In practical work, it often happens that a relationship is obtained between variable quantities, say x and y, in the course of experimentation and measurements. Ordinarily, such a relationship is given in the form of a table with each value of x at which the measurement was taken indicated and paired with the appropriate measured value of y. In this chapter we first examine the general rules for operating with tables and then certain new aspects that arise when processing the results of experiments.

2.1 Tables and differences

Very often one has to consider functions given in tabular form, such as mathematical tables, say, tables of logarithms, sines, squares, etc. There are also tables of physical quantities taken from handbooks, such as tables of the boiling point of a liquid versus pressure, and the like. Finally, a relationship between variable quantities can be obtained in the form of the unprocessed results of experiments or measurements. In all such cases, the numerical values of the dependent variable are given, in a table, for definite numerical values of the independent variable. The functions thus specified can be used in further operations; for instance, it may be required to differentiate or integrate such functions. There may be a need of function values for values of the independent variable not given in the table *(interpolation problem)* or for values of the independent variable that go beyond the limits of the table *(extrapolation problem)*.

For the sake of simplicity, we assume that the independent variable x takes on values that form an arithmetic progression, that is, $x = x_0$, $x = x_1 = x_0 + h$, $x = x_2 = x_0 + 2h$, ..., $x = x_n = x_0 + nh$. These values of the argument will be called *integral values* and h will be the *tabulation interval*. The corresponding values of the function entered in the table will be denoted by $y_0 = y(x_0)$, $y_1 = y(x_1)$, ..., $y_n = y(x_n)$.

The increments in the variable x are all the same and are equal to h. The increments in the variable y are, generally speaking, different. They are called *first-order differences* (or *differences of the first order*) and are denoted by $\delta y_{1/2} = y_1 - y_0$, $\delta y_{1+1/2} = y_2 - y_1$, $\delta y_{2+1/2} = y_3 - y_2, \dots,$ $\delta y_{n-1/2} = y_n - y_{n-1}$ since they are naturally associated with *half-integral* values of x, that is, the midpoints between adjacent integral values of the argument:

$$x_{1/2} = x_0 + h/2, \quad x_{1+1/2} = x_0 + 3h/2, \dots, \quad x_{n-1/2} = x_0 + (\mathrm{n} - 1/2)h$$ [*]

From these differences we can again take differences to obtain so-called *second-order differences*, again defined for integral values of x:

$$\delta^2 y_1 = \delta y_{1+1/2} - \delta y_{1/2}, \quad \delta^2 y_2 = \delta y_{2+1/2} - \delta y_{1+1/2},$$
$$\dots, \delta^2 y_{n-1} = \delta y_{n-1/2} - \delta y_{n-3/2}$$

(the superscript 2 indicates the order of the difference and is not an exponent), and so on.

To illustrate, we give a portion of a table of common logarithms with computed differences multiplied by 10^5.

Table 1

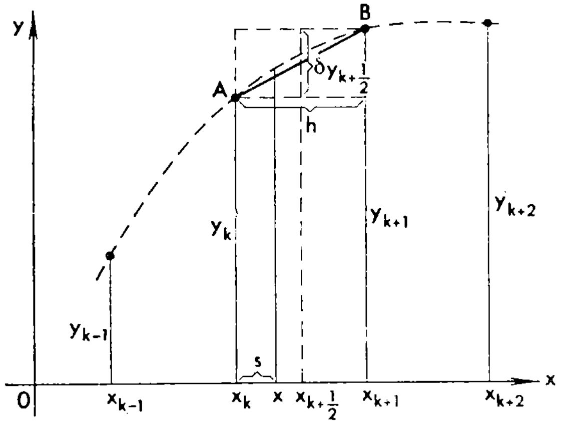
k	0	1/2	1	3/2	2	5/2	3	7/2
x_k	10.0	10.05	10.1	10.15	10.2	10.25	10.3	10.35
y_k	1.00000		1.00432		1.00860		1.01284	
$10^5\,\delta y_k$		432		428		424		419
$10^5\,\delta^2 y_k$			—4		—4		—5	
$10^5\,\delta^3 y_k$				0		—1		2

k	4	9/2	5	11/2	6	13/2	7
x_k	10.4	10.45	10.5	10.55	10.6	10.65	10.7
y_k	1.01703		1.02119		1.02531		1.02938
$10^5\,\delta y_k$		416		412		407	
$10^5\,\delta^2 y_k$	—3		—4		—5		
$10^5\,\delta^3 y_k$		—1		—1			

The smallness of the second-order differences compared with the first-order differences and their practical constancy (the third-order differences are of the order of rounding errors) in this example

[*] Sometimes the difference $y_{k+1} - y_k$ is associated not with the value $x = x_{k+1/2}$ but with $x = x_k$. Then it is ordinarily denoted as $\Delta y_k = \Delta y|_{x=x_k} = y_{k+1} - y_k$. With our notation, that is, $y_{k+1} - y_k = \delta y_{k+1/2} = \delta y|_{x=x_{k+1/2}}$, the differences are called *central differences*.

Fig. 4



points up the smoothness of the function and the absence of any random irregularities. Such regularity may appear in higher-order differences and always indicates a regularity in the variation of the function (cf. Exercise 2). Of course, if the tabulation interval is not small, and also close to discontinuities, and the like, the difference may not be small, but some sort of regularity is ordinarily apparent.

Differences are widely used in interpolation. Let it be required to find the value of y for a certain value of x between the tabulated values of x_k and x_{k+1}.The simplest approach is that of a *linear interpolation*, which consists in an approximate replacement of the function under study by a linear function and in a manner such that both functions are coincident for $x = x_k$ and $x=x_{k+1}$ (Fig. 4). Geometrically, this amounts to replacing the arc AB of the unknown graph shown dashed in Fig. 4 by the chord AB joining two known points A and B. Set $x - x_k = s$. Since a linear function is expressed by a first-degree equation, the desired value of y depends on s by the equation

$$y = a + bs \tag{1}$$

where a and b are certain coefficients. From the conditions at x_k and x_{k+1} we find that $y_k = a$, $y_{k+1} = a + bh$, whence

$$\delta y_{k+1/2} = y_{k+1} - y_k = bh$$

Expressing a and b from this equation and substituting into (1), we get the final formula for a linear interpolation:

$$y = y_k + \delta y_{k+1/2} \frac{s}{h} \tag{2}$$

(Derive this formula from the similarity of triangles in Fig. 4.) Formula (2) can be used if the function under study on the interval from x_k to x_{k+1} is only slightly different from a linear function, that is to say, if h is sufficiently small.* For $k = 0$ and $s < 0$, (2) effects a *linear extrapolation* of the function towards $x < x_0$, and for $k = n - 1$ and $s > h$, towards $x > x_n$. Of course, in the case of extrapolation, one should not go far beyond the tabulated values of x since our linear law of variation of the function is justified only over a small interval in the variation of x.

The formula of linear interpolation (2), like subsequent formulas as well, can be written in a form that does not contain differences. Substituting the expression $\delta y_{k+1/2} = y_{k+1} - y_k$ into (2), we get an equivalent formula:

$$y = y_k + \frac{y_{k+1} - y_k}{h} s = \left(1 - \frac{s}{h}\right) y_k + \frac{s}{h} y_{k+1}$$

It is quite evident that when s varies from 0 to h, the coefficient of y_k varies from 1 to 0, and the coefficient of y_{k+1} from 0 to 1. And so for $s = 0$ we get $y = y_k$ and for $s = h$ we have $y = y_{k+1}$.

More accuracy is obtained by *quadratic interpolation*, in which the function at hand is approximately replaced by a quadratic function in such a manner that both functions coincide for $x = x_k$, x_{k+1} and x_{k+2} (this was done, in a different notation, in Sec. 1.1 in the derivation of Simpson's rule). The indicated quadratic function is conveniently sought in the form

$$y = a + bs + cs(s - h) \tag{3}$$

By the hypothesis,

$$y_k = a, \quad y_{k+1} = a + bh, \quad y_{k+2} = a + b2h + c2h^2$$

whence

$$\delta y_{k+1/2} = y_{k+1} - y_k = bh, \quad \delta y_{k+3/2} = y_{k+2} - y_{k+1} = 2ch^2 + bh,$$
$$\delta^2 y_{k+1} = \delta y_{k+3/2} - \delta y_{k+1/2} = 2ch^2$$

Now expressing a, b, c in these terms and substituting into (3), we obtain *Newton's formula* for quadratic interpolation:

$$y = y_k + \delta y_{k+1/2} \frac{s}{h} + \frac{\delta^2 y_{k+1}}{2} \frac{s}{h} \left(\frac{s}{h} - 1\right) \tag{4}$$

As has already been mentioned, this formula may be used also for extrapolation. Equation (4) is not quite symmetric: we have used the values y_k, y_{k+1} and y_{k+2} where as x lies between x_k and x_{k+1}. If we reverse the direction of the x-axis and utilize the values y_{k+1}, y_k and y_{k-1} in similar fashion, then in place of (4) we get

$$y = y_{k+1} + (-\delta y_{k+1/2}) \frac{h - s}{h} + \frac{\delta^2 y_k}{2} \frac{h - s}{h} \left(\frac{h - 3}{h} - 1\right) \tag{5}$$

* From the more exact equation (4) it is readily seen that for small h the error in a linear interpolation is of the order of h^2, since the second-order differences are of that order.

Fig. 5

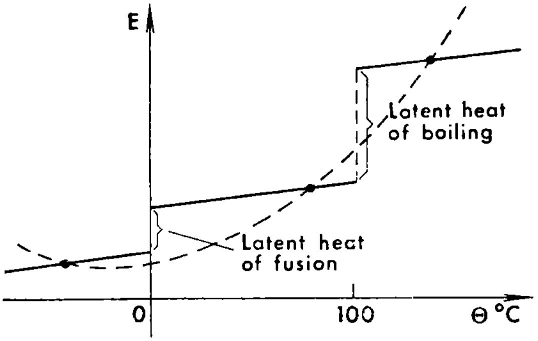
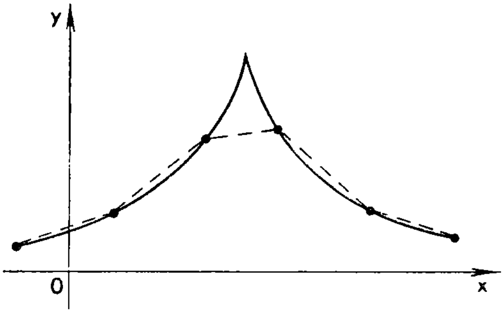


Fig. 6



which is nonsymmetric in the other direction. Now taking the half-sum of the right members of (4) and (5), we obtain the symmetric *Bessel formula*:

$$y = \frac{y_k + y_{k+1}}{2} + \delta y_{k+1/2}\left(\frac{s}{h} - \frac{1}{2}\right) + \frac{\delta^2 y_k + \delta^2 y_{k+1}}{4}\,\frac{s}{h}\left(\frac{s}{h} - 1\right)$$

which possesses a high degree of accuracy. We leave it to the reader to transform the Newton and Bessel formulas so that y is expressed directly in terms of the "mesh-point" values y_k and not in terms of their differences.

In that manner, it is possible to derive interpolation formulas of still higher degree. However, since the accuracy of experimental findings is limited — and the same can be said of tabulated values of functions — the use of differences of too high an order is ordinarily not justified. In most cases (with the exception of measurements and calculations carried out to particularly high degrees of accuracy), second-order and even first-order differences suffice.

If the function at hand is discontinuous, the interpolation may be carried out only over intervals that do not contain points of discontinuity. If this fact is lost sight of, the interpolation may give a distorted picture of the true behaviour of the function. For instance, in Fig. 5 we have the internal energy of a unit mass of water at normal pressure shown as a function of the temperature.* This function has discontinuities at the temperatures of phase transitions, that is, at the freezing and boiling points of water. The dashed line in the figure shows the result of quadratic interpolation when the chosen mesh points correspond to distinct phases; it is quite clear that this interpolation distorts the actual picture. A similar distortion results in the case of a discontinuity of the derivative (a salient point on the curve) of the function under study. The dashed line in Fig. 6 shows the result of a linear interpolation in the case of a "cuspidal" maximum. The error near this maximum stands out clearly.

Exercises

1. Let $y_0 = 1.00$, $y_1 = 1.25$, $y_2 = 1.65$, $y_3 = 2.34$. Find $y_{3/2}$ using formulas (2), (4), (5) and the Bessel formula.
2. Verify the fact that if a table has been compiled for an nth degree polynomial, then nth order differences are constant and differences of order $n + 1$ are equal to zero.

2.2 Integration and differentiation of tabulated functions

The integration of functions given in tabular form presents no special interest. Earlier, in the numerical integration of functions specified by formulas we first set up a table of the integrand (see Sec. 1.1).

When computing the derivative of a tabulated function, bear in mind that the best way to find the derivative $y'(x)$ from two values of the function is to take the values on the right and on the left at equal distances from that value of x for which we want to determine the derivative:

$$y'(x) \approx \frac{y\left(x + \frac{h}{2}\right) - y\left(x - \frac{h}{2}\right)}{h}$$

Thus, if we have values of y for values of x at equally spaced intervals (that is, in arithmetic progression), then it is convenient to compute the derivatives at the midpoints of the intervals. In other words, if the values of y were specified for "integral" values of x (see

* More frequently, use is made of the so-called *enthalpy (heat content)* $H = E + pV$, where p is the pressure and V is the volume. The heat capacity at constant temperature is precisely equal to the derivative of H with respect to Θ.

Sec. 2.1), then the values of y' will be computed for "half-integral" values of x. The same procedure is used to find the derivative of y', i.e. y'', from the values of y'. Here, the values of y'' will again be found for integral values of x. The derivative y'' may be expressed directly in terms of y:

$$y''(x) \approx \frac{y'\left(x+\frac{h}{2}\right)-y'\left(x-\frac{h}{2}\right)}{h}$$

$$\approx \frac{\dfrac{y(x+h)-y(x)}{h}-\dfrac{y(x)-y(x-h)}{h}}{h} = \frac{y(x+h)-2y(x)+y(x-h)}{h^2}$$

The formulas given here can be conveniently written down in the notation of Sec. 2.1:

$$y_{k+1/2} = y'(x_{k+1/2}) = y'\left(x_k+\frac{h}{2}\right) \approx \frac{y(x_k+h)-y(x_k)}{h} = \frac{\delta y_{k+1/2}}{h}$$

this equation is justification of the fact that we referred the difference $y_{k+1} - y_k$ to the value $x = x_{k+1/2}$.
Similarly

$$y_k'' \approx \frac{\delta^2 y_k}{h^2}$$

Here is an example (Table 2, in which half-integral values of the argument are not tabulated).

Table 2

x	y	y'	y''
1.00	1.6487		
		0.846	
1.10	1.7333		0,42
		0.888	
1.20	1.8221		0.46
		0.934	
1.30	1.9155		0.49
		0.983	
1.40	2.0138		0.49
		1.032	
1.50	2.1170		

Values of y' for other values of x may be obtained by interpolation. For instance, for integral x we obtain, via interpolation,

$$y_k' \approx \frac{1}{2}\left(y'_{k-1/2}+y'_{k+1/2}\right) \approx \frac{1}{2}\left(\frac{y_k-y_{k-1}}{h}+\frac{y_{k+1}-y_k}{h}\right) = \frac{y_{k+1}-y_{k-1}}{2h}$$

Thus, we can determine at once the values of y' for integral x via the values of y for integral x on the right and on the left (see Table 3).

Table 3

x	y	y'
1.00	1.6487	
1.10	1.7333	0.867
1.20	1.8221	0.911
1.30	1.9155	0.960
1.40	2.0138	1.002
1.50	2.1170	

However, using this method we obtain, firstly, the values of the derivatives less than in the first method by one and, secondly, less information on the behaviour of the derivative at the endpoints. Say, in Table 2 we know the derivative for $x = 1.05$ (near the beginning of the interval $x = 1$) and for $x = 1.45$ (near the end of the interval $x = 1.5$), but in Table 3 only for $x = 1.1$ and for $x = 1.4$, which is to say, at values of the argument farther away from the endpoints of the interval. Finally, when computing values of y' for nonintegral (say half-integral) values of x via interpolation, the values obtained from Table 2 turn out to be more reliable than those obtained from Table 3, since the inclination of a curve is more accurately reproduced by small chords than by large chords. The first method is therefore preferable.

Very important, fundamental questions arise in connection with the restricted accuracy and errors inherent in every measurement.

In computing an integral, each separate measured value of y is multiplied by the quantity Δx. Therefore as the number of separate measured values of the function y is increased, the coefficient with which each separate value enters the expression of the integral diminishes in inverse proportion to the number of subintervals Δx.

Consequently, there is also a reduction in the error in the integral that is due to the errors in each separate measurement of the quantity y.

In computing a derivative, the difference between two values of y is divided by Δx. The smaller the subinterval Δx, that is the smaller the denominator, the larger the error contributed to the derivative by the given error in each measured value of y. For this reason, the derivative of a function specified by experimental values proves to be known with an accuracy less than that of the function itself.

Fig. 7

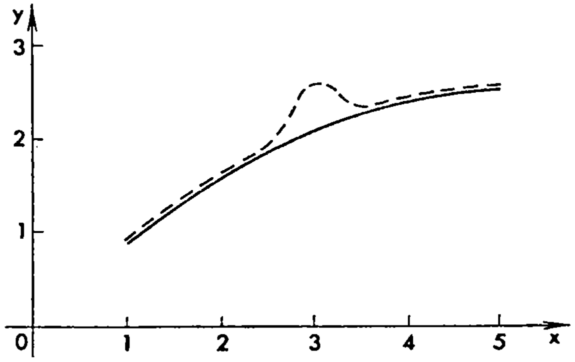
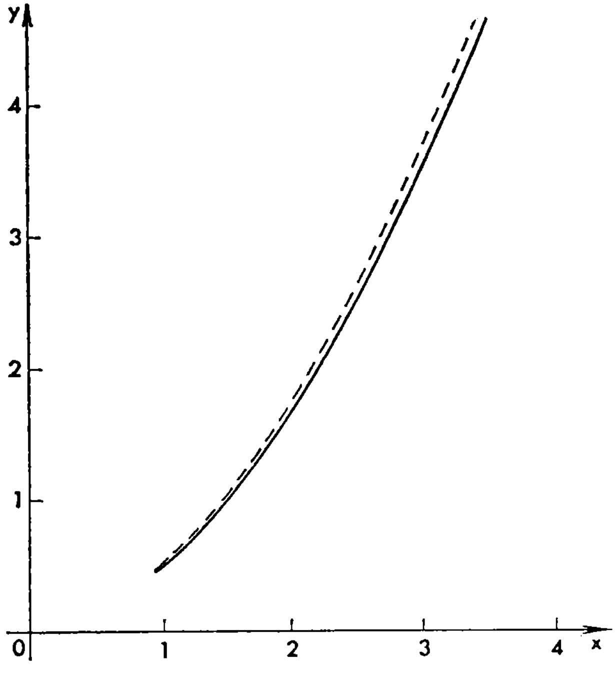


Fig. 8



We illustrate the difference between differentiation and integration with an example.

Fig. 7 depicts two curves, one solid, the other dashed. The solid curve is the graph of the function $y = x - 0.1x^2$, the dashed curve is the graph of the function $y_1 = x - 0.1x^2 + 0.5e^{-8(x-3)^2}$. From the figure it is clear that one curve differs perceptibly from the other only over a small range of x.

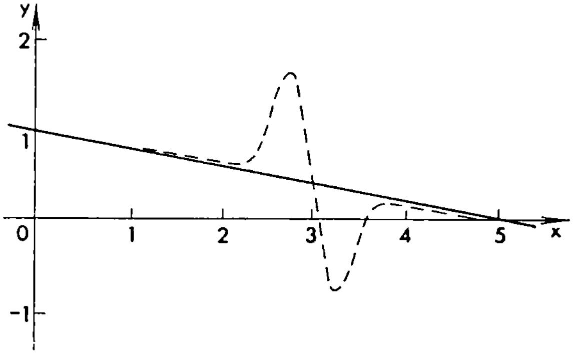


Fig. 9

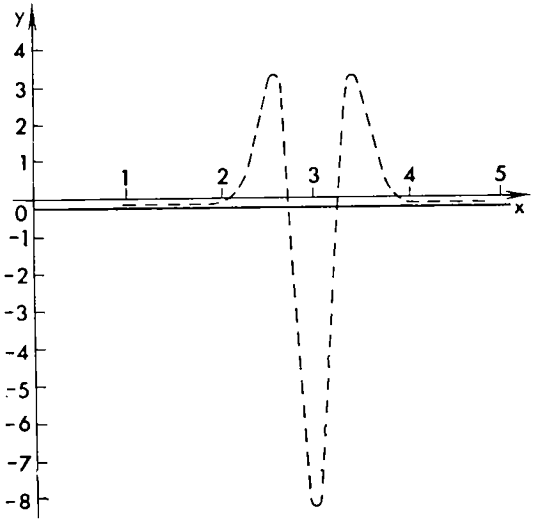


Fig. 10

Fig. 8 shows the graphs of $I(x) = \int_0^x y\,dx$ and $I_1(x) = \int_0^x y_1 dx$ (dashed curve). We see that the difference between the curves $y(x)$ and $y_1(x)$ produces a slight addition to the integral $I_1(x)$, which the graph is only apparent for $x > 2.8$. On the whole, there is only a slight difference between the curves $I(x)$ and $I_1(x)$.

Fig. 9 depicts the curves of the derivatives $y'(x)$ and $y_1'(x)$. We see that a slight change in the function over a small interval gave rise (over the same interval) to large changes in the derivative. The second derivatives exhibit a still greater difference. Their graphs are shown in Fig. 10, where the scale on the y-axis is half that of Figs. 7 to 9.

When a curve is obtained experimentally, a slight variation in the curve over some interval may be the result of an error in the experiment. It is clear from the preceding example that such isolated errors do not substantially affect the magnitude of the integral, but they strongly affect the value of the derivative (particularly higher derivatives).

In order to obtain reliable values of the derivative, it is first necessary to find a formula that gives a good description of the experimental data and then find the derivative using this formula.

Since the formula is constructed on the basis of all experimental data, the value of the derivative, for each value of x, will be found from the formula only if all the data are taken into account, and not only two or three close-lying values. It is therefore natural to expect that random errors in certain measurements will have a smaller effect on the value of the derivative.

Choosing a formula to describe the results of an experiment is, generally, an essential part of the treatment of experimental findings. This problem of curve fitting on the basis of experimental data is discussed in the next two sections.

Exercises

Use the conditions of Exercise 2, Sec. 1, to compute the values of y' for half-integral values of x, assuming $\Delta x = h = 0.5$. Interpolate the result linearly and also use formulas (4) and (5) for integral values of x.

2.3 Fitting of experimental data by the least-squares method

The fitting of formulas to experimental findings is called *curve fitting*. Actually, of course, a formula is the better, the more theoretical material has been put into it and the less empirical it is. In reality, one first has to specify the type of formula, and then, using the results of experiments, to determine the values of the various constants in the formula.

Before attempting to fit a formula, it is useful to plot the experimental findings on a graph and then draw freehand the most likely curve through the points obtained. In doing so, we will immediately perceive those findings that are particularly suspect as being erroneous. In drawing the curve, it is very important, besides using the experimentally found points, to reason generally about how the curve should behave at values of the argument very close to zero, at large values of the argument, whether the curve passes through the origin of coordinates, or intersects the coordinate axes, or is tangent to them, etc.

Now suppose all this preliminary work has been carried out and a type of equation has been chosen; we need to determine the values of the constants that enter into this equation.

How is this done?

Let us consider the simplest case.

Suppose that y is proportional to x, that is, we seek a formula of the form $y = kx$. The problem reduces to determining the coefficient k. Each experiment yields a specific value for k, namely,

$$k_n = \frac{y_n}{x_n}$$

where x_n, y_n are the values of the quantities x and y obtained in the nth experiment. The index n on k indicates the value corresponding to the nth experiment. We can take the mean of the values k_n, putting

$$\bar{k} = \frac{\sum_{n=1}^{p} k_n}{p}$$

where p is the total number of experiments. We obtain the formula $y = \bar{k}x$.

Observe that this is the simplest but not the best procedure for choosing k. Indeed, let x be a quantity that we specify exactly and that describes the conditions of the experiment; let y be the result of an experiment that contains a certain error of measurement. We will assume that the measurement error Δy is roughly the same for both small and large values of y. Then the error in the quantity k_n, equal to $\frac{\Delta y_n}{x_n}$, is the larger, the smaller x_n is. Hence, in determining the quantity k, it is best to rely on experiments with large x_n.

We now pose the problem of finding the value of k for which the function $y = kx$ fits the experimental data in the best way. (The meaning of this rather vague word "best" will become clear from what follows.) As a measure of the deviation of the function from the experimental data in the nth experiment we choose the quantity $(y_n - kx_n)^2$. Why do we choose $(y_n - kx_n)^2$ and not $y_n - kx_n$? It is clear that both signs of the deviation of kx_n from y_n are bad: it is bad if k is such that $y_n < kx_n$ and it is also bad if k is such that $y_n > kx_n$. If for the measure of the deviation we took the quantity $y_n - kx_n$ and then sought the sum of the deviations in several experiments, we might obtain a very small value due to mutual cancelling of individual large terms of opposite sign. But this would in no way indicate that the function $y = kx$ is good. Now if we take $(y_n - kx_n)^2$ for the deviation, no such mutual cancelling will occur, since all the quantities $(y_n - kx_n)^2$ are positive. Note that we could, in principle, take $|y_n - kx_n|$, $(y_n - kx_n)^4$, etc. instead of $(y_n - kx_n)^2$. But subsequent computations would be unduely complicated.

As a measure of the total error S in a description of experimental findings via the function $y = kx$ we take the sum of the deviations in all experiments, that is,

$$S = \sum_{n=1}^{p} (y_n - kx_n)^2 \tag{6}$$

The method of determining the constants in a formula by requiring that the total deviation S be least is called the *method of least squares.*

Observe that if one quantity $y_n - kx_n = 10$, that is, for some one $x = x_n$ the formula yields an error of 10 units, then this will introduce 100 units into S. On the other hand, if we have 10 errors of 1 each, then there will be only 10 units in S. It is therefore clear that S is mostly affected by the largest errors, whereas small errors, even if they occur frequently, have but a small influence. The method of least squares is aimed at reducing the largest deviations.

To find $k = \bar{\bar{k}}$ for which S is the least, we solve the equation $\frac{dS}{dk} = 0$. Using (6), we get

$$\frac{dS}{dk} = 2\sum_{n=1}^{p} (y_n - kx_n)(-x_n) = 0$$

whence

$$2k\sum_{n=1}^{p} x_n^2 - 2\sum_{n=1}^{p} x_n y_n = 0$$

which yields

$$k = \bar{\bar{k}} = \frac{\sum\limits_{n=1}^{p} x_n y_n}{\sum\limits_{n=1}^{p} x_n^2} = \frac{x_1 y_1 + x_2 y_2 + \dots + x_p y_p}{x_1^2 + x_2^2 + \dots + x_p^2} \tag{7}$$

If in every experiment we get $y_n = kx_n$ exactly, then from (7) we obtain

$$\bar{\bar{k}} = \frac{x_1 \cdot kx_1 + x_2 \cdot kx_2 + \dots + x_p \cdot kx_p}{x_1^2 + x_2^2 + \dots + x_p^2} = k$$

If the quantity $k_n = \frac{y_n}{x_n}$ differs in different experiments, then by substituting in (7) the value $k_n x_n$ for y_n we obtain

$$\bar{\bar{k}} = \frac{k_1 x_1^2 + k_2 x_2^2 + \dots + k_p x_p^2}{x_1^2 + x_2^2 + \dots + x_p^2} \tag{8}$$

Among the quantities $k_1, k_2, \dots, k_p$ obtained in distinct experiments there is a largest one $k_{\max}$ and a smallest one $k_{\min}$. If in the right member of (8) we replace all k_n by $k_{\max}$, then the fraction will only increase and we get

$$\bar{\bar{k}} < \frac{k_{\max}x_1^2 + k_{\max}x_2^2 + \dots + k_{\max}x_p^2}{x_1^2 + x_2^2 + \dots + x_p^2} = k_{\max}$$

In the very same way we can prove that $\bar{\bar{k}} > k_{\min}$.

Thus, the $\bar{\bar{k}}$ found from the minimum condition of S satisfies the inequalities $k_{\min} < \bar{\bar{k}} < k_{\max}$, which means it is indeed the mean of all the values $k_1, k_2, \dots, k_p$; but this mean is formed by a more complicated rule than

$$\bar{k} = \frac{k_1 + k_2 + \dots + k_p}{p}$$

In (8) each k_n enters into the numerator with the factor x_n^2, which is called the *weight.*[*]

It is clear that the greater the weight x_n^2, the greater the effect the measurement corresponding to the value $x = x_n$ will have on k. This confirms the idea expressed earlier that measurements with large x_n are more important for a correct determination of k.

If there is no reason to suppose that $y = 0$ for $x = 0$, then the simplest is the formula $y = kx + b$. Here too we can apply the method of least squares. For this case, S is given by

$$S = \sum_{n=1}^{p} (y_n - kx_n - b)^2 \tag{9}$$

The aim is to choose k and b so that S is a minimum.

We proceed as follows. If b were known, then only k would have to be changed in the right member of (9), and so it should be true that[**]

$$\frac{\partial S}{\partial k} = 2\sum_{n=1}^{p} (y_n - kx_n - b)(-x_n) = 0$$

[*] The name "weight" stems from the following mechanical analogy. Imagine a scale with points spaced at distances of $k_1, k_2, \dots, k_p$ and with weights at these points. If all weights are the same, the centre of gravity of the system (we ignore the weight of the scale itself) lies at the scale point $\bar{k} = \dfrac{k_1 + k_2 + \dots + k_p}{p}$.
But if at point k_1 we have a weight x_1^2, at k_2 a weight $x_2^2, \dots,$ and at k_p a weight x_p^2, then the position of the centre of gravity is given by formula (8). Thus, this formula corresponds to the idea of a different significance and of different weights of separate observations.

[**] When we consider a function of several variables, the derivative with respect to one of the variables, the others being held constant, is denoted by ∂ and not d (see, for example HM, Sec. 3.12). We will have more to say on this topic in Ch. 4.

On the other hand, if k were already known, then

$$\frac{\partial S}{\partial b} = -2\sum_{n=1}^{p}(y_n - kx_n - b) = 0$$

These two conditions give us the following system of equations for determining the numbers k and b:

$$\left.\begin{aligned} \sum_{n=1}^{p} x_n y_n - k\sum_{n=1}^{p} x_n^2 - b\sum_{n=1}^{p} x_n &= 0, \\ \sum_{n=1}^{p} y_n - k\sum_{n=1}^{p} x_n - bp \qquad &= 0 \end{aligned}\right\} \tag{10}$$

It is easy to find k and b from (10). For brevity set

$$\sigma_1 = \sum_{n=1}^{p} x_n, \quad \sigma_2 = \sum_{n=1}^{p} x_n^2, \quad r_0 = \sum_{n=1}^{p} y_n, \quad r_1 = \sum_{n=1}^{p} x_n y_n$$

Then the system (10) can be rewritten thus:

$$\left.\begin{aligned} \sigma_2 k + \sigma_1 b &= r_1, \\ \sigma_1 k + pb &= r_0 \end{aligned}\right\}$$

Solving it we obtain

$$k = \frac{pr_1 - r_0\sigma_1}{p\sigma_2 - \sigma_1^2}, \quad b = \frac{r_0\sigma_2 - r_1\sigma_1}{p\sigma_2 - \sigma_1^2}$$

But this method can be viewed from another angle.

By specifying a linear relationship

$$y = kx + b$$

between the quantities x and y, we obtain two unknown parameters k and b. By means of measurements we arrive at certain relations between these parameters:

$$\left.\begin{aligned} kx_1 + b &= y_1, \\ kx_2 + b &= y_2, \\ &\cdots\cdots\cdots\cdots \\ kx_p + b &= y_p \end{aligned}\right\}$$

In other words, we have a system of p equations in two unknowns. For $p > 2$, such a system is overdetermined since it is sufficient in principle to have two equations in order to find the unknowns. But if we note that the physical quantities x and y are measured with a definite error, we find that in the case of two measurements (when $p = 2$) the values of k and b may be essentially affected by random errors of measurement and so the accuracy of the result remains obscure. For this reason, it is dangerous to reduce the number of equations containing such random factors. On the contrary, the larger the number of measurements, that is, the more overdetermined the system, the better, because then the random errors of individual measurements cancel out and the solution found by the method of least squares becomes more reliable.

It is not difficult to generalize the method of least squares to the case of more complicated relationships between the quantities x and y. It is well to bear in mind, however, that the method of least squares frequently leads to rather unwieldy computations. When the desired parameters appear nonlinearly in the relations, the method leads to a system of nonlinear equations with a concomitant extreme buildup in computational difficulty. It is for this reason that in practice graphical methods of curve fitting have proved more effective. We shall consider them in the next section.

Exercises

1. Using the method of least squares, choose a formula of the type $y = kx$ based on the following experimental findings:

(a)

x	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
y	0.40	0.50	0.90	1.28	1.60	1.66	2.02	2.40

(b)

x	0.25	0.50	0.75	1.00	1.25
y	0.16	0.18	0.80	0.60	1.08

(c)

x	0.4	0.8	1.2	1.6	2.0
y	0.69	1.44	2.08	2.74	3.52

In each of the foregoing three cases, plot the tabulated points and the straight line obtained by the method of least squares. (Use cross-section paper).

2. Using the following table, choose the numbers k and b for the formula $y = kx + b$ by the method of least squares:

x	-0.20	0.20	0.40	0.60	0.70	0.80
y	0.96	1.40	1.56	1.74	1.92	2.04

3. Given two points (x_1, y_1) and (x_2, y_2). Use them to choose the numbers k and b for the equation of the straight line $y = kx + b$ via the method of least squares. Show that in doing so we obtain an absolutely exact result, that is, we get the equation of the straight line passing through the two indicated points.

2.4 The graphical method of curve fitting

Recall that the equation of a straight line has the form $y = kx + b$, where the numbers k and b have a simple geometric meaning (see HM, Sec. 1.4): b is the line segment intercepted on the y-axis and k is the slope of the straight line to the x-axis (Fig. 11).

Let us assume that y and x are related linearly, i.e. $y = kx + b$. Plot the experimental points on the graph. Now put a transparent rule on the graph and move it to obtain the straight line closest to the experimental points (Fig. 12).

By drawing that line we obtain b and $k = Y/X$ from the figure.

The big advantage of the graphical method is its pictorial nature. If the experimental points lie on the line, with the exception of a few that fall outside the line, these points are clearly exhibited and we can see what points have to be verified. If on the whole the experimental points do not lie on a straight line, this too is clear from the graph. In this case the relationship between x and y is more complicated than $y = kx + b$. An added advantage of the graphical method is that no messy calculations like those involved in the method of least squares are needed and the chances of an accidental calculation error are reduced.

The straight line occupies an exceptional place in the graphical method of curve fitting. No other line can be drawn so simply and reliably through the given points. Anyone who has ever made a practical comparison in determining the numbers k and b in the equation of a straight line on a graph by the least-squares method knows that the difference is always very slight.

How does one go about choosing the constants of a formula with the aid of the graph if the formula is more complicated than $y = kx + b$?

Let us consider an example.

Suppose we are investigating the relationship between the temperature T of a wire and the direct current i flowing in the wire. It is clear that a change in the direction of current flow does not change T, that is, $T(-i) = T(i)$. Therefore a relationship like $T = ai + b$ is not suitable. We will seek a formula like $T = ai^2 + b$. The graph of the function $T(i)$ is a parabola and it is not easy to draw a parabola by eye. We therefore introduce a new variable $z = i^2$. Then $T = az + b$ so that in terms of the coordinates z, T the desired relationship is given by a straight line. The temperature

Fig. 11

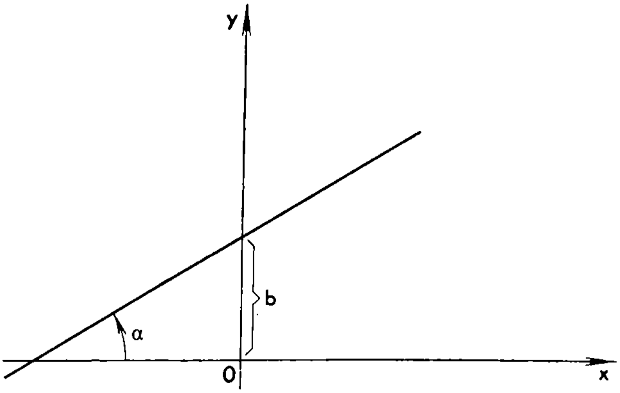
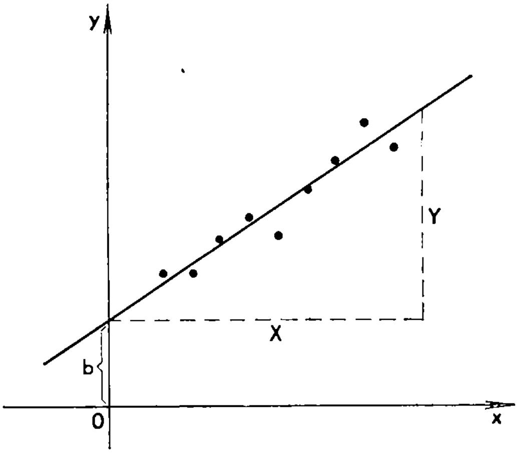


Fig. 12



$b = T_0$ in the absence of current may be taken as known, so that it remains to determine the coefficient a of i^2.

For heavy currents with accompanying high temperatures, the resistance of the wire cannot be assumed constant. For this reason, the heating capacity (the amount of heat released per unit time), which is equal to $W = Ri^2$, is not in reality merely proportional to i^2 since R varies. In the equation of the thermal system

$$W = Ri^2 = \alpha S(T - T_0)$$

where α is the heat transfer coefficient and S is the surface of the wire, the coefficient α is also variable at high temperatures. However we still have an equality of temperatures for currents i and $-i$. It is therefore natural to insert into the formula $T = ai^2 + b$, which may now prove to be inexact, the term ci^4 instead of ci^3.

And so we seek the formula in the form $T = ci^4 + ai^2 + b$. Note that $T = b$ for $i = 0$ so that b does not differ from the ambient temperature and is therefore known (see above). We rewrite the formula thus:

$$\frac{T - b}{i^2} = ci^2 + a$$

Introducing new variables $x = i^2$, $y = \frac{T - b}{i^2}$, we obtain $y = cx + a$, which means x and y are linearly related. It is easy to determine a and c by constructing a graph in terms of the coordinates x and y.

Thus, the general idea behind the graphical method is to introduce new variables so that the desired relationship in terms of these variables is linear.

Here are some other examples.

Frequently, the relationship between x and y is such that it is known for sure that y must be zero when $x = 0$, but the experimental findings on the graph do not lie on a straight line. The following formula may then be true:

$$y = ax + bx^2$$

Divide all terms by x to get $y/x = a + bx$. Putting $y/x = z$, we obtain z as a linear function of x:

$$z = a + bx$$

Another formula that may come in handy in this case is $y = ax^n$. The question is how to determine the exponent n. To do this we take logs of both sides of the equation:

$$\log y = n \log x + \log a$$

Introducing new variables $z = \log y$, $t = \log x$, we obtain the linear relationship

$$z = \log a + nt$$

The *law of radioactive decay* is given by the equation $n = n_0 e^{-\omega t}$, where n is the number of atoms that have not yet decayed at time t, n_0 is the total number of atoms, and ω is the probability of decay. Taking logs of both sides, we get

$$\ln n = \ln n_0 - \omega t$$

We thus obtain a straight line in terms of the coordinates t, $y = \ln n$. (*Radioactive decay* is discussed in detail in HM, Sec. 5.3.)

Investigations of some quantity x as a function of temperature T very often yield a formula like

$$x = Be^{-\frac{A}{kT}}$$

Such a formula results in cases* where a contribution is made only by those molecules (or electrons) whose energies exceed A, k is the Boltzmann constant ($k = 1.4 \times 10^{-16}$ erg/deg).

Taking logs, we get $\ln x = \ln B - \frac{A}{k}\frac{1}{T}$. The relationship becomes linear if we consider the quantities $y = \frac{1}{T}$ and $z = \ln x$. Indeed, $z = \ln B - \frac{A}{k} y$.

In the foregoing examples we chose a type of formula and then introduced new variables so as to make the relationship between the new variables linear. However, it may happen that the experimental points in the new variables do not lie on a straight line, which means that the type of formula was inaptly chosen and that a different type must be sought.

Suppose a series of experiments have been carried out in which for values of the argument $x_1, x_2, \ldots, x_p$ we have obtained the function values $y_1, y_2, \ldots, y_p$. Let the values of the argument be arranged in an increasing sequence $x_1 < x_2 < \ldots < x_p$. Determining an expected experimental value of y for a given value of x lying inside the investigated range of the argument ($x_1 < x < x_p$) constitutes an interpolation problem (cf. the beginning of Sec. 2.1).

Interpolation is a simple matter if an empirical formula has been found. And if the formula is a suitable one, the interpolation ordinarily yields good results and rarely leads to crude errors. A much more difficult problem is this: what value of y is to be expected from experiment for a certain value of x lying outside the experimental range of the argument, for example, when $x > x_p$. Determining such a value from experiment constitutes an extrapolation problem.

The solution of extrapolation problems requires a careful study of the matter in each specific instance. Such problems cannot be solved in a formal manner merely by using a selected formula. (This idea was put very neatly by Kretya Patachkuvna in the book *Pi*, when she said: "It is very hard to foresee anything, particularly if it's in the future.")

For example, if experimental findings yield a formula of the type $y = a + bx + cx^2 + px^3$, and this formula gives a good description of experimental results, then as a rule the terms cx^2, px^3 are introduced to describe the deviation of experimental points from the straight line over the range of x where the measurements were made. In this case, the terms cx^2 and px^3 are ordinarily in the form of small corrections to the principal term $a + bx$.

* Several such cases are discussed in HM, Ch. 7.

Now if we use such a formula and extrapolate y to values of x that are far outside the experimental range, then the terms cx^2 and px^3 will become dominant, yet this might not at all fit the essence of the matter. This situation is reminiscent of an Andersen fairy tale where a shadow separates itself from the man and begins an existence all its own, launching out on a career and finally forcing the man himself into submission.

If as x increases without bound y approaches a definite value y_∞, then it is useful to find that value. This problem is called *extrapolation to infinity*. In the solution of this problem it is frequently useful to introduce a new independent variable z that remains finite for $x = \infty$, say, $z = 1/x$. After such a transition, the interval (in z) over which the extrapolation is performed will then be finite.

Let us consider an example.

Suppose we have a long strip of explosive. At one end we set off an explosion that is propagated lengthwise along the strip. It is clear that if the strip is very long, its action on some obstacle will cease to be dependent on the length of the strip. This is quite evident since if we increase the length of the strip sufficiently, we thus also increase the quantity of explosive which is at some distance from the obstacle and which for this reason makes hardly any contribution at all. Suppose for example we denote by y the maximum thickness of a steel wall to be destroyed by an explosive charge of length x. Experimental findings have been plotted in Fig. 13. We can see here that as x increases, y approaches a definite value, y_∞. But one cannot find that value from the graph.

The question is how to find it. Suppose that for large x the formula is of the form $y = y_\infty - a/x$. Introducing a new variable $z = 1/x$, we get $y = y_\infty - a \cdot z$. Now y_∞ corresponds to $z = 0$. Constructing the experimental findings in terms of the coordinates z, y (in Figs. 13 and 14 the numbers in brackets label the same points), we can determine by eye the presumed value of y for $z = 0$ (Fig. 14).

The formula $y = y_\infty - a/x$ holds only for sufficiently large x. For $x = a/y_\infty$ it gives $y = 0$ and if $x < a/y_\infty$ we even get $y < 0$, which is completely meaningless. Therefore, the points in Fig. 14 obtained experimentally lie on a curve and not on a straight line; however, in terms of the coordinates z and y we can extrapolate this curve to $z = 0$, which corresponds to $x = \infty$.

From the physical meaning of the problem it is also clear that y must be 0 at $x = 0$. The simplest formula reflecting both these properties of the function y ($y = 0$ for $x = 0$ and y approaches the

Fig. 13

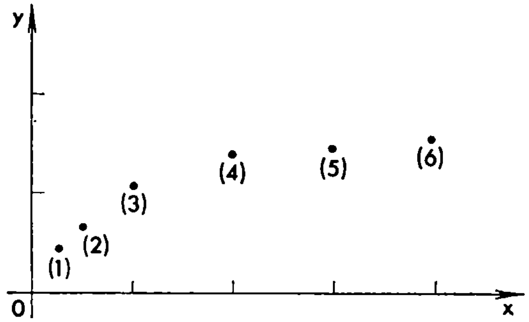
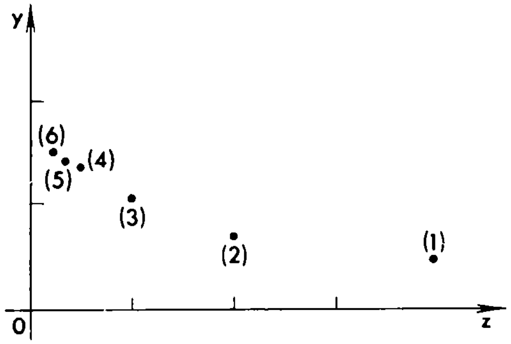


Fig. 14



value y_∞ without bound if x increases without bound) is of the form

$$y = \frac{y_\infty x}{x + b} \tag{11}$$

How can we determine the constants y_∞ and b by using experimental findings?

First of all, rewrite the formula as

$$\frac{1}{y} = \frac{x + b}{y_\infty x}$$

or

$$\frac{1}{y} = \frac{1}{y_\infty} + \frac{b}{y_\infty} \cdot \frac{1}{x}$$

We can therefore hope that if we carry out the construction on a graph of $1/y$ versus $1/x$, we will get a straight line and will be able to determine $1/y_\infty$ and b/y_∞. Even if the points do not fit the straight line closely, the extrapolation on this graph is more reliable than on the graph of Fig. 14 since formula (11) was constructed with account taken of two properties (instead of the earlier one property) of the function $y(x)$.

Exercises

1. Using the data of the following table, find a formula of the form $y = ax^2 + b$:

x	0.5	1.0	1.5	2.0	2.5	3.0
y	1.20	1.10	2.35	3.05	4.40	5.50

Solve this problem in two ways:
(a) via the method of least squares, and
(b) graphically.

2. Find graphically a formula of the form $y = ax^2 + bx$ if the experimental findings are:

x	0	0.5	1.0	1.5	2.0	2.5	3.0
y	0	1.7	3.1	3.8	3.9	3.8	3.0

3. Use the tabulated data

x	1	2	3	4
y	0.5	1.4	2.5	4

to find a formula of the form $y = Ax^b$. (Use the graphical method.)

4. Measurements yield the following data:

x	0.5	1.0	1.5	2.0	2.55	3.0	3.5	4.0
y	1.66	1.58	1.50	1.44	1.37	1.30	1.22	1.17

Besides, we know that:
(a) y approaches zero as x increases without bound;
(b) y has a very definite value at $x = 0$. These conditions are satisfied, for instance, by the following simple formulas:

$$y = \frac{1}{B + Cx} \quad \text{and} \quad y = Ae^{-kx}$$

Choose values for the parameters of these formulas. Using the formulas, obtain the values of y for $x = 1.25$, $x = 3.75$, $x = -1$, $x = -2$, $x = -3$.
Compare the results.

5. The following data have been obtained from experiment:

x	1.0	1.5	2.0	3.0	4.0	5.0
y	1.6	1.7	2.0	2.3	2.4	2.5

Furthermore, it is known that y approaches a value y_∞ as x increases without bound. Find this limiting value in two ways:

(a) by choosing a formula of the form $y = A + \frac{B}{x}$;

(b) by choosing a formula like $y = \frac{ax}{x+b}$.

ANSWERS AND SOLUTIONS

Sec. 2.1

1. By formula (2), $y_{3/2} = 1.45$; by (4), $y_{3/2} = 1.41$; by (5), $y_{3/2} = 1.43$; by the most accurate formula (Bessel's) $y_{3/2} = 1.42$.
2. If $y_i = y\,|_{x=x_i} = x_i^n$ and $\Delta x = h$, then by the binomial theorem,

$$\delta y_{i+1/2} = \delta y\,|_{x=x_i+h/2} = (x_i + h)^n - x_i^n$$
$$= nhx_i^{n-1} + \frac{n(n-1)}{2} h^2 x_i^{n-2} + \dots$$

Setting $x_i + h/2 = x_i'$, we get

$$\delta y\,|_{x=x_i'} = nh\left(x_i' - \frac{h}{2}\right)^{n-1} + \frac{n(n-1)}{2} h^2 \left(x_i' - \frac{h}{2}\right)^{n-2} + \dots$$

Removing the brackets in the right member, we see that in this example the differences form a sequence of values of a polynomial of degree $n - 1$. The same will hold true for the function $y = ax^n$ ($a =$ constant) since in the formation of differences the coefficient a is a common factor. Observing that when functions are added, their differences are also added, we conclude that for any polynomial of degree n the differences form a sequence of values of some polynomial of degree $n - 1$. Which means the second differences form a sequence of values of a polynomial of degree $n - 2$, and so on, while the nth differences constitute a sequence of values of a zero-degree polynomial, or constants. For this reason, the differences of order $n + 1$ are here equal to zero.

Sec. 2.2

$y'_{1/2} = 0.50$, $y'_{3/2} = 0.80$, $y'_{5/2} = 1.38$. By linear interpolation, $y'_1 = 0.65$, $y'_2 = 1.09$. By formula (4), $y'_1 = 0.62$.
By formula (5), $y'_2 = 1.06$.

Sec. 2.3

1. (a) $y = 1.18x$, (b) $y = 0.78x$, (c) $y = 1.75x$.

2. $y = 1.03x - 1.19$.

3. The equation of a straight line is of the form $y = kx + b$. The numbers k and b are found from the conditions $y = y_1$ for $x = x_1$ and $y = y_2$ for $x = x_2$. We obtain the system of equations

$$\left.\begin{aligned} kx_1 + b &= y_1, \\ kx_2 + b &= y_2 \end{aligned}\right\}$$

from which we find

$$k = \frac{y_1 - y_2}{x_1 - x_2}, \quad b = \frac{x_1y_2 - x_2y_1}{x_1 - x_2} \tag{12}$$

We will show that the values of k and b are obtained by using the method of least squares.

In our case

$$S = (y_1 - kx_1 - b)^2 + (y_2 - kx_2 - b)^2$$

Therefore

$$\left.\frac{\partial S}{\partial k}\right|_{b=\text{constant}} = -2[(y_1 - kx_1 - b)\, x_1 + (y_2 - kx_2 - b)x_2]$$

$$\left.\frac{\partial S}{\partial b}\right|_{k=\text{constant}} = -2[(y_1 - kx_1 - b) + (y_2 - kx_2 - b)]$$

Equating these derivatives to zero, we get the following system of equations:

$$\left.\begin{aligned} (y - kx_1 - b)\, x_1 + (y_2 - kx_2 - b)\, x_2 = 0, \\ y_1 - kx_1 - b + y_2 - kx_2 - b = 0 \end{aligned}\right\}$$

Solving this system yields the same values as in (12). Incidentally, the coincidence of results may be obtained from the simple argument that one straight line can be drawn between any two points.

Sec. 2.4

1. Form a sum in order to take advantage of the method of least squares.

$$S = \sum_{k=1}^{6} (y_k - ax_k^2 - b)^2$$

Proceeding further in the usual way, we find $a = 0.48$, $b = 1.23$, that is, the desired formula is

$$y = 0.48x^2 + 1.23$$

To solve this problem graphically, we introduce a new variable $t = x^2$; then $y = at + b$. In terms of the coordinates (t, y) we obtain a straight line. Plotting the points $(t_k = x_k^2, y_k)$ on the graph, we find $a = 0.49$, $b = 1.35$, or $y = 0.49x^2 + 1.35$.

2. Set $\frac{y}{x} = z$, then $z = ax + b$. Plotting points in terms of the coordinates (x, z), we find $a = -1.02$, $b = 4$, so that $y = -1.02x^2 + 4x$.

3. In this case, lay off $\log x$ and $\log y$ along the coordinate axes to obtain $A = 0.5$, $b = 1.5$, $y = 0.5x^{1.5}$.

4. Using the graphical method, we find $y = \frac{1}{0.56 + 0.07x}$, $y = 1.74e^{-0.1x}$. The values of y found by the first and second formulas for several of the indicated values have been tabulated as follows:

x	Value of y by first formula	Value of y by second formula
1.25	1.54	1.54
3.75	1.22	1.20
-1	2.01	2.00
-2	2.38	2.12
-3	2.86	2.35

The meaning of the foregoing calculation is this: as indicated in the text, interpolation rarely leads to considerable errors. In our case, both formulas yield close results for $x = 1.25$ and for $x = 3.75$. Extrapolation, on the contrary, is unreliable. It is evident from the table that the farther x is from the tabulated values, the bigger the difference in the values of y obtained by the different formulas. Observe that for $x = -8$ the first formula is meaningless, while for $x < -8$, the first formula yields $y < 0$ and the second one gives $y > 0$.

5. (a) $y_\infty = 2.8$; (b) $y_\infty = 2.9$.

Chapter 3

MORE ON INTEGRALS AND SERIES

3.1 Improper integrals

In the usual definition of an integral (see, for example, HM, Sec. 2.8) it is assumed that the interval of integration is finite and that the integrand function does not vanish on the interval. Such integrals are termed *proper integrals*. If at least one of these two conditions does not hold true, then the integral is said to be *improper*. Such integrals occur even in rather simple problems of integral calculus (see, for instance, HM, Secs. 3.16, 4.3, 6.2).

We first consider an integral of the form

$$I = \int_a^\infty f(x)\,dx \tag{1}$$

where the lower limit a and the integrand $f(x)$ for $a \leqslant x < \infty$ are assumed to be finite. Such an integral is improper because the upper limit is infinite. We say it has a *singularity* at the upper limit.

Suppose that integral (1) was obtained in the solution of a physical problem and the variable x has a definite physical meaning (length, time, and the like). Then in actuality x does not go to infinity but to some very large but finite limit, which we denote by N. Thus, instead of (1) we have to consider the integral

$$I_N = \int_a^N f(x)\,dx \tag{2}$$

It may turn out that the integral (2) does depend on N but for sufficiently large N remains practically unaltered. Then this value is taken for the value of the integral (1). To be more exact, it is then assumed that

$$\int_a^\infty f(x)\,dx = \lim_{N\to\infty} \int_a^N f(x)\,dx$$

and (1) is called a *convergent integral*. From this limit and from the equation

$$\int_a^\infty f(x)\,dx = \int_a^N f(x)\,dx + \int_N^\infty f(x)\,dx$$

it is evident that the basic contribution to a convergent integral is made by its "finite" ("proper") part, whereas the contribution of the singularity, for sufficiently large N, is arbitrarily small. In other words, if (1) is convergent, then the "actual" integral (2), for large N, which is often not known exactly, may be replaced by the "limit" integral (1), which ordinarily is simpler in theoretical investigations.

If, as N increases, the value of the integral (1) does not become steady but tends to infinity or oscillates without having a definite limit, then (1) is said to be a *divergent integral*. In that case, the value of (2) for large N depends essentially on N, and (2) cannot be replaced by (1). Then the question can arise of a more detailed description of the behaviour of (2) as N increases, that is, of obtaining *asymptotic formulas* for the integral. (Incidentally, such a question also arises for the convergent integrals (1), since it is often not enough merely to establish the fact of convergence or divergence or even to find the numerical value in the case of convergence, for the law of convergence itself may be needed.)

From the foregoing it follows that the fact of convergence or divergence of the integral (1) depends only on the behaviour of the function $f(x)$ "at the singularity of the integral", that is to say, as $x \to \infty$. This fact is most often perceived by comparing $f(x)$ with the power function C/x^p, the integral of which can be taken with ease. Let us consider the integral

$$I = \int_{x_0}^\infty \frac{C}{x^p}\,dx \quad (C \text{ constant}) \tag{3}$$

where x_0 is any positive number (if we take x_0 negative, then the integral can have a singularity also at $x = 0$, where the integrand becomes infinite). We can readily evaluate this integral:

$$\begin{aligned}\int_{x_0}^\infty Cx^{-p}\,dx &= \left.\frac{Cx^{-p+1}}{-p+1}\right|_{x_0}^\infty = -C\left.\frac{1}{(p-1)\,x^{p-1}}\right|_{x_0}^\infty = \\ &= \frac{C}{(p-1)\,x_0^{p-1}} - \frac{C}{(p-1)\,\infty^{p-1}}\end{aligned} \tag{4}$$

Two cases have to be distinguished here: namely, if $p > 1$, then $p - 1 > 0$, $\infty^{p-1} = \infty$ and the last term on the right of (4) is equal to zero. Hence, in this case the integral (3) converges. But if $p < 1$, then $\infty^{p-1} = \frac{1}{\infty^{1-p}} = 0$, and for this reason the last term in (4) is infinite, which means that in this case (3) diverges to infinity. In other words, the appropriate integral I_N taken from x_0 to N tends to infinity as N increases. We obtain an expression for I_N if N is substituted into the right member of (4) in place of ∞ (in other cases too, it is very easy to obtain an expression for I_N if the corresponding indefinite integral can be evaluated). It is quite evident that in this expression, for large N, the principal term for $p > 1$ is the first term and for $p < 1$ the second term.

For $p = 1$, the integral (3) is

$$\int_{x_0}^{\infty} \frac{C}{x}\,dx = C\ln x \Big|_{x_0}^{\infty} = C\ln\infty - C\ln x_0 = \infty$$

Again the integral diverges to infinity. Thus, (3) converges when $p > 1$ and diverges when $p \leqslant 1$.

On the basis of this result we can conclude, for example, that the integral

$$\int_0^{\infty} \frac{1}{\sqrt[3]{x^2+1}}\,dx \tag{5}$$

which has a singularity at the upper limit, diverges to infinity since for large x the integrand

$$\frac{1}{\sqrt[3]{x^2+1}} = \frac{1}{x^{2/3}}\,\frac{1}{\sqrt[3]{1+x^{-2}}} \tag{6}$$

is asymptotically equal to $1/x^{2/3}$, which is to say, in this case $p = 2/3 < 1$. On the contrary, the integral

$$\int_0^{\infty} \frac{1}{\sqrt{x^3+1}}\,dx \tag{7}$$

is convergent since the integrand is asymptotically equal to $1/x^{3/2}$ as $x \to \infty$; here, $p = 3/2 > 1$. The integral

$$\int_0^{\infty} e^{-x^2}\,dx \tag{8}$$

is also convergent since the integrand tends to zero faster than any power of x as $x \to \infty$. In all these examples, the corresponding indefinite integrals are not expressible in terms of elementary functions, so that it would be rather difficult to establish the convergence by evaluating the indefinite integral.

It is not hard to obtain asymptotic expressions of the last three integrals taken from 0 to N as N increases. The following device is used for a divergent integral of type (1): a function $f_1(x)$ is selected whose integral is easily found, and such that it is asymptotically (as $x \to \infty$) nearly equal to $f(x)$; then in the right member of

$$\int_a^N f(x)\,dx = \int_a^N f_1(x)\,dx + \int_a^N [f(x) - f_1(x)]\,dx$$

the first integral (principal term) is readily investigated, while the second integral may prove to be convergent as $N \to \infty$, or the same device may be applied. For the integral (5) it is natural to assume $f_1(x) = x^{-2/3}$, that is, to write

$$\int_0^N \frac{dx}{\sqrt[3]{x^2+1}} = \int_0^a \frac{dx}{\sqrt[3]{x^2+1}} + \int_a^N \frac{dx}{\sqrt[3]{x^2+1}}$$

$$= C_1 + \int_a^N \frac{1}{\sqrt[3]{x^2}}\,dx + \int_a^N \left(\frac{1}{\sqrt[3]{x^2+1}} - \frac{1}{\sqrt[3]{x^2}}\right) dx$$

$$= C_1 + 3\sqrt[3]{N} - 3\sqrt[3]{a} + \int_a^N \left(\frac{1}{\sqrt[3]{x^2+1}} - \frac{1}{\sqrt[3]{x^2}}\right) dx \qquad (9)$$

Here we passed from $\int_0^N$ to $\int_a^N$, where a is a positive number (although this is not obligatory in the given case), in order to avoid the improper integral $\int_0^N x^{-2/3}\,dx$, which has a singularity at the lower limit. It can be verified that as $N \to \infty$ the last integral in (9) is convergent and for this reason the entire expression (9), for large N, has the asymptotic representation $3\sqrt[3]{N} + C +$ an infinitesimal, where C is a constant. In order to find the value of the constant C, we have to take advantage of

$$C \approx \int_0^N \frac{dx}{\sqrt[3]{x^2+1}} - 3\sqrt[3]{N}$$

for some N, evaluating the integral in the right member via one of the formulas of numerical integration.

In similar fashion we investigate the asymptotic behaviour of the integrals (7) and (8). A useful transformation for a convergent integral is often

$$\int_a^N f(x)\,dx = \int_a^\infty f(x)\,dx - \int_N^\infty f(x)\,dx$$

For the integral (7) we get

$$\int_0^N \frac{dx}{\sqrt{x^3+1}} = \int_0^\infty \frac{dx}{\sqrt{x^3+1}} - \int_N^\infty \frac{1}{\sqrt{x^3}}(1+x^{-3})^{-1/2}\,dx$$

$$\approx \int_0^\infty \frac{dx}{\sqrt{x^3+1}} - \int_N^\infty \frac{dx}{\sqrt{x^3}} = D - \frac{2}{\sqrt{N}}$$

where the constant D, equal to the value of (7), can be computed as C was in the preceding paragraph.

Apply integration by parts to (8):

$$\int_0^N e^{-x^2}\,dx = \int_0^\infty e^{-x^2}\,dx - \int_N^\infty e^{-x^2}\,dx = E + \int_N^\infty \frac{1}{2x}\,de^{-x^2}$$

$$= E - \frac{1}{2N}e^{-N^2} + \frac{1}{2}\int_N^\infty \frac{e^{-x^2}}{x^2}\,dx \approx E - \frac{1}{2N}e^{-N^2}$$

The constant E, that is, the value of the integral (8), is equal to $\sqrt{\pi}/2$, as we shall see in Sec. 4.7.

By way of another illustration we consider the integral

$$\int_0^\infty \sin x\,dx \tag{10}$$

In this case the integral over a finite interval is

$$I_N = \int_0^N \sin x\,dx = -\cos x\Big|_0^N = 1 - \cos N \tag{11}$$

As N increases, the value of $\cos N$ oscillates and does not have a definite limit, which means (10) is a divergent integral that diverges in oscillatory fashion.

It is easy to verify that bringing a declining factor $e^{-\alpha x}$ (α a constant > 0) under the integral sign in (10) leads to the convergent integral

$$\int_0^\infty e^{-\alpha x} \sin x \, dx$$

We can also prove convergent an integral of a more general type,

$$\int_0^\infty f(x) \sin x \, dx$$

where $f(x)$ is any decreasing function that tends to zero as $x \to \infty$.

Improper integrals different from (1) are considered in a manner similar to (1). For example, suppose we have the integral

$$\int_a^b f(x) \, dx \tag{12}$$

for which the limits of integration are finite but the integrand goes to infinity as $x \to a$, that is, the integral has a singularity at $x = a$. This singularity is then "cut off", and instead of (12) we consider

$$\int_{a+\varepsilon}^b f(x) \, dx \tag{13}$$

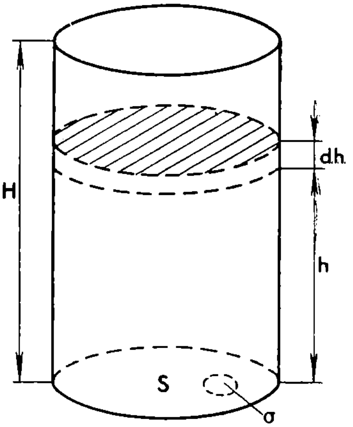
where ε is a small positive number. If for a sufficiently small ε the integral (13) for all practical purposes no longer depends on ε, then (12) is said to be a convergent integral and we put

$$\int_a^b f(x) \, dx = \lim_{\varepsilon \to 0} \int_{a+\varepsilon}^b f(x) \, dx$$

In this case we can pass from the integral (13) (which frequently appears in the solution of physical problems since all physical quantities are finite) to the simpler integral (12), that is to say, we ignore the contribution to the integral (12) of the singularity. Now if the integral (13) depends essentially on ε for small ε, that is, if it does not have a finite limit as $\varepsilon \to 0$, but tends to infinity or oscillates without a definite limit, then (12) is called a divergent integral; in this case we cannot pass from (13) to (12).

The convergence or divergence of an improper integral of type (12) is ordinarily established by comparing the integrand $f(x)$ with the power function $\frac{C}{(x-a)^p}$, which is also equal to infinity at $x = a$

Fig. 15



and is readily integrable. We leave it for the reader to verify that the improper integral

$$\int_b^a \frac{C}{(x-a)^p}\,dx \quad (C \text{ constant}) \tag{14}$$

converges for $p < 1$ and diverges for $p \geqslant 1$.

To illustrate, let us consider the problem of the flow of liquid from a cylindrical vessel, in the bottom of which is an opening of area σ (Fig. 15). The level h of the liquid depends on the time t, $h = h(t)$. If the liquid is not viscous and we can disregard the forces of surface tension, the rate v of discharge from the vessel can, to a sufficient degree of accuracy, be given by Torricelli's law:

$$v = \sqrt{2gh}$$

Therefore the volume discharged in time dt is

$$\sigma v\,dt = \sigma\sqrt{2gh}\,dt$$

On the other hand, the same volume is equal to $-\,Sdh$ (we take into account that h decreases and therefore $dh < 0$). Equating both expressions, we get

$$\sigma\sqrt{2gh}\,dt = -\,Sdh, \text{ that is, } dt = -\frac{S}{\sigma\sqrt{2g}}\frac{dh}{\sqrt{h}}$$

In order to obtain the total discharge time we have to integrate:

$$T = -\frac{S}{\sigma\sqrt{2g}}\int_H^0 \frac{dh}{\sqrt{h}} = -\frac{S}{\sigma\sqrt{2g}}\left.\frac{h^{1/2}}{\frac{1}{2}}\right|_H^0 = \frac{S}{\sigma}\sqrt{\frac{2H}{g}} \tag{15}$$

Actually, the discharge does not reach $h = 0$ but only $h = \varepsilon$, where ε is a small quantity comparable to the roughness of the bottom or to the thickness of wetting film; formula (15) ought to be rewritten thus:

$$T = -\frac{S}{\sigma\sqrt{2g}}\int_H^{\varepsilon}\frac{dh}{\sqrt{h}} \tag{16}$$

However, since the improper integral (15) turned out convergent $\left(\text{this was evident from the computations of (15); what is more, it is an integral of type (14) for } p = \frac{1}{2}\right)$, the integral (16) may be replaced by (15). It is evident here that ε is not known exactly and also that it is not essential because for a convergent integral the important thing to know is that ε is small.

In numerical integration (cf. Sec. 1.1) improper integrals require particular attention. It often happens that the given integral is represented in the form of the sum of a proper integral, obtained by excluding the interval about the singularity from the interval of integration, and the improper integral taken over the interval near the singularity. The former is found numerically and the latter by a series expansion, or the integrand is merely replaced approximately by some other function (a power function, for example) whose integral can be readily evaluated. To illustrate, let us evaluate the integral $\int_0^{\pi}\frac{dx}{\sqrt{\sin x}}$. Here the integrand increases without bound as x approaches 0 and as x approaches π. We partition the interval of integration into three subintervals: from 0 to $\pi/6$, from $\pi/6$ to $5\pi/6$, and from $5\pi/6$ to π. In the first subinterval we can assume that $\sin x \approx x$, since x is small. Therefore

$$\int_0^{\pi/6}\frac{dx}{\sqrt{\sin x}} \approx \int_0^{\pi/6}\frac{dx}{\sqrt{x}} = 2\sqrt{x}\,\Big|_0^{\pi/6} = 2\sqrt{\frac{\pi}{6}} = 1.45 \quad *$$

* A series expansion can be applied for greater accuracy:

$$\int_0^{\alpha}\frac{dx}{\sqrt{\sin x}} = \int_0^{\alpha}\frac{1}{\sqrt{x}}\left(\frac{\sin x}{x}\right)^{-1/2}dx = \int_0^{\alpha}\frac{1}{\sqrt{x}}\left(1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \dots\right)^{-1/2}dx$$

$$= \int_0^{\alpha}\frac{1}{\sqrt{x}}\left(1 + \frac{x^2}{12} + \frac{x^4}{160} + \dots\right)dx = \sqrt{\alpha}\left(2 + \frac{\alpha^2}{30} + \frac{\alpha^4}{720} + \dots\right)$$

However, for the accuracy used here the correction is negligible (we get 1.46). The only useful thing is that we get some idea of the degree of reliability of the result.

In the third subinterval, $5\pi/6 < x < \pi$, we take advantage of the formula $\sin x = \sin(\pi - x)$, and since the quantity $\pi - x$ is small, we can put $\sin(\pi - x) \approx \pi - x$. In this subinterval we have $\sin x \approx \approx \pi - x$. We obtain

$$\int_{5\pi/6}^{\pi} \frac{dx}{\sqrt{\sin x}} \approx \int_{5\pi/6}^{\pi} \frac{dx}{\sqrt{\pi - x}} = -2\sqrt{\pi - x}\,\Big|_{5\pi/6}^{\pi} = 2\sqrt{\frac{\pi}{6}} = 1.45$$

We compute the integral over the second subinterval using Simpson's rule and dividing the subinterval into two pieces to get

$$\int_{\pi/6}^{5\pi/6} \frac{dx}{\sqrt{\sin x}} \approx \frac{\pi}{9}[1.41 + 4.1 + 1.41] = 2.38$$

Consequently

$$\int_{0}^{\pi} \frac{dx}{\sqrt{\sin x}} \approx 1.45 + 2.38 + 1.45 = 5.28$$

The exact value of this integral, to two decimal places, is 5.25.

Exercise

Evaluate the integral $\int_0^{\pi/2} \frac{dx}{\sqrt[8]{\sin x}}$ carrying out the calculations to within slide-rule accuracy.

3.2 Integrating rapidly varying functions

When integrating numerically, it is useful to be able to estimate the order of magnitude of the integral beforehand. From the geometric meaning of the integral

$$I = \int_a^b y(x)\,dx$$

there follows immediately the obvious estimate

$$I < \int_a^b y_{\max}\,dx = y_{\max}\cdot(b - a)$$

where $y_{\max}$ is the maximum value of the integrand $y(x)$ on the interval of integration. If this function is positive and varies but slightly over the interval of integration, we can take $y \approx y_{\max}$, or

$$I \lesssim y_{\max}\cdot(b - a) \qquad (17)$$

(this estimate occurred in HM, Sec. 3.16).

It is well to observe from the start that (17) and also the subsequent estimates of this section are inconvenient in case of alternating functions $y(x)$. For alternating functions the interval of integration may be divided into several parts so that within each part the sign remains unchanged; then an estimate is made of the integrals over these parts. However, the overall estimate will be satisfactory only if the contribution of integrals of one sign essentially exceeds the contribution of integrals of the opposite sign.

For this reason, throughout this section we will regard the integrand as positive over the interval of integration.

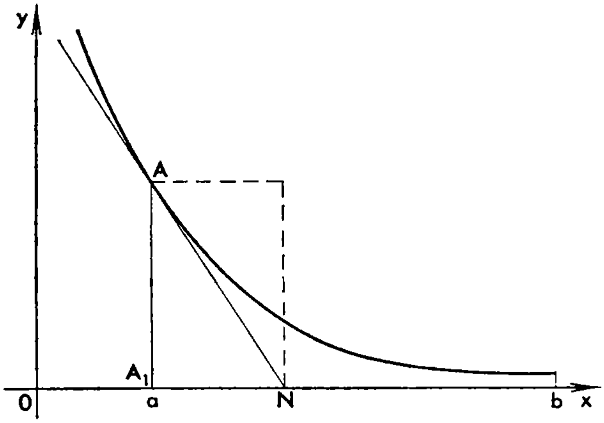
If the function $y(x)$ over the interval of integration is positive but decreases rapidly and b is comparatively large, then the estimate (17) may lead to substantial errors. This is so because when we use (17), we replace the function with its maximum value. But if the function varies rapidly, then its values are close to maximum only over a small portion of the region of integration. To illustrate we consider $I = \int\limits_a^b e^{-x}\,dx$ $(b > 0)$. The maximum value of the integrand over the integration interval is obtained at $x = 0$. It is equal to 1. The estimate (17) yields $I \approx b$. But in this instance it is easy to obtain an exact formula for the integral: $I = 1 - e^{-b}$. Form a table of the exact value of I as a function of b:

b	0	0.1	0.2	0.5	1	2	3	5	10
I	0	0.095	0.18	0.39	0.63	0.86	0.95	0.993	0.99996

It is clear from this table that the estimate (17) is fairly decent as long as b is small (in which case the function varies but slightly over the interval of integration). But if b is great, then the approximation $I \approx b$ becomes very poor.

Let a function $y(x)$ be decreasing rapidly over the interval of integration. Then the maximum value of the function is attained at the left endpoint of the interval, that is, at $x = a$. (Observe that this does not imply the equation $\left.\frac{dy}{dx}\right|_{x=a} = y'(a) = 0$, see Fig. 16.) Since for the rapidly decreasing function $y(x)$ the integral $I \int\limits_a^b y\,dx$ cannot substantially vary as b increases, a rough estimate of the integral I should not include b; we assume, as it were, $b = \infty$. It is natural to assume that in this case the integral is roughly equal to the product of $y_{\max} = y(a)$ into the length Δx (independent of b) of the integration interval. For a given $y(a)$, this length should be the smaller, the faster the function is decreasing, that is to say, the greater $|y'(a)|$.

Fig. 16



The quantity Δx having the same dimensions as x can be constructed in unique fashion by proceeding from $y(a)$ and $|y'(a)|$:

$$\Delta x = \frac{y(a)}{|y'(a)|} m$$

where m is a nondimensional constant of proportionality. We then get

$$I \approx y(a)\,\Delta x = \frac{y^2(a)}{|y'(a)|} m \tag{18}$$

If the function $y(x)$ is increasing over the interval of integration, then it reaches a maximum at the right endpoint of the interval, that is, at $x = b$. Then (18) becomes

$$I \approx \frac{y^2(b)}{y'(b)} m$$

A typical example of a rapidly varying function is the exponential function $y = Ce^{-kx}$ $(C > 0,\ k > 0)$. Choose a value m in (18) so that (18) is absolutely exact for $\int\limits_a^\infty Ce^{-kx}\,dx$.

Since $y = Ce^{-kx}$, $y' = -kCe^{-kx}$, it follows that $y(a) = Ce^{-ka}$ and $|y'(a)| = kCe^{-ka}$, and so (18) yields

$$I = m\frac{C^2e^{-2ka}}{Cke^{-ka}} = m\frac{Ce^{-ka}}{k} = \frac{m}{k}y(a)$$

The exact value of the integral at hand is

$$I = \int\limits_a^\infty Ce^{-kx}\,dx = -\frac{C}{k}e^{-kx}\Big|_a^\infty = \frac{1}{k}y(a)$$

Comparing results, we get $\frac{m}{k}y(a) = \frac{1}{k}y(a)$, whence $m = 1$. And so the formula (18) takes the form

$$I \approx \frac{y^2(a)}{|y'(a)|} \tag{19}$$

Generally speaking, this formula is not exact in the case of an infinite interval for a rapidly varying function of a different type and also in the case of a finite interval, but it yields fairly decent results.

In order to make the meaning of (19) pictorial, draw a tangent to the curve $y = y(x)$ at the point A (Fig. 16) and find the length of A_1N. The equation of the tangent is $y - y(a) = y'(a)(x - a)$. Setting $y = 0$, we get the point of intersection of the tangent and the x-axis. This produces $x = a - \frac{y(a)}{y'(a)}$ or, noting that $y'(a) < 0$ since $y(x)$ is a decreasing function, we get $x = a + \frac{y(a)}{|y'(a)|}$, and so

$$A_1N = a + \frac{y(a)}{|y'(a)|} - a = \frac{y(a)}{|y'(a)|} = \Delta x$$

The estimate of the integral by means of (19) amounts to replacing the area under the curve $y = y(x)$ by the area of the rectangle in Fig. 16.

Let us consider an example. Use (19) to evaluate $\int\limits_1^\infty \frac{dx}{x^8}$. Here, $y = \frac{1}{x^8}$, $y' = -\frac{8}{x^9}$, and so $y(1) = 1$, $|y'(1)| = 8$,

$$\int\limits_1^\infty \frac{dx}{x^8} \approx \frac{1}{8} = 0.125$$

The exact value of this integral is $I = \int\limits_1^\infty \frac{dx}{x^8} = -\left.\frac{dx}{7x^7}\right|_1^\infty = \frac{1}{7} = 0.143.$ The error comes to 13%.

Another frequently occurring type of integral is that in which the integrand $y(x)$ reaches a maximum at $x = x_m$ somewhere inside the interval of integration (Fig. 17a). Here $y'(x) = 0$ at the maximum. We can split the integral into two: one from a to x_m and the other from x_m to b. Then in each of them the integrand reaches a maximum at an endpoint of the respective interval. It might appear that this problem then reduces to the preceding one. Actually this is not so because splitting the integral into two gives nothing since at $x = x_m$ we still have $y' = 0$ and the estimate (19) is inapplicable. Thus, this is indeed a new case and we have to find some other way to isolate

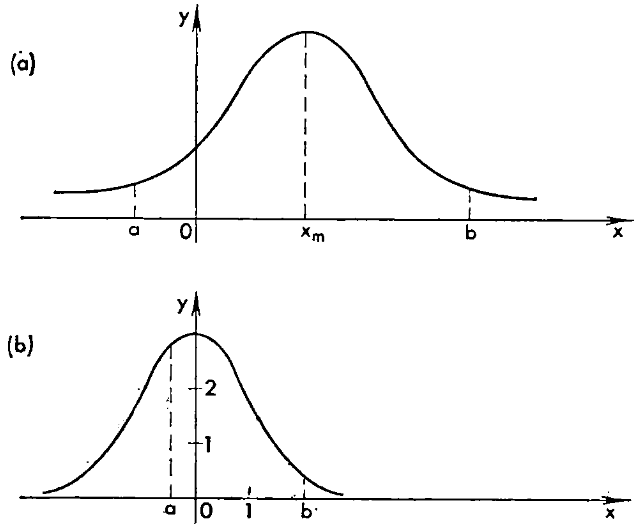


Fig. 17

the necessary part Δx from the interval of integration. The idea here is that the quantity Δx is determined by the value of $y''(x_m)$, that ts, it is determined by the magnitude of the curvature at the maximum point. From the figure it is clear that the steeper the curve, the smaller Δx must be. The dimension of the second derivative coincides with that of the quantity $\frac{y}{x^2}$, and so a quantity of the same dimension as Δx is obtained from the quantities $y(x_m)$ and $y''(x_m)$ thus:

$$\Delta x = l\sqrt{\frac{y(x_m)}{|y''(x_m)|}}\ ^*$$

[*] We can also arrive at an expression of this type for Δx in the following manner: expand $y(x)$ in a Taylor series near the maximum, that is, in powers of $x - x_m$, and take the first two terms that do not vanish. We get $y(x) = y(x_m) + \frac{1}{2}(x - x_m)^2 \cdot y''(x_m)$. Then find the value of the difference $x - x_m = \Delta x$ at which this approximate expression for $y(x)$ vanishes:

$$y(x_m) + \frac{1}{2}(x - x_m)^2 \cdot y''(x_m) = 0$$

whence

$$\Delta x = x - x_m = \sqrt{-\frac{2y(x_m)}{y''(x_m)}} = \sqrt{\frac{2y(x_m)}{|y''(x_m)|}}$$

Accordingly, the approximate value of I is

$$I \approx \int\limits_{x_m-\Delta x}^{x_m+\Delta x}\left[y(x_m) + \frac{1}{2}(x - x_m)^2 y''(x_m)\right]dx = \frac{4}{3}\,y(x_m)\,\Delta x = \sqrt{\frac{32y^3(x_m)}{9|y''(x_m)|}}$$

(We write $|y''(x_m)|$ instead of $y''(x_m)$ because $y''(x_m) < 0$, since for $x = x_m$ the function $y(x)$ has a maximum.) The quantity l is a dimensionless coefficient. For the integral we have the estimate

$$I \approx y(x_m) \cdot \Delta x = l \cdot y(x_m) \sqrt{\frac{y(x_m)}{|y''(x_m)|}} = l \sqrt{\frac{y^3(x_m)}{|y''(x_m)|}} \tag{20}$$

We determine the value of the coefficient l from the condition that formula (20) be absolutely exact for $\int\limits_{-\infty}^{+\infty} y\,dx$, where

$$y(x) = Ce^{-kx^2}, \quad k > 0,\ C > 0$$

(The graph of the function $y = Ce^{-kx^2}$ for the case $C = 3$, $k = 0.5$ is shown in Fig. 17*b*).

Here, $x_m = 0$, $y(x_m) = C$, $y''(x_m) = -2Ck$. By (20) we get

$$I = l \sqrt{\frac{C^3}{2Ck}} = \frac{lC}{\sqrt{2k}} \tag{21}$$

To find the exact value of the integral $\int\limits_{-\infty}^{+\infty} Ce^{-kx^2}\,dx$ make the change of variable $z = x\sqrt{k}$, $dz = \sqrt{k}\,dx$. We then get

$$\int\limits_{-\infty}^{+\infty} Ce^{-kx^2}\,dx = \frac{C}{\sqrt{k}} \int\limits_{-\infty}^{+\infty} e^{-z^2}\,dz$$

The value of the integral $\int\limits_{-\infty}^{\infty} e^{-z^2}\,dz$ may be found exactly: in Sec. 4.7 we will show that this integral is equal to $\sqrt{\pi}$. Therefore

$$\int\limits_{-\infty}^{+\infty} Ce^{-kx^2}\,dx = C\sqrt{\frac{\pi}{k}}$$

Comparing this formula with (21), we get

$$\frac{lC}{\sqrt{2k}} = C\sqrt{\frac{\pi}{k}}$$

whence $l = \sqrt{2\pi}$. Formula (20) takes the form

$$I \approx \sqrt{2\pi \frac{y^3(x_m)}{|y''(x_m)|}} \tag{22}$$

Thus, for the types of rapidly varying functions we obtained two formulas, (19) and (20), and the coefficients in these formulas are chosen so that the formulas are exact for integrals of typical functions over an infinite interval: (19) is exact for $\int\limits_{a}^{\infty} Ce^{-kx}\,dx$ and (22) is exact for $\int\limits_{-\infty}^{+\infty} Ce^{-kx^2}\,dx$. For a different choice of typical functions the formulas would contain different coefficients. For instance, let us determine the value of the coefficient l from the condition that formula (20) be exact for $I = \int\limits_{-\infty}^{+\infty} \frac{C}{1+kx^2}\,dx$, $C > 0$, $k > 0$. (The reader will readily see that the function $y = \frac{C}{1+kx^2}$ has a maximum at $x_m = 0$.) Since $y = \frac{C}{1+kx^2}$, $y'' = 2Ck\frac{3kx^2-1}{(1+kx^2)^3}$, (20) gives

$$I = l\sqrt{\frac{C^3}{2Ck}} = \frac{lC}{\sqrt{2k}}$$

To compute the integral exactly, set $\sqrt{k}x = z$, $dz = \sqrt{k}\,dx$; then

$$I = \frac{C}{\sqrt{k}}\int\limits_{-\infty}^{+\infty}\frac{dz}{1+z^2} = \frac{C\pi}{\sqrt{k}}$$

Therefore $\frac{lC}{\sqrt{2k}} = \frac{C\pi}{\sqrt{k}}$, whence $l = \pi\sqrt{2}$. In this case we get the formula

$$I \approx \pi\sqrt{\frac{2y^3(x_m)}{|y''(x_m)|}}$$

However, preference is given precisely to formulas (19) and (22). The reason is this. It turns out that if a rapidly varying function is obtained by raising to the power n some given function $f(x)$ for which $|f(x)| < f(x_m)$ when $x \neq x_m$, then for sufficiently large n the relative error of (19) and (22) becomes arbitrarily small.

We illustrate with a couple of examples.

1. We will integrate successive powers of the function $u = \frac{1}{1+x}$ from $x = 0$ to $x = \infty$. Set $u_n = u^n = \frac{1}{(1+x)^n}$. These are rapidly varying functions of the first type with maximum at an endpoint,

and the greater the exponent n, the steeper the function u_n falls as x increases from 0. We find $I^{(n)} = \int_0^{\infty} \frac{dx}{(1+x)^n}$ with the aid of (19).

Since $u_n' = \frac{-n}{(1+x)^{n+1}}$, it follows that $|u_n'(0)| = n$ and so the approximate value of this integral is

$$I_{\text{ap}}^{(n)} = \frac{1}{n}$$

The exact value of the integral is

$$I_{\text{ex}}^{(n)} = \int_0^{\infty} \frac{dx}{(1+x)^n} = \frac{1}{-n+1} \cdot \frac{1}{(1+x)^{n-1}} \Big|_0^{\infty} = \frac{1}{n-1}$$

The ratio

$$\frac{I_{\text{ap}}^{(n)}}{I_{\text{ex}}^{(n)}} = \frac{n-1}{n}$$

Since the fraction $\frac{n-1}{n}$ gets closer to unity as n increases, we have

$$\frac{I_{\text{ap}}^{(n)}}{I_{\text{ex}}^{(n)}} \approx 1$$

for very large values of n, which is what we asserted. If in (18) we chose $m \neq 1$, the value of m would have remained in the right-hand member of the last formula, that is, for large n we would have a systematic error.

2. Suppose $z = \frac{1}{1+x^2}$ with $-\infty < x < \infty$. We form the rapidly varying functions $z_n = z^n = \frac{1}{(1+x^2)^n}$. These are functions of the second kind with maximum inside the region (here, $x_m = 0$). We approximate $I^{(n)} = \int_{-\infty}^{+\infty} \frac{dx}{(1+x^2)^n}$ using (22) and get $I_{\text{ap}}^{(n)} = \sqrt{\frac{\pi}{n}}$.

The integral $I^{(n)}$ can be computed exactly:

$$I_{\text{ex}}^{(n)} = \frac{1 \cdot 3 \cdot 5 \dots (2n-3)}{2 \cdot 4 \cdot 6 \dots (2n-2)} \pi \quad *$$

* Prove this by integrating the equation

$$\left[\frac{x}{(1+x^2)^{n-1}}\right]' = -\frac{2n-3}{(1+x^2)^{n-1}} + \frac{2n-2}{(1+x^2)^n}$$

The ratio

$$\frac{I_{\mathrm{ap}}^{(n)}}{I_{\mathrm{ex}}^{(n)}}=\frac{2\cdot 4\cdot 6\ldots(2n-2)}{\sqrt{\pi n}\,1\cdot 3\cdot 5\ldots(2n-3)} \tag{23}$$

Let us compute the values of the right member of (23) for several values of n. For $n = 2$ we get 0.797. For $n = 4$ we have 0.904; for $n = 6$ we have 0.938; for $n = 8$ we get 0.962; and finally for $n = 10$ we obtain 0.978. Thus, as n increases, the value of the right member of (23) approaches 1*.

We now pass to the general case. Let us consider $y = [f(x)]^n$ where the function $f(x) > 0$ has a unique maximum point. Let $\ln f(x) = \varphi(x)$, then $f(x) = e^{\varphi(x)}$. Therefore

$$y = e^{n\varphi(x)} \tag{24}$$

The integral $I = \int\limits_a^b y\,dx$ is mainly determined by the values of the integrand in the region where y does not differ substantially from its maximum value.

It is clear that y attains a maximum at the same time as $\varphi(x)$, that is, for one and the same value $x = x_m$. For y to decrease e times as compared with $y_{\max}$, it is necessary that $n\varphi(x)$ be less than $n\varphi(x_m)$ by unity, that is, that $n\varphi(x) = n\varphi(x_m) - 1$, whence

$$\varphi(x) = \varphi(x_m) - \frac{1}{n}$$

The greater n, the less $\varphi(x)$ differs from $\varphi(x_m)$, and so the greater n, the more exact is the replacement of $\varphi(x)$ by the first two nonzero terms in the Taylor expansion. Writing down the two terms of the expansion, we get $\varphi(x) = \varphi(x_m) + (x - x_m)\,\varphi'(x_m)$ (for a function having a maximum at an endpoint) or $\varphi(x) = \varphi(x_m) + \frac{1}{2}(x - x_m)^2\varphi''(x_m)$ (for a function with a maximum inside the interval).

In the former case, using (24) we get

$$y = e^{n\varphi(x_m)+n(x-x_m)\varphi'(x_m)} = Ae^{-b(x-x_m)}$$

where $A = e^{n\varphi(x_m)}$, $b = -n\varphi'(x_m) = n|\varphi'(x_m)|$.

In the latter case we obtain

$$y = e^{n\varphi(x_m)+\frac{1}{2}n(x-x_m)^2\cdot\varphi''(x_m)} = Ae^{-c(x-x_m)^2}$$

where $c = -\frac{1}{2}n\varphi''(x_m) = \frac{1}{2}n|\varphi''(x_m)|$.

* This can be proved in a more rigorous fashion with the aid of Stirling's formula (see the exercise in Sec. 3.3).

Thus, the function $y(x)$ may be approximately replaced with either a function for which (19) is absolutely exact or a function for which (22) is absolutely exact, the replacement being the more exact, the greater n is.

For the sake of simplicity we assume that $x_m = 0$. Then in order to construct a rapidly varying function we could go over to $f(nx)$ instead of $[f(x)]^n$. But in that case, both members of (19), or, respectively, (22), are merely divided by n, that is to say, the relative error of both formulas does not change. This is due to the fact that in the indicated transition the relative portion of the contribution to the integral of large and small values of the function $f(x)$ remains unchanged (whereas when we pass from $f(x)$ to $[f(x)]^n$, the portion of small values tends to zero with increasing n).

In conclusion, we recommend *Approximate Methods in Quantum Mechanics* by Migdal and Krainov [11], which contains a variety of methods for estimating integrals and other mathematical expressions.

Exercises

1. Find the integral $I = \int\limits_0^\infty \frac{x\,dx}{1+e^x}$ by splitting it into a sum of two integrals: $I = \int\limits_0^3 \frac{x\,dx}{1+e^x} + \int\limits_3^\infty \frac{x\,dx}{1+e^x}$. Evaluate the first integral by Simpson's rule, the second integral by the formulas of this section. Carry the computations to within slide-rule accuracy.

2. Evaluate the integral $I = \int\limits_0^\infty \sqrt{x}e^{-x^2}\,dx$ by splitting it into a sum of two integrals: $I = \int\limits_0^a \sqrt{x}e^{-x^2}\,dx + \int\limits_a^\infty \sqrt{x}e^{-x^2}\,dx$. Find the first integral using the formulas of Sec. 1.1, the second using the formulas of Sec. 3.2. Consider the cases $a = 1$ and $a = 2$. Carry the calculations to three decimal places.

3. Estimate the integral $\int\limits_N^\infty e^{-x^2}\,dx$ considered on p. 69.

3.3 Stirling's formula

As an interesting example in the use of the formulas of Sec. 3.2 we obtain a convenient formula for approximating $n! = n(n-1) \times \times (n-2) \dots 3 \cdot 2 \cdot 1$ for large n's. This formula will come in handy

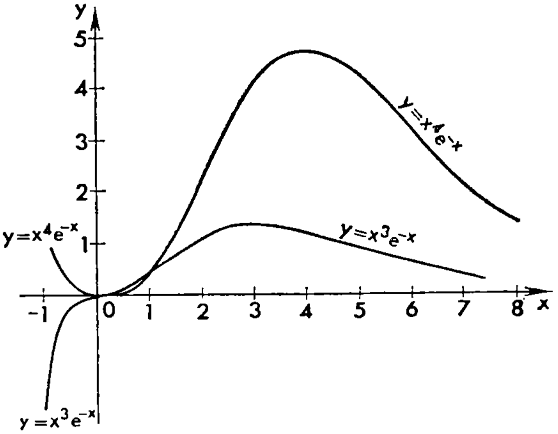


Fig. 18

in the study of probability theory. By means of integration by parts it is easy to establish (see, for example, HM, Sec. 4.3) that the following equation holds true for n a positive integer:

$$\int_0^\infty x^n e^{-x}\,dx = n!$$

Let us estimate this integral by the method of the preceding section.

In our case $y = x^n e^{-x}$, $y' = (nx^{n-1} - x^n)\,e^{-x}$. Equating the first derivative to zero, we get two values: $x = 0$ and $x = n$. It is easy to see that the function $y(x)$ has a maximum at $x = n$ and is zero at $x = 0$. (Fig. 18 depicts the graphs of $y = x^n e^{-x}$ for $n = 3$ and $n = 4$.) Thus, to evaluate the integral we have to consider the region of the maximum of the function, $x = n$, which means using formula (22).

To find y'', we get $y'' = [n(n-1)\,x^{n-2} - 2nx^{n-1} + x^n]\,e^{-x}$ and so $y''(n) = -\,n^{n-1}e^{-n} = -\,\frac{1}{n}\left(\frac{n}{e}\right)^n\cdot$ Therefore

$$I \approx \sqrt{\frac{2\pi y^3(n)}{|\,y''(n)\,|}} = \sqrt{2\pi\left(\frac{n}{e}\right)^{3n} n\left(\frac{n}{e}\right)^{-n}} = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Thus,

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

This is *Stirling's formula*. Its relative error tends to zero with increasing n (this follows from the reasoning of Sec. 3.2 if we represent

$x^n e^{-x}$ as $n^n\left(\frac{x}{n}e^{-\frac{x}{n}}\right)^n$. But even for small n it yields very good results. For example,

$$n = 1,\ n! = 1, \qquad \sqrt{2\pi}\left(\frac{1}{e}\right)^1 = 0.92, \text{ error } 8\%;$$

$$n = 2,\ n! = 2, \qquad \sqrt{4\pi}\left(\frac{2}{e}\right)^2 = 1.92, \text{ error } 4\%;$$

$$n = 3,\ n! = 6, \qquad \sqrt{6\pi}\left(\frac{3}{e}\right)^3 = 5.84, \text{ error } 2.7\%;$$

$$n = 4,\ n! = 24, \qquad \sqrt{8\pi}\left(\frac{4}{e}\right)^4 = 23.5, \text{ error } 2.1\%;$$

$$n = 5,\ n! = 120, \qquad \sqrt{10\pi}\left(\frac{5}{e}\right)^5 = 118, \text{ error } 1.7\%$$

Exercise

Prove that the ratio (23) tends to 1 as $n \to \infty$.

Hint. Multiply the numerator and the denominator of the fraction by the numerator.

3.4 Integrating rapidly oscillating functions

In the study of rapidly oscillating actions on physical systems one had to consider integrals of rapidly oscillating functions, that is, functions which change sign many times over a finite interval of integration. Such integrals have peculiarities.

Suppose we have to compute the integral

$$I = \int_a^b F(x)\,dx \tag{25}$$

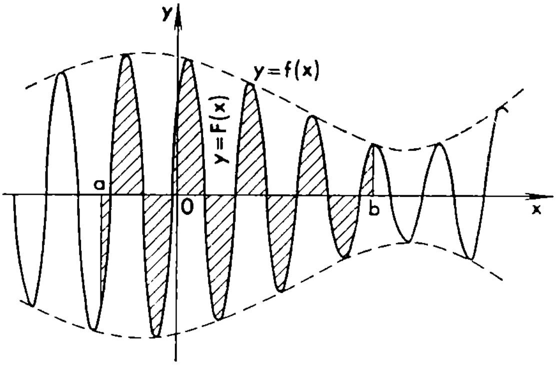
where the graph of $F(x)$ has the form shown in Fig. 19. At first we assume ω (frequency) to be constant and the amplitude to be varying in accordance with the law $y = f(x)$; in other words, we assume the integral (25) to have the form

$$I = \int_a^b f(x)\sin(\omega x + \alpha)\,dx \tag{26}$$

where ω is great.

From Fig. 19 it is evident that (26) is small for large ω since the positive part is almost neutralized by the negative part. To

Fig. 19



obtain a more accurate estimate, perform an integration by parts to get

$$I = \frac{1}{\omega}\,[f(a)\cos(\omega a + \alpha) - f(b)\cos(\omega b + \alpha)] + \frac{1}{\omega}\int_a^b f'(x)\cos(\omega x + \alpha)\,dx \qquad (27)$$

This integral has the form of (26) and so the entire last term, for large ω, is of higher order than $1/\omega$. Dropping this term, we get the approximate formula

$$I \approx \frac{1}{\omega}\,[f(a)\cos(\omega a + \alpha) - f(b)\cos(\omega b + \alpha)] \qquad (28)$$

This result can also be expressed in terms of the integrand $F(x)$ of the original integral (25): since

$$F'(x) = f'(x)\sin(\omega x + \alpha) + \omega f(x)\cos(\omega x + \alpha) \approx \omega f(x)\cos(\omega x + \alpha)$$

we can also write

$$I \approx \frac{1}{\omega^2}[F'(a) - F'(b)] = -\frac{1}{\omega^2}F'(x)\Big|_a^b \qquad (29)$$

For more precision, we can integrate by parts once more in the right member of (27); and after dropping the resulting integral we arrive at the approximate formula

$$I \approx \frac{1}{\omega}\,[f(a)\cos(\omega a + \alpha) - f(b)\cos(\omega b + \alpha)] + \frac{1}{\omega^2}[f'(b)\sin(\omega b + \alpha) - f'(a)\sin(\omega a + \alpha)] \qquad (30)$$

Here too we can write a formula similar to (29). To do this, from expressions for $F'(x)$ and $F'''(x)$ compute $f(x)\cos(\omega x + \alpha)$ and $\frac{1}{\omega}f'(x)\sin(\omega x + \alpha)$ to within terms of the order of $1/\omega^2$ and substitute the result into (30). We leave the manipulations to the reader and write down the final formula:

$$I \approx -\frac{1}{\omega^2}\left[2F'(x) + \frac{1}{\omega^2}F'''(x)\right]\Bigg|_a^b$$

Like (30), this formula is true to within terms of the order of $\frac{1}{\omega^3}$.

Using the foregoing procedure it is possible to refine the asymptotic formulas for I, but the resulting formulas will be progressively more unwieldy and inconvenient for practical use. In practice, the most frequently used formulas are (28) and (29). It is interesting to note that in all these formulas the values of the functions f or F and their derivatives participate only at the endpoints of the interval of integration. Incidentally, this is in accord with the approximate formulas for alternating sums obtained in Sec. 1.2.

Note that in the foregoing formulas we had to assume that the function $f(x)$ and its derivatives of the orders under consideration are continuous within the interval of integration. If $f(x)$ has a finite jump there at $x = c$, then we have to pass to the sum of integrals from a to c and from c to b, and only then apply the indicated transformations to these integrals, as a result of which the point $x = c$ will make its contribution to the asymptotic formulas. This contribution is particularly substantial in the important case where $a = -\infty$, $b = \infty$ and $f(x)$ together with all its derivatives vanishes at $x = \pm\infty$, for then the right sides of (28) and (30) are zero. This question will be dealt with in more detail in Sec. 14.4.

As an example we consider the integral $I = \int\limits_0^b e^{-x}\sin\omega x\,dx$. Its exact value is

$$I_{ex} = \frac{\omega}{1+\omega^2} - \frac{e^{-b}}{1+\omega^2}(\sin\omega b + \omega\cos\omega b)$$

Formula (28) yields the approximate value

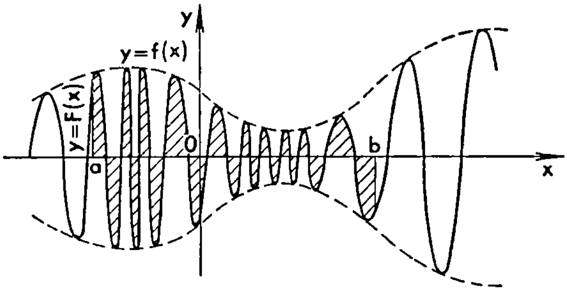
$$I_{(28)} = \frac{1}{\omega}(1 - e^{-b}\cos\omega b)$$

and (29) the value

$$I_{(29)} = \frac{1}{\omega^2}[\omega - e^{-b}(\omega\cos\omega b - \sin\omega b)]$$

Both approximate formulas have an error of the order of $1/\omega^2$.

Fig. 20



Now consider a case where the rapidly oscillating function $F(x)$ under the sign of the integral (25) has a graph as shown in Fig. 20; that is, it varies in both amplitude and frequency. In this case the integral (25) can often be written in the form

$$I = \int_a^b f(x) \sin(\omega\varphi(x) + \alpha)\, dx \tag{31}$$

where $\varphi(x)$ is an increasing (but not at a constant rate!) function.

The integral (31) can be reduced to (26) by means of a change of variable: $\varphi(x) = s$. If we denote the inverse function by $x = g(s)$, we get

$$I = \int_{\varphi(a)}^{\varphi(b)} f(g(s))\, g'(s) \sin(\omega s + \alpha)\, ds$$

Using the approximate formula (28) gives

$$I \approx -\frac{1}{\omega}[f(g(s))\, g'(s) \cos(\omega s + \alpha)]\Big|_{s=\varphi(a)}^{\varphi(b)}$$

$$= -\frac{1}{\omega}\left[\frac{f(x)}{\varphi'(x)} \cos(\omega\varphi(x) + \alpha)\right]\Bigg|_{x=a}^{b} \tag{32}$$

(here we made use of the formula $g'(s) = 1/\varphi'(x)$ for the derivative of the inverse function). If $\varphi(x) \equiv x$, then (32) becomes (28). The modified formula (29), which we suggest that the reader derive, becomes

$$I \approx -\frac{F'(x)}{\omega^2[\varphi'(x)]^2}\Bigg|_a^b$$

Exercises

1. Apply (28), (29) and (30) to the integral $\int_{-a}^{a} \frac{\cos \omega x}{1 + x^2} dx$.

2. Write an analogue of formula (30) for the integral (31).

3. Evaluate the integral $I = \int\limits_1^2 \frac{\sin \omega x}{x} dx$ at $\omega = 1$ and $\omega = 10$, using the approximate formulas (28), (29) and (30). Compare the results with the exact values taken from tables of the sine integral.

3.5 Numerical series

A numerical series is an "infinite sum" of numbers

$$a_1 + a_2 + a_3 + \dots + a_n + a_{n+1} + \dots \text{*} \tag{33}$$

This is not a real sum, of course, since only a finite number of numbers can be added. It is a sum with a singularity, similar to the singularities of improper integrals considered in Sec. 3.1. For this reason, the approach to the concept of the sum of the series (33) is similar to that applied in Sec. 3.1; namely, first the singularity is cut off, as it were, and one considers the *partial sums* of the series (33):

$$S_1 = a_1, \quad S_2 = a_1 + a_2, \quad S_3 = a_1 + a_2 + a_3 \dots,$$
$$S_n = a_1 + a_2 + a_3 + \dots + a_n \tag{34}$$

Now if we increase n so as to "exhaust the singularity" and watch the behaviour of the partial sum (34), two possible cases emerge.

(1) The partial sum may approach a definite finite limit S so that for large n it is practically equal to S. In this case (33) is said to be a *convergent series*, and the sum of the series is set equal to S. Thus, in the case of convergence we can go from a partial sum with a large n to the complete sum of the series and conversely, that is, the contribution of the singularity to the total sum of the series is not essential and is arbitrarily small for large n.

(2) The partial sum may tend to infinity or may oscillate without having a definite limit. In this case, (33) is said to be a ***divergent series***.

The simplest example of a convergent series is given by the sum of an infinite decreasing geometric progression:

$$a + aq + aq^2 + \dots + aq^{n-1} + aq^n + \dots \ (|q| < 1) \tag{35}$$

Here, as we know,

$$S_n = a\frac{1 - q^n}{1 - q}$$

and since the power q^n can be neglected for large n, the sum of the series (35) is, in the limit,

$$S = \lim_{n\to\infty} S_n = \frac{a}{1 - q}$$

* An elementary exposition of series is given in HM, Sec. 3.17. Here we continue the discussion, which makes use of the results of Sec. 1.2.

The series $1 + 1 + 1 + \ldots$ is an instance of a series diverging to infinity, and the series $1 - 1 + 1 - 1 + 1 - 1 + \ldots$ is an instance of a series diverging in oscillatory fashion, since its partial sums are equal successively to 1, 0, 1, 0, 1, ... and do not have a definite limit.

Since in the case of a convergent series the partial sums with large n are nearly the same, terms with large n are nearly equal to zero: to be more precise, if the series (33) converges, then its "general term" a_n tends to zero as n increases. However, in the case of a divergent series the general term may tend to zero too. For example, the divergent series

$$1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \ldots + \frac{1}{\sqrt{n}} + \ldots$$

(The divergence of this series can be demonstrated thus: $S_n = 1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \ldots + \frac{1}{\sqrt{n}} > \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n}} + \ldots + \frac{1}{\sqrt{n}} = n\frac{1}{\sqrt{n}} = \sqrt{n} \xrightarrow[n\to\infty]{} \infty$). Hence, this test alone is not enough to establish the convergence of a series. Nevertheless, if the dependence of the general term a_n on n is known and is not very complicated, then the convergence or divergence of (33) can ordinarily be decided on the basis of other tests which we give below. But if it is not possible to establish such a simple relationship, one merely computes the terms in succession and if they do not go beyond the limits of the accepted accuracy and there are no grounds to expect subsequent terms to make substantial contributions, then all subsequent terms are dropped, the series is stated to be convergent and its sum is regarded as equal to the partial sum of the computed terms.

The first test for the convergence of series (33) is what is known as *d'Alembert's ratio test*, which is based on an analogy with the sum of the infinite geometric progression (35). For the "pure" progression (35), the ratio of every term to the preceding one is a constant (it is equal to the common ratio q of the progression). Now suppose that for (33) the ratio

$$\frac{a_{n+1}}{a_n}$$

of a term to the preceding term is no longer a constant but tends to some limit q as n increases. Then for large n this ratio is approximately equal to q and (33), like (35), converges if $|q| = \lim\limits_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| < 1$; (33) diverges if $|q| > 1$. And only when $|q| = 1$ is it impossible to establish, via the ratio test, whether (33) converges or not, so other tests have to be invoked.

Consider the series

$$\frac{a}{1^p} + \frac{a^2}{2^p} + \frac{a^3}{3^p} + \ldots + \frac{a^n}{n^p} + \ldots \tag{36}$$

Applying the ratio test, we get

$$\lim_{n\to\infty}\frac{a_{n+1}}{a_n}=\lim_{n\to\infty}\left[\frac{a^{n+1}}{(n+1)^p}:\frac{a^n}{n^p}\right]=\lim_{n\to\infty}\frac{a}{\left(1+\dfrac{1}{n}\right)^p}=a$$

Thus, for $|a|<1$ the series (36) converges like a geometric progression, for $|a|>1$, it diverges; when $a=\pm 1$, the ratio test cannot be applied to (36).

Let us now consider a stronger test called *Cauchy's integral test* for convergence that is applicable to a series with positive terms. Suppose we know the expression of the general term a_n of (33) as a function of n, that is, $a_n=f(n)$, and the function $f(n)$ is positive and decreasing with increasing n. Then, by virtue of formula (9) of Ch. 1, we can take

$$S_n=a_1+a_2+\ldots+a_n=f(1)+f(2)+\ldots+f(n)$$
$$\approx\int_1^n f(x)\,dx+\frac{1}{2}f(1)+\frac{1}{2}f(n) \tag{37}$$

for the approximate value of the partial sum (34). Thus, if the integral

$$\int_1^\infty f(x)\,dx \tag{38}$$

converges, then the right side of (37) remains finite as $n\to\infty$, that is, the series (33) converges. But if the integral (38) diverges to infinity, then the series (33) diverges.

We consider, for example, the series

$$\frac{1}{1^p}+\frac{1}{2^p}+\frac{1}{3^p}+\ldots+\frac{1}{n^p}+\ldots \tag{39}$$

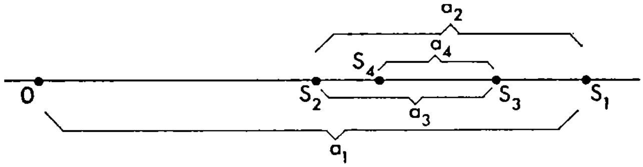
which is a consequence of (36) for $a=1$, when d'Alembert's ratio test fails. To apply Cauchy's integral test we have to consider the integral

$$\int_1^\infty\frac{dx}{x^p}$$

This integral was considered in Sec. 3.1 (formula (3)), where we knew that it converges for $p>1$ and diverges for $p\leqslant 1$. Hence, the series (39) too converges for $p>1$ and diverges for $p\leqslant 1$. In particular, when $p=1$ we get the so-called *harmonic series*

$$1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}+\ldots=\infty$$

Fig. 21



In the case of series with terms of arbitrary sign, frequent use is made of the *Leibniz test*, according to which the series

$$a_1 - a_2 + a_3 - a_4 + a_5 - a_6 + \dots \tag{40}$$

(all the a_i are considered to be positive so that two adjacent terms are of opposite sign) converges if

$$a_1 > a_2 > a_3 > \dots > a_n > \dots \to 0 \tag{41}$$

Indeed, if (Fig. 21) we depict on an auxiliary axis the partial sums (32), then from the condition (41) it follows that the transition from S_1 to S_2 and from S_2 to S_3, from S_3 to S_4 and so forth is in the form of damped oscillations, that is, the partial sums tend to a definite limit.

Thus, for example, the series

$$1 - \frac{1}{2^p} + \frac{1}{3^p} - \frac{1}{4^p} + \dots \tag{42}$$

converges for any $p > 0$.

If, as in the preceding test, the terms of the series (40) have the form $f(n)$, then an approximate value of the sum can be computed by the methods of Sec. 1.2. Procedures for refining that value will be discussed later on.

Convergent series with terms of arbitrary sign (not necessarily alternating as in (40)) are of two types.

(1) It may turn out that both the "positive part" of the original series (that is, a series made up solely of the positive terms of the original series) and the "negative part" converge. Then the original series is termed an *absolutely convergent* series, since a series made up of the absolute values of its terms also converges.

(2) It may turn out that both the positive and the negative parts of the original series diverge to infinity, but the series itself converges due to a compensation of these infinities. Such a series is called *conditionally convergent* since a series composed of the absolute values of its terms diverges.

For example, recalling the series (39), we conclude that the series (42) is absolute convergent for $p > 1$ and conditionally convergent for $0 < p \leqslant 1$.

We can operate with convergent series just as we do with finite sums, since, practically speaking, the sum of a convergent series is simply equal to a partial sum having a sufficiently large number of terms. A complication arises, though it is not so obvious, when we rearrange the terms of a convergent series. Such a rearrangement does not affect the sum of an absolutely convergent series, but a conditionally convergent series may, after such a rearrangement, alter its sum or even become divergent because such a rearrangement may alter or even upset the "compensation of infinities" mentioned above. Consider, for instance, the convergent series

$$1-\frac{1}{\sqrt{2}}+\frac{1}{\sqrt{3}}-\frac{1}{\sqrt{4}}+\frac{1}{\sqrt{5}}-\frac{1}{\sqrt{6}}+\dots \tag{43}$$

and regroup the terms so that one negative term follows two positive ones:

$$1+\frac{1}{\sqrt{3}}-\frac{1}{\sqrt{2}}+\frac{1}{\sqrt{5}}+\frac{1}{\sqrt{7}}-\frac{1}{\sqrt{4}}+\frac{1}{\sqrt{9}}+\frac{1}{\sqrt{11}}-\frac{1}{\sqrt{6}}+\dots \tag{44}$$

The partial sum S_{3n} of this series (the number of the term is $3n$) consists of a group of positive terms,

$$1+\frac{1}{\sqrt{3}}+\frac{1}{\sqrt{5}}+\frac{1}{\sqrt{7}}+\dots+\frac{1}{\sqrt{4n-3}}+\frac{1}{\sqrt{4n-1}}$$

and a group of negative terms,

$$-\frac{1}{\sqrt{2}}-\frac{1}{\sqrt{4}}-\dots-\frac{1}{\sqrt{2n}}$$

But the first sum exceeds

$$\frac{1}{\sqrt{2}}+\frac{1}{\sqrt{4}}+\frac{1}{\sqrt{6}}+\frac{1}{\sqrt{8}}+\dots+\frac{1}{\sqrt{4n-2}}+\frac{1}{\sqrt{4n}}$$

and so the general sum

$$S_{3n}>\frac{1}{\sqrt{2n+2}}+\frac{1}{\sqrt{2n+4}}+\dots+\frac{1}{\sqrt{4n-2}}+\frac{1}{\sqrt{4n}}$$
$$=\frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{n+1}}+\frac{1}{\sqrt{n+2}}+\dots+\frac{1}{\sqrt{2n-1}}+\frac{1}{\sqrt{2n}}\right)$$

By virtue of (11), Ch. 1, the right side is approximately equal to

$$\frac{1}{\sqrt{2}}\int\limits_{n+\frac{1}{2}}^{2n+\frac{1}{2}}\frac{1}{\sqrt{x}}\,dx=\sqrt{2}\left(\sqrt{2n+\frac{1}{2}}-\sqrt{n+\frac{1}{2}}\right)$$
$$=\sqrt{2n}\left(\sqrt{2+\frac{1}{2n}}-\sqrt{1+\frac{1}{2n}}\right)\xrightarrow[n\to\infty]{}\infty$$

To summarize, then, the two series (43) and (44) differ solely in the order of their terms but the former converges and the latter diverges to infinity.

If, as is often done in practice, we want to replace the sum of a series by the partial sum of a few terms, then the series must not merely converge but converge rapidly so that only a small number of terms almost exhausts the total sum to the accuracy we want. For slowly converging series (ordinarily, these are conditionally convergent series) one cannot drop the remainder of the series but must estimate it by using the methods of Sec. 1.2.

It may be essential, both for convergent and divergent series, to find the asymptotic law of variation of the partial sum as n increases. This can be done with the aid of the methods of Sec. 1.2, that is, using formulas (9), (11), (14), or (15) of Ch. 1, although this results in a definite error. For the sake of simplicity, we confine ourselves to series of positive terms. Ordinarily, the partial sum of a series is of the form

$$f(a)+f(a+h)+f(a+2h)+\dots+f(a+nh)$$

where the number of terms increases due to increasing n, while h remains constant. Thus, for example, the sum $S_3=1+\frac{1}{2}+\frac{1}{3}$ yields the sum

$$S_5=1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}$$

or

$$S_8=1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}$$

In such cases, as a rule, the absolute value of the error does not diminish with increasing number of terms in the sum.

Let us consider some examples

1. $S_n=1+\frac{1}{2^2}+\frac{1}{3^2}+\dots+\frac{1}{n^2}\cdot$ By (11) of Ch. 1 we get

$S_n\approx\int\limits_{\frac{1}{2}}^{n+\frac{1}{2}}\frac{dx}{x^2}=2-\frac{1}{n+\frac{1}{2}}\cdot$ By (9) of Ch. 1 we obtain

$$S_n\approx\int\limits_1^n\frac{dx}{x^2}+\frac{1}{2}+\frac{1}{2n^2}=1.5-\frac{1}{n}+\frac{1}{2n^2}=1.5-\frac{2n-1}{2n^2}$$

We know that if n increases without bound, the value of S_n approaches without bound the number $S=\frac{1}{6}\pi^2\approx1.665$. (We omit the proof here.) For very large n, formula (11) of Ch. 1 yields $S_n = 2$ (error 20%), formula (9), Ch. 1, $S_n = 1.5$ (error 10%).

Observe that in this case it is easy to refine the computation. The reason for the rather considerable error lies in the fact that the first terms of the sum vary rapidly and, hence, over the interval of integration corresponding to the distance between two successive terms of the sum the function $f(x)$ varies much too nonuniformly and so the trapezoidal formula, on which formulas (9) and (11) of Ch. 1 are based, gives a poor result.

It is therefore possible to find the sum of a few terms by direct addition and then to apply approximate formulas to the remaining sum. In our example, we start by finding the sum of the first three terms directly:

$$S_3 = 1 + \frac{1}{2^2} + \frac{1}{3^2} = 1 + 0.25 + 0.111 = 1.361$$

Let

$$S'_{n-3} = \frac{1}{4^2} + \frac{1}{5^2} + \dots + \frac{1}{n^2}$$

By formula (11), Ch. 1, we get $S'_{n-3} \approx 0.286 - \dfrac{1}{n + \frac{1}{2}}$, and by (9), Ch. 1, we obtain $S'_{n-3} \approx 0.281 - \frac{2n-1}{2n^2}$. Therefore

$$S_n \approx 1.647 - \frac{1}{n + \frac{1}{2}}$$

by (11) of Ch. 1 and

$$S_n \approx 1.642 - \frac{2n-1}{2n^2}$$

by (9) of Ch. 1. For n increasing without bound, formula (11), Ch. 1, gives $S \approx 1.647$, and formula (9), Ch. 1, yields $S \approx 1.642$. In each case the error is less than 2%.

2. Consider the sum of the decreasing geometric progression

$$S_n = 1 + \frac{1}{z} + \frac{1}{z^2} + \dots + \frac{1}{z^{n-1}}$$

where $z > 1$. The exact formula, it will be recalled, is

$$S_n = \frac{1 - \frac{1}{z^n}}{1 - \frac{1}{z}} = \frac{z - \frac{1}{z^{n-1}}}{z - 1}$$

whence, for unlimited growth of n, we get $S=\frac{z}{z-1}\cdot$ By formula (11) of Ch. 1 we have

$$S_n \approx \int\limits_{-\frac{1}{2}}^{n-\frac{1}{2}} \frac{dx}{z^x} = \int\limits_{-\frac{1}{2}}^{n-\frac{1}{2}} e^{-x\ln z}\,dx$$

$$= \frac{1}{\ln z}\left[e^{\frac{1}{2}\ln z} - e^{-\left(n-\frac{1}{2}\right)\ln z}\right] = \frac{\sqrt{z}}{\ln z}\left(1-\frac{1}{z^n}\right)$$

When n increases without bound, we have $S \approx \frac{\sqrt{z}}{\ln z}\cdot$ From the table below it is evident that for z close to unity both formulas yield almost the same results:

z	1.2	1.5	2.0	3.0	6.0	20.0
$\frac{z}{z-1}$	6.00	3.00	2.00	1.5	1.20	1.05
$\frac{\sqrt{z}}{\ln z}$	6.00	3.01	2.04	1.57	1.36	1.49

But if $z \gg 1$, then adjacent terms of the series differ substantially and so the approximate formula produces poor results.

In some cases the sum may increase without bound as the number of terms increases, despite the fact that the terms of the sum diminish without bound (cases of the divergence of the corresponding infinite series).

Let us consider some examples.

1. $S_n = 1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \ldots + \frac{1}{\sqrt{n}}\cdot$

By formula (11), Ch. 1, we get

$$S_n \approx \int\limits_{\frac{1}{2}}^{n+\frac{1}{2}} \frac{dx}{\sqrt{x}} = 2\sqrt{x}\,\Bigg|_{\frac{1}{2}}^{n+\frac{1}{2}} = 2\sqrt{n+\frac{1}{2}} - 2\sqrt{\frac{1}{2}}$$

By (9), Ch. 1, we obtain

$$S_n \approx \int\limits_{1}^{n} \frac{dx}{\sqrt{x}} + \frac{1}{2} + \frac{1}{2\sqrt{n}} = 2\sqrt{n} - 1.5 + \frac{1}{2\sqrt{n}}$$

For large n we find from (11), Ch. 1,

$$S_n \approx 2\sqrt{n} - 1.41$$

and from (9), Ch. 1, $S_n = 2\sqrt{n} - 1.50$. $\left(\text{Here we drop the term proportional to } \frac{1}{\sqrt{n}}\cdot\right)$

A more exact formula (obtained by straightforward addition of the first few terms) is

$$S_n \approx 2\sqrt{n} - 1.466$$

2. In many problems we encounter the sum

$$S_n = 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n}$$

For large n, using (11) of Ch. 1, we find $S_n \approx \ln n + \ln 2 = \ln n + 0.69$, using (9) of Ch. 1 we obtain $S_n \approx \ln n + 0.50$. The limit of the difference $S_n - \ln n$ for unlimited increase in n is denoted by C and is called *Euler's constant.*[1] Thus we can write the formula $S_n = \ln n + C + \alpha_n$, where $\alpha_n \to 0$ as $n \to \infty$. Therefore the asymptotically exact formula is of the form $S_n \approx \ln n + C$. We obtained very rough values for the Euler constant, but formulas (9) and (11) of Ch. 1 enable us to obtain a more exact value for C by straightforward summing of the first few terms. It turns out that $C = 0.5772\dots$

Now let us examine sums whose terms increase with increasing n. The sum here increases without bound with the growth of n (i.e. with increase in the number of terms). Two cases are possible for increasing n.

1. The error of the approximate formulas (9) and (11) of Ch. 1 decreases in absolute value with increasing n or if it increases, it does so more slowly than the sum so that the relative error diminishes. This case results if the terms of the sum increase more slowly than a geometric series, for instance, like powers.

2. With increasing n, the relative error (and all the more so the absolute error) does not decrease.

The second case is obtained when the terms of the sum increase in geometric progression, that is, when the sum is of the form $S_n = a + ay + ay^2 + ay^3 + \dots + ay^{n-1}$, where $|y| > 1$, and also if the terms of the sum grow faster than a progression, for example, $S_n = y + y^4 + y^9 + \dots + y^{n^2}$. In this case the last term constitutes the major portion of the entire sum. To illustrate, in the case of the sum $S_n = y + y^4 + y^9 + \dots + y^{n^2}$ we give a table of the values of $\frac{S_n}{y^{n^2}}$ when $y = 2$:

n	1	2	3	4	5
2^{n^2}	2	16	512	65 536	33 500 000
S_n	2	18	530	66 066	33 600 000
$\frac{S_n}{2^{n^2}}$	1	1.12	1.03	1.01	1.003

It is evident from the table that for large n the magnitude of the entire sum is practically determined by the last term alone.

The situation is similar for an ascending geometric series. Indeed, in the formula

$$S_n = 1 + z + \dots + z^{n-1} = \frac{z^n - 1}{z - 1} \ (|z| > 1)$$

we disregard unity in comparison with z^n and get

$$S_n \approx \frac{z^n}{z - 1}$$

And so

$$\frac{S_n}{z^{n-1}} \approx \frac{z^n}{z^{n-1}(z - 1)} = \frac{z}{z - 1} = \text{constant}$$

(for sufficiently large n). Thus, in this case the portion of the contribution of the last term approaches a constant, and for large $|z|$ this portion is close to unity. Clearly, then, there is no need of summation formulas in this case, for a high degree of accuracy can be obtained merely by taking the sum of the last few terms.

Summation formulas are useful when the ratio of the sum to the last term increases with increasing n. Then the formula makes it possible to reduce computations for large n and we always have Case 1, that is, the relative error in the formulas (9) and (11) of Ch. 1 definitely decreases. Here is an example.

From elementary algebra we have the familiar formula

$$S_n = 1^2 + 2^2 + 3^2 + \dots + n^2 = \frac{n(n + 1)(2n + 1)}{6} = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}$$

Apply the approximate formula (11), Ch. 1, to the sum S_n to get

$$S_n \approx S'_n = \int_{\frac{1}{2}}^{n+\frac{1}{2}} x^2\,dx = \frac{x^3}{3}\Bigg|_{\frac{1}{2}}^{n+\frac{1}{2}} = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{4}$$

The absolute error is $S'_n - S_n = \frac{n}{12}$ and hence increases with increasing n. The relative error

$$\frac{S'_n - S_n}{S_n} = \frac{n}{12\left(\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right)} \approx \frac{n}{12\,\frac{n^3}{3}} = \frac{1}{4n^2}$$

It falls off rapidly with increase n.

Consider the sum

$$S_n^{(p)} = 1^p + 2^p + 3^p + \ldots + n^p \quad (p > -1)$$

Applying (10), Ch. 1, we get

$$S_n^{(p)} \approx \int_1^n x^p\,dx = \frac{n^{p+1}}{p+1} - \frac{1}{p+1}$$

Thus, for large n the following simple but rough formula holds true:

$$S_n^{(p)} \approx \frac{n^{p+1}}{p+1}\ ^{*}$$

Exercises

1. Refine the value of Euler's constant C (see page 96) by finding the sum of the first five terms and the first ten terms directly.
2. Suppose we have a sum

$$S_n = u_1 + u_2 + \ldots + u_n$$

such that

$$0 < u_1 < u_2 < u_3 < \ldots < u_n$$

Form the sum

$$\sigma_n = 1 + \frac{u_{n-1}}{u_n} + \frac{u_{n-2}}{u_n} + \ldots + \frac{u_1}{u_n}$$

[*] When p is a positive integer, elementary algebra permits writing an exact formula for $S_n^{(p)}$ (incidentally, we made use of such a formula for $p = 2$). However, for large p the formulas become unwieldy and so the rough formula given here may prove useful for positive integers as well.

It is clear that the terms of this sum decrease. How does σ_n behave when n increases if S_n is an increasing sum of the first type? of the second type?

3.6 Integrals depending on a parameter

Consider an integral of the form

$$I = \int_a^b f(x, \lambda)\, dx \tag{45}$$

where, under the integral sign, we have the variable of integration x and also a parameter (arbitrary constant) λ, that is, a quantity that remains constant throughout the integration but can in general assume distinct values. Then the result of the integration will, generally, depend on λ, that is, $I = I(\lambda)$. Such integrals are often encountered in applications when the integrand includes certain masses, dimensions, and the like, which remain constant throughout the integration process. Here are some simple instances:

$$\int_0^1 (x^2 + \lambda x)\, dx = \frac{1}{3} + \frac{\lambda}{2}, \quad \int_0^1 \sin \alpha x\, dx = \frac{1 - \cos \alpha}{\alpha},$$

$$\int_0^1 (s + 1)\, x^s\, dx = 1 \quad (s > -1)$$

In the case of proper integrals we have the same properties as are encountered in the consideration of finite sums of functions. We know, for instance, that the derivative of a sum of functions is equal to the sum of the derivatives. Similarly, the derivative of integral (45) with respect to the parameter is equal to the integral of the derivative with respect to that parameter: $\frac{dI}{d\lambda} = \int_a^b \frac{\partial f(x, \lambda)}{\partial \lambda}\, dx$. Under the integral sign we have here the derivative of the function $f(x, \lambda)$, with respect to λ, taken for fixed x. A similar rule holds for integration with respect to a parameter:

$$\int_\alpha^\beta I(\lambda)\, d\lambda = \int_a^b \left(\int_\alpha^\beta f(x, \lambda)\, d\lambda \right) dx$$

To verify these simple rules we would have to write them out for the integral sums and then pass, in the limit, from sums to integrals, but we will not go into that here.

For improper integrals dependent on parameters there can be complications due first of all to the possibility of their divergence. The simplest case is that of "regularly convergent" improper integrals. Thus, an integral of the type

$$I(\lambda) = \int_a^\infty f(x, \lambda)\, dx \tag{46}$$

where the function f itself is finite, is called a *regularly convergent integral* if it is dominated by a convergent integral not dependent on the parameter, that is, if $|f(x, \lambda)| \leqslant F(x)$ where $\int_a^\infty F(x)\, dx < \infty$.

For example, the integral

$$\int_1^\infty \frac{\sin \lambda x}{x^2}\, dx$$

is regularly convergent since

$$\left|\frac{\sin \lambda x}{x^2}\right| \leqslant \frac{1}{x^2}, \quad \int_1^\infty \frac{1}{x^2}\, dx = 1 < \infty$$

The properties of regularly convergent integrals are the same as those of proper integrals.*

In the study of irregularly convergent (and also divergent integrals of type (40)) integrals, one often does as follows: one cuts off the singularity, that is, one passes to a proper integral:

$$I_N(\lambda) = \int_a^N f(x, \lambda)\, dx$$

and then considers the asymptotic behaviour of this integral as $N \to \infty$. In this way we can justify operations on improper and even divergent integrals.

It is very important that the result of performing various operations — subtraction, differentiation with respect to a parameter and the like — on divergent integrals may prove to be a finite quan-

* Here the requirement of regular convergence may be replaced by the somewhat less restrictive requirement of *uniform convergence*, which signifies that

$$\max_\lambda \left| \int_N^\infty f(x, \lambda)\, dx \right| \xrightarrow[N \to \infty]{} 0.$$

tity, in particular, it may be expressed in terms of convergent integrals (the converse can also occur). To illustrate, let us consider the divergent integral

$$I(\lambda) = \int_0^\infty \frac{1}{x+\lambda}\,dx = \infty \quad (\lambda > 0)$$

After differentiating with respect to the parameter we arrive at the convergent integral

$$\frac{dI}{d\lambda} \int_0^\infty \frac{\partial}{\partial\lambda}\left(\frac{1}{x+\lambda}\right) dx = \int_0^\infty -\frac{1}{(x+\lambda)^2}\,dx = \frac{1}{x+\lambda}\bigg|_{x=0}^{\infty} = -\frac{1}{\lambda}$$

To get to the meaning of this equation, we cut off the singularity to obtain

$$I_N(\lambda) = \int_0^N \frac{1}{x+\lambda}\,dx = \ln(x+\lambda)\bigg|_{x=0}^{N} = \ln(N+\lambda) - \ln\lambda$$

Differentiating we get

$$\frac{dI_N}{d\lambda} = \frac{1}{N+\lambda} - \frac{1}{\lambda}$$

Now, if $N \to \infty$, then $I_N(\lambda) \to \infty$ and $\frac{dI_N}{d\lambda} \to -\frac{1}{\lambda}$, that is, we obtain the foregoing result. Verify that the following equation has a similar meaning:

$$I(\lambda_1) - I(\lambda_2) = \int_0^\infty \left[\frac{1}{x+\lambda_1} - \frac{1}{x+\lambda_2}\right] dx = \ln\lambda_2 - \ln\lambda_1\,{}^*$$

We take another example to illustrate the complication that may arise in the case of irregularly convergent integrals. Consider the integral

$$I(\lambda) = \int_0^\infty \frac{\sin\lambda x}{x}\,dx$$

[*] In the 1950's physicists developing quantum electrodynamics began to make extensive use of an imperfect theory containing divergent integrals. It has been common practice to compute quantities like the derivatives and differences of these integrals. Such things turn out to be finite and are in excellent agreement with experiment, which is the ultimate criterion of the truth.

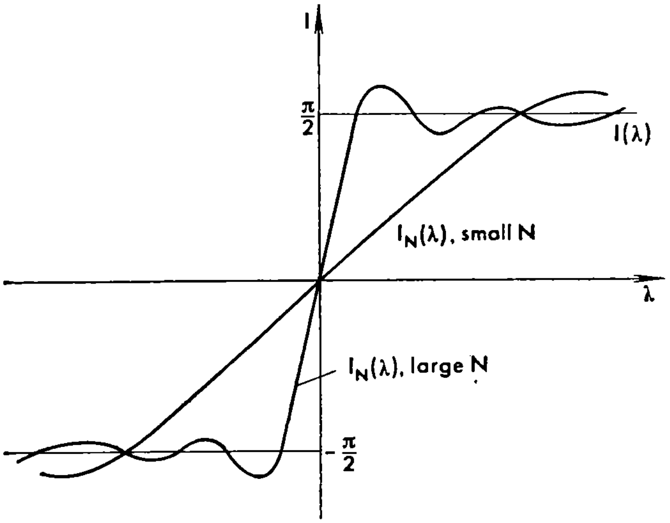


Fig. 22

It can be verified (see Exercise 2) that for $\lambda = 1$ the integral converges and is equal to $\frac{1}{2}\pi$. From this, using the substitution $\lambda x = s$ for $\lambda > 0$, we immediately get

$$I(\lambda) = \int_{0}^{\infty} \frac{\sin s}{\frac{s}{\lambda}} \frac{ds}{\lambda} \int_{0}^{\infty} \frac{\sin s}{s} ds = \frac{\pi}{2}$$

At the same time, when $\lambda = 0$, we get $I = 0$ and for $\lambda < 0$, taking -1 outside the integral sign, we obtain $I = -\frac{1}{2}\pi$. Thus, in this case $I(\lambda)$ has a jump discontinuity at $\lambda = 0$. This might appear to be strange since for a small variation of λ the integrand varies to an arbitrarily small degree. However, a small change in the integrand over an infinite interval can lead to a rather substantial change in the integral! Fig. 22 shows the discontinuous graph of $I(\lambda)$ and the graphs of

$$I_N(\lambda) = \int_{0}^{N} \frac{\sin \lambda x}{x} dx$$

for different N. Although these latter graphs do not have discontinuities, for large N the transition from $-\frac{1}{2}\pi$ to $\frac{1}{2}\pi$ takes place over a small interval of λ, and the larger the N, the smaller the interval.

In the limit, when $N = \infty$, this transition is accomplished over an infinitely small interval of λ, which is to say we have a discontinuity.

We shall return (in different notations) to the functions

$$I(\lambda) = \int_0^\infty \frac{\sin \lambda x}{x}\, dx \text{ and } \frac{\partial I}{\partial \lambda} = \int_0^\infty \cos \lambda x\, dx$$

in Sec. 6.3 in connection with the theory of discontinuous functions and in Sec. 14.2 in connection with the Fourier transformation.

When considering series whose terms depend on a parameter, the very same questions arise as in the study of improper integrals dependent on a parameter. The properties are completely analogous and so we will not dwell on them here.

Exercises

1. Starting with the integral $\int_0^\infty e^{-\lambda x}\, dx$ $(\lambda > 0)$, use differentiation with respect to the parameter for $\lambda = 1$ to obtain the value of the integral $\int_0^\infty x^n e^{-x}\, dx$ (cf. Sec. 3.3), and use integration with respect to the parameter to evaluate the integral $\int_0^\infty \frac{e^{-x} - e^{-\lambda x}}{x}\, dx$ (note that in the latter example the indefinite integral is not expressible in terms of elementary functions).

2. Starting with the integral $\int_0^\infty e^{-\lambda x} \sin x\, dx$ $(\lambda > 0)$, integrate with respect to the parameter from α to β and, setting $\beta \to \infty$, obtain the formula

$$\int_0^\infty e^{-\alpha x} \frac{\sin x}{x}\, dx = \frac{\pi}{2} - \arctan \alpha \quad (\alpha > 0) \tag{47}$$

(The value of this integral for $\alpha = 0$ was mentioned in the text.)

ANSWERS AND SOLUTIONS

Sec. 3.1

Straightforward use of the trapezoidal rule or Simpson's rule is impossible since the integrand blows up at $x = 0$, and so we split the interval of integration into two parts; from

$x = 0$ to $x = \frac{1}{6}\pi$ and from $x = \frac{1}{6}\pi$ to $x = \frac{1}{2}\pi$. For $0 \leqslant x \leqslant \frac{1}{6}\pi$ the relation $\sin x \approx x$ holds true and so

$$I_1 = \int_0^{\frac{\pi}{6}} \frac{dx}{\sqrt[3]{\sin x}} \approx \int_0^{\frac{\pi}{6}} \frac{dx}{\sqrt{x}} = \frac{3}{2} x^{\frac{2}{3}} \Big|_0^{\frac{\pi}{6}} = 1.5 \times 0.65 = 0.98$$

Evaluate the integral $I_2 = \int_{\frac{\pi}{6}}^{\frac{\pi}{2}} \frac{dx}{\sqrt[3]{\sin x}}$ by Simpson's rule. Split the interval into two parts to obtain $I_2 \approx 1.13$, whence

$$\int_0^{\frac{\pi}{2}} \frac{dx}{\sqrt[3]{\sin x}} \approx 0.98 + 1.13 = 2.11$$

Sec. 3.2

1. The value of the first of these integrals has already been obtained (see Problem 2 in the Exercises of Sec. 1.1). Let us find $\int_3^{\infty} \frac{x\,dx}{e^x + 1}$. The integrand is $f(x) = \frac{x}{e^x + 1}$ and so $f'(x) = \frac{e^x(1 - x) + 1}{(e^x + 1)^2}$. Since $f'(3) \neq 0$, we apply formula (19) to get

$$\int_3^{\infty} \frac{x\,dx}{e^x + 1} \approx 0.23, \quad \int_0^{\infty} \frac{x\,dx}{e^x + 1} \approx 0.63 + 0.23 = 0.86$$

Observe, for the sake of comparison, that the exact value of the desired integral (to two decimal places) is 0.82.

2. 0.67, 0.59. (The exact value to three places is 0.612.)

3. Here, $y = e^{-x^2}$, $y' = -2xe^{-x^2}$, $a = N$; by (19) we find that

$$\int_N^{\infty} e^{-x^2}\,dx \approx (e^{-N^2})^2 : 2Ne^{-N^2} = \frac{1}{2N} e^{-N^2}$$

as was indicated on page 69.

Sec. 3.3

The fraction D is, after the indicated transformation, equal to

$$\frac{[2\cdot 4\cdot 6 \ldots (2n-2)]^2}{\sqrt{\pi n}\,1\cdot 2\cdot 3\cdot 4 \ldots (2n-3)\,(2n-2)}$$

$$= \frac{[2^{n-1}\cdot 1\cdot 2\cdot 3 \ldots (n-1)]^2}{\sqrt{\pi n}\,(2n-2)!} = \frac{2^{2n-2}[(n-1)!]^2}{\sqrt{\pi n}(2n-2)!}$$

Using Stirling's formula we obtain, for large n,

$$D \approx \frac{2^{2n-2}\left[\sqrt{2\pi(n-1)}\left(\frac{n-1}{e}\right)^{n-1}\right]^2}{\sqrt{\pi n}\sqrt{2\pi(2n-2)}\left(\frac{2n-2}{e}\right)^{2n-2}}$$

$$= \frac{2^{2n-2}2\pi\,(n-1)\,(n-1)^{2n-2}\,e^{-(2n-2)}}{2\pi\sqrt{n(n-1)}\,2^{2n-2}\,(n-1)^{2n-2}\,e^{-(2n-2)}} = \sqrt{\frac{n-1}{n}} \approx 1$$

Sec. 3.4

1. $I_{(28)} = \dfrac{2\sin\omega a}{\omega(1+a^2)}$; $I_{(29)} = I_{(28)} + \dfrac{4a\cos\omega a}{\omega^2(1+a^2)^2}$; $I_{(30)} = I_{(28)} - \dfrac{4a\cos\omega a}{\omega^2(1+a^2)^2}$.

2. $I \approx I_{(32)} + \dfrac{1}{\omega^2}\left[\dfrac{1}{\varphi'(x)}\left(\dfrac{f(x)}{\varphi'(x)}\right)'\sin(\omega\varphi(x)+\alpha)\right]\Bigg|_{x=a}^{b}$

3. For $\omega = 1$ the exact value is $I = \mathrm{Si}\,2\omega - \mathrm{Si}\,\omega = 0.66$; approximate values are $I_{(28)} = 0.75$, $I_{(29)} = 0.13$, $I_{(30)} = 1.36$. It is evident that the accuracy is quite unsatisfactory. For $\omega = 10$ the exact value is $I = -0.110$, the approximate values are $I_{(28)} = -0.104$; $I_{(29)} = -0.115$, $I_{(30)} = -0.112$ with accuracy of the order of a few percent.

Sec. 3.5

1. $S_n = S_5 + \bar{S}_{n-5}$, where $\bar{S}_{n-5} = \frac{1}{6} + \frac{1}{7} + \ldots + \frac{1}{n}$. Using formula (9), Ch. 1, we obtain $\bar{S}_{n-5} \approx \int_6^n \frac{dx}{x} - \frac{1}{12} + \frac{1}{2n} \approx \ln n - \ln 6 + 0.083$.

Since

$$S_6 = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} = 2.283$$

it follows that

$$S_n \approx \ln n - \ln 6 + 0.083 + 2.283 = \ln n + 0.575$$

whence $C = 0.575$. Summing the first 10 terms, we obtain $C = 0.576$.

2. In the first case, σ_n increases without bound as n increases without bound, in the second case, σ_n approaches 1 without bound.

Sec. 3.6

1. We have

$$\int_0^\infty e^{-\lambda x}\,dx = \frac{e^{-\lambda x}}{-\lambda}\Big|_{x=0}^{\infty} = \frac{1}{\lambda} \quad (\lambda > 0) \tag{48}$$

This integral converges regularly on any interval $\alpha \leqslant \lambda \leqslant \beta$ where $0 < \alpha < \beta < \infty$ since on such an interval $e^{-\lambda x} \leqslant e^{-\alpha x}$ and $\int_0^\infty e^{-\alpha x}\,dx < \infty$. Differentiating with respect to the parameter λ, we successively obtain

$$\int_0^\infty (-x)\,e^{-\lambda x}\,dx = (-1)\,\lambda^{-2},\quad \int_0^\infty (-x)^2 e^{-\lambda x}\,dx = (-1)\,(-2)\,\lambda^{-3},\ \ldots,$$

$$\int_0^\infty (-x)^n e^{-\lambda x}\,dx = (-1)\,(-2)\,\ldots\,(-n)\,\lambda^{-(n+1)}$$

Putting $\lambda = 1$ and cancelling out $(-1)^n$, we get

$$\int_0^\infty x^n e^{-x}\,dx = n!$$

Integrating formula (48) from 1 to λ with respect to the parameter, we obtain

$$\int_0^\infty \left(\int_1^\lambda e^{-\lambda x}\,d\lambda\right) dx = \int_0^\infty \frac{e^{-x} - e^{-\lambda x}}{x}\,dx = \ln \lambda$$

2. The integral

$$\int_0^\infty e^{-\lambda x} \sin x\,dx = \frac{1}{\lambda^2 + 1} \quad (\lambda > 0)$$

is evaluated in elementary fashion with the aid of the indefinite integral. As in Exercise 1, the integral converges regularly on any interval $\alpha \leqslant \lambda \leqslant \beta$ $(0 < \alpha < \beta < \infty)$.
Integrating with respect to λ from α to β, we get

$$\int_0^\infty (e^{-\alpha x} - e^{-\beta x})\,\frac{\sin x}{x}\,dx = \arctan \beta - \arctan \alpha$$

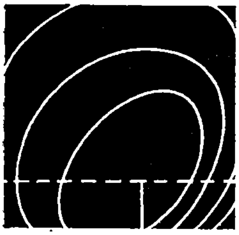
Passing to the limit as $\beta \to \infty$, we get formula (47). For $\alpha = 0$ this formula yields

$$\int_0^\infty \frac{\sin x}{x}\, dx = \frac{\pi}{2}$$

It is to be noted that formula (47) was derived for $\alpha > 0$ and its validity for $\alpha = 0$ requires special substantiation, which we will not dwell on here.

Chapter 4

FUNCTIONS OF SEVERAL VARIABLES



Up to now we have considered only functions of one independent variable. Functions of several variables have appeared from time to time (see HM, Sec. 3.12). The theory of functions dependent on a number of independent variables contains many new features, and, what is more, nearly all the fundamentally new features are revealed in functions of two variables:

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

For the sake of simplicity we will as a rule consider precisely this case.

The concept itself of a function of two variables is simple enough: a quantity z is specified via a formula or table so that to every pair of values x and y there corresponds a definite value of z.

4.1 Partial derivatives

In the case of a function of one variable, $g = g(x)$, a small change dx in the variable x leads to a small change in the function g, and

$$g(x + dx) - g(x) = g'(x)\, dx$$

In this formula, the right-hand side of which is denoted by dg, we neglect terms of order $(dx)^2$ since dx is a very small quantity. In the case of a function of two variables the variation of the function occurs as a result of changes in both variables x and y and is equal to $f(x + dx, y + dy) - f(x, y)$.

We now show * that this change consists of two parts, one part being proportional to dx and the other to dy; that is,

$$f(x + dx, y + dy) - f(x, y) = a\, dx + b\, dy \tag{1}$$

Here, we disregard terms of order $(dx)^2$, $(dy)^2$, $dx \cdot dy$ since dx and dy are extremely small quantities.

* With certain alterations, this is a repetition of the reasoning given in HM, Sec. 3.12.

We write the left part of (1) thus:

$$f(x + dx,\ y + dy) - f(x, y) = f(x + dx,\ y + dy) - f(x + dx,\ y) + f(x + dx, y) - f(x, y)$$

Let us consider the difference of the last two terms, $f(x + dx,\ y) - - f(x, y)$. For every concrete fixed y, this difference, to within terms of the order of $(dx)^2$, is the differential of the function depending solely on x:

$$f(x + dx,\ y) - f(x, y) = \frac{\partial f(x, y)}{\partial x}\bigg|_{y} dx$$

where $\frac{\partial f(x, y)}{\partial x}\Big|_{y}$ denotes the derivative of the function $f(x,. y)$ calculated on the supposition that y is constant. It is called a *partial derivative* of the function $f(x, y)$ with respect to x with y held constant. It can also be denoted by $f'_x(x, y)$. Similarly, to within terms of the order of $(dy)^2$,

$$f(x + dx,\ y + dy) - f(x + dx,\ y) = \frac{\partial f(x + dx,\ y)}{\partial y}\bigg|_{x+dx} dy$$

where $\frac{\partial f(x + dx,\ y)}{\partial y}\Big|_{x+dx}$ is the partial derivative with respect to y with the first argument, equal to $x + dx$, held constant. It is clear that

$$\frac{\partial f(x + dx,\ y)}{\partial y}\bigg|_{x+dx} - \frac{\partial f(x, y)}{\partial y}\bigg|_{x}$$

is the smaller, the smaller dx is; more precisely,

$$\frac{\partial f(x + dx,\ y)}{\partial y}\bigg|_{x+dx} - \frac{\partial f(x, y)}{\partial y}\bigg|_{x} = \alpha\, dx$$

where α is a bounded quantity. With the indicated simplifications, the left side of (1) is

$$\frac{\partial f(x, y)}{\partial x}\bigg|_{y} dx + \left[\frac{\partial f(x, y)}{\partial y}\bigg|_{x} + \alpha\, dx\right] dy$$

$$= \frac{\partial f(x, y)}{\partial x}\bigg|_{y} dx + \frac{\partial f(x, y)}{\partial y}\bigg|_{x} dy + \alpha\, dx\, dy$$

Finally, noting that the term $\alpha\, dx\, dy$ may be disregarded, we find that, to within quantities of the order of $(dx)^2$, $(dy)^2$ and $dx\, dy$, the left-hand side of (1) is equal to a sum, which is denoted by df and is called the *total differential* of the function f:

$$df = \frac{\partial f(x, y)}{\partial x}\bigg|_{y} dx + \frac{\partial f(x, y)}{\partial y}\bigg|_{x} dy \qquad (2)$$

Comparing this with (1), we get

$$a = \left.\frac{\partial f(x, y)}{\partial x}\right|_y = f'_x(x, y), \quad b = \left.\frac{\partial f(x, y)}{\partial y}\right|_x = f'_y(x, y)$$

If from the context it is clear what quantity is assumed to be constant in a calculation of a partial derivative, it is not indicated and instead of $\left.\frac{\partial f}{\partial x}\right|_y$ we write $\frac{\partial f}{\partial x}$. However, since the variables in the problem as a whole are x and y, the derivative is written with a round d, ∂, in order to distinguish it from the ordinary derivative. * From the foregoing it follows that the quantity $\frac{\partial f}{\partial x}$ is to be found as if y remains unchanged in the expression $f(x, y)$. In exactly the same way, we find $\frac{\partial f}{\partial y}$ on the assumption that only y varies and x is held constant. For example, if

$$f(x, y) = x^2y^3 + xe^y \tag{3}$$

then

$$\frac{\partial f}{\partial x} = 2xy^3 + e^y, \quad \frac{\partial f}{\partial y} = 3x^2y^2 + xe^y$$

To get a better picture of formula (2), consider the "plane of independent variables", that is, the xy-plane. In this plane, every point M has definite coordinates x and y and therefore this point is associated with a definite value of the function $f(x, y)$; we can say that the function assumes a definite value at every point of the plane. If we give a small increment to x or to y or to both variables, then in the xy-plane we get points of a small rectangle $MNPQ$, which is shown in Fig. 23 enlarged. The coordinates of the vertices are indicated inside the rectangle, and the function values (with second-order infinitesimals dropped) given outside the rectangle; f is to be understood as the value of $f(x, y)$.

A formula similar to (2) holds true for any number of independent variables; for example $df(x, y, z, u) = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial z}dz + \frac{\partial f}{\partial u}du$. Here $\frac{\partial f}{\partial x}$ is calculated on the assumption that y, z, u are held constant, $\frac{\partial f}{\partial y}$ is calculated with x, z, u held constant, and so on.

* In the formula $\frac{d}{dx}(e^{kx}) = ke^{kx}$ we also actually have to do with a partial derivative calculated with k held constant. However, there is no need to use the partial derivative sign since k remained constant throughout the discussion of the problem.

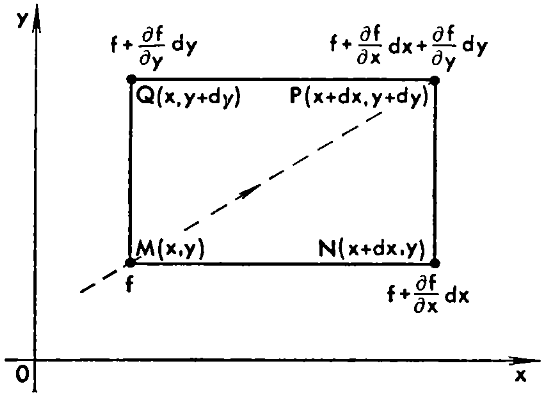

y
$f+\frac{\partial f}{\partial y}dy$
$f+\frac{\partial f}{\partial x}dx+\frac{\partial f}{\partial y}dy$
Q(x,y+dy)
P(x+dx,y+dy)
M(x,y)
N(x+dx,y)
f
$f+\frac{\partial f}{\partial x}dx$
0
x

Fig. 23

We give an example to show that the value of a partial derivative is essentially dependent on how the other fixed variables are chosen. From physics we know that the energy of a capacitor is $W = \frac{C\varphi^2}{2} = \frac{q^2}{2C}$, where $q = C\varphi$; here, C is the capacity of the capacitor, φ is the potential difference on the faces, and q is the quantity of electricity, or the charge. Considering the relationship $W = W(C, \varphi)$, we get $\left.\frac{\partial W}{\partial C}\right|_{\varphi} = \frac{\varphi^2}{2} = \frac{W}{C}$. Considering $W = W(C, q)$, we get $\left.\frac{\partial W}{\partial C}\right|_{q} = -\frac{q^2}{2C^2} = -\frac{W}{C}$ and so $\left.\frac{\partial W}{\partial C}\right|_{\varphi} = -\left.\frac{\partial W}{\partial C}\right|_{q}$.

Up to now we have regarded the variables x and y as varying independently of one another. Now suppose they depend on a certain variable t, that is,

$$x = x(t), \quad y = y(t) \tag{4}$$

These equations "parametrically" specify a certain line in the xy-plane with t the parameter (see, for example HM, Sec. 1.8). The parameter t may have different physical meanings, but it is most convenient to regard it as the time, that is, to assume that we are considering the path of a particle moving in the xy-plane. Then $z = f(x, y) = f(x, (t), y(t))$ so that in reality z is a function of the single variable t. However it is specified in a complicated fashion. Let us find its derivative $\frac{dz}{dt}$ (the *total derivative* along the curve (4)).[*] Since from (2) and (4) we get

$$dz = \frac{\partial z}{\partial x}dx + \frac{\partial z}{\partial y}dy = \frac{\partial z}{\partial x}\frac{dx}{dt}dt + \frac{\partial z}{\partial y}\frac{dy}{dt}dt = \left(\frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}\right)dt$$

[*] The reader is advised to review the derivation of the formula for the derivative of a composite function of one variable (see HM, Sec. 3.3).

it follows that

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}$$

In particular, if $z = f(x, y)$, where $y = y(x)$, then $\frac{dz}{dx} = \frac{\partial z}{\partial x} + $ $+ \frac{\partial z}{\partial y}\frac{dy}{dx}$. We see that the value $\frac{dz}{dx}$ of the total derivative along the curve for given x, y depends not only on the type of function f but also on $\frac{dy}{dx}$, that is to say, on the slope of the curve to the x-axis at the given point.

Both these formulas are easy to understand with the aid of Fig. 23. If during time dt the point moved from M to P along the dashed line, the rate of change of the function is

$$\frac{df}{dt} = \frac{\frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}$$

To obtain the second formula, it is necessary to put df/dx instead of df/dt.

To illustrate, let us examine a case where a certain quantity u, say the pressure or temperature of a gas flow, is determined at every instant of time t at every point (x, y, z) of space. Here u is a function of four variables, x, y, z, t: $u = u(x, y, z, t)$. Furthermore, suppose we have a law of motion $x = x(t)$, $y = y(t)$, $z = z(t)$ of a particle M. If we examine the value of u in M in the course of the motion, then this value will be a composite function of the time:

$$u = u(x(t), y(t), z(t), t)$$

The rate of change of this value of u, that is the rate of change of u along the "path", is equal, by virtue of the formula for the derivative of a composite function, to

$$\frac{du}{dt} = \frac{\partial u}{\partial x}\frac{dx}{dt} + \frac{\partial u}{\partial y}\frac{dy}{dt} + \frac{\partial u}{\partial z}\frac{dz}{dt} + \frac{\partial u}{\partial t} \qquad (5)$$

If u does not depend explicitly on t (we then say that "the field u is stationary"), then $\frac{\partial u}{\partial t} = 0$, and we have only three terms on the right. Thus they yield the rate of change of u obtained solely from the movement of point M along the path from one value of u to another (for example, if u is the temperature, then it will be due to the motion from a cooler portion of space to a hotter portion, and the like). This is termed the *convective* rate. The last summand gives the rate of change of the field at a fixed point resulting from the nonstationarity of the field; this is the *local* rate. In the general case, both factors operate, and the rate of change of the field along the path is compounded of the convective and the local rates of change of the field.

Let us return to the case of the function $z = f(x, y)$ of two independent variables. It is clear that the partial derivatives $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$ of this function depend on x and y themselves. We can therefore find their partial derivatives. These are called *partial derivatives of the second order*, or *second partial derivatives*; their partial derivatives are *partial derivatives of the third order*, or *third partial derivatives*, and so on. We can form the following second partial derivatives:

$$\frac{\partial}{\partial x}\left(\frac{\partial z}{\partial x}\right), \quad \frac{\partial}{\partial y}\left(\frac{\partial z}{\partial x}\right), \quad \frac{\partial}{\partial x}\left(\frac{\partial z}{\partial y}\right), \quad \frac{\partial}{\partial y}\left(\frac{\partial z}{\partial y}\right)$$

or, in different notation,

$$z''_{xx} = \frac{\partial^2 z}{\partial x^2}, \quad z''_{xy} = \frac{\partial^2 z}{\partial x \partial y}, \quad z''_{yx} = \frac{\partial^2 z}{\partial y \partial x}, \quad z''_{yy} = \frac{\partial^2 z}{\partial y^2}$$

The derivatives $\frac{\partial^2 z}{\partial x \partial y}$ and $\frac{\partial^2 z}{\partial y \partial x}$ are called *mixed derivatives.*

For example, in (3)

$$\frac{\partial^2 z}{\partial x^2} = \frac{\partial}{\partial x}(2xy^3 + e^y) = 2y^3, \quad \frac{\partial^2 z}{\partial x \partial y} = \frac{\partial}{\partial y}(2xy^3 + e^y) = 6xy^2e^y,$$

$$\frac{\partial^2 z}{\partial y \partial x} = \frac{\partial}{\partial x}(3x^2y^2 + xe^y) = 6xy^2 + e^y,$$

$$\frac{\partial^2 z}{\partial y^2} = \frac{\partial}{\partial y}(3x^2y^2 + xe^y) = 6x^2y + xe^y$$

We see that in this example, $\frac{\partial^2 z}{\partial x \partial y} = \frac{\partial^2 z}{\partial y \partial x}$, that is, the mixed derivatives do not depend on the order in which the differentiation is performed.

We will show that mixed derivatives are equal in the general case as well. Note that by the definition of a derivative (see Sec. 2.2)

$$\left.\frac{\partial^2 z}{\partial x \partial y}\right|_{\substack{x=x_0\\ y=y_0}} = \left.\frac{\partial}{\partial y}\left(\frac{\partial z}{\partial x}\right)\right|_{\substack{x=x_0\\ y=y_0}} = \frac{\left.\frac{\partial z}{\partial x}\right|_{\substack{x=x_0\\ y=y_0+k}} - \left.\frac{\partial z}{\partial x}\right|_{\substack{x=x_0\\ y=y_0-k}}}{2k}$$

(to any degree of accuracy) provided that k is sufficiently small.*

* To be precise, the derivative is the limit to which the right member tends as k tends to zero. Similar formulations apply to the formulas that follow.

In the same way,

$$\left.\frac{\partial z}{\partial x}\right|_{\substack{x=x_0\\ y=y_0+k}} = \frac{f(x_0+h,\ y_0+k)-f(x_0-h,\ y_0+k)}{2h},$$

$$\left.\frac{\partial z}{\partial x}\right|_{\substack{x=x_0\\ y=y_0-k}} = \frac{f(x_0+h,\ y_0-k)-f(x_0-h,\ y_0-k)}{2h}$$

Substituting these expressions into the formula for the mixed derivative, we get

$$\left.\frac{\partial^2 z}{\partial x\partial y}\right|_{\substack{x=x_0\\ y=y_0}} = \frac{f(x_0+h,\ y_0+k)-f(x_0-h,\ y_0+k)}{4hk}$$
$$-\frac{f(x_0+h,\ y_0-k)-f(x_0-h,\ y_0-k)}{4hk} \tag{6}$$

In the same manner we now obtain

$$\left.\frac{\partial^2 z}{\partial y\partial x}\right|_{\substack{x=x_0\\ y=y_0}} = \left.\frac{\partial}{\partial x}\left(\frac{\partial z}{\partial y}\right)\right|_{\substack{x=x_0\\ y=y_0}} = \frac{\left.\frac{\partial z}{\partial y}\right|_{\substack{x=x_0+h\\ y=y_0}} - \left.\frac{\partial z}{\partial y}\right|_{\substack{x=x_0-h\\ y=y_0}}}{2h}$$

where

$$\left.\frac{\partial z}{\partial y}\right|_{\substack{x=x_0+h\\ y=y_0}} = \frac{f(x_0,+h,\ y_0+k)-f(x_0+h,\ y_0-k)}{2k},$$

$$\left.\frac{\partial z}{\partial y}\right|_{\substack{x=x_0-h\\ y=y_0}} = \frac{f(x_0-h,\ y_0+k)-f(x_0-h,\ y_0-k)}{2k}$$

Finally we get

$$\left.\frac{\partial^2 z}{\partial y\partial x}\right|_{\substack{x=x_0\\ y=y_0}} = \frac{f(x_0+h,\ y_0+k)-f(x_0+h,\ y_0-k)}{4hk}$$
$$-\frac{f(x_0-h,\ y_0+k)-f(x_0-h,\ y_0-k)}{4hk} \tag{7}$$

Comparing (6) and (7), we see that

$$\left.\frac{\partial^2 z}{\partial y\partial x}\right|_{\substack{x=x_0\\ y=y_0}} = \left.\frac{\partial^2 z}{\partial x\partial y}\right|_{\substack{x=x_0\\ y=y_0}}$$

and since the point (x_0, y_0) is quite arbitrary, these derivatives are equal for all values of x and y.

Similarly, for derivatives of any order of functions of any number of variables, the essential thing is the number of differentiations to be performed with respect to a given variable and not the order in which they are performed.

To get a better idea of the meaning of a mixed derivative, locate in the plane of independent variables those four points associated with the values of the function that enter into the right members of (6) and (7). Compare (6) and (7) with similar expressions for $\frac{\partial^2 z}{\partial x^2}$ and $\frac{\partial^2 z}{\partial y^2}$.

Exercises

1. $z = x^2 + y^2$. Find $\frac{\partial z}{\partial x}$ for $x = 1$, $y = 1$, and $\frac{dz}{\partial y}$ for $x = 2$, $y = 0.5$. Find the partial derivatives of the following functions:
2. $z = e^{-(x^2+y^2)}$. **3.** $z = xe^y + ye^x$. **4.** $z = x \sin y$. **5.** $z = \sin (xy)$.
6. $z = \sqrt{x^2 + y^2}$.
Find the total derivatives of the following functions:
7. $z = x^2 - y^2$, $x = t + \frac{1}{t,}$, $y = t + \sqrt{t}$.
8. $z = e^{x-y}$, $x = \sin t$, $y = t^2$.
9. $z = x^3 + 3xy^2$, $x = t^2$, $y = e^{-t}$.
10. Find $\frac{dz}{dx}$ if $z = \ln(x + e^y)$, $y = x^2$.
11. Verify the equation

$$\frac{\partial^2 z}{\partial x \partial y} = \frac{\partial^2 z}{\partial y \partial x}$$

for all functions of Exercises 2 to 6.

4.2 Geometrical meaning of a function of two variables

A function of two variables, $z = f(x, y)$, is conveniently visualized geometrically as an equation of a surface, $z = f(x, y)$, where z is the altitude and x and y are coordinates of a point in the horizontal plane (Fig. 24). Since it is difficult to depict a surface on a flat drawing, we will represent a function of two variables graphically by constructing a set of curves representing sections of the surface $z = f(x, y)$ made by planes parallel to the xz-plane. These planes are perpendicular to the y-axis and the y-coordinate is constant in each plane. The intersection of a given plane with the plane $z = f(x, y)$ yields a curve, $z = f(x, y = \text{constant})$. To make this pictorial, plot several such curves in one figure in coordinates x and z. We thus obtain a family of curves. Fig. 25 shows a number of such curves of a family for the hemisphere $z = \sqrt{16 - x^2 - y^2}$. Each curve is labelled with the value, $y = y_n$, that corresponds to the curve. If we have a family of curves corresponding to $y = \text{constant}$

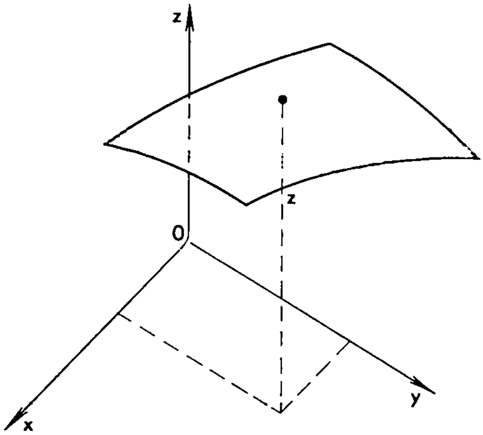


Fig. 24

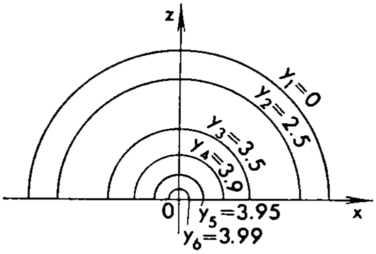


Fig. 25

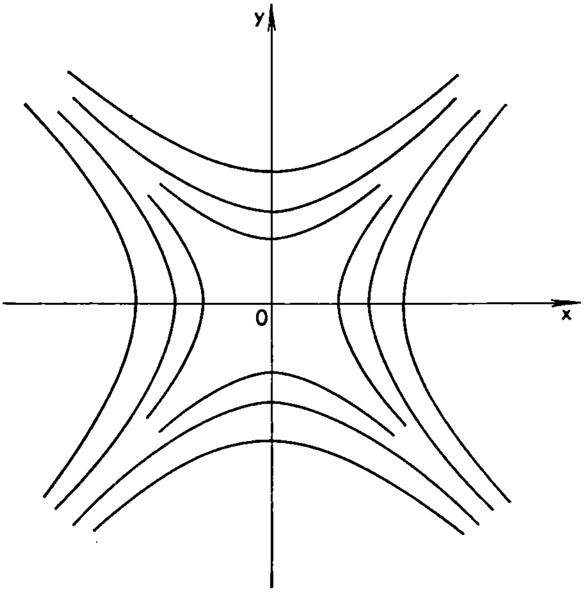


Fig. 26

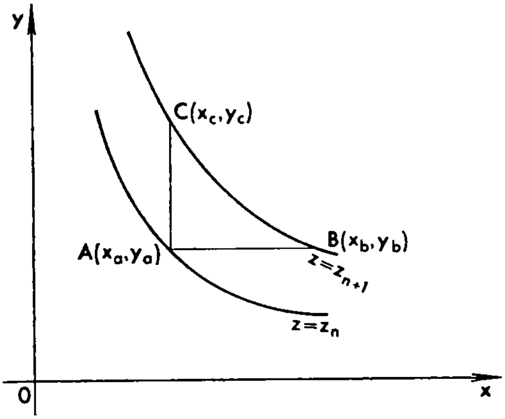


Fig. 27

in the xz-plane, then the derivative $\frac{\partial z}{\partial x}$ signifies geometrically (exactly like $\frac{\partial z}{\partial x}$ does in the case of one variable) the slope of the tangent line to the x-axis. The derivative $\frac{\partial z}{\partial y}$ may be found by forming the ratio $\frac{z_{n+1}(x) - z_n(x)}{y_{n+1} - y_n}$; here, two adjacent curves are needed.

Naturally, the function $z = f(x, y)$ can be represented graphically if we construct the graphs of $z(y)$ for $x =$ constant in the yz-plane.

A fundamentally different way of describing a function $z = f(x, y)$ is obtained if we construct a section of the surface $z = f(x, y)$ by horizontal planes $z =$ constant and then plot the resulting curves (so-called *level lines*) on the xy-plane. This is the method used for indicating relief on maps. Fig. 26 illustrates such a representation for the function $z = x^2 - y^2$. Each curve corresponds to a definite constant value of z.

How does one find derivatives on a graph of this kind? A portion of the graph is shown enlarged in Fig. 27. It is clear that if the level lines are drawn close enough together, then, to any degree of accuracy,

$$\frac{\partial z}{\partial x} = \frac{z_{n+1} - z_n}{x_b - x_a}, \quad \frac{\partial z}{\partial y} = \frac{z_{n+1} - z_n}{y_c - y_a}$$

The closer together the lines with distinct values of z, the greater are the partial derivatives. (It is assumed that these lines are drawn for equal intervals with respect to z.)

Exercises

Construct a family of curves corresponding to $y =$ constant for the following functions:

1. $z = xy$. 2. $z = x^2 - y^2$. 3. $z = \sqrt{x^2 + y^2}$.
4. Construct a family of curves corresponding to $z =$ constant for the function $z = x^2 - y^2$.

4.3 Implicit functions

It is often useful to apply the concepts of the theory of functions of several variables when investigating functions of one variable. (This was illustrated in HM, Sec. 3.12, but here we will discuss it in more detail.) Such a function is usually specified by means of an "explicit" formula, for example, $y = ax^2$ or $y = be^{ax}$ and the like, in which it is directly indicated how to evaluate y for a given value of x. Such formulas are best suited for performing mathematical operations.

An alternative method of defining y as a function of x is called defining the function *implicitly*, thus:

$$f(x, y) = 0 \tag{8}$$

For example,

$$c^3 + y^3 + 3axy = 0 \tag{9}$$

or

$$(x + y)^3 - b^2(x - y) = 0 \tag{10}$$

and so on.

In this case, to find y for a given x it is necessary to solve (8) for y. The solution of the equation is often in a much more complicated form than formula (8). Take (9) for instance:

$$y = \sqrt[3]{-\frac{c^3}{2} + \sqrt{\frac{c^6}{4} + a^3x^3}} + \sqrt[3]{-\frac{c^3}{2} - \sqrt{\frac{c^6}{4} + a^3x^3}}$$

It often happens that the solution cannot even be written as a formula and the equation (8) can only be solved numerically.

Still, there are cases where it is sufficient to resort to a function of only one variable in order to investigate the functional relationship at hand. Such is the case when the equation is solved or is readily solvable for x, that is, when it can be written as

$$x = \varphi(y) \tag{11}$$

and also when the equation is solvable parametrically (this will be discussed later on). To illustrate, take example (9). Here y is expressed in terms of x in an extremely complicated fashion but it is easy to find

$$x = -\frac{c^3 + y^3}{3ay}$$

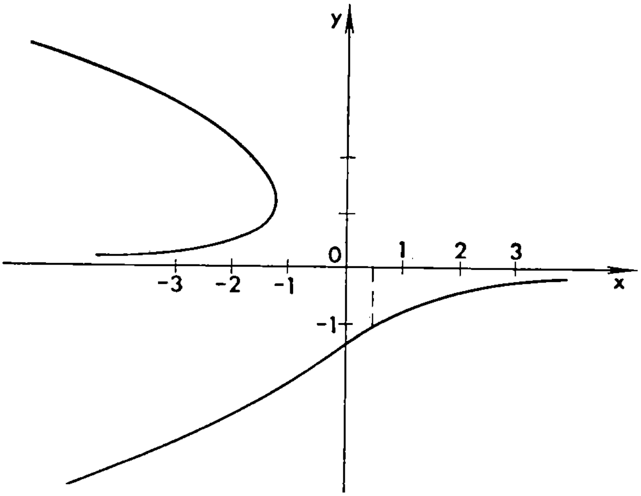


Fig. 28

Then by specifying various values of y, it is easy to find the corresponding values of x, which can be tabulated and plotted as a graph in terms of x- and y-coordinates. (For $a = 1$, $c = 1.44$, see Table 4 and the graph in Fig. 28).

Table 4

y	x	y	x	y	x
-4.0	-5.08	-0.5	1.91	1.5	-1.42
-3.0	-2.66	-0.25	3.98	2.0	-1.83
-2.0	-0.83	0.25	-4.02	2.5	-2.48
-1.5	-0.084	0.5	-2.02	3.0	-3.34
-1.0	-0.67	1.0	-1.33	4.0	-5.58

Having constructed the curve, we can find y from a given x. To do this (see Fig. 28), draw a vertical line corresponding to the required x and find the desired value of y.* If we have a table of values of x for certain values of y, then for a y that corresponds

* From Fig. 28 it is evident that certain values of x are associated with one value of y, and some values of x are associated with three values of y, and there is one value of x that corresponds to two values of y. This is due to the fact that an equation of the third degree can have one, two or three real distinct roots (if there are two, one of them is a double root, and it corresponds to tangency of the straight line $x =$ constant and the curve).

to a value of x not in the table we can use linear interpolation (cf. Sec. 2.1).

Here is how this is done. Take two pairs of numbers from the table, (y_1, x_1) and (y_2, x_2), such that $x_1 < x < x_2$. We assume that when x varies from x_1 to x_2, the graph of the function differs only slightly from the straight line $y = kx + b$. The numbers k and b are readily determined from the fact that $y = y_1$ when $x = x_1$ and $y = y_2$ when $x = x_2$. Indeed, $y_1 = kx_1 + b$, $y_2 = kx_2 + b$, whence $k = \frac{y_2 - y_1}{x_2 - x_1}$, $b = \frac{x_2 y_1 - x_1 y_2}{x_2 - x_1}$. For this reason

$$y = \frac{y_2 - y_1}{x_2 - x_1} x + \frac{x_2 y_1 - x_1 y_2}{x_2 - x_1}$$

The values of y for all x located between x_1 and x_2 may be computed from the formula thus obtained. The closer together the numbers x_1 and x_2 and also y_1 and y_2, that is, the closer together the points are in the table, the more exact is the formula for y.

How does one find the derivative $\frac{dy}{dx}$ in that case? There is no need to solve the equation and find $y = f(x)$. If the function is given in the form $x = \varphi(y)$, then $dx = \frac{d\varphi}{dy} dy$, whence

$$\frac{dy}{dx} = \frac{1}{\frac{d\varphi}{dy}} = \frac{1}{\varphi'(y)}$$

The only drawback to this expression is that the derivative is given as a function of y. And so if it is necessary to find $\frac{dy}{dx}$ for a given x, then again we first have to find the y corresponding to a given x, and only then substitute the y into the expression

$$\frac{dy}{dx} = \frac{1}{\varphi'(y)}$$

At first glance it would seem that it is easier to determine the derivative numerically, as in Sec. 2.2, for if we can determine y for a given x, then we can also find $y(x + \Delta x)$ corresponding to the quantity $x + \Delta x$, after which we can approximate the derivative by, say, the formula

$$\frac{y(x + \Delta x) - y(x)}{\Delta x}$$

Actually, this procedure is much worse: we would have to find y for two distinct values of x, that is, we would have to solve equation (11) twice. Also, it is necessary to determine y to a high degree of accuracy because we need the difference of two close-lying values of y, and a small error in each of them can produce a substantial relative error in their difference. In the expression for the derivative, the denominator contains the small quantity Δx and therefore even a slight error in the numerator can lead to an appreciable error in the derivative. Hence it is better to use the formula

$$\frac{dy}{dx}=\frac{1}{\varphi'(y)}$$

It is well to observe that the formula for the second derivative is much more complicated. Don't think that

$$\frac{d^2y}{dx^2}=\frac{1}{\dfrac{d^2x}{dy^2}}=\frac{1}{\dfrac{d^2\varphi}{dy^2}}$$

Actually, this formula is incorrect dimensionally: d^2y/dx^2 has the dimensions of y/x^2, while $1/\dfrac{d^2\varphi}{dy^2}$ has the dimensions of y^2/x. The proper formula is obtained thus:

$$\frac{d^2y}{dx^2}=\frac{d}{dx}\left(\frac{dy}{dx}\right)=\frac{dy}{dx}\,\frac{d}{dy}\left(\frac{dy}{dx}\right)=\frac{1}{\varphi'(y)}\,\frac{d}{dy}\left(\frac{1}{\varphi'(y)}\right)=-\frac{\varphi''(y)}{[\varphi'(y)]^3}=-\frac{\dfrac{d^2x}{dy^2}}{\left[\dfrac{dx}{dy}\right]^3}$$

It is easy to verify the dimensionality of this expression: it is the same as that of the ratio $\frac{x/y^2}{(x/y)^3}=\frac{y}{x^2}$, which is as it should be.

When considering the implicit functional relationship (8), we can arrive at a more general case where both variables x and y are expressible as a function of one and the same auxiliary variable t:

$$\left.\begin{aligned}x&=x(t),\\ y&=y(t)\end{aligned}\right\}$$

For instance, if in (10) we put

$$x+y=t \qquad (12)$$

then from (10) we get $x-y=t^3/b^2$, whence, invoking (12), we find

$$\left.\begin{aligned}x&=\frac{t}{2}+\frac{t^3}{2b^2},\\ y&=\frac{t}{2}-\frac{t^3}{2b^2}\end{aligned}\right|$$

This is the so-called *parametric* representation of a function (the variable t is the *parameter*). This representation can be obtained without involving relations like (8). It is considered in elementary calculus text books (see, for example, HM, Secs. 1.8, 3.3, 6.14). It is shown there how to find the derivative of such a function: write dx as $dx = \frac{dx}{dt}\,dt = x'(t)\,dt$, and similarly, $dy = \frac{dy}{dt}\,dt = = y'(t)\,dt$, whence $\frac{dy}{dx} = \frac{y'(t)}{x'(t)}$.

Let us find the second derivative $\frac{d^2y}{dx^2}$. We take advantage of the relation

$$\frac{dz}{dx} = \frac{dz}{dt}\cdot\frac{dt}{dx} = \frac{dz}{dt}\cdot\frac{1}{\frac{dx}{dt}}$$

Therefore

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d}{dx}\left(\frac{y'(t)}{x'(t)}\right) = \frac{d}{dt}\left(\frac{y'(t)}{x'(t)}\right)\frac{dt}{dx}$$

whence

$$\frac{d^2y}{dx^2} = \frac{1}{x'(t)}\cdot\frac{d}{dt}\left(\frac{y'(t)}{x'(t)}\right) = \frac{y''(t)}{[x'(t)]^2} - \frac{y'(t)\,x''(t)}{[x'(t)]^3}$$

where the primes indicate derivatives with respect to t.

In the general case of implicit representation, the function $f(x, y)$ may be such that the equation (8) is not solvable either in the form (11) or in parametric form. Let us, for instance, consider y as a function of x specified by the formula

$$y^5 + xy + x^5 - 7 = 0 \tag{13}$$

(Geometrically, this formula specifies a curve in the xy-plane.) From this it is impossible to obtain the expression $y = f(x)$. To find the derivative $\frac{dy}{dx}$, equate the derivatives of both members of (13), assuming y to be a function of x, $y = y(x)$, determined from this equation. With such an interpretation it is an identity and therefore admits differentiation:

$$5y^4y' + (1\cdot y + x\cdot y') + 5x^4 = 0$$

From this we get

$$y' = \frac{dy}{dx} = -\frac{y + 5x^4}{5y^4 + x} \tag{14}$$

We stress that the x and y on the right are not independent but are connected by the relation (13), so there is only one independent variable. To find y' for a given x, solve (13) numerically and find y for a given x and substitute this y into (14).

We now take up the general case of an implicit function of several (say two) variables.

If z is given as a function of x and y, then, when solving the equation $z = f(x, y)$ for x, we can obtain x as a function of y and z: $x = \varphi(y, z)$; and when solving the equation $z = f(x, y)$ for y, we can find $y = \psi(x, z)$. However, irrespective of whether we have solved the equation $z = f(x, y)$ for x or not, this equation yields x as a function of y and z. This is called an implicit representation of the function x.

The derivatives $\frac{\partial x}{\partial y}$ and $\frac{\partial x}{\partial z}$ may be found without expressing x explicitly. This is how it is done. From the relation $z = f(x, y)$ we find $dz = \frac{\partial z}{\partial x} dx + \frac{\partial z}{\partial y} dy$, whence

$$dx = \left(\frac{\partial z}{\partial x}\right)^{-1} dz - \left(\frac{\partial z}{\partial y}\right)\left(\frac{\partial z}{\partial x}\right)^{-1} dy \tag{15}$$

On the other hand, if $x = \varphi(y, z)$, then.

$$dx = \frac{\partial x}{\partial z} dz + \frac{\partial x}{\partial y} dy \tag{16}$$

Comparing expressions (15) and (16), we get

$$\left.\frac{\partial x}{\partial z}\right|_y = \frac{1}{\left.\frac{\partial z}{\partial x}\right|_y}, \tag{17}$$

$$\left.\frac{\partial x}{\partial y}\right|_z = -\frac{\left.\frac{\partial z}{\partial y}\right|_x}{\left.\frac{\partial z}{\partial x}\right|_y} \tag{18}$$

Formula (17) appears to be quite natural. In (18) we find, surprisingly, a minus sign. Geometrical reasoning will convince us of the truth of this formula.

Consider Figs. 26 and 27. In these figures are plotted the curves $z = \text{constant}$, so that $\left.\frac{\partial x}{\partial y}\right|_z$ is the derivative $\frac{dx}{dy}$ taken in ordinary fashion along these curves. Therefore

$$\left.\frac{\partial x}{\partial y}\right|_z = \frac{dx}{dy} = \frac{x_b - x_c}{y_b - y_c}$$

Noting that $y_b = y_a$, $x_c = x_a$, we get

$$\left.\frac{\partial x}{\partial y}\right|_z = \frac{x_b - x_a}{y_a - y_c}$$

Recalling the values of $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$, we see that

$$\frac{\partial z}{\partial y}\bigg/\frac{\partial z}{\partial x} = \frac{x_b - x_a}{y_c - y_a} = -\left.\frac{\partial x}{\partial y}\right|_z$$

Formula (18) may be written thus:

$$\left.\frac{\partial x}{\partial y}\right|_z \cdot \left.\frac{\partial z}{\partial x}\right|_y = -\left.\frac{\partial z}{\partial y}\right|_x$$

Noting that $\left.\frac{\partial z}{\partial y}\right|_x = 1\bigg/\left.\frac{\partial y}{\partial z}\right|_x$, we find that $\left.\frac{\partial x}{\partial y}\right|_z \cdot \left.\frac{\partial y}{\partial z}\right|_x \cdot \left.\frac{\partial z}{\partial x}\right|_y = -1.$

Like (18), the latter formula shows that in contrast to the case of a composite function of one variable we cannot cancel ∂x, ∂y, ∂z in the numerator and denominator. The point is that in this formula the three partial derivatives are computed under different conditions (z is held constant in the first case, x in the second, and y in the third).

In Sec. 4.1 we saw that the value of the derivative $\frac{\partial W}{\partial C}$ depends on which argument is regarded as being constant during the computation. This alone shows that ∂W and ∂C cannot be dealt with like numbers and cannot be cancelled with ∂W and ∂C in other formulas without taking into account the conditions under which the appropriate partial derivatives are evaluated.

Here is an example to illustrate the calculation of derivatives of implicit functions.

Let $z = x^2 + px + y^2 + qy + kxy$.

Find the derivatives $\frac{\partial x}{\partial z}$ and $\frac{\partial x}{\partial y}$. In order to express x in terms of y and z, it is necessary to solve a quadratic equation. There are no fundamental difficulties here, but the expression is cumbersome involving roots. Since $\frac{\partial z}{\partial x} = 2x + p + ky$, $\frac{\partial z}{\partial y} = 2y + q + kx$, it follows, by (17) and (18), that

$$\frac{\partial x}{\partial z} = \frac{1}{2x + p + ky}, \quad \frac{\partial x}{\partial y} = -\frac{2y + q + kx}{2x + p + ky}$$

To determine the values of $\frac{\partial x}{\partial z}$ and $\frac{\partial x}{\partial y}$ for concretely given y and z, one has to know the numerical value of x for these y and z, but one does not necessarily have to have the analytic expression

$x = \psi(y, z)$. It is quite obviously much simpler to numerically solve the equation $z = f(x, y)$ for x, given definite values of y and z, than to construct the general formula $x = \psi(y, z)$. Besides, in the case of an equation of higher than fourth degree and also an equation involving transcendental functions, the construction of a general formula is at times impossible.

The most general method for implicitly representing a function of two variables is by specifying it via a relation of the form $F(x, y, z) = 0$. We leave it to the reader to differentiate this relation, obtain expressions for the partial derivatives $\frac{\partial z}{\partial x}$, $\frac{\partial y}{\partial x}$, and so on and apply the result to the relation considered above, $z = f(x, y)$, by rewriting it as $z—f(x, y) = 0$.

The foregoing devices are conveniently used in the case of a function of one variable when it is defined implicitly with the aid of an equation of type (8). To do this, consider a function of two variables, $z = f(x, y)$, in other words, consider the relation (8) as the equation of the zero level line (that is, one corresponding to the value $f = 0$) of the function $z = f(x, y)$. Then the problem of evaluating $\frac{dy}{dx}$ reduces to the foregoing problem of computing $\left.\frac{\partial y}{\partial x}\right|_{z=const.}$

To illustrate, in (13) we have to put

$$z = y^5 + xy + x^5$$

Then by formula (18)

$$\frac{dy}{dx} = \left.\frac{\partial y}{\partial x}\right|_z = -\left.\frac{\partial z}{\partial x}\right|_y \bigg/ \left.\frac{\partial z}{\partial y}\right|_x = -\frac{y + 5x^4}{5y^4 + x}$$

which is to say we again arrive at (14).

Exercises

1. Find $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ for the following functions given parametrically:

   (a) $x = \frac{1}{2}t$, $y = t^2 + t$; (b) $x = 2\sin^3 t$, $y = 2\cos^3 t$;

   (c) $x = \cos t + t\sin t$, $y = \sin t - t\cos t$.

2. $x = \sin t$, $y = \cos 2t$. Find $\frac{dy}{dx}$ for $x = \frac{1}{2}$.

3. $z = x^3 + y^3 + xy$. Find $\frac{\partial x}{\partial z}$, $\frac{\partial x}{\partial y}$ for $y = 0$, $z = 1$.

4. $z = x^5 + xy^2 + y^5$. Find $\frac{\partial x}{\partial z}$, $\frac{\partial x}{\partial y}$.

5. $x^2 + y^2 - 4x - 10y = -4$. Find $\frac{\partial y}{\partial x}$.

6. $x^4y + xy^4 - x^2y^2 - 1 = 0$. Find $\frac{dy}{dx}$.

4.4 Electron tube

An interesting example of a function of two variables is the current j passing through the anode (plate) of a three-electrode (grid-controlled) electron tube (Fig. 29).

The flow of electrons to the anode is influenced by the grid potential u_G and the potential of the anode (the cathode is grounded so that it has potential zero). We thus have to do with a function of two variables $j = j(u_G, u_A)$. Ordinarily the graph is in the form of a family of curves in terms of the coordinates u_G, j with constant values of u_A on every curve. Fig. 30 illustrates such a family for the Soviet tube 6 C1Zh. The current j is given in milliamperes, the voltages u_G and u_A in volts. The derivative $\left.\frac{\partial j}{du_G}\right|_{u_A}$ is proportional to the slope of the tangent lines to the curves in Fig. 30. Over rather considerable ranges of u_G, these curves differ only slightly from straight lines. Their slope over this interval is termed the *transconductance* of the tube and is denoted by S (other symbols are also used, for example g_m, so that $S = \frac{\partial j}{\partial u_G}$. S is expressed in milliamperes per volt.

We form $\left.\frac{\partial j}{\partial u_A}\right|_{u_G}$. The reciprocal quantity,

$$R = \frac{1}{\left.\frac{\partial j}{\partial u_A}\right|_{u_G}} = \left.\frac{\partial u_A}{\partial j}\right|_{u_G}$$

is called the *anode resistance.* The meaning of this term stems from the fact that if the tube in our circuit were replaced by a fixed resistor R_1, then by Ohm's law $j = \frac{u_A}{R_1}$, whence

$$\frac{\partial j}{\partial u_A} = \frac{1}{R_1}$$

or

$$R_1 = \frac{1}{\frac{\partial j}{\partial u_A}}$$

When the current is expressed in milliamperes and the potential in volts, then R is obtained in kilohms. Finally, a very important characteristic of an electron tube is the quantity $\left.\frac{\partial u_A}{\partial u_G}\right|_j$. This is a

Fig. 29

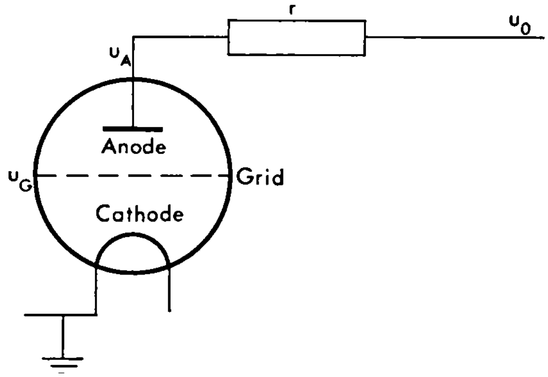
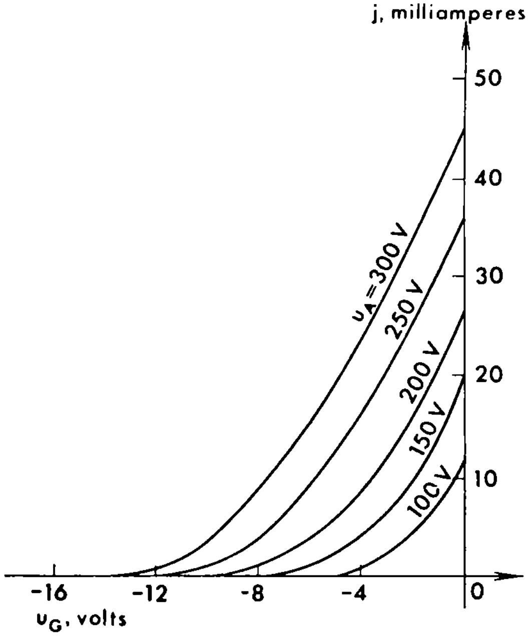


Fig. 30



dimensionless negative quantity. The number $\mu = -\left.\frac{\partial u_A}{\partial u_G}\right|_j$ is called the *amplification factor* of the tube. By formula (18)

$$\mu = \left.\frac{\partial j}{\partial u_G}\right|_{u_A}\left(\left.\frac{\partial j}{\partial u_A}\right|_{u_G}\right)^{-1} = S \cdot R$$

Let us find the relation between the variation of grid potential and that of the anode potential. Regarding u_A as a function of the variables u_G and j, we find

$$du_A = \left.\frac{\partial u_A}{\partial j}\right|_{u_G} dj + \left.\frac{\partial u_A}{\partial u_G}\right|_j du_G, \text{ whence } du_A = R\, dj - \mu\, du_G$$

If the anode current is kept constant, then $dj = 0$, and, hence, $du_A = -\mu\, du_G$. Thus, the change in anode potential is μ times that of the grid potential. We can therefore say that the tube amplifies the grid potential μ times, whence the name *amplification factor* for μ.

The relation $\frac{du_A}{du_G} = -\mu$ refers to the ideal case when j remains constant. Actually, however, a change in u_G ordinarily causes a change in the current, and the change in u_A is less than in the case of $j =$ constant. Consider the circuit shown in Fig. 29 where we have a resistance r in the anode circuit and a constant voltage u_0. Then by Ohm's law

$$j = \frac{u_0 - u_A}{r} = j(u_A, u_G)$$

where the right-hand member is a function that we examined at the beginning of this section. Let us take the total differential of the right and left sides of the equation, assuming u_0 and r to be constant:

$$-\frac{1}{r}\, du_A = \frac{\partial j}{\partial u_A}\, du_A + \frac{\partial j}{\partial u_G}\, du_G$$

or

$$-\frac{1}{r}\, du_A = \frac{1}{R}\, du_A + S\, du_G$$

whence

$$du_A = -\frac{S \cdot Rr}{r + R}\, du_G = -\mu\frac{r}{r + R}\, du_G$$

From this formula it is clear that in the circuit of Fig. 29 the absolute value of the ratio $\frac{du_A}{du_G}$ is less than μ since $\frac{r}{r + R} < 1$. If $r \gg R$, then the current j is almost constant and $\left|\frac{du_A}{du_G}\right|$ is very close to μ. But a high voltage u_0 is needed to drive the given current through the large resistance r, with a substantial portion of the electric power going to heat up the resistor r. It turns out that in all cases, except for amplifying a constant or slowly varying voltage, an inductor (coil) is better than a resistor.

The facts presented in this section show that a precise mathematical statement of the laws governing the operation of an electron tube is connected with functions of two variables and the main quantities describing a tube are partial derivatives.

Exercise

Use Fig. 30 to find the values of S, R and μ for a 6 C1Zh radio tube.

4.5 Envelope of a family of curves

Another illustration of the use of partial derivatives is the problem of finding the envelope of a one-parameter family of curves. Let us examine this problem.

Consider a collection of flight paths (trajectories) of shells fired from one point (the coordinate origin) with one and the same initial velocity v_0 but at different angles φ ranging from 0° to 180° (Fig. 31).

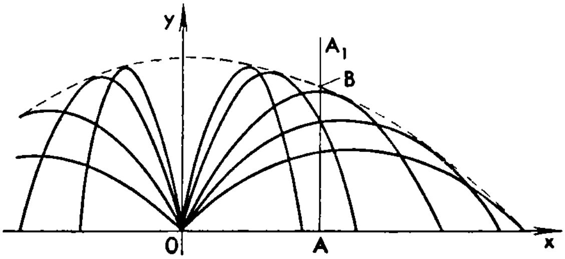
Each trajectory is a curve in the xy-plane, that is, it is described by a specific relationship $y(x)$. Writing down the relations $x = x(t)$ and $y = y(t)$, disregarding air resistance and eliminating t, we easily find (see, for example, HM, Sec. 6.14) y as a function of x:

$$y = x \tan \varphi - \frac{g}{2v_0^2 \cos^2 \varphi} x^2 \tag{19}$$

There is a definite curve for each concrete value of φ. Regarding a variety of values of the parameter φ, we obtain a family of curves. We can take it that the height y of the trajectory of a shell is a function of two variables: the horizontal distance x and the angle of departure φ, or $y = y(x, \varphi)$. Then separate trajectories yield y as a function of x for constant φ. (The family of trajectories is depicted in Fig. 31.) Considering this family of trajectories, it is easy to see that they fill one portion of the xy-plane and do not appear at all in the other portion. Thus, there is a safe zone in which no shells fall, no matter what the initial velocity or the angle of departure.

Let us try to define the boundary of the effective (killing) zone. (In Fig. 31 the boundary is shown dashed.) Each point of the boundary is also a point of one of the trajectories, otherwise the boundary would not be reached. It will be seen that for the boundary shown in Fig, 31, each point is a point of a trajectory corresponding to an angle of departure $\varphi \geqslant 45°$. (It can be demonstrated that for $\varphi = 45°$ we attain the greatest possible firing range. Trajectories corresponding to $\varphi < 45°$ are tangent — as will be seen below — to the boundary of the effective zone below the x-axis. This portion of the boundary is of practical interest if the fire is aimed at targets located below the gun level.) At the same time, the trajec-

Fig. 31



tory cannot cross the boundary but must touch it. If a trajectory crossed the boundary it would go outside the limits of the zone, but this runs counter to the fact that the boundary separates the effective zone from the safe zone.

The boundary of the region filled with the family of curves touching this boundary is called the *envelope* of the family of curves. Let us find the equation of the envelope of the family of trajectories. To do this, draw a vertical line AA_1 and locate the point B at which the vertical line cuts the envelope. We have taken a definite value of x by drawing this vertical line. The point B corresponds to the greatest height y at which a shell can be in covering the horizontal distance x for any angle of departure φ. And so we have to find the maximum of $y(\varphi)$ for a given fixed x. We get the condition

$$\left.\frac{\partial y(x, \varphi)}{\partial \varphi}\right|_x = 0 \tag{20}$$

Condition (20) gives us an equation that connects x and φ. For every value of x there is a value of φ defined by equation (20) so that we get $\varphi = \varphi(x)$. Substituting $\varphi = \varphi(x)$ into the equation of the family (19), we find the equation of the envelope.

Let us carry out the necessary manipulations. For the sake of convenience we introduce the variable $\theta = \tan\varphi$ in place of the variable φ. Noting that by a familiar formula of trigonometry $\frac{1}{\cos^2\varphi} = $ $= \tan^2\varphi + 1 = \theta^2 + 1$, we rewrite the equation (19) as

$$y = \theta x - \frac{g}{2v_0^2}(\theta^2 + 1)\, x^2$$

Finally, introducing the notation $\frac{v_0^2}{g} = l$ (as will be seen from (22), l is the maximum horizontal range of fire), we get

$$y = \theta x - \frac{x^2}{2l}(\theta^2 + 1) \tag{21}$$

whence

$$\frac{\partial y}{\partial \theta} = x - \frac{x^2}{l}\theta$$

The condition (20) * gives $\theta = \frac{l}{x}$. Substituting this value into (21), we get the equation of the envelope,

$$y = \frac{l}{2} - \frac{x^2}{2l} \tag{22}$$

We advise the reader to carry through the entire derivation without passing to the variable θ.

Exercise

A shell leaves a gun with the velocity $v_0 = 100$ m/s. Can we hit a target at a horizontal distance of 500 metres from the gun and at a height of 300 metres? at 500 metres?

4.6 Taylor's series and extremum problems

Taylor's power series, which is well known for functions of one variable, can also be useful for functions of many variables. Recall (see, for example, HM, Sec. 3.17) that for functions of one variable this series is of the form

$$f(x) = f(a) + \frac{f'(a)}{1!}(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \frac{f'''(a)}{3!}(x - a)^3 + \dots \tag{23}$$

Now let us consider a function $f(x, y)$ of two variables. Let x be close to a constant value a, and y to a constant value b. Then the crudest approximation, the formula of the zero-order approximation that does not take into account changes in x and y, is of the form

$$f(x, y) = f(a, b)$$

A more exact first-order approximation that takes into account first-order terms is

$$f(x, y) = f(a, b) + A(x - a) + B(y - b) \tag{24}$$

where A and B are certain numerical coefficients. To find them, we take the partial derivatives of both sides of (24) with respect to x:

$$f'_x(x, y) = A$$

(the first and third terms in the right-hand member of (24) do not depend on x, and so their partial derivatives with respect to x are equal to zero). Since the coefficient A must be a constant, we get

* In making the substitution $\varphi = \varphi(\theta)$ we get $\frac{\partial y}{\partial \theta} = \frac{\partial y}{\partial \varphi} \cdot \frac{\partial \varphi}{\partial \theta}$ so that $\frac{\partial y}{\partial \theta} \neq \frac{\partial y}{\partial \varphi}$. However, the condition $\frac{\partial y}{\partial \theta} = 0$ coincides with the condition $\frac{\partial y}{\partial \varphi} = 0$.

$A = f'_x(a, b)$. In similar fashion, differentiating (24) with respect to y, we find $B = f'_y(a, b)$, that is to say, (24) is indeed of the form

$$f(x, y) = f(a, b) + f'_x(a, b)(x - a) + f'_y(a, b)(y - b) \qquad (25)$$

(Derive this formula in straightforward fashion from the results of Sec. 4.1.)

The more exact formula of the "second approximation", which takes into account second-order terms as well, is of the form

$$\begin{aligned} f(x, y) = f(a, b) &+ [A(x - a) + B(y - b)] \\ &+ [C(x - a)^2 + D(x - a)(y - b) + E(y - b)^2] \end{aligned} \qquad (26)$$

where A, B, C, D, E are numerical coefficients. To find them, we compute the first- and second-order partial derivatives:

$$\begin{aligned} f'_x(x, y) &= A + 2C(x - a) + D(y - b), \\ f'_y(x, y) &= B + D(x - a) + 2E(y - b), \\ f''_{xx}(x, y) &= 2C, f''_{xy}(x, y) = f''_{yx}(x, y) = D, f''_{yy}(x, y) = 2E \end{aligned}$$

Putting $x = a$, $y = b$ in these equations, we get

$$A = f'_x(a, b), \quad B = f'_y(a, b), \quad C = \frac{1}{2} f''_{xx}(a, b),$$

$$D = f''_{xy}(a, b), \quad E = \frac{1}{2} f''_{yy}(a, b)$$

Substituting these values into (26), we obtain the final formula for the second approximation:

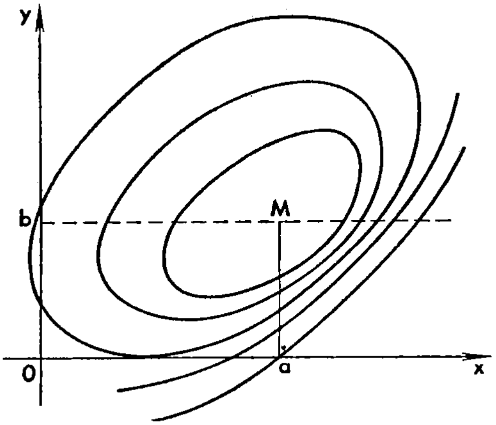
$$\begin{aligned} f(x, y) = f(a, b) &+ [f'_x(a, b)(x - a) + f'_y(a, b)(y - b)] \\ &+ \frac{1}{2}[f''_{xx}(a, b)(x - a)^2 + 2f''_{xy}(a, b)(x - a)(y - b) \\ &+ f''_{yy}(a, b)(y - b)^2] \end{aligned} \qquad (27)$$

Although we derived this formula independently of (25), we see that the right side has the same linear part as in (25) and, besides, second-order corrections have appeared. We leave it to the reader to derive, in similar fashion, the formula for the third approximation (incidentally, this formula is rarely used); added third-order corrections will be on the right side of (27).

(As a check we will assume that the function f does not depend on y; then in the right members of (25) and (27) all the partial derivatives in which differentiation with respect to y occurs will drop out and we will get the partial sums of the series (23) for a function of one variable.)

As in the case of functions of one variable, the formulas (25) and (27) are most convenient if $|x - a|$ and $|y - b|$ are very small.

Fig. 32



But if these differences are great, the formulas fail to hold and can lead to erroneous conclusions. Formulas for functions of more than two independent variables are similar in form but we will not dwell on them.

The formulas just obtained can be used, say, for investigating extremum points of the function $f(x, y)$. Suppose that this function has an extremum (maximum or minimum) at $x = a$, $y = b$. The rough arrangement of the level lines of the function f in the plane of the arguments near the extremum point M is shown in Fig. 32. It is then clear that if we put $y = b$ and change x alone, then the resulting function $f(x, b)$ of one variable x has an extremum at $x = a$. Geometrically, this means that if we go along the straight dashed line in Fig. 32, we get an extremum at M. But then, as we know from the calculus of functions of one variable, the derivative at the extremum point must be zero:

$$\left[\frac{d}{dx} f(x, b)\right]\bigg|_{x=a} = 0$$

In the square brackets we have the derivative with respect to x with fixed $y = b$, that is, a partial derivative with respect to x. In similar fashion we consider the case of a fixed $x = a$ and a varying y. We arrive at the necessary conditions for the extremum of a function of two variables:

$$f'_x(a, b) = 0, \quad f'_y(a, b) = 0 \tag{28}$$

(We made use of similar reasoning in Sec. 2.3 when we considered the method of least squares.)

As we know, for a function $f(x)$ of one variable the necessary condition $f'(a) = 0$ of an extremum is "almost sufficient": for example, if $f''(a) \neq 0$, then there must be an extremum at the point $x = a$, that is, a maximum for $f''(a) < 0$, and a minimum for $f''(a) > 0$. It might be expected that also in the case of a function of two variables there must be an extremum at the point (a, b), provided condition (28) holds if at that point the second-order partial derivatives are different from zero. But this is not so, the sufficient condition turns out to be more complicated.

Let us first consider some examples. Let

$$z = f(x, y) = x^2 + y^2$$

The condition (28) yields $2x = 0$, $2y = 0$, that is, the suspected extremum is the origin of coordinates. Indeed, here at the origin we have a minimum since $z = 0$ and at other points $z > 0$. As for the second-order partial derivatives, they are constant in this example, and $f''_{xx} = 2, f''_{xy} = 0, f''_{yy} = 2$. In similar fashion it is easy to verify that the function

$$z = Ax^2 + Cy^2 \tag{29}$$

has a minimum at the origin for $A > 0$, $C > 0$ and a maximum for $A < 0$, $C < 0$.

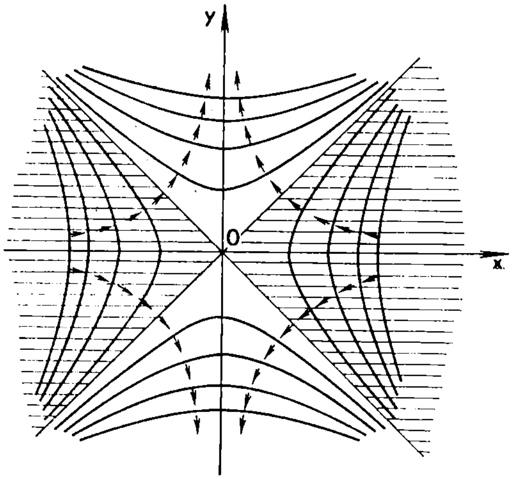
A completely new picture emerges if in (29) A and C are of different sign, for instance, if we consider the function

$$z = x^2 - y^2$$

The appropriate level lines are shown in Fig. 33. The portion of the plane where $z > 0$ is shown hatched in the figure; the small arrows indicate the direction of fall of the level (on a map, this would mean the direction of water flow). For $y = 0$ we get $z = x^2$, that is, the function increases in both directions from the origin along the x-axis and has a minimum at the origin itself. But if $x = 0$, then $z = -y^2$, that is, the function decreases in both directions along the y-axis and has a maximum at the origin. If we consider other straight lines passing through the origin, the function has a maximum at the origin of some lines and a minimum at the origin of others. This case is called a *minimax*, and there is no extremum at the origin, although the necessary conditions (28) hold and the second-order partial derivatives are not all zero. In geography, a minimax is called a *saddle*; it is observed, for instance, at the highest point of a mountain pass. Another example is a simple horse saddle. For example, take a saddle corresponding to Fig. 33; one can sit on it hanging his legs along the y-axis and facing the x-axis. Then in front and behind (along the x-axis) the saddle rises.

We have a similar minimax for the function $z = xy$ at the origin; the corresponding pattern of level lines is shown in Fig. 26. (Where is $z > 0$ and $z < 0$ in Fig. 26?)

Fig. 33



Now let us examine the general case. If the necessary conditions (28) are fulfilled, then formula (27) becomes

$$f(x, y) = f(a, b) + \frac{1}{2}[f''_{xx}(a, b)(x-a)^2 + 2f''_{xy}(a, b)(x-a)(y-b) + f''_{yy}(a, b)(y-b)^2] \quad (30)$$

In deriving this formula, we of course disregard third-order terms, but when investigating the extremum we must be near the values $x = a$, $y = b$, where these terms are less essential than the second-order terms that have been written down. For brevity put

$$f''_{xx}(a, b) = A, \quad f''_{xy}(a, b) = B, \quad f''_{yy}(a, b) = C,$$

$$x - a = \xi, \quad y - b = \eta$$

Then formula (30) shows that everything depends on the behaviour of the *quadratic form* (this is a homogeneous polynomial of second degree) $P(\xi, \eta) = A\xi^2 + 2B\xi\eta + C\eta^2$ since, approximately, near the values $x = a$, $y = b$, $f(x, y) = f(a, b) + \frac{1}{2} P(\xi, \eta)$. If it is positive for all ξ, η (for example, if it is of the form $\xi^2 + \eta^2$), then $f(x, y) > f(a, b)$ near the point (a, b) and thus the function f has a minimum at this point. If this form is negative, then f has a maximum at (a, b). But if the form can assume values of both signs (say, if it is of the form $\xi^2 - \eta^2$), then there will be a minimax at the point (a, b), in other words there will not be any extremum.

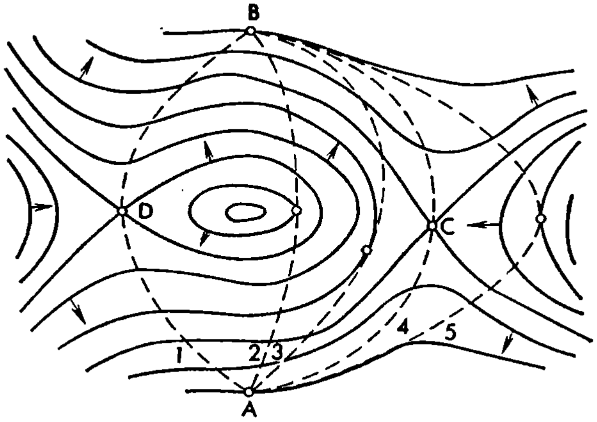


Fig. 34

How do we know, from the coefficients, A, B, C, which of these cases occurs? Write

$$P(\xi, \eta) = \eta^2\left[A\left(\frac{\xi}{\eta}\right)^2 + 2B\frac{\xi}{\eta} + C\right] = \eta^2(At^2 + 2Bt + C) \qquad (31)$$

where t stands for ξ/η. From elementary mathematics we know that if the discriminant

$$B^2 - AC > 0 \qquad (32)$$

the polynomial in brackets has two real zeros where it changes sign. Hence this is a minimax case. But if

$$B^2 - AC < 0 \qquad (33)$$

then the indicated polynomial has imaginary zeros and so there is no change of sign (a polynomial can change sign only when it passes through a zero). Hence this is an extremum case. To find out what sign the right member of (31) has, put $t = 0$. We see that if in addition to (33) we have $C > 0$, then the right side of (31) is positive for all t and therefore, by virtue of the preceding paragraph, the function f has a minimum at the point (a, b). If in addition to (33) we have $C < 0$, then f has a maximum. (Verify all these conditions using the examples discussed above.)

Minimax points are of great importance in the solution of an important class of problems which we will illustrate with a vivid example. Suppose we have two villages A and B in a region of rolling country with the level lines as indicated in Fig. 34. It is required to construct a road between them. This can be done in a variety of ways, several of which are shown by dashed lines in Fig. 34. (For pictorialness, arrows have been added to indicate lines of water flow.)

Taking any road (l) from A to B, we have to go uphill to the point indicated in Fig. 34 and then downhill. If the uphill portion is difficult for transport, the natural thing is to require that it be a minimum. To be more precise in stating this requirement, we denote by $z(M)$ the altitude at any point M on the map. Then the uphill portion along the line (l) is equal to $\max\limits_{M \text{ on } (l)} z(M) - z(A)$, where $\max\limits_{M \text{ on } (l)} z(M)$ stands for the greatest value of altitude z on the line (l). But the value of $z(A)$ for all lines being compared is the same; thus, from among all lines (l) joining A and B it is required to find that line along which $\max\limits_{M \text{ on } (l)} z(M)$ is a minimum, that is to say, the line that realizes $\min\limits_{(l)} \max\limits_{M \text{ on } (l)} z(M)$. It is clear that the desired line passes through the point C of the mountain pass which is precisely the minimax point of the function $z(M)$. True, there is another such point D, but it is higher and therefore worse. (Verify the fact that if the level lines in Fig. 34 are drawn every 100 metres, the intermediate rise in altitude along the roads indicated in the drawing is, respectively, 500, 600, 300, 200, and 400 metres.)

The difference between the cases (32) and (33) serves as a basis for a classification of points on an arbitrary surface in space. Suppose we have a surface (S) and any point N on that surface. We choose a system of coordinates so that the x-axis and y-axis are parallel to the plane (P) tangent to (S) at the point N. Then, near N, the surface (S) may be represented by the equation $z = f(x, y)$, and, because of the choice of axes, equations (28) hold at point N; here, a and b are the values of the coordinates x and y at the point N so that $N = (a, b, f(a, b))$. Then, depending on whether inequality (33) or (32) holds, N is called an *elliptic* or *hyperbolic point* of the surface (S). In the former case, the surface (S) near N is convex and is located to one side of (P). In the latter case, the surface (S) near N is saddle-like and is located on both sides of (P); the tangent plane (P) intersects (S) in two lines that intersect in the point N.

It can be demonstrated that the conditions (32) and (33) are not violated under a rotation of the coordinate axes in space. Therefore if the equation of the surface (S) is given as $z = f(x, y)$, then in order to determine the type of some point of the surface there is no need to choose new axes in order to satisfy the condition (28). One merely has to verify that the conditions (32) or (33) hold in the original system of coordinates x, y, z.

There are surfaces (spheres, ellipsoids, paraboloids of revolution, etc.) where all points are elliptic; such surfaces are convex in the large. There are surfaces, for instance, the surface with equation $z = x^2 - y^2$ that corresponds to Fig. 33, in which all points are hyperbolic. On the other hand, there are surfaces with points of both kinds. The surface of a *torus* (a doughnut of ideal shape) is one such example. Here, the portions filled with elliptic points are separated from the portions filled with hyperbolic points by lines at the points of which $B^2 - AC = 0$. These are called *parabolic points*. On the torus these are points at which the tangent plane is perpendicular to the axis of the torus; they fill two circles. (Try to figure out where on your own body there are lines of parabolic points). Incidentally there are surfaces — cylindrical or conic — that are completely filled with parabolic points.

It is clear that the type of point of surface does not change under any translation of the surface in space, that is, the indicated classification of the points of the surface is geometrically invariant. In contrast, the notion of a highest or lowest point of a surface, which is essential in the study of extrema, is not invariantly linked to the surface itself, since the highest point ceases to be such under a rotation of the surface.

Similarly, the points of the graph (L) of a function $y = \varphi(x)$ corresponding to its extremal values are not connected invariantly with the line (unlike, say, the inflection points of this line). The points of the line with equation $f(x, y) = 0$, in which $\frac{dy}{dx} = 0$ or $\frac{dy}{dx} = \infty$, are not invariant either.

The foregoing reasoning can be applied to a study of the structure of a line (L) with the equation $f(x, y) = 0$ on an xy-plane in the vicinity of a point $N(a, b)$ of that line. In this connection, it is useful to consider the auxiliary surface (S) with the equation $z = f(x, y)$ so that (L) results from an intersection of (S) with the (P) plane: $z = 0$. Here, the line (L) turns out to be included in the family of curves with equation $f(x, y) = C$ for different values of C, that is to say, level lines of the function f resulting from the intersection of (S) with the planes $z = C$ parallel to (P). If $f'_x(a, b) \neq 0$ or $f'_y(a, b) \neq 0$ — in that case N is called an *ordinary point* of the line (L) — then (P) is not tangent to (S) at the point $(a, b, 0)$ and so (L) has the form of a smooth arc near N. But if $f'_x\,(a, b) = f'_y\,(a, b) = 0$, then N is called a *singular point* of the line (L) and (L) consists of points common to the surface (S) and the plane (P) tangent to (S). From the foregoing it follows that if inequality (33) holds, then N is an *isolated point* of the line (L), that is, there are no other points of (L) in the vicinity of N. If the inequality (32) holds true, then N is a *point of self-intersection* of (L), that is, in some vicinity of N the curve (L) consists of two arcs intersecting at N. But if $B^2 - AC = 0$, then the structure of (L) in the vicinity of N can be substantially more intricate. (To illustrate, pass to polar coordinates and construct the curves $x^2 - y^2 = (x^2 + y^2)^2$ and $x^2 + y^2 = (x^2 + y^2)^2$. What is the coordinate origin for each of them?)

Exercise

Test the following functions for extrema:

(a) $f(x, y) = x^2 - xy + y^2 - 2x + 4y - 1$;

(b) $f(x, y) = x^2 + 3xy - 2y + 2$;

(c) $f(x, y, z) = x^2 + y^2 - z^2 + 2z$.

4.7 Multiple integrals

We start with an example. Consider a solid (Ω) whose density ρ g/cm^3 is known and is nonhomogeneous; that is, the density differs at different points. Suppose we want to compute the mass m of the solid. A similar problem for a linear distribution of mass is known to be solved with the aid of ordinary integrals: if the mass is distributed along a line segment a, b with linear density α g/cm, then

$$m = \int_a^b \alpha(x)\,dx$$

The spatial case is studied in exactly the same way as the linear case. We partition (mentally) (Ω) into subvolumes $(\Delta\Omega_1)$, $(\Delta\Omega_2)$, ..., $(\Delta\Omega_n)$ and choose in each of them a point: M_1, M_2, ...,M_n (Fig. 35 depicts a solid with subvolumes numbered in arbitrary fashion). If the subvolumes are small enough, we can regard the density in each as being constant. Then the mass $m_{(\Delta\Omega_1)}$ of the first subvolume can be computed as the product of the density by the numerical value of the volume, or $\rho(M_1)\,\Delta\Omega_1$, the mass of the second subvolume is found in the same way, and so on. We then have

$$m_{(\Omega)} \approx \rho(M_1)\,\Delta\Omega_1 + \rho(M_2)\,\Delta\Omega_2 + \dots + \rho(M_n)\,\Delta\Omega_n = \sum_{k=1}^{n} \rho(M_k)\,\Delta\Omega_k$$

where $\Delta\Omega_k$ is to be understood as the numerical value of the volume $(\Delta\Omega_k)$ (in cm^3).

This is an approximate equation since the density inside each subvolume is not precisely constant. However, the finer the partition, the more exact it is, and in the limit, when the fineness of the partition becomes infinite, we obtain the exact equation

$$m_{(\Omega)} = \lim \sum_{k=1}^{n} \rho(M_k)\,\Delta\Omega_k \tag{34}$$

Arguing in similar fashion, we can conclude that if a charge with density σ is distributed in a solid (Ω), then the charge q can be computed from the formula

$$q_{(\Omega)} = \lim \sum_{k=1}^{n} \sigma(M_k)\,\Delta\Omega_k \tag{35}$$

with the designations having the same meaning.

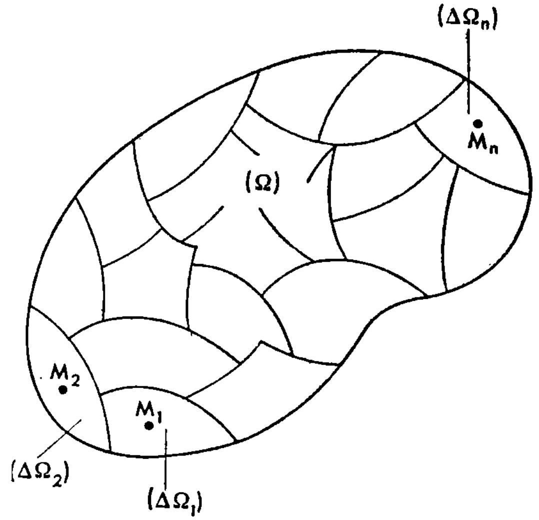


Fig. 35

The uniformity of formulas (34) and (35) serves as a basis for the general definition of the concept of a volume integral.

Suppose we have a finite portion of space or, as we more commonly say, a finite *region* (Ω) and in it (that is to say, at every point M) a function $u = f(M)$ that takes on finite values. (In our first example the region was a solid (Ω) and the function was its density, which at every point assumed a value.) To form the integral sum, we partition the region (Ω) into subregions $(\Delta\Omega_1)$, $(\Delta\Omega_2)$, ..., $(\Delta\Omega_n)$ and in each subregion we take an arbitrary point, M_1, M_2, ..., M_n, and then form the sum

$$\sum_{k=1}^{n} u_k\Delta\Omega_k = \sum_{k=1}^{n} f(M_k)\,\Delta\Omega_k \tag{36}$$

where $\Delta\Omega_k$ stands for the numerical value of a subregion $(\Delta\Omega_k)$. This sum has a high degree of arbitrariness: its value depends both on the mode of partition of the region (Ω) into subregions $(\Delta\Omega_k)$ and also on the choice of points M_k inside each subregion. However, with increasing fineness of partition of the region (Ω) the arbitrary nature hardly affects the sum and, in the limit, ceases to affect it altogether. The limit of the sum for an infinitely fine partition of the region (Ω) is called the (*volume*) *integral* of the function f over the region (Ω):

$$\int\limits_{(\Omega)} u\,d\Omega = \int\limits_{(\Omega)} f(M)\,d\Omega = \lim \sum_{k=1}^{n} f(M_k)\,\Delta\Omega_k \tag{37}$$

Now the formulas (34) and (35) may be written thus:

$$m_{(\Omega)} = \int\limits_{(\Omega)} \rho\,d\Omega, \qquad q_{(\Omega)} = \int\limits_{(\Omega)} \sigma\,d\Omega$$

The product $\rho\, d\Omega$, which corresponds to an infinitely small volume ("element of volume") $(d\Omega)$, is called the *element* (or *differential*) of mass and is denoted

$$dm = \rho\, d\Omega \tag{38}$$

where ρ is the density at any one of the points $(d\Omega)$. In setting up this expression, we can drop higher-order infinitesimals (which result from the fact that the density even in a small volume is variable), so that it is directly proportional to the volume $d\Omega$. Summing all elements (38) over the volume (Ω), we get the total mass

$$m_{(\Omega)} = \int\limits_{(\Omega)} dm = \int\limits_{(\Omega)} \rho\, d\Omega$$

The basic properties of volume integrals are similar to the corresponding properties of "simple" integrals.

The integral of a sum is equal to the sum of the integrals, that is,

$$\int\limits_{(\Omega)} (u_1 \pm u_2)\, d\Omega = \int\limits_{(\Omega)} u_1\, d\Omega \pm \int\limits_{(\Omega)} u_2\, d\Omega$$

A constant factor may be taken outside the integral sign:

$$\int\limits_{(\Omega)} Cu\, d\Omega = C \int\limits_{(\Omega)} u\, d\Omega \quad (C \text{ constant})$$

For any partition of the region (Ω) into parts, say (Ω_1) and (Ω_2), it will be true that

$$\int\limits_{(\Omega)} u\, d\Omega = \int\limits_{(\Omega_1)} u\, d\Omega + \int\limits_{(\Omega_2)} u\, d\Omega$$

The integral of unity is equal to the volume of the region of integration:

$$\int\limits_{(\Omega)} d\Omega = \Omega$$

If the variables at hand are dimensional, the dimensionality of the integral is equal to the product of the dimensions of the integrand into the dimensions of the volume:

$$\left[\int\limits_{(\Omega)} u\, d\Omega\right] = [u]\cdot[\Omega]$$

where the square brackets denote the dimensions of a quantity.

Example: $[\rho] = \text{g cm}^{-3}$; $\left[\int \rho\, d\Omega\right] = \text{g cm}^{-3}\ \text{cm}^3 = \text{g} = [m]$.

The *mean value* ("integral mean", "arithmetical mean") *of the function* $u(M)$ over the region (Ω) is introduced as a constant $\bar{u}$, the integral of which over the region (Ω) is equal to the integral of the function u over this region. Thus

$$\int_{(\Omega)} \bar{u}\, d\Omega = \int_{(\Omega)} u\, d\Omega$$

whence

$$\int_{(\Omega)} u\, d\Omega = \bar{u}\Omega \quad\text{and}\quad \bar{u} = \frac{1}{\Omega}\int_{(\Omega)} u\, d\Omega$$

As in the case of "simple" integrals, we derive $\min\limits_{(\Omega)} u \leqslant \bar{u} \leqslant \max\limits_{(\Omega)} u$.

At the beginning of this section we regarded the concept of density as straightforward and clear and expressed the mass in terms of the density with the aid of integration. Conversely, we can proceed from the mass and express the density in terms of the mass. To do this we have to form the ratio

$$\rho_{mean} = \frac{m_{(\Delta\Omega)}}{\Delta\Omega}$$

of the mass corresponding to a small region $(\Delta\Omega)$ to the volume of this region. This ratio is the mean density in the region $(\Delta\Omega)$. To obtain the density at a point M, we have to infinitely shrink $(\Delta\Omega)$ to this point and take the limit of the indicated ratio:

$$\rho(M) = \lim_{(\Delta\Omega)\to M} \frac{m_{(\Delta\Omega)}}{\Delta\Omega} = \frac{dm}{d\Omega} \tag{39}$$

This process is similar to differentiation.

If one takes into account the discrete molecular structure of matter, then in (39) the volume $(\Delta\Omega)$ cannot be infinetely shrunk to a point even in one's imagination. Instead of this formula we have to write

$$\rho(M) = \frac{m_{(\Delta\Omega)}}{\Delta\Omega}$$

where $(\Delta\Omega)$ is a "practically infinitesimal region" containing the point M, that is, a region sufficiently small relative to macroscopic bodies and at the same time sufficiently large if compared with molecular dimensions. Here we pass, as it were, from the discrete picture of a material body to its continuous model, the density of which is obtained by averaging, that is, via a computation of the mean density of the original picture over volumes of the indicated "practically infinitesimal" dimensions.

From now on, when we consider continuous media, we will abstract from the discrete structure of matter and assume that such a transition to the continuous model of the medium and its density has already been carried out.

It may turn out that one of the dimensions of a portion of space occupied by a mass or charge is substantially less than the other two dimensions (say the thickness is much less than the length and width). We then assume that the mass or charge is distributed over the surface. Similarly, if two dimensions of this portion are appreciably less than the third, then we assume that the mass or charge is distributed along a line. In that case, formulas (34) and (35) remain valid if by density we have in mind the surface density (referred to a unit surface) or linear density (referred to a unit length) and by $\Delta\Omega_k$ the area or length of the subregion $(\Delta\Omega_k)$, respectively. In the general case we say that $\Delta\Omega_k$ is the *measure* of the region $(\Delta\Omega_k)$, and is to be understood as the volume, area, or length, depending on whether we are discussing volume, surface or linear regions.

The definition of an integral over a surface (plane or curved) and also of an integral along a line is given in the same way as the volume integral, that is, via formula (37). Naturally, in place of the volume of a subportion one has to take the area or the length. Volume and surface integrals are called *multiple integrals* for reasons that will emerge later, surface integrals being termed *double integrals*, and volume integrals, *triple integrals*. The integral along a line is called a *line integral* (actually it is a curvilinear integral). The properties of all these integrals are completely analogous and will be widely used in this text in Chs. 10 and 11, where we discuss the theory of vector fields.

Multiple integrals are calculated with the aid of ordinary integrals. Let us consider a double integral, for example

$$I = \int\limits_{(\Omega)} u \, d\Omega$$

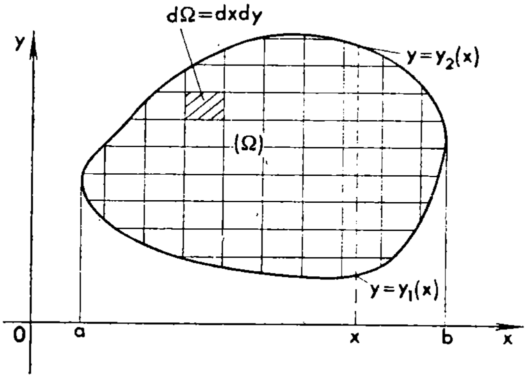
taken over the region (Ω) in a plane. To go over to ordinary integrals, we have to introduce coordinates, say the Cartesian coordinates x, y (Fig. 36). Then u may be regarded as a function of x, y and we have

$$I = \int\limits_{(\Omega)} u(x, y) \, \mathrm{d}\Omega$$

Since in defining an integral we can partition the region (Ω) in arbitrary fashion, we can take a partition by the straight lines $x =$ constant and $y =$ constant, which fits integration in terms of Cartesian coordinates. Then all subregions $(\Delta\Omega)$ will be in the form of rectangles with sides dx and dy, so that $\mathrm{d}\Omega = dx\, dy$, or

$$I = \iint\limits_{(\Omega)} u(x, y) \, dx \, dy$$

Fig. 36



Here we have two integral signs since the summation is naturally carried out in two stages: for instance, we first sum over all rectangles of a column corresponding to a certain fixed x and then we sum the results over all x. The first summation yields

$$\int_{y_1(x)}^{y_2(x)} u(x, y)\, dx\, dy = \left(\int_{y_1(x)}^{y_2(x)} u(x, y)\, dy\right) dx$$

where $y = y_1(x)$ and $y = y_2(x)$ (see Fig. 36) are the coordinates of the lower and upper points of the region (Ω) for a given x.
After the second summation we finally get

$$I = \int_a^b \left(\int_{y_1(x)}^{y_2(x)} u(x, y)\, dy\right) dx$$

or, as it is more common to write,

$$I = \int_a^b dx \int_{y_1(x)}^{y_2(x)} u(x, y)\, dy \tag{40}$$

Thus, computing a double integral reduces to twofold ordinary integration. First the *inner integral* is evaluated:

$$\int_{y_1(x)}^{y_2(x)} u(x, y)\, dy$$

After integration and substitution of the limits we get a result that, generally, depends on x (since x was held constant in the process of

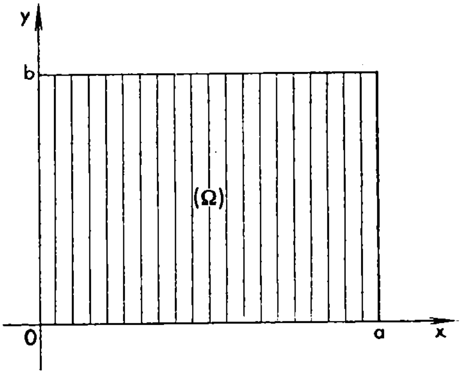


Fig. 37

integration), that is, we have a certain function $f(x)$ that must be integrated with respect to x to yield the final result:

$$I = \int_a^b f(x)\,dx$$

The integration could have been performed in the reverse order: first (*inner*) with respect to x, and then (*outer*) with respect to y. The result is the same, although in a practical evaluation one method may prove to be more involved than the other.

It is simplest to set the limits of integration if the region (Ω) is a rectangle with sides parallel to the coordinate axes: then not only the limits of the outer integral but also those of the inner integral will be constant. And if, besides, the integrand is a product of a function dependent on x alone by a function dependent on y alone, then the entire double integral can be decomposed into a product of two ordinary (single) integrals:

$$\int_a^b dx \int_c^d [f(x)\ \varphi(y)]\,dy = \int_a^b f(x)\,dx \left(\int_c^d \varphi(y)\,dy\right)$$
$$= \int_c^d \varphi(y)\,dy \cdot \int_a^b f(x)\,dx \qquad (41)$$

By way of an illustration let us evaluate the mass of a flat lamina shown in Fig. 37, the surface density of which varies by the law

$$\sigma = \rho_0(h + \alpha x + \beta y) \qquad (42)$$

where α, β are certain constant coefficients. This law is valid if the lamina is homogeneous but bounded from below by the xy-plane and above by a plane with equation $z = h + \alpha x + \beta y$ oblique to the first plane. Then the surface density at any point (x, y) is obtained by multiplying the density ρ_0 of the material of the lamina by the altitude z of the lamina at the given point, which means we arrive at formula (42). (At the same time we see that the mass distributed over the plane can be obtained merely by projecting onto this plane the mass distributed over the volume.)

By formula (40) we get

$$m = \int_{(\Omega)} \sigma \, d\Omega = \iint_{(\Omega)} \sigma(x, y) \, dx \, dy = \int_0^a dx \int_0^b \rho_0(h + \alpha x + \beta y) \, dy$$

The inner integration is performed with respect to y for x fixed, that is to say, along the parallel strips shown in Fig. 37. Evaluate the inner integral:

$$\tau = \int_0^b \rho_0(h + \alpha x + \beta y) \, dy = \rho_0 \left[(h + \alpha x) \, y + \frac{\beta y^2}{2} \right]_{y=0}^{b}$$

$$= \rho_0 \left[(h + \alpha x) \, b + \frac{\beta b^2}{2} \right]$$

It now remains to take this result which depends on x and integrate it with respect to x:

$$m = \int_0^a \tau(x) \, dx = \int_0^a \rho_0 \left[(h + \alpha x) \, b + \frac{\beta b^2}{2} \right] dx$$

$$= \rho_0 \left(hbx + \alpha b \frac{x^2}{2} + \frac{\beta b^2}{2} x \right) \Bigg|_{x=0}^{a} = \rho_0 \left(abh + \frac{\alpha a^2 b + \beta a b^2}{2} \right)$$

The physical meaning of these computations consists in our projecting, as it were, the lamina on the x-axis to obtain the mass distributed along a line, namely a material segment with linear density

$$\tau(x) = \int_a^b \sigma(x, y) \, dy$$

(this is the inner integral). The meaning of linear density is immediately apparent here: it is the mass of the lamina per unit length, so that the mass of the lamina on the interval from x to $x + dx$ is $dm = \tau(x) \, dx$. Integrating this result over the segment of the x-axis, we get the mass of the lamina. (Verify the expression just found for the mass by carrying out the integration in reverse order, first with respect to x and then with respect to y, and also by decomposing (42) into three terms and applying (41) to each one of the appropriate integrals.)

Now suppose the lamina is of triangular shape (Fig. 38) and the surface density varies by the same law (42), with inner integration being carried out with respect to y, for x fixed, from $y = 0$ to $y = bx/a$ (Fig. 38). The computations proceed as follows:

$$m = \int_0^a dx \int_0^{bx/a} \rho_0(h + \alpha x + \beta y)\,dy$$

$$= \int_0^a \rho_0\left(h\frac{b}{a}x + \alpha\frac{b}{a}x^2 + \beta\frac{b^2}{a^2}\frac{x^2}{2}\right)dx$$

$$= \rho_0\left(\frac{ab}{2}h + \alpha\frac{a^2b}{3} + \beta\frac{ab^2}{6}\right)$$

As a check, perform the integration in the reverse order. If the inner integration is performed with respect to x with y constant, then the limits of integration are (Fig. 39) $x = ay/b$ and $x = a$, whence

$$m = \int_0^b dy \int_{ay/b}^a \rho_0(h + \alpha x + \beta y)\,dx$$

$$= \int_0^b \rho_0\left[h\left(a - \frac{a}{b}y\right) + \frac{\alpha}{2}\left(a^2 - \frac{a^2}{b^2}y^2\right) + \beta y\left(a - \frac{a}{b}y\right)\right]dy$$

$$= \rho_0\left(abh - \frac{a}{b}\frac{b^2}{2}h + \frac{\alpha a^2}{2}b - \frac{\alpha a^2}{2b^2}\frac{b^3}{3} + \beta a\frac{b^2}{2} - \frac{\beta a}{b}\frac{b^3}{3}\right)$$

$$= \rho_0\left(\frac{ab}{2}h + \frac{\alpha a^2 b}{3} + \frac{\beta ab^2}{6}\right)$$

The result is the same.

Other types of integrals are evaluated in the same manner. One always has to express $d\Omega$ in terms of the differentials of the coordinates, set the limits in accord with these coordinates, and then evaluate the resulting integrals by the usual rules of integral calculus. Integrals over areas are double, integrals over volumes are triple (with three integral signs).

If after passing to the coordinates the integrand turns out to be independent of one of the coordinates, then the integration with respect to that coordinate is performed in trivial fashion, and we straightway pass from a double integral to a single integral, and from a triple integral to a double or even to a single integral. For instance, the area of the figure (Ω) depicted in Fig. 36 may be found as a double integral of unity (see the properties of an integral):

$$\Omega = \int_{(\Omega)} d\Omega = \iint_{(\Omega)} dx\,dy$$

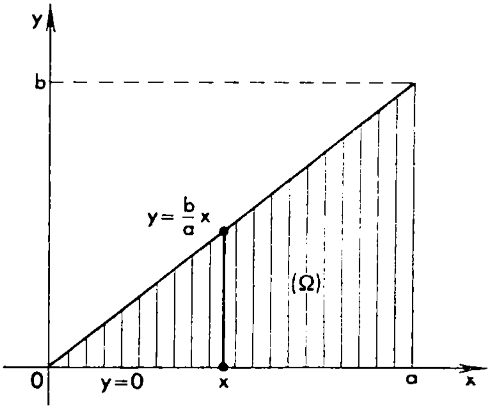


Fig. 38

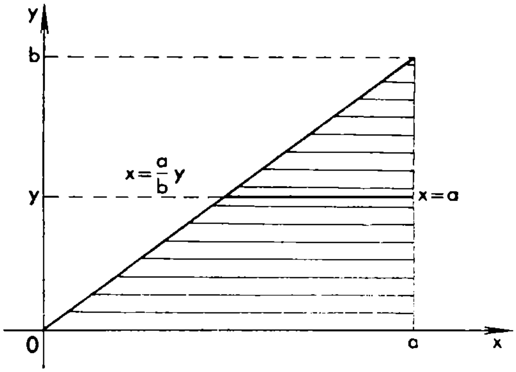


Fig. 39

To pass to a single integral we set the limits

$$\Omega = \int_a^b dx \int_{y_1(x)}^{y_2(x)} dy = \int_a^b [y_2(x) - y_1(x)]\, dx$$

We have an obvious formula if we recall the geometric meaning of a definite integral. Similarly, when computing a volume we can pass from a triple integral straightway to a double integral. Actually, that is what we did when we calculated the mass of the lamina.

To illustrate the application of double integrals let us evaluate the integral

$$K = \int_{-\infty}^{\infty} e^{-x^2}\, dx$$

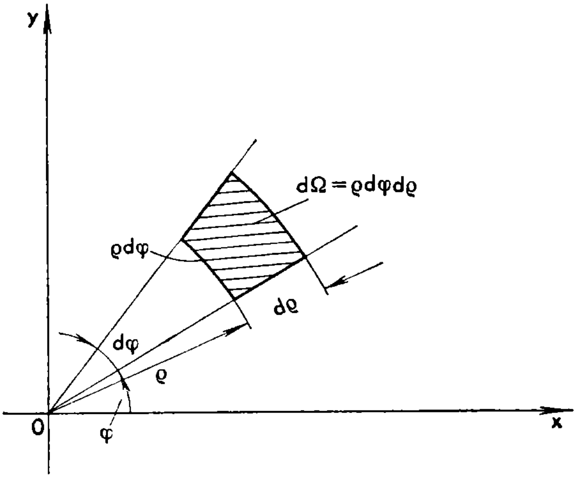


Fig. 40

that was mentioned in Sec. 3.2. To do this we have to consider the auxiliary integral

$$I = \int\limits_{(\Omega)} e^{-x^2-y^2}\, d\Omega$$

where (Ω) is the complete xy-plane. Setting the limits in terms of the coordinates x and y when $d\Omega = dx\, dy$, we get (taking into account formula (41))

$$I = \int\limits_{-\infty}^{\infty} dx \int\limits_{-\infty}^{\infty} e^{-x^2-y^2}\, dy = \int\limits_{-\infty}^{\infty} dx \int\limits_{-\infty}^{\infty} e^{-x^2} e^{-y^2}\, dy$$

$$= \int\limits_{-\infty}^{\infty} e^{-x^2}\, dx \int\limits_{-\infty}^{\infty} e^{-y^2}\, dy = K^2$$

On the other hand, that same integral I may be computed with the aid of the polar coordinates ρ, φ. Then it is more natural to partition the plane (Ω) into subregions by means of circles $\rho =$ constant and rays $\varphi =$ constant. The area of the resulting subregions (one is shown in Fig. 40) is equal to $d\Omega = \rho\, d\varphi\, d\rho$. Since $x^2 + y^2 = \rho^2$, we obtain (think about how the limits are to be set)

$$I = \iint\limits_{(\Omega)} e^{-\rho^2} \rho\, d\varphi\, d\rho = \int\limits_{0}^{2\pi} d\varphi \int\limits_{0}^{\infty} e^{-\rho^2} \rho\, d\rho$$

$$= \int\limits_{0}^{2\pi} d\varphi \cdot \int\limits_{0}^{\infty} e^{-\rho^2} \rho\, d\rho = \varphi \Big|_{0}^{2\pi} \cdot \frac{e^{-\rho^2}}{-2} \Big|_{0}^{\infty} = 2\pi \cdot \frac{1}{2} = \pi$$

Equating the results obtained we get

$$K = \sqrt{\pi}, \text{ i.e. } \int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$$

Exercises

1. Compute the integral of the function $x \sin xy$ over the rectangle $0 \leqslant x \leqslant 1$, $0 \leqslant y \leqslant \pi$.
2. Compute the mean value of the function $u = e^{xy}$ in the square $0 \leqslant x \leqslant 1$, $0 \leqslant y \leqslant 1$.
   Hint. For the second integration, expand the integrand in a Taylor series.
3. Evaluate the integral $I = \int_0^1 dx \int_x^1 e^{y^2}\,dy$.
   Hint. Reverse the order of integration.

4.8 Multidimensional space and number of degrees of freedom

In the case of a function of two variables we considered, in Sec. 4.2, a "plane of arguments". We can do the same for a function of three variables and consider a "space of arguments". These notions are very pictorial and it is therefore desirable to retain the conception of a space of arguments for the case of functions of any number of independent variables exceeding three. This is done in the following manner. Suppose we have a function of four variables $u = f(x, y, z, t)$. Then we agree to say that every set of values x, y, z, t defines a "point" in the "four-dimensional space of the arguments x, y, z, t". There is no way to picture such a point or such a space geometrically. Strictly speaking, the point is nothing but the set of values x, y, z, t and the space is merely the collection of all such sets. For instance, the set (or quadruple) $(-3, 0, 2, 1.3)$ is one such point and the quadruple $(0, 0, 0, 0)$ is another point (the coordinate origin, in our case), and so on. We can now say that the function f is defined throughout the four-dimensional space x, y, z, t or in a portion (region) of it.

For the function $z = f(x, y)$ of two variables, the "argument space" is the xy-plane, while the space of the arguments and function in which the graph of the function is located is our ordinary three-dimensional space x, y, z. In this case the graph is a two-dimensional surface in three-dimensional space. Reasoning in this manner, we see that the "graph" of the function $u = f(x, y, z, t)$ requires a five-dimensional space of arguments and function x, y, z, t, u. In order to find the points of this graph we have to assign to x, y, z, t arbitrary values and then find the corresponding values of u. For example, verify that the graph of the function $u = xz - 2y^2t$ passes through the points $(1, 1, 2, 0, 2)$, $(-1, 2, 0, -2, 16)$, and so on. Thus, the relationship $u = f(x, y, z, t)$ defines a four-dimensional surface in five-dimensional space.

Directly connected with this terminology is the concept of the *number of degrees of freedom*. We know that the position of a point in (ordinary) space may be described by the Cartesian coordinates x, y, z. There are other coordinate systems, but what they have in common is that the position of a point in space is determined by three coordinates (whereas a point in a plane is determined by two coordinates, and a point on a line by a single coordinate).

This fact can be expressed as follows: there are three degrees of freedom when choosing a point in physical space, or, what is the same thing, when a point is in motion in space. There are two degrees of freedom for choosing a point in a plane or on any surface, and one degree of freedom for a point on a line. In other words, space is three-dimensional, whereas surfaces are two-dimensional and lines are one-dimensional.

In the general case, the concept of the number of degrees of freedom is introduced as follows. Let there be a collection of entities (in the foregoing example this was a collection of points in space), each of which may be described by indicating the numerical values of certain continuous parameters (they were coordinates in the above case). Let these parameters be:

(1) independent, that is, capable of assuming arbitrary values: for instance, if we fix all parameters except one, then this one can possibly be made to vary in an arbitrary manner within certain limits;

(2) essential, which means that the object under study actually changes for any change in the parameters. Then if there are k such parameters, we say that there are k degrees of freedom for choosing an entity from the given collection. The collection itself is termed a (generalized) *k-dimensional space* or a *k-dimensional manifold*. The parameters are then termed (generalized) *coordinates* in the given space. As in the case of ordinary coordinates in ordinary space, they may be chosen in a variety of ways, one way being more convenient than another for any given investigation. In this way, multidimensional space is interpreted in a very concrete manner.

For example, in physics one constantly has to do with collections of "events", each one of which is described by answering the questions "where"? and "when"? To the first query we respond with the Cartesian coordinates x, y, z, to the second, with the instant of time t. These parameters are independent (they may be altered in arbitrary fashion) and they are essential (any variation in them gives rise to a new event). Thus, the space of events is a four-dimensional space where the generalized coordinates may be x, y, z, t.

Another illustration. Let us consider a sequence (system) of gears, each of which is meshed with the preceding one. There is only one degree of freedom here, and for the generalized coordinate we can take the angle of rotation of the first gear wheel, since any specification of this angle completely defines the position of all the other gear wheels. (How many degrees of freedom would there be if the gear wheels were not engaged?)

Let us determine the number of degrees of freedom that a line segment of a given length l has when in motion over a plane. Each such line segment is fully determined by the coordinates (x_1, y_1) and (x_2, y_2) of its endpoints. These coordinates may be taken as the parameters defining the position of the segment. These parameters are clearly essential but they are not independent, for they are connected by the relation

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

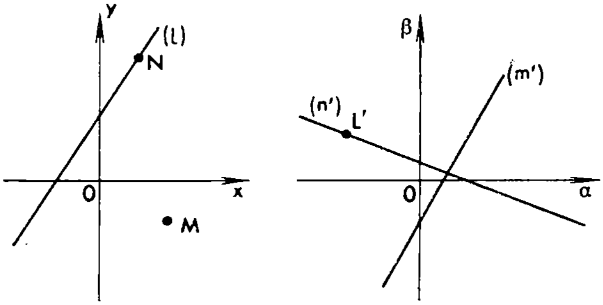
(why?). Thus, only three parameters can be considered independent, while the fourth parameter is expressed in terms of these via the given relation, which means that a line segment of given length has three degrees of freedom when in motion on a plane.

In the general case, if there are n parameters and they are essential but are connected by m independent equations (equations such that none follows from the others), then $n - m$ parameters can be assumed independent and the remaining m will be expressed in terms of them; in other words, we have $n - m$ degrees of freedom.

Now finally let us count the number of degrees of freedom in choosing an infinite straight line in a plane. We can reason as follows: choose two arbitrary points A and B in the plane (each has two coordinates) and draw through them a straight line P, which will thus be defined by four parameters. Since these parameters are independent, there would seem to be four degrees of freedom. This reasoning, however, is faulty because under any change of the parameters (coordinates) the points A and B will change but the straight line P may still remain unchanged. Hence the requirement that the parameters be essential is not fulfilled. Since the line P does not change if A slides along it (one degree of freedom) or B slides along it (another degree of freedom), it follows that we have two extra degrees of freedom, and the actual number of degrees of freedom is $4 - 2 = 2$. For the independent and essential parameters we can take, say, the coefficients k and b in the equation $y = kx + b$; true, straight lines parallel to the y-axis are not described by such equations, but these special cases cannot have any effect on counting the number of degrees of freedom.

The number of degrees of freedom of an infinite line in a plane has turned out to have the same number of degrees of freedom as a point in the plane, and so we can associate straight lines in a plane with points in that plane or some other plane. This is most

Fig. 41



conveniently done as follows: the last equation is divided by b and written as

$$\alpha x + \beta y = 1 \tag{43}$$

Then a point with coordinates (α, β) is associated with the straight line having this equation. It is then natural to consider two planes: with coordinates x, y and with coordinates α, β (Fig. 41). Here, the straight line (l) $y = 2x + 1$, i.e. $-2x + y = 1$ is associated with the point L' of the (α, β)-plane having coordinates $(-2, 1)$. This same formula (43) associates with the point (x, y) of the first plane a straight line with equation (43) in the second plane. For example, the point $M(2, -1)$ is associated with the straight line (m') having the equation $2\alpha - \beta = 1$, the point $N(1, 3)$ is associated with the straight line (n') with equation $\alpha + 3\beta = 1$ (see Fig. 41). It can readily be proved in the general form as well that if a straight line (l) in the first plane passes through a point N, then in the second plane the straight line (n') corresponding to N will pass through the point L' that corresponds to the line (l). We conclude that any assertion referring to an arbitrary combination of points and lines in a plane implies the "dual" assertion, in which the lines are replaced by points and the points by lines. There is a special branch of geometry in which such assertions and dual relationships are studied (it is called projective geometry). Note that the equations of the lines passing through the coordinate origin cannot be written as (43); in other words, such lines are not associated in the indicated manner with any points, there is no straight line that corresponds to the origin. For this reason, in projective geometry one introduces so-called "points at infinity" (ideal points) and "a line at infinity" (ideal line). Then this relationship has no exceptions. We will not dwell further on this subject here.

Exercise

How many degrees of freedom have the following objects when in motion in space:

(a) A line segment of given length?

(b) A triangle with given sides (a rigid triangle)?
(c) A rigid body with a fixed point?
(d) A free rigid body?

ANSWERS AND SOLUTIONS

Sec. 4.1

1. $\dfrac{\partial z}{\partial x} = 2x$ and therefore $\left.\dfrac{\partial z}{\partial x}\right|_{\substack{x=1\\y=1}} = 2,\ \left.\dfrac{\partial z}{\partial y}\right|_{\substack{x=2\\y=0.5}} = 1.$

2. $\dfrac{\partial z}{\partial x} = -2xe^{-(x^2+y^2)},\ \dfrac{\partial z}{\partial y} = -2ye^{-(x^2+y^2)}.$

3. $\dfrac{\partial z}{\partial x} = e^y + ye^x,\ \dfrac{\partial z}{\partial y} = xe^y + e^x.$

4. $\dfrac{\partial z}{\partial x} = \sin y,\ \dfrac{\partial z}{\partial y} = x \cos y.$

5. $\dfrac{\partial z}{\partial x} = y \cos (xy),\ \dfrac{\partial z}{\partial y} = x \cos (xy).$

6. $\dfrac{\partial z}{\partial x} = \dfrac{x}{\sqrt{x^2 + y^2}},\ \dfrac{\partial z}{\partial y} = \dfrac{y}{\sqrt{x^2 + y^2}}.$

7. $\dfrac{\partial z}{\partial t} = \dfrac{\partial z}{\partial x}\dfrac{dx}{dt} + \dfrac{\partial z}{\partial y}\dfrac{dy}{dt} = 2x\dfrac{dx}{dt} - 2y\dfrac{dy}{dt}.$

Noting that $\dfrac{dx}{dt} = 1 - \dfrac{1}{t^2}$ and $\dfrac{dy}{dt} = 1 + \dfrac{1}{2\sqrt{t}}$, we obtain

$$\frac{dz}{dt} = 2\left(t + \frac{1}{t}\right)\left(1 - \frac{1}{t^2}\right) - 2(t + \sqrt{t})\left(1 + \frac{1}{2\sqrt{t}}\right)$$
$$= -\left(\frac{2}{t^3} + 3\sqrt{t} + 1\right).$$

8. $\dfrac{dz}{dt} = e^{x-y}\cdot(\cos t - 2t).$

9. $\dfrac{dz}{dt} = 6(x^2 + y^2)\,t - 6xye^{-t}.$

10. $\dfrac{dz}{dx} = \dfrac{1 + 2xe^y}{x + e^y}.$

Sec. 4.2

1. A family of straight lines passing through the origin.
2. $z = x^2 - c^2$, a family of parabolas (Fig. 42).
3. A family of hyperbolas located in the upper half-plane (Fig. 43).
4. A family of hyperbolas (Fig. 43).

Fig. 42

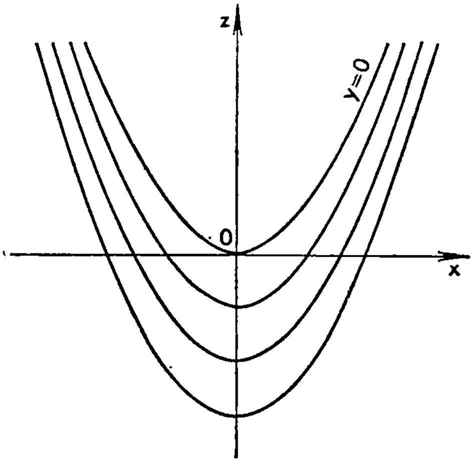
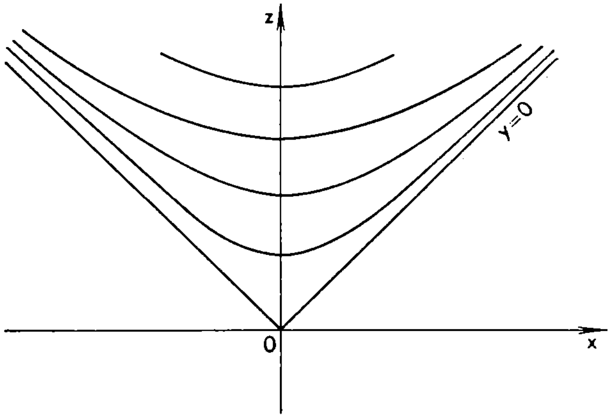


Fig. 43



Sec. 4.3

1. (a) $\dfrac{dy}{dx}=\dfrac{(2t+1)\,dt}{\dfrac{1}{2}\,dt}=4t+2,\ \dfrac{d^2y}{dx^2}=\dfrac{d}{dx}\left(\dfrac{dy}{dx}\right)=\dfrac{d(4t+2)}{dx}$

$$=\frac{d(4t+2)}{dt}\cdot\frac{dt}{dx}=4\,\frac{dt}{dx}$$

Note that $\dfrac{dx}{dt}=\dfrac{1}{2}$; therefore $\dfrac{dt}{dx}=2$ so that $\dfrac{d^2y}{dx^2}=8.$

(b) $\dfrac{dy}{dx}=-\cot t,\ \dfrac{d^2y}{dx^2}=\dfrac{1}{6\sin^4 t\cos t}\cdot$

(c) $\dfrac{dy}{dx}=\tan\ t,\ \dfrac{d^2y}{dx^2}=\dfrac{1}{t\cos^3 t}\cdot$

2. $\frac{dy}{dx} = -- 4 \sin t$; the value of t corresponding to the value $x = \frac{1}{2}$ is obtained from the equation $\frac{1}{2} = \sin t$, whence $t = \frac{\pi}{6}$, $\left.\frac{dy}{dx}\right|_{x=\frac{1}{2}} = -4 \cdot \frac{1}{2} = -2$. The other value $t = 2\pi - \frac{\pi}{6}$ yields $\frac{dy}{dx} = 2$.

3. $\frac{\partial x}{\partial z} = \frac{1}{3x^2 + y,}$, $\frac{\partial x}{\partial y} = -\frac{3y^2 + x}{3x^2 + y}$. Setting $y = 0$, $z = 1$ in the formula $y = x^3 + y^3 + xy$, we get $x^3 = 1$, whence $x = 1$. For this reason,

$$\left.\frac{\partial x}{\partial z}\right|_{\substack{y=0\\ z=1}} = \frac{1}{3}, \quad \left.\frac{\partial x}{\partial y}\right|_{\substack{y=0\\ z=1}} = -\frac{1}{3}.$$

4. $\frac{\partial x}{\partial y} = \frac{1}{5x^4 + y^2}$, $\frac{\partial x}{\partial y} = -\frac{2xy + 5y^4}{5x^4 + y^2}$.

5. $\frac{dy}{dx} = \frac{2 - x}{y - 5}$.

6. $\frac{dy}{dx} = \frac{2xy^2 - y^4 - 4x^3y}{x^4 - 4xy^3 - 2x^2y}$.

Sec. 4.4

$S = 4.5$ milliampere/volt, $R = 6.7$ kohms, $\mu = 30$.

Sec. 4.5

Since $l = \frac{v_0^2}{g}$, in the case that interests us $l = 1020$.

The equation of the envelope is of the form $y = 510 - \frac{x^2}{2040}$.

For $x = 500$ m we get $y = 390$ m. Hence, a target can be hit at 300 metres altitude but not at 500 metres altitude.

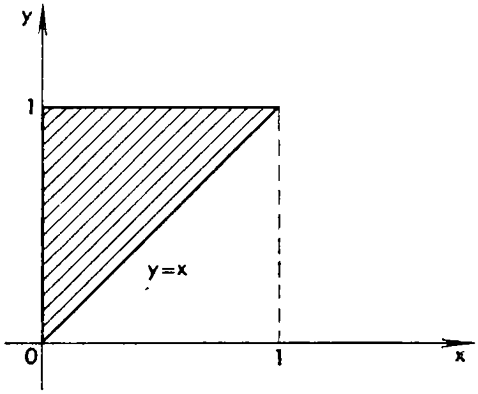
Sec. 4.6

The function has: (a) a minimum at the point $x = 0$, $y = -2$; (b) a minimax at the point $x = \frac{2}{3}$, $y = -\frac{4}{9}$; (c) a minimax at $x = 0$, $y = 0$, $z = 1$.

Sec. 4.7

1. $$\int_0^1 dx \int_0^\pi \sin xy\, dy = \int_0^1 dx(-\cos xy)\Big|_{y=0}^{\pi} = \int_0^1 (1 - \cos \pi x)\, dx$$
$$= \left(x - \frac{\sin \pi x}{\pi}\right)\Big|_0^1 = 1.$$

Fig. 44



Note that if the integration were reversed (first with respect to x and then with respect to y), we would have to apply integration by parts.

2. Since the area of a square is 1, it follows that

$$\bar{u} = \int_0^1 dx \int_0^1 e^{xy}\,dy = \int_0^1 dx\left(\frac{e^{xy}}{x}\right)\Bigg|_{y=0}^{1} = \int_0^1 \frac{e^x - 1}{x}\,dx$$

$$= \int_0^1 \frac{1}{x}\left(1 + \frac{x}{1} + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} + \dots - 1\right)dx$$

$$= \int_0^1 \left(\frac{1}{1} + \frac{x}{2} + \frac{x^2}{6} + \frac{x^3}{24} + \frac{x^4}{120} + \dots\right)dx$$

$$= 1 + \frac{1}{2\cdot 2} + \frac{1}{3\cdot 6} + \frac{1}{4\cdot 24} + \frac{1}{5\cdot 120} + \dots = 1.318.$$

3. In the inner integral the integration with respect to y is carried out from $y = x$ to $y = 1$, and so the entire double integral is extended over the triangle shown in Fig. 44. Reversing the order of integration, we get

$$I = \int_0^1 dy \int_0^y e^{y^2}\,dx = \int_0^1 e^{y^2} y\,dy = \frac{e^{y^2}}{2}\Bigg|_0^1 = \frac{e-1}{2} = 0.859$$

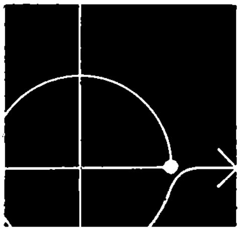
It is interesting to note that in this example integration cannot be carried out in the original order with the aid of elementary functions, but reversing the order of integration permits this to be done.

Sec. 4.8

(a) 5, (b) 6, (c) 3, (d) 6.

Chapter 5

FUNCTIONS OF A COMPLEX VARIABLE



5.1 Basic properties of complex numbers

In algebra we have what is called the *imaginary unit*, which is denoted by i and is defined by the condition $i^2 = -1$.

Quantities of the form $x + iy$ (where x and y are real numbers) are called *complex numbers*, and are obtained, for example, in the solution of algebraic equations. The imaginary unit itself is clearly the square root of the quadratic equation $x^2 = -1$.

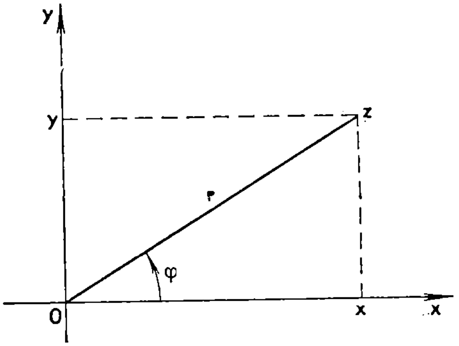
If we confine ourselves to real roots, then a quadratic equation has either two roots or none, depending on its coefficients. However, if we consider all the roots, real and imaginary, then a quadratic equation always has two roots. In the same way, a third-degree equation always has three roots, and so forth. Thus, the study of complex numbers enables us to establish a general theorem on the number of roots of an algebraic equation. In the process we perform algebraic operations with complex quantities. If we examine more complicated functions of complex quantities (raising to a complex power, for instance), we get many important and elegant results. A great variety of relationships between real quantities are conveniently obtained through the use of complex numbers.* It is precisely for this reason that complex quantities are so important to physics and other natural sciences, despite the fact that all measurements and the results of all experiments are expressed as *real* numbers.

Let us recall some of the properties of complex numbers from the school course of algebra.

A complex number $z = x + iy$ can be represented as a point in a plane (Fig. 45), where x is laid off on the axis of abscissas (this is the *real part* of the number), and y on the axis of ordinates (for the *imaginary part* of the number z; sometimes the product iy is also termed the imaginary part of the complex number; this might appear

* According to the celebrated French mathematician Jacques Hadamard (1865-1963), "The shortest path between two truths in the real domain passes through the complex domain."

Fig. 45



to be more natural but it is more convenient to adhere to the definition given above). To every z there corresponds a definite point in the plane. Instead of a point z we can speak of a vector* z, which is a directed line segment with origin at the origin of the coordinate system, and with terminus at the point z. The position of a point in a plane may be described by its distance r from the coordinate origin and the angle φ between the vector z and the x-axis. We can thus speak of the length of the vector z (also called the *absolute value*, or *modulus*, of the complex number, which is denoted by $|z|$) and we can speak of the direction of the vector z (it is indicated by the angle φ). The angle φ is termed the *argument* (*amplitude* or *phase*) of the complex number. It is reckoned from the positive x-axis counterclockwise. For a positive real number, $\varphi = 0$, for a pure imaginary number (the number $2i$, for example), $\varphi = \frac{\pi}{2}$ or (say for the number $-2i$) $\varphi = \frac{3\pi}{2}\cdot$

As can be seen in Fig. 45, $x = r\cos\varphi$, $y = r\sin\varphi$, so that we can express z in terms of r and φ:

$$z = x + iy = r(\cos\varphi + i\sin\varphi)$$

The notation $z = r(\cos\varphi + i\sin\varphi)$ is called the *trigonometric* (or *modulus-argument*) *form of a complex number.* Also observe that if we know x and y, it is simple (see Fig. 45) to find r and φ from the formulas

$$r = \sqrt{x^2+y^2}, \quad \sin\varphi = \frac{y}{\sqrt{x^2+y^2}}, \quad \cos\varphi = \frac{x}{\sqrt{x^2+y^2}}$$

Algebraic operations with complex numbers are performed by the ordinary rules of algebra with the convention that $i^2 = -1$. It

* See Ch. 9 for more on the theory of vectors. Here, any basic school course of physics will suffice as an introduction to vectors.

Fig. 46

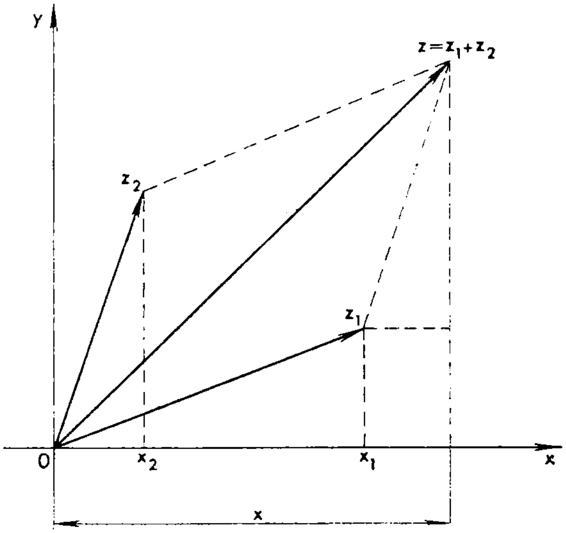
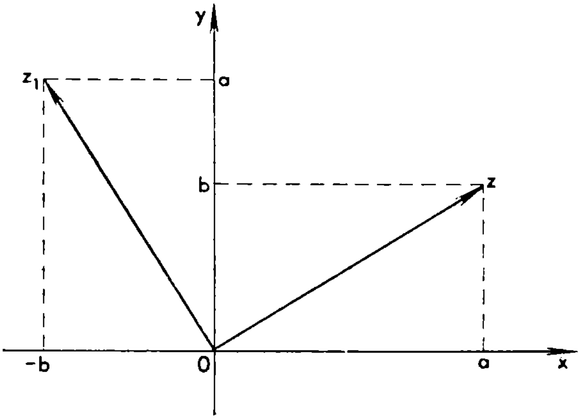


Fig. 47



is useful to take a good look at the geometric picture of algebraic operations. For the addition of two numbers $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ we have

$$z = z_1 + z_2 = (x_1 + x_2) + i(y_1 + y_2)$$

It is readily seen that in the plane the number z is represented by a vector obtained from the vectors z_1 and z_2 by adding via the parallelogram rule, that is, in the same way, say, as the result of two forces (Fig. 46).

The product of a complex number z by a positive real number k, $z_1 = kz$, is a vector in the same direction as z but with modulus (length) equal to kr, where r is the modulus of z. But if k is a negative real number, then the vector kz has modulus $|k|r$, but with direction opposite that of z. It is also easy to construct the product z_1 of the complex number $z = x + iy$ by the imaginary unit i. We have $z_1 = = i(x + iy) = -y + ix$. From the drawing (see Fig. 47 where for pictorialness we have used a and b instead of x and y) it is easy to see that the vector z_1 is perpendicular to z. It has been rotated counterclockwise from z through a right angle. The modulus of z_1 is equal to that of z.

The product of two complex numbers is found from the simple formula

$$(x_1 + iy_1)(x_2 + iy_2) = x_1x_2 + x_1iy_2 + iy_1x_2 + i^2y_1y_2$$
$$= (x_1x_2 - y_1y_2) + i(x_1y_2 + y_1x_2)$$

In algebra there is a simple proof (carry it out!) which makes use of the formulas for the sine and cosine of a sum. It states that if

$$z_1 = r_1(\cos\varphi_1 + i\sin\varphi_1), \quad z_2 = r_2(\cos\varphi_2 + i\sin\varphi_2)$$

then

$$z = z_1z_2 = r_1r_2[\cos(\varphi_1 + \varphi_2) + i\sin(\varphi_1 + \varphi_2)]$$

When multiplying complex numbers, multiply their moduli and add the arguments. We will obtain the same result in Sec. 5.3 in a different way.

Note in conclusion that complex numbers cannot be connected by an inequality sign: one complex number cannot be greater or less than another. It might appear to be natural to say that $z_1 > z_2$ if $|z_1| > |z_2|$, that is, to compare complex numbers via their moduli. But then we would have to write, say, $-3 > -1$ for real numbers, but this is a contradiction.

Exercise

Write the following numbers in trigonometric form: $1 - i$, $3 + 4i$, $-2i$, -3, 1, 0.

5.2 Conjugate complex numbers

Two complex numbers that differ only in the sign of the imaginary part are called *conjugate* complex numbers. The conjugate of z will be denoted by z^*. If $z = x + iy$, then $z^* = x - iy$. In particular, for real numbers, and only for them, it is true that $A^* = A$, since in this case the imaginary part is zero. For pure imaginaries (that is, numbers with real part zero), and only for them, we have $A^* = -A$.

Suppose we have two complex numbers $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$. Adding them, we get

$$z_1 + z_2 = (x_1 + x_2) + i(y_1 + y_2)$$

Now form the sum of the conjugates

$$z_1^* + z_2^* = (x_1 + x_2) - i(y_1 + y_2)$$

We thus have a complex number conjugate to the sum $z_1 + z_2$.

Thus, if $z_1 + z_2 = w$, then $z_1^* + z_2^* = w^*$; that is, if the terms in the sum are replaced by their conjugates, then the sum becomes conjugate as well.

We now show that a similar property holds true for the product of complex numbers:

$$\text{if } z_1z_2 = w, \text{ then } z_1^*z_2^* = w^* \qquad (1)$$

Indeed,

$$z_1z_2 = (x_1 + iy_1)(x_2 + iy_2) = (x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1)$$
$$z_1^*z_2^* = (x_1 - iy_1)(x_2 - iy_2) = (x_1x_2 - y_1y_2) - i(x_1y_2 + x_2y_1)$$

Comparing these two equations, we see that (1) holds true.

Putting $z_1 = z_2$ in (1), we see that if $z_1^2 = w$, then $(z_1^*)^2 = w^*$; that is, squaring conjugate numbers results in conjugates. It is clear that this is valid for arbitrary positive integer powers. Suppose, say, $z^3 = w$. We write this as $z^2z = w$. By (1) it follows that $(z^2)^*z^* = w^*$ but $(z^2)^* = (z^*)^2$ so that $(z^*)^3 = w^*$. Thus, $z^3 = w$ implies $(z^*)^3 = w^*$.

It is easy to establish that if $\frac{z_1}{z_2} = w$, then $\frac{z_1^*}{z_2^*} = w^*$ (see Exercise 1). Combining these results, we obtain the following general proposition.

Suppose the complex numbers $z_1, z_2, \ldots, z_n$ are involved in a certain number of arithmetical operations (additions, multiplications, divisions, raising to integral powers) and the result has proved equal to w. Then if these same operations (in the same order) are performed on the numbers $z_1^*, z_2^*, \ldots, z_n^*$, the result will be w^*. In other words, if all complex numbers in an equation are replaced by their conjugates, then equation holds true. It can be shown that this rule is also valid for nonarithmetic operations (taking roots and logarithms, etc.). From this it follows that any relation involving complex numbers holds true if i is everywhere replaced by $-i$, for this means that we have merely passed from an equation of complex numbers to one of their conjugates. Thus, the numbers i and $-i$ are essentially indistinguishable.* (This of course does not mean that $2 + 3i = 2 - 3i$!)

* We can obtain this result in straightforward fashion if it is observed that $(-i)^2 = -1$ and also $i^2 = -1$, whence $-i$ and i have the same properties.

Let us consider a quadratic equation with real coefficients:

$$az^2 + bz + c = 0 \qquad (2)$$

Let $z_0 = x_0 + iy_0$ be a root of this equation. Then

$$az_0^2 + bz_0 + c = 0$$

Replacing all numbers here by their conjugates and recalling that $a^* = a$, $b^* = b$, $c^* = c$ because a, b, c are real numbers, we obtain

$$a(z_0^*)^2 + bz_0^* + c = 0^* = 0$$

which is to say that z_0^* is also a root of this quadratic equation. Hence if a quadratic equation with real coefficients has imaginary (nonreal,* that is) roots, then these roots are conjugate complex numbers. Incidentally, this is immediately apparent from the quadratic formula. Equation (2) has the roots

$$z_{1,\,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

The roots are imaginary if $b^2 - 4ac < 0$, and

$$z_1 = -\frac{b}{2a} + i\frac{\sqrt{4ac - b^2}}{2a}, \quad z_2 = -\frac{b}{2a} - i\frac{\sqrt{4ac - b^2}}{2a}$$

These are conjugate complex numbers.

The value of this device lies in the fact that it applies not only to quadratic equations but to equations of any degree. Indeed, suppose the equation

$$a_0z^n + a_1z^{n-1} + \dots + a_{n-1}z + a_n = 0 \qquad (3)$$

with real coefficients $a_0, a_1 \dots, a_n$ has the imaginary root $z_0 = x_0 + iy_0$. Then

$$a_0z_0^n + a_1z_0^{n-1} + \dots + a_{n-1}z_0 + a_n = 0$$

Replacing all the numbers by their conjugates in this equation and noting that $a_0^* = a_0$, $a_1^* = a_1$, ..., $a_n^* = a_n$, we get

$$a_0(z_0^*)^n + a_1(z_0^*)^{n-1} + \dots + a_{n-1}z_0^* + a_n = 0^* = 0$$

Therefore $z_0^* = x_0 - iy_0$ is also a root of (3). Hence in the case of a real polynomial (that is, a polynomial with real coefficients) imaginary roots can only appear as conjugate pairs: if there exists a root $x_0 + iy_0$, then there is also a root $x_0 - iy_0$.

From this it also follows that a real polynomial can have only an even number of imaginary roots. In higher algebra, proof is given that a polynomial of degree n has exactly n roots. For this reason,

* Do not confuse the notion *imaginary* (which means nonreal) with *purely imaginary* (which means a real part equal to zero).

a real polynomial of odd degree must have at least one real root. (Prove this in another way by considering the behaviour of the graph of a polynomial $P(x)$ of odd degree for large $|x|$.) A polynomial of even degree may not have real roots.

From school algebra we know that if the equation (3) has the roots $z_1, z_2, \ldots, z_n$, then the polynomial on the left of (3) can be decomposed into factors as follows:

$$a_0z^n + a_1z^{n-1} + \ldots + a_{n-1}z + a_n = a_0(z - z_1)(z - z_2) \ldots (z - z_n) \quad (4)$$

Therefore if $x_0 + iy_0$ is a root of the polynomial, then in the factorization of this polynomial there is a complex factor $(z - x_0 - iy_0)$. Since in that case the polynomial also has the root $x_0 - iy_0$, the factorization likewise involves the factor $(z - x_0 + iy_0)$. In order to avoid dealing with imaginary numbers in (4), these two factors can conveniently be combined into one:

$$(z - x_0 - iy_0)(z - x_0 + iy_0) = z^2 - 2x_0z + x_0^2 + y_0^2$$

In place of two factors with imaginary coefficients we obtain one factor with real coefficients, but of second degree in the variable z. Thus, from the fundamental theorem of algebra (which we did not prove but merely took on trust) it follows that any real polynomial can be decomposed into real factors involving the variable z to the first and second powers but to no higher power.

Exercises

1. Show that if $\frac{z_1}{z_2} = w$, then $\frac{z_1^*}{z_2^*} = w^*$.
2. Find all the roots of the polynomial $z^4 - 6z^3 + 11z^2 - 2z - 10$ if we know that one of the roots is equal to $2 - i$.
3. Demonstrate that $z^{**} = z$. **4.** Let $z = x + iy$ and then find zz^*.

5.3 Raising a number to an imaginary power. Euler's formula

We now take up the very important question of what raising a number to an imaginary power means.

How does one approach such a question? First let us see how the simpler problem of negative and fractional powers is dealt with in algebra. The definition of positive integral powers is the only straightforward and pictorial one:

$$a^1 = a, \; a^2 = a \cdot a, \; a^3 = a \cdot a \cdot a, \ldots, a^n = \underbrace{a \cdot a \ldots a}_{n \text{ times}}$$

From this definition stem the following rules:

$$\frac{a^n}{a^m} = a^{n-m} (\text{if } n > m), \; (a^n)^m = a^{nm}$$

It is then assumed that these rules hold true for all exponents. From this follows the definition for fractional and negative powers. For example, from the formula $\left(a^{\frac{1}{n}}\right)^n = a^{\frac{1}{n}\cdot n} = a^1 = a$ it follows that $a^{\frac{1}{n}}$ is a number which, when raised to the power n, yields a, or $a^{\frac{1}{n}}$ is $\sqrt[n]{a}$. In the same way, we have $a^0 = 1$, $a^{-n} = \frac{1}{a^n}\cdot$

With this approach there is no way of defining a^i. Let us try to do so using what we know about the derivative of an exponential function, that is to say, using the tools of higher mathematics. The simplest and most convenient formulas, with no extra coefficients, are the formulas for differentiating e^t, e^{kt}:

$$\frac{de^t}{dt} = e^t, \quad \frac{de^{kt}}{dt} = ke^{kt}\,*$$

These formulas serve as a starting point for defining the notion of raising a number to an imaginary power:

$$\frac{de^{it}}{dt} = ie^{it}$$

Set $e^{it} = z(t)$, then $\frac{dz}{dt} = iz$. The relationship between z and dz is shown in Fig. 48. Since $dz = iz\,dt$, it follows that for real t and dt, dz is perpendicular to z. Hence, geometrically the change in $z = e^{it}$, with increasing t, by the amount dt amounts to a *rotation* of the vector z through the angle $d\varphi$. From the figure it is clear that since dz is perpendicular to z, then $d\varphi$ is equal to the ratio of the length of the line segment dz (which is equal to $r\,dt$) to the length of the line segment z (which is equal to r). Therefore $d\varphi = dt$. Of course the angles here (φ, $d\varphi$ and so on) must be expressed in natural units (radians and not degrees).

For $t = 0$ we have $e^{it}|_{t=0} = e^0 = 1$, or $z(0)$ is a horizontal vector whose length is equal to unity. Since the change in t by dt is associated with a rotation of the vector z through the angle $d\varphi = dt$, the variation of t from 0 to the given value t_1 is associated with the rotation through the angle $\varphi = t_1$.

Thus, $z = e^{i\varphi}$ is a vector obtained from $z(0) = 1$ by a rotation through the angle φ. Set $z = e^{i\varphi} = x + iy$. From Fig. 49 it is clear that $x = 1\cdot\cos\varphi = \cos\varphi$, $y = 1\cdot\sin\varphi = \sin\varphi$ and so

$$e^{i\varphi} = \cos\varphi + i\sin\varphi \qquad (5)$$

* It is precisely the condition that the formulas have a simple aspect is what defines the number e (see, for example, HM, Sec. 3.8).

Fig. 48

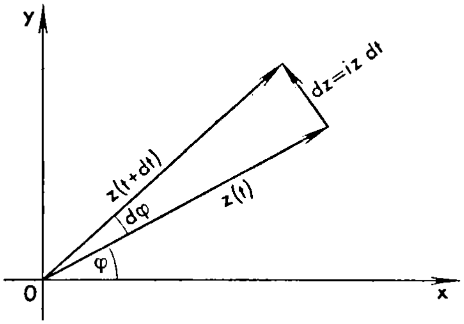
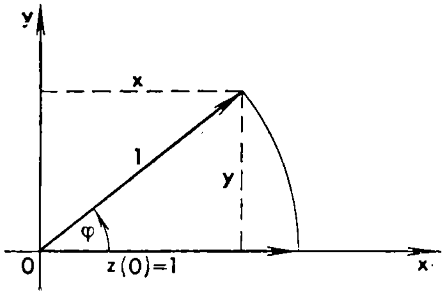


Fig. 49



This is *Euler's formula.* It is easy to verify that we then have $\frac{d}{d\varphi}(e^{i\varphi}) = ie^{i\varphi}$. Indeed,

$$\frac{d}{d\varphi}(e^{i\varphi}) = \frac{d}{d\varphi}(\cos\varphi) + i\frac{d}{d\varphi}(\sin\varphi) = -\sin\varphi + i\cos\varphi$$

$$= ii\sin\varphi + i\cos\varphi = i(\cos\varphi + i\sin\varphi) = ie^{i\varphi}$$

Euler's formula was obtained in a different manner in HM, Sec. 3.18: via an expansion of the left and right members in a Taylor series. But the definition itself of the functions e^t, e^{kt} in HM was based on expressing the derivative of the exponential function, which means both proofs are fundamentally the same.

Using Euler's formula, we can write any complex number z with modulus r and argument φ thus:

$$z = r(\cos\varphi + i\sin\varphi) = re^{i\varphi}$$

A complex number in this notation, $z = re^{i\varphi}$, is said to be *in exponential form.* Here, the rule of adding arguments when multiplying complex numbers becomes a simple consequence of the addition of exponents in the multiplication of powers.

Indeed, let $z_1 = r_1 e^{i\varphi_1}$, $z_2 = r_2 e^{i\varphi_2}$; then

$$z = z_1 z_2 = r_1 e^{i\varphi_1} \cdot r_2 e^{i\varphi_2} = r_1 r_2 \cdot e^{i(\varphi_1+\varphi_2)} = re^{i\varphi}$$

where

$$r = r_1 r_2, \quad \varphi = \varphi_1 + \varphi_2$$

From Euler's formula it is easy to obtain formulas for the sine and cosine of a sum:

$$e^{i(\varphi+\psi)} = \cos(\varphi + \psi) + i \sin(\varphi + \psi)$$

On the other hand,

$$\begin{aligned} e^{i(\varphi+\psi)} = e^{i\varphi} e^{i\psi} &= (\cos\varphi + i\sin\varphi)(\cos\psi + i\sin\psi) \\ &= \cos\varphi\cos\psi - \sin\varphi\sin\psi \\ &\quad + i(\sin\varphi\cos\psi + \cos\varphi\sin\psi) \end{aligned}$$

Comparing the real and imaginary parts in these two expressions for $e^{i(\varphi+\psi)}$ we get

$$\cos(\varphi + \psi) = \cos\varphi\cos\psi - \sin\varphi\sin\psi$$

$$\sin(\varphi + \psi) = \sin\varphi\cos\psi + \sin\psi\cos\varphi$$

Of course, in this simple case we obtained the familiar formulas that are readily proved in trigonometry without resorting to complex numbers. But if, say, it is necessary to express $\cos 5\varphi$ and $\sin 5\varphi$ in terms of $\cos\varphi$ and $\sin\varphi$, then the use of Euler's formula is more convenient, practically speaking, than ordinary trigonometric transformations.

This is done as follows:

$$(e^{i\varphi})^5 = e^{i5\varphi} = \cos 5\varphi + i \sin 5\varphi$$

On the other hand,

$$e^{i\varphi} = \cos\varphi + i\sin\varphi$$

and so

$$\begin{aligned} (e^{i\varphi})^5 &= (\cos\varphi + i\sin\varphi)^5 \\ &= \cos^5\varphi + i5\cos^4\varphi\sin\varphi - 10\cos^3\varphi\sin^2\varphi \\ &\quad - i10\cos^2\varphi\sin^3\varphi + 5\cos\varphi\sin^4\varphi + i\sin^5\varphi\ * \end{aligned}$$

(The powering was done by the binomial theorem.) Comparing the real and imaginary parts, we get

$$\cos 5\varphi = \cos^5\varphi - 10\cos^3\varphi\sin^2\varphi + 5\cos\varphi\sin^4\varphi$$

$$\sin 5\varphi = \sin^5\varphi - 10\sin^3\varphi\cos^2\varphi + 5\sin\varphi\cos^4\varphi$$

* Replace the exponent 5 by any integer n and we get the identity $(\cos\varphi + i\sin\varphi)^n = \cos n\varphi + i\sin n\varphi$ which is known as the *de Moivre formula*.

Here is another example of the use of Euler's formula. Suppose it is required to find $\int e^{ax} \cos bx \, dx$. We consider the equation

$$\int e^{(a+ib)x} \, dx = \frac{e^{(a+ib)x}}{a+ib} + C \tag{6}$$

where $C = C_1 + iC_2$. Note that by Euler's formula

$$\int e^{(a+ib)x} \, dx = \int e^{ax} \cos bx \, dx + i \int e^{ax} \sin bx \, dx$$

By means of the following manipulations we can separate the real part of the right-hand member of (6) from the imaginary part:

$$e^{(a+ib)x} = e^{ax} e^{ibx} = e^{ax}(\cos bx + i \sin bx)$$

$$\frac{1}{a+ib} = \frac{a-ib}{(a+ib)(a-ib)} = \frac{a}{a^2+b^2} - i \frac{b}{a^2+b^2}$$

Multiplying, we find that the quotient in the right-hand member of (6) is

$$\frac{e^{ax}}{a^2+b^2}(a \cos bx + b \sin bx) + i \frac{e^{ax}}{a^2+b^2}(a \sin bx - b \cos bx)$$

Then equation (6) becomes

$$\int e^{ax} \cos bx \, dx + i \int e^{ax} \sin bx \, dx$$

$$= \frac{e^{ax}}{a^2+b^2}(a \cos bx + b \sin bx) + i \frac{e^{ax}}{a^2+b^2}(a \sin bx - b \cos bx) + C$$

Comparing the real and imaginary parts, we obtain

$$\int e^{ax} \cos bx \, dx = \frac{e^{ax}(a \cos bx + b \sin bx)}{a^2+b^2} + C_1$$

$$\int e^{ax} \sin bx \, dx = \frac{e^{ax}(a \sin bx - b \cos bx)}{a^2+b^2} + C_2$$

Exercises

1. Write down the following numbers in exponential form: (a) $1+i$, (b) $1-i$, (c) -1, (d) $3i$.
2. Using Euler's formula, find (a) $(1+i)^{16}$, (b) $\left(\frac{1}{2} + i\frac{\sqrt{3}}{2}\right)^9$.
3. Express $\cos 3\varphi$, $\sin 4\varphi$ in terms of $\sin \varphi$ and $\cos \varphi$.
4. Prove the formulas

$$\cos \varphi = \frac{e^{i\varphi} + e^{-i\varphi}}{2}, \quad \sin \varphi = \frac{e^{i\varphi} - e^{-i\varphi}}{2i}$$

5.4 Logarithms and roots

It is a remarkable fact that an exponential function with imaginary exponents becomes periodical. From Euler's formula it follows that $e^{2\pi i} = \cos 2\pi + i \sin 2\pi = 1$ and so $e^{i(t+2\pi)} = e^{it}e^{2\pi i} = e^{it}$, that is to say, the function $z(t) = e^{it}$ is periodic with period 2π.

The periodic properties of a power are evident in the following simple instances:

1. $(-1)^0 = 1,\ (-1)^1 = -1,\ (-1)^2 = 1,$
   $(-1)^3 = -1, \ldots, (-1)^{2n+g} = (-1)^g$
2. $i^0 = 1,\ i^1 = i,\ i^2 = -1,\ i^3 = -i,\ i^4 = 1,\ i^5 = i,$
   $i^6 = -1, \ldots, i^{4n+s} = i^s.$

From these examples we conclude that $(-1)^g$ as a function of g is of period 2, and i^s as a function of s is of period 4.

We will now demonstrate that the periodicity of these functions is a consequence of the periodicity of the function e^{it}. Using Euler's formula, write -1 in exponential form: $-1 = e^{\pi i}$. Therefore $(-1)^g = e^{i\pi g}$, where g is any number. Now let us determine the period of the function $e^{i\pi g}$ for real g. If G is the period, then $e^{i\pi(g+G)} = e^{i\pi g}$. Then it must be true that $e^{i\pi G} = 1$, whence $\pi G = 2\pi$, $G = 2$. Similarly, $i = e^{i\frac{\pi}{2}}$ and so $i^s = e^{i\frac{\pi s}{2}}$. The period of the function $e^{i\frac{\pi s}{2}}$ is 4.

The periodicity of the exponential function has exciting consequences for logarithms. The *logarithm* of a complex number z is a complex number w such that $e^w = z$.

Let $z = re^{i\varphi}$. Write r as $r = e^\rho$, where ρ is the natural logarithm of the positive number r. Then $z = e^{\rho + i\varphi}$ and so

$$w = \ln z = \rho + i\varphi = \ln r + i\varphi$$

But we already know that $e^{2\pi ki} = 1$ if k is an integer, and so we can write $z = e^{\rho + i\varphi + 2\pi ki}$, whence $\ln z = \ln r + i(\varphi + 2k\pi)$.

Thus, the logarithm of a complex number has infinitely many values. This situation is reminiscent of that existing between trigonometric and inverse trigonometric functions. For example, since $\tan \varphi$ has a period equal to π, that is, $\tan (\varphi + k\pi) = \tan \varphi$ if k is integral, then $\arctan x$ has an infinity of values. Indeed, if $\varphi = \arctan x$ is one of the values of the arctangent, then $\varphi + k\pi$ is also a value of the arctangent (k an integer).

The periodicity of the exponential function is also important in *root extraction*, that is, when raising a number to a fractional power.

Suppose we have $z = re^{i\varphi}$; as we have seen, this can be written as $z = re^{i\varphi + 2\pi ki}$. Then

$$z^n = r^n e^{(i\varphi + 2\pi ki)n} = r^n e^{in\varphi} e^{2\pi kin} = ce^{2\pi kni} \tag{7}$$

where c stands for $r^n e^{in\varphi}$. If n is an integer, then $e^{2\pi kni} = 1$, and for this reason we get only one value of z^n, namely the number c. Thus, raising to an integral power is a unique operation. The situation changes if n is fractional, $n = \frac{p}{q}$, where p and $q > 0$ are prime to each other (that is, they do not have any common divisors other than unity). Then $e^{2\pi kni}$ can be different from 1 and by (7) we get new values of z^n that differ from c. It can be proved that the total number of distinct values of z^n is equal to q.

Consider an example. Find all the values of $1^{1/3} = \sqrt[3]{1}$. Since $1 = e^{2\pi ki}$, it follows that $1^{1/3} = e^{\frac{2\pi ki}{3}}$. Putting $k = 0$ in the last equation, we get $1^{1/3} = 1$; putting $k = 1$, we get

$$1^{1/3} = e^{\frac{2\pi i}{3}} = \cos\frac{2\pi}{3} + i\sin\frac{2\pi}{3} = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$$

Putting $k = 2$, we obtain

$$1^{1/3} = e^{\frac{4\pi i}{3}} = \cos\frac{4\pi}{3} + i\sin\frac{4\pi}{3} = -\frac{1}{2} - i\frac{\sqrt{3}}{3}$$

Thus, we have three values for $\sqrt[3]{1}$. (Verify by means of straightforward involution that $\left(-\frac{1}{2} \pm i\frac{\sqrt{3}}{2}\right)^3 = 1$.) It is easy to see that by putting $k = 3, 4, \ldots$, or $k = -1, -2, \ldots$, we do not get new values of the root.

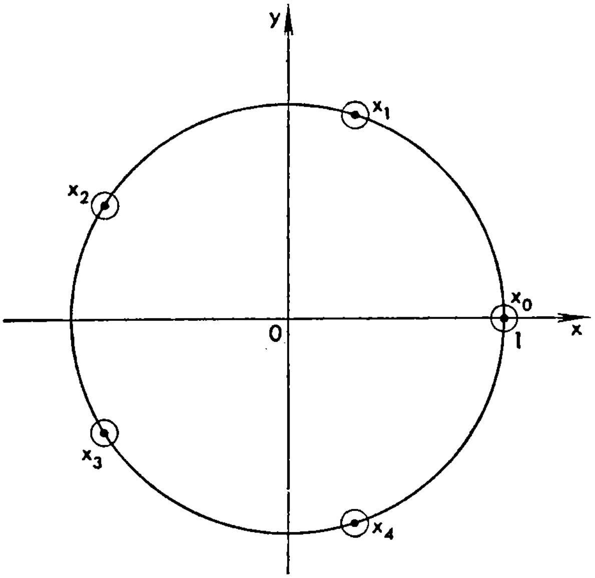
If n is an integer, then the roots of the equation $x^n = 1$ are n numbers of the form

$$x_k = \cos\frac{2k\pi}{n} + i\sin\frac{2k\pi}{n}, \quad \text{where } k = 0, 1, 2, \ldots, n-1$$

Here, substituting $k = 0$, we get $x_0 = \cos 0 + i\sin 0 = 1$, which is a definitely known value. The numbers x_k are depicted in the plane by points located at the vertices of a regular n-gon inscribed in a circle of radius 1. Indeed, the modulus of any one of the numbers x_k is equal to $r_k = \sqrt{\cos^2\frac{2k\pi}{n} + \sin^2\frac{2k\pi}{n}} = 1$, the argument of the number x_0 is zero, and the argument increases by $\frac{2\pi}{n}$ as k is increased by unity. Fig. 50 shows the roots of the equation $x^5 = 1$.

For the simplest type of equation of degree n (n a positive integer), $x^n = 1$, we saw that the total number of real and imaginary roots is equal to n. Actually (but we do not give the proof here) *every algebraic equation of degree n has n complex roots,* some of which may be real, for it is always assumed that real numbers constitute a particular case of complex numbers. The number of distinct roots

Fig. 50



may be less than n, since some roots may be coincident (so-called *multiple roots*). For example, the quartic equation

$$(x-1)^3(x+1)=0$$

has a triple root $x_{1,2,3}=1$ and a simple root $x_4=-1$, making a total of four roots.

We conclude this section with a survey of the process of increasing complication of mathematical operations and the concomitant development of the number notion. The starting point was the class (set) of positive integers (also called natural numbers). Using these numbers alone, we can always carry out addition, inasmuch as the result of adding two natural numbers is again a natural number; subtraction however is not always possible. To make subtraction possible we have to consider negative integers and zero.

Let us consider the next two operations: multiplication and division. The result of multiplying two integers is an integer, but the operation of dividing integers is not always possible in whole numbers (integers). For division to be possible in all cases, we have to introduce fractions. Together, integers and fractions constitute the class of rational numbers. But we know that the limit of a sequence of rational numbers may be a nonrational, that is, an irrational number (these numbers are further subdivided into "algebraic" and "transcendental" numbers, but we will not go any further into this matter here). The rational and irrational numbers together form the class of real numbers in which the operation of passing to a limit is always possible. However, in this last class it is not always possible to carry out algebraic operations: for example, it is possible to take the roots of any power of positive numbers but it is not possible to take the roots of even powers (square roots, say) of negative numbers.

Taking the square root of a negative number becomes possible only when we introduce complex numbers. It is a remarkable fact that the introduction of complex numbers is the *final* extension of the number system. Indeed, if we have complex numbers at our disposal, then we can take the roots (of any degree) not only of negative but also of any complex numbers. The results are always complex numbers. But in mathematics there are operations that cannot be performed if we remain within the class of real numbers only. For instance, raising a positive number to any real power always results in a positive number, so that there are no logarithms of negative numbers. Another impossible operation: there is no real number φ such that, say, $\cos \varphi = 2$. The question then arises: perhaps it is necessary, in order to find the logarithms of negative numbers, to introduce a new "imaginary unit" that differs from the i introduced earlier in order to be able to take the roots of even powers of negative numbers? Will the solution of the equation $\cos \varphi = 2$ require some other third "imaginary unit"? It turns out that this is not the case: with the introduction of complex numbers we are now able to take the logarithms of negative and even complex numbers (see above) and solve equations of the form $\cos \varphi = k$, where k is any number (see Exercise 4 below), and so on. Thus, any operations involving complex numbers are possible and the results of such operations are again complex numbers. Therefore there is no need to introduce any new numbers.

Exercises

1. Find the logarithms of the following numbers:
(a) -1, (b) i, (c) $-i$, (d) $1 + i$.

2. Find all the values of $\sqrt[3]{-1}$.

3. Find all the values of $\sqrt[6]{1}$.
In Problems 2 and 3 obtain values of the roots in trigonometric and algebraic forms. Verify the values of the roots by raising to the appropriate power using the binomial theorem.

4. Solve the equation $\cos \varphi = 2$. Indicate all solutions.

5. State all solutions of the equation $\sin \varphi = 2$.
Hint. In Problems 4 and 5 use formulas that express the sine and cosine in terms of the exponential function. (See Exercise 4 of Sec. 5.3.)

5.5 Describing harmonic oscillations by the exponential function of an imaginary argument

Knowing that the exponential function is periodic for imaginary exponents, we can use it to solve problems involving mechanical vibrations and the theory of electric circuits.

Let us consider, for example, *harmonic oscillations* (which are oscillations that obey the sinusoidal law) of electric current in a circuit in which the current obeys the law

$$j = j_0 \sin(\omega t + \alpha) \tag{8}$$

Here, j_0 is the amplitude of the oscillations (maximum current), ω is the frequency, and α is the initial phase (the value of the phase, $\omega t + \alpha$, at $t = 0$). It appears to be convenient to introduce, in addition to (8), the notion of a *complex current*

$$J = j_0 e^{i(t\omega+\alpha)} \tag{9}$$

for which the "real" current (8) serves as the imaginary part, since by Euler's formula (5),

$$j_0 e^{i(\omega t+\alpha)} = j_0 \cos(\omega t + \alpha) + i j_0 \sin(\omega t + \alpha)$$

The complex current (9) is represented in the complex plane by a vector (Fig. 51) of length j_0 that forms with the positive real axis an angle $\omega t + \alpha$; thus, at $t = 0$ this vector has a slope of α, and as t increases it rotates uniformly with angular velocity ω. The current (8) is obtained by projecting the terminus of the vector J on the imaginary axis. (If instead of (8) the law of current were given by $j = j_0 \cos(\omega t + \alpha)$, then it would be the real part of the complex current (9) and would be obtained by projecting J on the real axis.)

Expression (9) is an example of a complex-valued function of a real independent variable. In the general case, any such function may be written as

$$f(t) = g(t) + ih(t)$$

where $g(t)$ and $h(t)$ are ordinary real functions of a real independent variable. Such complex-valued functions have the following obvious properties.

If the complex functions are added, so also are their real and imaginary parts.

If a complex function is multiplied by a real constant or a real function, the real and imaginary parts are multiplied by that factor.

If a complex function is differentiated or integrated, the same operations are performed on its real and imaginary parts.

These properties make it possible to perform the indicated operations on the whole complex function (instead of on the real or imaginary part) and then take the real or the corresponding

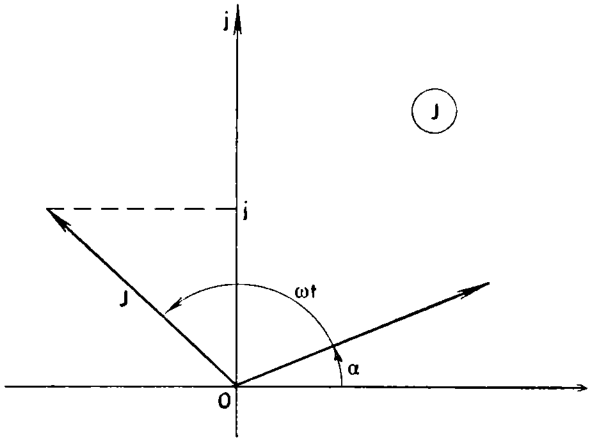


Fig. 51

imaginary part of the result. It is a remarkable thing that such a transition to complex quantities together with the reverse transition to the required real quantities may turn out to be simpler and more pictorial than straightforward operations performed on real quantities.

Let us consider, for example, the superposition of oscillations having the same frequency. Suppose we have to add currents:

$$j = j_1 \sin(\omega t + \alpha_1) + j_2 \sin(\omega t + \alpha_2)$$

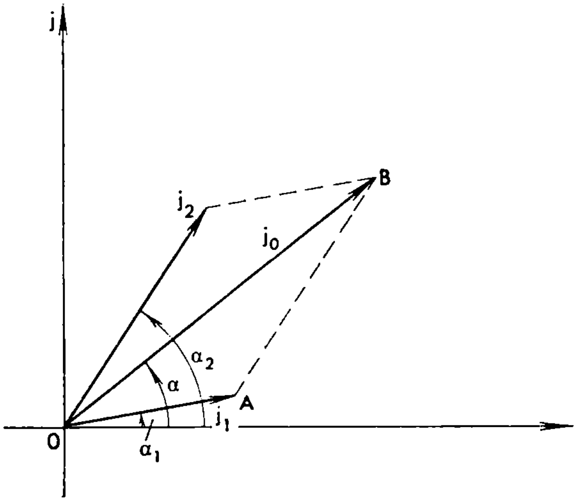
We have seen that the appropriate complex currents J_1 and J_2 are represented in the plane of the complex variable by vectors uniformly rotating with angular velocity ω. And in Sec. 5.1 we demonstrated that such vectors are added by the parallelogram rule. This means that the total complex current J will uniformly rotate with angular velocity ω, that is, it may be written as (9), where j_0 and α are readily obtained geometrically. In Fig. 52 we have the position of a rotating parallelogram at time $t = 0$, which is the time we need for determining j_0 and α. These parameters may be obtained via a geometric construction, as in Fig. 52, but they are also obtainable analytically. To find j_0 we can apply the cosine theorem to the triangle OAB to get

$$\begin{aligned} j_0^2 = OB^2 &= AO^2 + AB^2 - 2AO \cdot AB \cos(\widehat{AO, AB}) \\ &= j_1^2 + j_2^2 - 2j_1j_2 \cos[180^\circ - (\alpha_2 - \alpha_1)] \\ &= j_1^2 + j_2^2 + 2j_1j_2 \cos(\alpha_2 - \alpha_1) \end{aligned}$$

To find α, note that the projections of the vector $J|_{t=0}$ on the coordinate axes are respectively equal to

$$j_1 \cos\alpha_1 + j_2 \cos\alpha_2 \text{ and } j_1 \sin\alpha_1 + j_2 \sin\alpha_2$$

Fig. 52



whence

$$\tan \alpha = \frac{j_1 \sin \alpha_1 + j_2 \sin \alpha_2}{j_1 \cos \alpha_1 + j_2 \cos \alpha_2}$$

Similar results are obtained in superposing any number of oscillations that have the same frequency. It is clear at the same time that superposition of oscillations occurring with different frequencies produces a total current of a complicated nature that will not vary by the harmonic law.

Still more pictorial is the differentiation of a complex current. From (9) we get

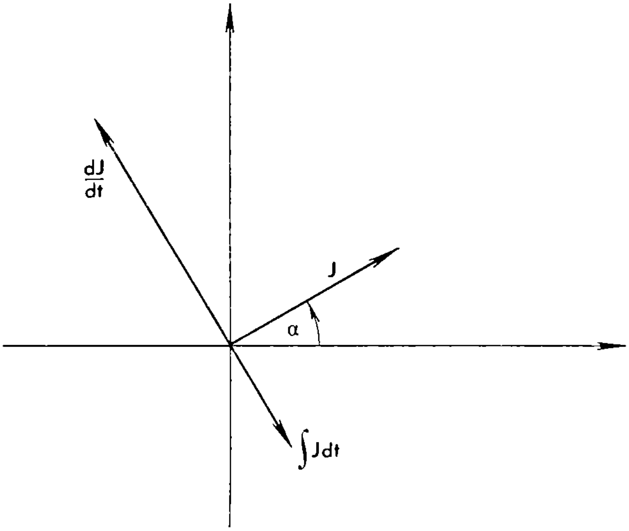
$$\frac{dJ}{dt} = j_0 e^{i(\omega t+\alpha)} i\omega = i\omega J \tag{10}$$

By virtue of Sec. 5.1, this multiplication reduces to a ω-fold stretching of the vector J and counterclockwise rotation of it through 90°. Which means the vector $\frac{dJ}{dt}$ is also in uniform rotation with velocity ω. Similarly, dropping the arbitrary constant, we get

$$\int J\, dt = j_0 \frac{e^{i(\omega t+\alpha)}}{i\omega} = \frac{J}{i\omega} = -i\frac{J}{\omega} \tag{11}$$

This vector is obtained from J by a ω-fold compression and a 90° rotation in the negative direction. The position of these vectors at $t = 0$ is shown in Fig. 53 on what is known as a *phase diagram.*

Fig. 53



These results can be applied to working out oscillations in an electric circuit containing any combinations of resistors, inductors, and capacitors. For this purpose we take advantage of Fig. 54, which shows the relationship between the voltage drop on a circuit element and the current flow in that element. (The appropriate formulas are derived in electricity theory courses; see also HM, Sec. 8.1.) Also, use is made of the *Kirchhoff laws*, which state that the algebraic sum of all currents flowing toward any point in a network is equal to zero; and the algebraic sum of the voltage drops on any sequence of elements forming a closed circuit is equal to zero. These laws are needed to form complicated, extended networks, since for simple circuits they do not offer anything beyond the obvious initial relationships.

Consider, for example, the *RL* circuit shown in Fig. 55. It has a voltage source varying by the harmonic law

$$\varphi = \varphi_0 \sin(\omega t + \beta) \tag{12}$$

This gives rise to a current that varies by the harmonic law (8), though j_0 and α are not known beforehand. Equating φ to the sum of the voltage drops on R and L on the basis of the formulas of Fig. 54, we get (here, in implicit form, we make use of both the Kirchhoff laws)

$$Rj + L\frac{dj}{dt} = \varphi$$

Fig. 54

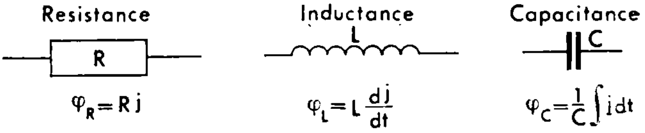
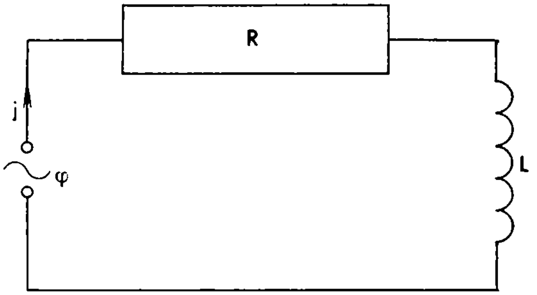


Fig. 55



Passing to the complex current (9) and the complex voltage,

$$\Phi = \varphi_0 e^{i(\omega t + \beta)}$$

we get

$$RJ + L\frac{dJ}{dt} = \Phi \tag{13}$$

or, using (10),

$$RJ + i\omega LJ = \Phi \tag{14}$$

whence

$$J = \frac{\Phi}{R + i\omega L} = \frac{\varphi_0 e^{i\beta}}{R + i\omega L} e^{i\omega t}$$

We see that the inductance L may be interpreted as a resistance numerically equal to $i\omega L$; it is called the *impedance* of element L. Now, to obtain the desired current, it remains to take the imaginary part of the last expression. It is simpler, however, to write the coefficient $\frac{\varphi_0 e^{i\beta}}{R + i\omega L}$ in the exponential form $j_0 e^{i\alpha}$, which produces the desired values j_0 and α at once. From this we see that

$$j_0 = \left|\frac{\varphi_0 e^{i\beta}}{R + i\omega L}\right| = \frac{\varphi_0}{\sqrt{R^2 + \omega^2 L^2}},$$

$$\alpha = \arg\frac{\varphi_0 e^{i\beta}}{R + i\omega L} = \beta + \arg\frac{1}{R + i\omega L} = \beta + \arg\frac{R - i\omega L}{R^2 + \omega^2 L^2}$$
$$= \beta + \arg\,(R - i\omega L)$$

(here we make use of the fact that the argument arg of a complex number remains unchanged when this number is multiplied by a positive integer). Hence the phase of the current in an RL circuit is delayed relative to the phase of the voltage source, which is of course due to the presence of an inductance.

The desired current can also be obtained with the aid of the geometrical construction shown in Fig. 56. Suppose for a moment that we have the current (8) and seek the voltage (12). Then it is easy to construct the vector $J|_{t=0} = j_0 e^{i\alpha}$ and, together with it, the mutually perpendicular vectors $RJ|_{t=0}$ and $i\omega LJ|_{t=0}$. But by virtue of (14) the vector $\Phi|_{t=0} = \varphi_0 e^{i\beta}$ is a diagonal of a rectangle constructed on the last two vectors, which means that the last vector is readily constructed.

Now let us return to the original problem of finding the current from the voltage. Depict the vector $\Phi|_{t=0} = \varphi_0 e^{i\beta}$ of the outer complex voltage at time $t = 0$ (its modulus φ_0 and the argument β are given). This vector must serve as a diagonal of the rectangle constructed on the vectors $RJ|_{t=0}$ and $i\omega LJ|_{t=0}$, but vector $J|_{t=0}$ is not given, it is sought. But this rectangle is similar to the rectangle constructed on the vectors R and $i\omega L$ (the smaller rectangle in Fig. 56), which means we have to proceed as follows: construct the latter rectangle, then increase or decrease it (observing similarity relations) so that its diagonal is equal to φ_0, then rotate the diagonal to coincidence with the vector $\Phi|_{t=0}$; it will then have the position of the large rectangle in Fig. 56. Finally, divide the side $RJ|_{t=0}$ by R to get $J|_{t=0}$.

Now let us consider the LC circuit shown in Fig. 57. Here, instead of (13), we get

$$L\frac{dJ}{dt} + \frac{1}{C}\int J\,dt = \Phi$$

whence, using (10) and (11), we get

$$i\omega LJ - i\frac{J}{\omega C} = \Phi$$

that is,

$$J = \frac{\Phi}{i\left(\omega L - \dfrac{1}{\omega C}\right)} = -i\frac{\omega\varphi_0}{L\left(\omega^2 - \dfrac{1}{LC}\right)}e^{i\beta}e^{i\omega t} \tag{15}$$

$\left(\text{Thus, the impedance of the capacitor is the quantity } -\dfrac{i}{\omega C}.\right)$ Different cases are possible here. If the frequency ω of the external voltage source is sufficiently great, to be more exact, if $\omega^2 > \dfrac{1}{LC}$, then the bracketed quantity in the denominator is positive. Representing $-i = e^{-i\frac{\pi}{2}}$, we can write

$$J = \frac{\omega\varphi_0}{L\left(\omega^2 - \dfrac{1}{LC}\right)}e^{i\left(\beta - \frac{\pi}{2}\right)}e^{i\omega t}$$

Fig. 56

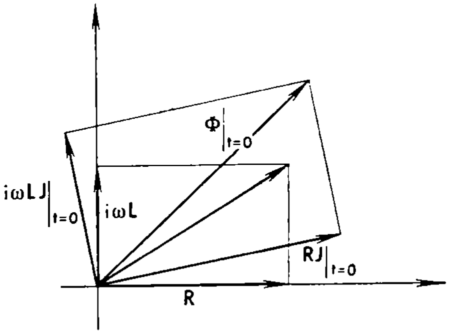
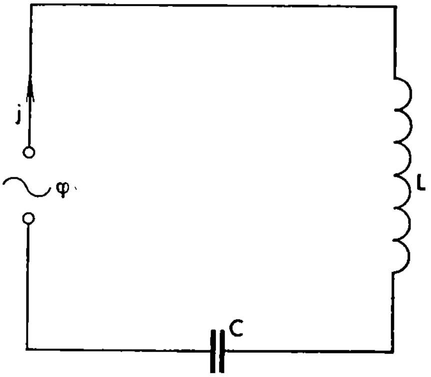


Fig. 57



A current is thus established in the circuit with amplitude

$$j_0 = \frac{\omega\varphi_0}{L\left(\omega^2 - \frac{1}{LC}\right)}$$

and a phase lag of 90° in comparison with the voltage source. If the frequency ω is small, that is, $\omega^2 < \frac{1}{LC}$, then it can be verified in similar fashion that the phase of the current being established is 90° in advance of the phase of the external voltage. Of particular interest is the intermediate case where

$$\omega^2 = \frac{1}{LC}$$

or

$$\omega = \frac{1}{\sqrt{LC}} \tag{16}$$

and the solution (15) is not suitable, since there is a zero in the denominator on the right. In this case, for all j_0 and α the expression (9) satisfies the relation

$$L\frac{dJ}{dt}+\frac{1}{C}\int J\,dt=i\omega LJ-i\frac{J}{\omega C}=\frac{iL}{\omega}\left(\omega^2-\frac{1}{LC}\right)J=0$$

which is to say that in a circuit without an external voltage source (when the voltage source is short-circuited), undamped harmonic oscillations via law (8) are possible. Such oscillations that occur in the absence of any external source are called *free*, or *natural*, oscillations. Thus, if the frequency of the external voltage satisfies the relation (16), then it is equal to the frequency of the natural undamped oscillations in the circuit. Under these conditions, as we know from physics, resonance sets in and instead of the periodic harmonic oscillations that we are considering here, there appear oscillations with increasing amplitude. Resonance will be examined in Ch. 7.

Bear in mind that this method enables one to obtain the current that is set up in a circuit after a certain "transitory period". The transient process is also described by the methods given in Ch. 7.

Exercises

1. Consider a series-connected circuit with resistance R, inductance L and capacitance C. What occurs in the particular case where $\omega^2=\frac{1}{LC}$?
2. Consider a parallel-connected circuit with resistance R and inductance L.

5.6 The derivative of a function of a complex variable

The variable $w=f(z)$ is called a function of the complex number z if to each value of z there is associated a definite value of $f(z)$. Since $z=x+iy$, where x is the real part and y is the imaginary part, it follows that specification of z signifies specification of two real numbers x and y. Here, $f(z)=u(x,y)+iv(x,y)$, where $u(x,y)$ and $v(x,y)$ are real functions. To each z there correspond definite x and y and, hence, definite u and v and, consequently, a definite value of $f(z)$. However, we will now consider as a function $f(z)$ not every expression $u(x,y)+iv(x,y)$ but only such a quantity as depends on z via such formulas as, say, $f(z)=1+z^2$, $f(z)=z$, $f(z)=e^z$, $f(z)=\sin z$, and so on. These formulas can involve algebraic operations or nonalgebraic operations on z, but they must be such as can be expressed with the aid of a Taylor series in z, for example, e^z, $\sin z$, and so forth.

We thus consider formulas involving z but not its real or imaginary part separately. (From this viewpoint, $z^*=x-iy$ or $|z|$

are not regarded as functions of z, although if we know z, it is easy to find z^* and $|z|$.) With this definition of $f(z)$, all functions $f(z)$ have one property in common: we can find the derivative

$$f'(z) = \frac{df}{dz}$$

by the ordinary rules for differentiating functions. A function $f(z)$ that has a derivative is called an *analytic function*. We are thus going to consider only analytic functions.

For functions given by simple formulas the computation of derivatives is no more complicated than for functions of a real variable. Consider, say, the function $w = z^2$. Giving z the increment Δz, we get the increment of the function:

$$\Delta w = (z + \Delta z)^2 - z^2 = 2z\,\Delta z + (\Delta z)^2$$

whence, passing to differentials and dropping (the moduli of) higher infinitesimals, we get

$$dw = 2z\,dz, \quad \frac{dw}{dz} = 2z, \quad \text{that is,} \quad \frac{d(z^2)}{dz} = 2z$$

It is thus clear that in these computations it is immaterial whether the independent variable assumes complex or only real values.

However in the general case of a complex variable, the existence of a derivative is not at all so obvious and simple as in the case of a real variable. Indeed, $z = x + iy$ and we can give the increment dz by changing x and y at the same time: $dz = dx + i\,dy$; we can give the increment dz by changing only x: $dz = dx$; finally, we can change only y: $dz = i\,dy$. If the derivative $f'(z)$ exists, then this means that for distinct modes of changing z the corresponding changes in f are such that $\frac{df}{dz}$ are the same in all cases.

Write $f(z)$ thus: $f(z) = u(x, y) + iv(x, y)$. Suppose only x changes; then $dz = dx$ and

$$\frac{df}{dz} = \frac{\frac{\partial f}{\partial x}\,dx}{dx} = \frac{\partial u}{\partial x} + i\,\frac{\partial v}{\partial x}$$

If it is only y that is changed, then $dz = i\,dy$, and so

$$\frac{df}{dz} = \frac{\frac{\partial f}{\partial y}\,dy}{i\,dy} = \frac{\left(\frac{\partial u}{\partial y} + i\,\frac{\partial v}{\partial y}\right)dy}{i\,dy} = \frac{1}{i}\,\frac{\partial u}{\partial y} + \frac{\partial v}{\partial y} = -i\,\frac{\partial u}{\partial y} + \frac{\partial v}{\partial y}$$

Equating the two expressions for the derivative obtained by different methods, we get

$$\frac{\partial u}{\partial x} + i\,\frac{\partial v}{\partial x} = -\,i\,\frac{\partial u}{\partial y} + \frac{\partial v}{\partial y}$$

and hence

$$\frac{\partial u}{\partial x}=\frac{\partial v}{\partial y}, \quad \frac{\partial v}{\partial x}=-\frac{\partial u}{\partial y} \tag{17}$$

These formulas are called the *Cauchy-Ricmann conditions*. Thus, the real and imaginary parts of an analytic function $f(z)$ are connected by definite relations.

Consider the following example. Suppose

$$f(z)=z^2=(x+iy)^2=x^2-y^2+i\cdot 2xy$$

Here, $u=x^2-y^2$, $v=2xy$,

$$\frac{\partial u}{\partial x}=\frac{\partial v}{\partial y}=2x, \quad \frac{\partial v}{\partial x}=-\frac{\partial u}{\partial y}=2y$$

And the Cauchy-Riemann conditions (17) hold.

Consider an example of an opposite nature. Let $f(z)=z^*=$ $=x-iy$, where $z=x+iy$. We noted above that although such a function $f(z)$ may be, in a certain sense, called a function of z (to every z there corresponds a definite z^*), it is not an analytic function, since $u=x$, $v=-y$ and so

$$\frac{\partial u}{\partial x}=1, \ \frac{\partial v}{\partial x}=0, \ \frac{\partial u}{\partial y}=0, \ \frac{\partial v}{\partial y}=-1$$

and the first of the Cauchy-Riemann conditions does not hold true.

Exercise

Verify the Cauchy-Riemann conditions for the function $f(z)=z^3$.

5.7 Harmonic functions

Let us return to analytic functions. We take the derivatives with respect to x of the first equation in (17) and with respect to y of the second to get

$$\frac{\partial^2 u}{\partial x^2}=\frac{\partial^2 v}{\partial x\,\partial y,}, \quad \frac{\partial^2 v}{\partial x\,\partial y}=-\frac{\partial^2 u}{\partial y^2}$$

whence

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0 \tag{18}$$

Similarly, differentiating the first equation in (17) with respect to y and the second with respect to x, we get $\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}=0$. Equation (18) is called the *Laplace equation*. By specifying different $f(z)$, we can obtain distinct solutions of this equation, which are called *harmonic functions*.

To summarize, then, for the real and imaginary parts of an analytic function we cannot take just any functions $u(x, y)$ and $v(x, y)$

but only harmonic functions. Also, these functions must be connected by the relations (17); such functions are called *conjugate harmonic functions*. It can be verified that if one of two conjugate functions is chosen, then by virtue of these relations the other one is completely defined up to an arbitrary additive constant. Let the function v be chosen and let the relations (17) be satisfied by two functions: $u = u_1$ and $u = u_2$. Then $\frac{\partial u_1}{\partial x} = \frac{\partial u_2}{\partial x}$, since both the derivatives are equal to $\frac{\partial v}{\partial y}$, and so $\frac{\partial(u_1 - u_2)}{\partial x} = 0$, which is to say that $u_1 - u_2$ does not depend on x. Similarly we obtain $\frac{\partial(u_1 - u_2)}{\partial y} = 0$ that is, the difference $u_1 - u_2$ does not depend on y either, and so it is a constant.

The Laplace equation is of great importance in mathematical physics. For example, the electrostatic potential in the space between long conductors stretching perpendicular to the xy-plane is a function of x, y alone and satisfies equation (18). At points of the xy-plane where it is punctured by the conductors, the Laplace equation for the potential fails. In particular, the potential has a singularity (becomes infinite) where the plane is intersected by a conductor whose cross-section is of an infinitely small diameter. Thus, to put it more precisely, we must say that the potential is a harmonic function in that portion of the plane where there are no charges. This will be discussed in more detail in Sec. 10.5.

The relationship between u and v has a simple geometric meaning. Construct the line $u(x, y) =$ constant. Since $\frac{\partial u}{\partial x}\,dx + \frac{du}{\partial y}\,dy = 0$, it follows that the slope of this line to the x-axis is

$$\tan \alpha_1 = \left.\frac{dy}{dx}\right|_{u=\text{const.}} = -\frac{\partial u}{\partial x}\Big/\frac{\partial u}{\partial y}$$

Similarly, the slope of the line $v(x, y) =$ constant to the x-axis is

$$\tan \alpha_2 = \left.\frac{dy}{dx}\right|_{v=\text{const.}} = -\frac{\partial v}{\partial x}\Big/\frac{\partial v}{\partial y}$$

Using the Cauchy-Riemann conditions, we get

$$\tan \alpha_2 = \frac{\partial u}{\partial y}\Big/\frac{\partial u}{\partial x} = -\frac{1}{\tan \alpha_1} \tag{19}$$

What we have is the condition of perpendicularity of tangent lines to the lines $u(x, y) =$ constant and $v(x, y) =$ constant.[*]

[*] For perpendicularity, it must be true that $\alpha_2 = \alpha_1 \pm 90°$, i.e.

$$\tan \alpha_2 = \tan(\alpha_1 \pm 90°) = -\cot \alpha_1 = -\frac{1}{\tan \alpha_1}$$

which is the condition (19).

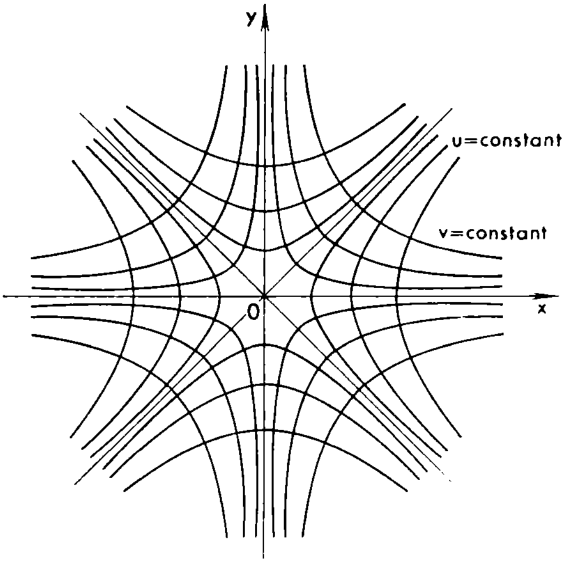


Fig. 58

Thus, taking any analytic function $f(z) = u(x, y) + iv(x, y)$, we get two families of curves intersecting at right angles at every point. These curves are shown in Fig. 58 for the case $f(z) = z^2$ given above.

If u is the electrostatic potential, then $u(x, y) =$ constant represents the lines of constant potential and $v(x, y) =$ constant, the lines of field intensity (force lines); at every point, the line of field intensity is along the normal to the line $u(x, y) =$ constant passing through this point.

Exercises

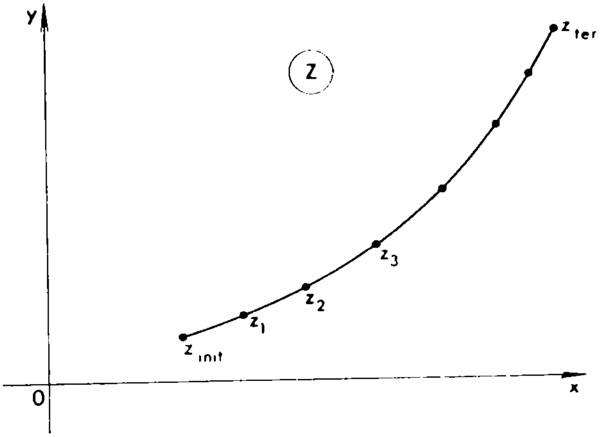
1. Let $f(z) = u(x, y) + iv(x, y)$. Knowing that $u(x, y) = x + x^2 - y^2$, determine $v(x, y)$ if $f(0) = 0$.
2. Knowing that $v(x, y) = -2\,xy$, $f(0) = 1$, determine $u(x, y)$.

5.8 The integral of a function of a complex variable

We now define the integral of the complex function $f(z)$. Take two points — the initial point z_{init} and the terminal point z_{ter} — in the plane and join them with some curve (l) (Fig. 59). Partition this curve into p subsections and number the subdivision points thus:

$$z_{\text{init}} = z_0, z_1, \dots, z_{\text{ter}} = z_p$$

Fig. 59



Form the sum

$$S = \sum_{j=1}^{p} f(\xi_j) \cdot (z_j - z_{j-1}) \tag{20}$$

where the point ξ_j is chosen arbitrarily on a subsection of the line (l) between z_{j-1} and z_j. The reader should bear in mind that the values of the function $f(\xi_j)$ and the quantities $z_j - z_{j-1}$ are complex numbers, and so when forming the sum (20) we perform operations with complex numbers, and the result, that is S, is also a complex number.

We use the term *integral* for the sum S, provided that the line (l) is subdivided into such small subsections that any further refinement in the partition does not alter the sum S (to put this more precisely, the limit of the sum for an infinite refinement of the partition of the curve). We denote the integral thus:

$$I = \int_{z_{\text{init}}}^{z_{\text{ter}}} f(z)\,dz$$

From the definition it follows that the integral is multiplied by -1 when the direction of integration is reversed, for then all differences $z_j - z_{j-1}$, and therefore the sum (20) as well, is multiplied by -1.

The following question arises. It is clear that the points z_{init} and z_{ter} may be joined in a variety of ways (Fig. 60). In forming the sum (20) for different curves joining the points z_{init} and z_{ter} we will have to do with distinct values of $f(\xi_j)$ and $z_j - z_{j-1}$. Does the integral I depend on the choice of path or does it depend only on the initial (z_{init}) and terminal (z_{ter}) points?

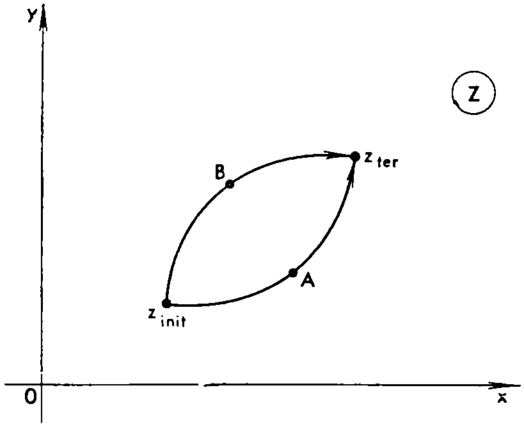


Fig. 60

It turns out that if $f(z)$ is an analytic function and if, in a region bounded by different paths, $f(z)$ never becomes infinite, then the integral does not depend on the choice of path.*

To prove this, denote the closed contour $z_{\text{init}}Az_{\text{ter}}Bz_{\text{init}}$ by the letter (L) and the integral $\oint f(z)\,dz$ by the letter I(the symbol $\oint$ denotes an integral over a closed contour). Since

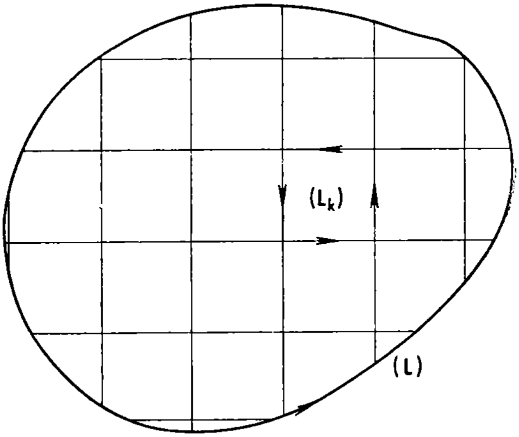
$$\begin{aligned} I &= \int\limits_{z_{\text{init}}Az_{\text{ter}}} f(z)\,dz + \int\limits_{z_{\text{ter}}Bz_{\text{init}}} f(z)\,dz \\ &= \int\limits_{z_{\text{init}}Az_{\text{ter}}} f(z)\,dz - \int\limits_{z_{\text{init}}Bz_{\text{ter}}} f(z)\,dz \end{aligned}$$

it suffices to verify that $I = 0$. Conversely, if integrals of $f(z)$ over the contours $z_{\text{init}}Az_{\text{ter}}$ and $z_{\text{init}}Bz_{\text{ter}}$ are equal, then $I = 0$. This means that the fact that the integral of an analytic function is independent of the path of integration is equivalent to the following assertion, which is known as *Cauchy's integral theorem:* The integral, extended over a closed curve, of a function that is analytic throughout the interior of the curve (contour) and on it is equal to zero.

To prove Cauchy's integral theorem, we partition the portion of the plane bounded by the contour (L) into small squares with contours (L_k) and traverse each contour counterclockwise, which

* Besides the condition that $f(z)$ not go to infinity, it is also necessary that $f(z)$ be a single-valued function or, at least, that when moving from one path to another we always make use of one branch of the function $f(z)$ (see the end of this section).

Fig. 61



is the sense of traversal of (L). (Fig. 61 depicts one of the contours.) Then

$$\oint_{(L)} f(z)\,dz = \sum_k \oint_{(L_k)} f(z)\,dz \qquad (21)$$

since in the right member all the integrals around the inner sides of the squares cancel out. If within (L_k) we take a point z_k, then on (L_k) we have

$$\frac{f(z)-f(z_k)}{z-z_k} = \frac{\Delta f}{\Delta z} \approx f'(z_k), \text{ or } \frac{f(z)-f(z_k)}{z-z_k} = f'(z_k) + \alpha$$

where α is an infinitesimal of the order of the length h of a side of a small square. From this we have

$$\oint_{(L_k)} f(z)\,dz = \oint_{(L_k)} [f(z_k) + f'(z_k)(z-z_k) + \alpha(z-z_k)]\,dz$$

The integral of the first two terms is easily taken; and since the integration is performed over a closed curve, it is equal to zero. This leaves only the integral of the third term, which has the order h^3, since the length of the contour of integration is of the order of h and the factor $z - z_k$ has the same order. But in the right-hand member of (21) the number of terms in the sum has the order $\frac{1}{h^2}$ and thus the entire sum has the order h. This means that the right side tends to zero, as $h \to 0$, when the squares become infinitely small. But it must be equal to the constant left member, that is, this constant is equal to zero, which is what we set out to prove.

Later on we will give a different proof of this important theorem via the methods of vector analysis (see Exercise 2 of Sec. 11.7).

Thus, for z_{init} fixed, $\int\limits_{z_{\text{init}}}^{z} f(z)\,dz$ depends only on the terminal point z of the path, which is to say, it is a function of z. Let us denote this function by $\Phi(z)$, then $\int\limits_{z_{\text{init}}}^{z} f(z)\,dz = \Phi(z)$.

Let us find the derivative of this function:

$$\frac{d\Phi(z)}{dz} = \frac{\Phi(z+dz) - \Phi(z)}{dz} = \frac{\int\limits_{z_{\text{init}}}^{z+dz} f(z)\,dz - \int\limits_{z_{\text{init}}}^{z} f(z)\,dz}{dz} = \frac{\int\limits_{z}^{z+dz} f(z)dz}{dz}$$

Consider $\int\limits_{z}^{z+dz} f(z)\,dz$. Since the numbers z and $z + dz$ are almost the same and $f(z)$ is a continuous function, $f(z)$ hardly changes at all as z varies within these limits. Therefore, $\int\limits_{z}^{z+dz} f(z)\,dz \approx f(z)\,(z + dz - z) = f(z)\,dz$ and this equation can be made as exact as desired by decreasing dz. Hence

$$\frac{d\Phi(z)}{dz} = \frac{f(z)dz}{dz} = f(z) \tag{22}$$

Formula (22) shows that the relationship between the integrand function $f(z)$ and the integral $\Phi(z)$ remains the same as in the case of a function of a real variable.

Also, we will show that the ordinary formula for computing an integral is retained:

$$\int\limits_{z_{\text{init}}}^{z_{\text{ter}}} f(z)\,dz = \Phi(z_{\text{ter}}) - \Phi(z_{\text{init}}) \tag{23}$$

where $\Phi(z)$ is *any* function satisfying relation (22).

Indeed, now $\int\limits_{z_{\text{init}}}^{z} f(z)\,dz = \Phi(z) + C$, where C is a constant. Here, setting $z = z_{\text{init}}$, we get $0 = \Phi(z_{\text{init}}) + C$ whence $C = -\Phi(z_{\text{init}})$. And so $\int\limits_{z_{\text{init}}}^{z} f(z)\,dz = \Phi(z) - \Phi(z_{\text{init}})$. Putting $z = z_{\text{ter}}$, we get (23).

The formulas (22) and (23) show that all the rules for finding integrals that hold for ordinary real integrals (see, for example, HM, Ch. 3) are also applicable to integrals of complex functions.

In the theory of analytic functions we often encounter multiple-valued functions. Consider for instance the function $w = \sqrt{z}$. In Sec. 5.4 we showed that this function has two values: if $z = re^{i\varphi}$, then

$$w_1 = \sqrt{r}e^{i\frac{\varphi}{2}}, \quad w_2 = \sqrt{r}e^{i\left(\frac{\varphi}{2}+\pi\right)}$$

If we choose one of these values (or, as we say, one branch of the function), we have the following interesting property. Let z go round the point $z = 0$ in a positive sense* and return to the original position. Then to φ is added 2π and the value w_1 becomes

$$\sqrt{r}e^{i\frac{\varphi+2\pi}{2}} = \sqrt{r}e^{i\left(\frac{\varphi}{2}+\pi\right)} = w_2$$

In the same way, w_2 will then go to

$$\sqrt{r}e^{i\left(\frac{\varphi+2\pi}{2}+\pi\right)} = \sqrt{r}e^{i\frac{\varphi}{2}} \cdot e^{2\pi i} = \sqrt{r}e^{i\frac{\varphi}{2}} \cdot 1 = w_1$$

Thus, the two branches of the function $w = \sqrt{z}$ continuously pass one into the other in circling the point $z = 0$, and if the point $z = 0$ is encircled twice, we return to the original branch.

The point $z = z_0$, which when encircled has one branch of a multiple-valued function replacing another, is called a *branch point*. Thus, for a function $w = \sqrt{z}$ the point $z = 0$ is a branch point of the second order (since there are two branches). The "infinite point" $z = \infty$ is generally said to be another branch point of this function. More than two branches can alternate at a branch point; for example, the function $w = \sqrt[n]{z}$ has n branches that continuously replace one another in circular order when encircling the branch point $z = 0$.

Another important example of a multivalued function is the function $w = \ln z$. In Sec. 5.4 we saw that this function has an infinite number of values: $w_k = \ln r + i(\varphi + 2k\pi)$ $(k = 0, \pm 1, \pm 2, \ldots)$. If the point z encircles the origin and 2π is added to φ, then the value w_0 goes into w_1, the value w_1 goes into w_2, and so on. If we again encircle the origin, we will pass to ever new branches and will never return to the original branch. Such a branch point is termed a branch point of infinite order.

In order to be able to consider a single branch independently of any other one, it is necessary in some way to prohibit the point z from making circuits about the branch points of the function at

* We assume the direction of traversal to be positive if the point $z = 0$ always remains on the left during the traversal.

hand. Ordinarily, to do this, one or several lines, called *branch cuts*, are drawn in the plane joining the branch points; these cuts cannot be crossed. For example, in considering the function $w = \sqrt{z}$ we can draw a branch cut along the positive real axis from the point $z = 0$ to infinity. If the point z varies in arbitrary fashion in the plane outside this cut, it cannot make a circuit about the branch point $z = 0$ and therefore that one branch cannot give way to another. After such a cut has been made, each branch may be regarded as a single-valued analytic function (although the branch has a discontinuity along the cut, it — the branch — takes on different values on different sides of the cut). In particular, we can apply the Cauchy integral theorem to such a branch.

From now on, for integrands we will consider only single-valued analytic functions. Incidentally, in Sec. 5.9 we will see that this does not save us from the necessity of considering multivalued functions that result from integration.

Exercise

Find the integrals of the following functions over the upper and lower semicircles going from $z = -1$ to $z = 1$ with centre at the point $z = 0$: (a) z^2, (b) $\frac{1}{z}$, (c) $\sqrt{z}$ (for the branch equal to i when $z = -1$). Explain the coincidence of results in (a) and the noncoincidence in (b) and (c).

5.9 Residues

To summarize: if an integral is independent of a path, then an integral over a closed circuit is equal to zero. Above we noted that an integral is not dependent on the path if the integrand does not become infinite. Let us consider an example in which the integrand becomes infinite.

Let $I = \oint \frac{dz}{z}$. Here, $f(z) = \frac{1}{z}$ becomes infinite when $z = 0$. We evaluate the integral over the closed path that makes a circuit of the point $z = 0$ in the positive direction (counterclockwise), say, around a circle (C) of radius r with centre at the coordinate origin (Fig. 62). On this circle, $z = re^{i\varphi}$, where r is the radius of the circle and the variable φ ranges from 0 to 2π. Then $dz = re^{i\varphi}\, i\, d\varphi$ and $\oint\limits_C \frac{dz}{z} = \int\limits_0^{2\pi} \frac{re^{i\varphi} i\, d\varphi}{re^{i\varphi}} = 2\pi i$. The integral round the closed circuit turned out not equal to zero.

We know that $\int \frac{dz}{z} = \ln z$. The fact that the integral $\oint \frac{dz}{z}$ round the closed circuit is not equal to zero is in remarkable agreement

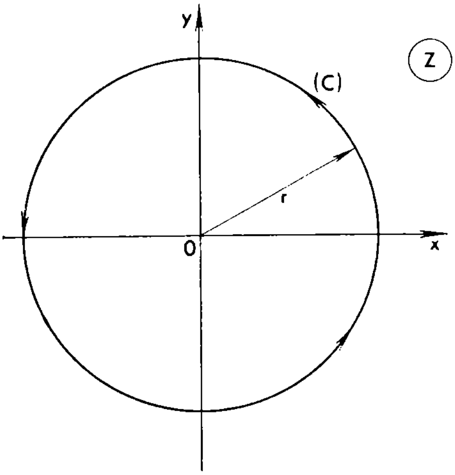


Fig. 62

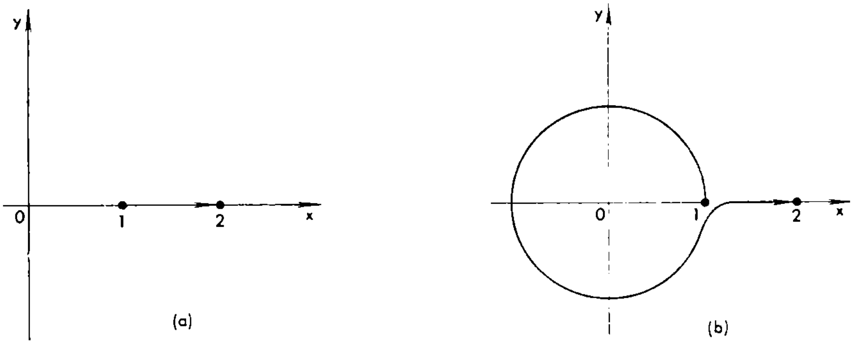


Fig. 63

with the multivalued nature of the function ln z. Consider, for example, $I = \int_1^2 \frac{dz}{z}$. Going from $z = 1$ to $z = 2$ via the shortest route (Fig. 63a), we find $I = \ln 2 - \ln 1 = \ln 2 = 0.69$. If we take a longer route: first one circuit about the origin and then to the desired point (Fig. 63b), we get $I_1 = 2\pi i + 0.69$. If we make n circuits about the origin, then we have $I_n = 2\pi i \cdot n + 0.69$.

In Sec. 5.4 we found out that, true enough, the quantities $2\pi i \cdot n + 0.69$ for all integral n serve as the logarithms of the number 2, since $e^{2\pi i \cdot n} = 1$. Thus, the multivalued nature of the logarithm is the result of a choice of distinct routes with different numbers of circuits about the point $z = 0$, where $1/z$ becomes infinite.

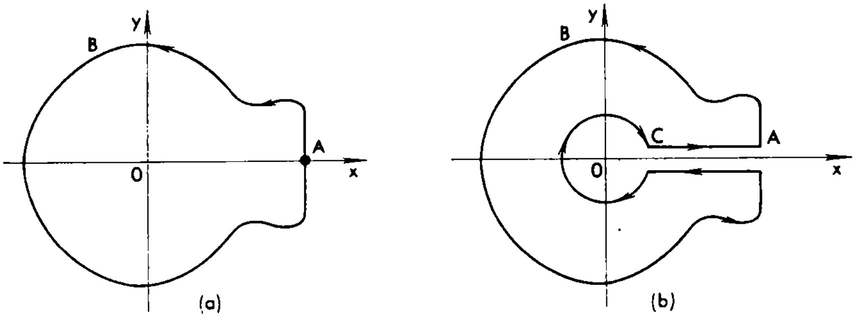


Fig. 64

The value of the integral depends on how many times and in which direction we made the circuit of the origin, but does not depend on the path traversed in any circuit. We now prove the last assertion. Let us try to find $I_A = \oint dz/z$ around some path ABA (Fig. 64a). Consider the integral $I_0 = \oint \frac{dz}{z}$ around the path shown in Fig. 64b. This path consists of ABA, two close-lying straight lines AC and CA and a circle of radius OC centred at the origin. $I_0 = 0$, since this is an integral around a closed circuit inside which $1/z$ does not become infinite anywhere.

The integral I_0 consists of I_A, two integrals $\int\limits_C^A dz/z$ and $\int\limits_A^C dz/z$ that mutually cancel out, and the integral around the circle of radius OC. Since the integration around the circle is in a direction opposite to that in which the angles are reckoned, the appropriate integral is equal to $-2\pi i$. Therefore $I_0 = 0 = I_A - 2\pi i$ or $I_A = 2\pi i$, which means that I_A coincides with the value of the integral around a circle of arbitrary radius.

By similar means it is possible to reduce integrals over curves that make calculation awkward to integrals around small circles about points that make the integrand go to infinity. Here, one should not think that the integral must definitely be different from zero. For example, in the integral

$$\oint \frac{1}{z^m}\,dz \quad (m = 2, 3, 4, \ldots) \tag{24}$$

the integrand has a singularity (it becomes infinite) at $z = 0$. However, this integral is equal to zero around any closed circuit whether it encloses the point or not (but does not pass through the point!). Indeed, in the given example, the indefinite integral is equal to $\frac{z^{-m+1}}{-m+1}+C$, which means it is a single-valued function; now the increment of a single-valued function around a closed circuit is equal to zero (why?).

For any m, the integral (24) around the circle $|z|=r$ can be evaluated in the following manner. Set $z=re^{i\varphi}$, then a few simple manipulation brings the integral to the form

$$ir^{1-m}\int_0^{2\pi} e^{i(1-m)\varphi}\,d\varphi$$

A straightforward evaluation shows that it is equal to zero for any integral $m \neq 1$. We excluded the case of a nonintegral m, because then the integrand is a multivalued function.

Let us consider another example. Suppose we have to evaluate the integral

$$\oint \frac{\cos z}{z^3}\,dz \tag{25}$$

around a closed circuit about the origin $z=0$ in the positive sense; in this example, the origin is a singular point for the integrand because this function becomes infinite at the origin.

Recalling the expansion of the function $\cos z$ about the point $z=0$ in a Taylor power series,

$$\cos z = 1-\frac{z^2}{2!}+\frac{z^4}{4!}-\frac{z^6}{6!}+\ldots$$

we can write

$$\frac{\cos z}{z^3}=\frac{1}{z^3}-\frac{1}{2!z}+\frac{z}{4!}-\frac{z^3}{6!}+\ldots \tag{26}$$

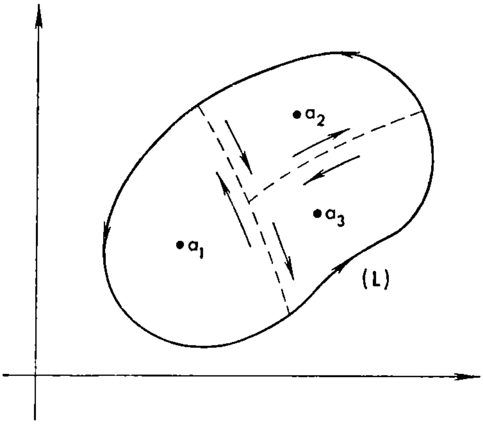
In this example, the integrand tends to infinity at the rate of $\frac{1}{|z|^3}$ as $z\to 0$. Such a singular point is termed a *pole of the third order.*

To evaluate the integral (25), carry out a term-by-term integration of the series (26). The indefinite integral of any term, except the second, yields a single-valued function (a power with integral exponent) and so the corresponding integral around a closed circuit is equal to zero. (For one thing, the integral of the first, principal, term of the expansion (26) is equal to zero.) By virtue of the foregoing, the integral of the second term is

$$\oint\left(-\frac{1}{2!z}\right)dz=-\frac{1}{2}\oint\frac{dz}{z}=-\frac{1}{2}2\pi i=-\pi i$$

Hence in this example the whole integral (25) is equal to $-\pi i$.

Fig. 65



Now let us consider a pole of general form. If a (single-valued) function $f(z)$ has a *pole of order n* at some point $z = a$, then, about this point, it can be expanded in what is known as a *Laurent series:*

$$f(z) = c_{-n}(z - a)^{-n} + c_{-n+1}(z - a)^{-n+1} + \dots + c_{-1}(z - a)^{-1}$$
$$+ c_0 + c_1(z - a) + c_2(z - a)^2 + \dots = \frac{c_{-n}}{(z - a)^n} + \frac{c_{-n+1}}{(z - a)^{n-1}} +$$
$$\dots + \frac{c_{-1}}{z - a} + c_0 + c_1(z - a) + c_2(z - a)^2 + \dots \qquad (27)$$

in terms of positive and negative integer powers of $z - a$, beginning with the nth power. Suppose it is required to evaluate the integral

$$\oint f(z)\,dz \qquad (28)$$

around a contour enclosing the point $z = a$ in the positive direction and not containing within it any other singular points except this one. As we have mentioned, it is possible to go from the given integral to an integral around a small circle centred at a, and near this point we can take advantage of the expansion (27). As in the preceding example, the integrals of all terms become zero after integration around the closed contour; that is, with the exception of $\oint \frac{c_{-1}}{z - a}\,dz = 2\pi i c_{-1}$. That is the value of the whole integral (28). The coefficient c_{-1} of the (-1)th power of $z - a$ in the Laurent expansion has a special name: the *residue* of the function $f(z)$ at the point a. Thus, the integral (28) is

$$2\pi i \operatorname{Res}_{z=a} f(z) \qquad (29)$$

Now let it be required to evaluate an integral of type (28) around a certain contour (L) (Fig. 65), where the integrand $f(z)$ is single-valued and analytic everywhere on the contour (L) and within it, with the exception of a certain number of singular points. (In Fig. 65 we have three such points: a_1, a_2, and a_3.) We draw auxiliary lines (shown dashed in Fig. 65) so that the region bounded by (L) is divided into parts, in each of which there is one singular point. Denote the contours of these parts that are traversed in the positive direction by (L_1), (L_2), and (L_3). It is then easy to verify that

$$\oint_{(L)} f(z)\,dz = \oint_{(L_1)} f(z)\,dz + \oint_{(L_2)} f(z)\,dz + \oint_{(L_3)} f(z)\,dz \tag{30}$$

because in the right-hand member the integrals taken along the auxiliary lines cancel out. Each of the contours (L_1), (L_2), (L_3) contains within it only one singular point and so each of the integrals on the right of (30) is evaluated by formula (29), and we get

$$\begin{aligned}\oint_{(L)} f(z)\,dz &= 2\pi i \operatorname{Res}_{z=a_1} f(z) + 2\pi i \operatorname{Res}_{z=a_2} f(z) + 2\pi i \operatorname{Res}_{z=a_3} f(z)\\ &= 2\pi i[\operatorname{Res}_{z=a_1} f(z) + \operatorname{Res}_{z=a_2} f(z) + \operatorname{Res}_{z=a_3} f(z)]\end{aligned} \tag{31}$$

To summarize: the integral (28) is equal to the product of $2\pi i$ by the sum of the residues of the integrand at all singular points located inside the contour of integration.

We now show how to compute a residue for the important case of a pole of the first order. A first-order pole ordinarily results if the integrand $f(z)$ is a ratio of two finite functions: $f(z) = g(z)/h(z)$, and at some point $z = a$ the numerator is different from zero and the denominator has a zero of the first order, that is, the expansion of the denominator in powers of $z - a$ begins with a first-degree term. Writing down the Taylor-series expansion of the numerator and denominator about $z = a$, we get

$$f(z) = \frac{g(a) + g'(a)\,(z-a) + \dfrac{g''(a)}{2}(z-a)^2 + \dots}{h'(a)\,(z-a) + \dfrac{h''(a)}{2}(z-a)^2 + \dots}$$

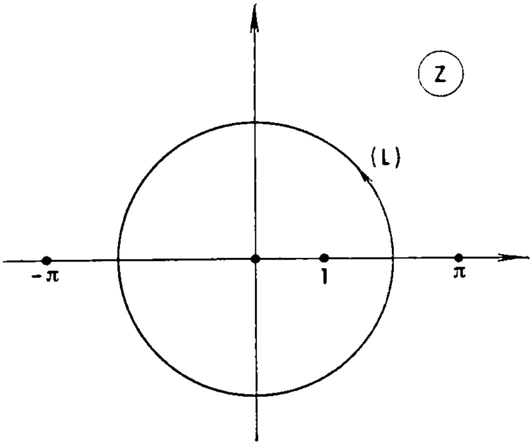
Near the point $z = a$ we can replace the right member by $\dfrac{g(a)}{h'(a)\,(z-a)}$. For this reason, the residue, that is, the coefficient of $(z-a)^{-1}$, is here equal to

$$\operatorname{Res}_{z=a} f(z) = \frac{g(a)}{h'(a)} \tag{32}$$

Let us consider an example. Suppose it is required to compute the integral

$$I = \oint_{(L)} \frac{z+1}{(z-1)\sin z}\,dz$$

Fig. 66



where (L) is a circle of radius 2 with centre at the coordinate origin (Fig. 66). In this case the indefinite integral cannot be expressed in terms of elementary functions, yet we will easily find the integral around the closed curve. Note that the integrand has singularities where its denominator vanishes, i.e. for $z = 1$ and $z = k\pi$ (k any integer). Of these points shown in Fig. 66, only two are inside (L): $z = 0$ and $z = 1$. Therefore, by (31),

$$I = 2\pi i[\text{Res}_{z=0}f(z) + \text{Res}_{z=1}f(z)] \tag{33}$$

Since at each of these points the denominator has a first-order zero and the numerator is nonzero, we have two poles of the first order and the residues at those points can be computed from formula (32). In the given example

$$g(z) = z + 1, \quad h(z) = (z - 1)\sin z, \quad h'(z) = \sin z + (z - 1)\cos z$$

whence $\text{Res}_{z=0}f(z) = \dfrac{g(0)}{h'(0)} = \dfrac{1}{-1} = -1, \ \text{Res}_{z=1}f(z) = \dfrac{g(1)}{h'(1)} = \dfrac{2}{\sin 1}$.

Substituting into (33), we get

$$I = 2\pi i\left(-1 + \frac{2}{\sin 1}\right) = 8.65i$$

Residue theory is also used to compute certain real-valued integrals with the aid of an artificial reduction of them to complex-valued ones. Consider, for example, the integral

$$I_1 = \int_{-\infty}^{\infty} \frac{\cos \omega x}{1 + x^2}\,dx \quad (\omega > 0) \tag{34}$$

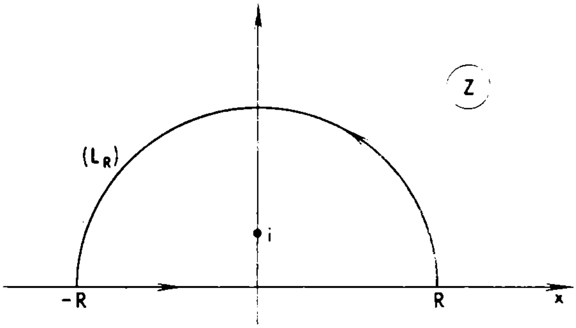


Fig. 67

Here again, the indefinite integral cannot be expressed in terms of elementary functions so that computing I_1 by any standard method is out of the question. To evaluate I_1, first of all note that

$$\int_{-\infty}^{\infty} \frac{\sin \omega x}{1+x^2}\, dx = 0 \tag{35}$$

as the integral of any odd integrand within limits that are symmetric about zero. From (34) and (35) it follows that

$$\int_{-\infty}^{\infty} \frac{e^{i\omega x}}{1+x^2}\, dx = \int_{-\infty}^{\infty} \frac{\cos \omega x + i \sin \omega x}{1+x^2}\, dx = I_1 \tag{36}$$

Now consider the auxiliary integral

$$I_R = \int_{(L_R)} \frac{e^{i\omega z}}{1+z^2}\, dz \tag{37}$$

where the contour (L_R) consists of the semicircle and its diameter shown in Fig. 67. Since the integrand has poles of the first order at the points $z = \pm i$, of which there is only inside (L_R), from formulas (31) and (32) we get

$$I_R = 2\pi i \operatorname{Res}_{z=i} \frac{e^{i\omega z}}{1+z^2} = 2\pi i \frac{e^{i\omega i}}{2i} = \pi e^{-\omega}$$

On the other hand, the integral (37) may be represented as a sum of the integral along the segment of the real axis equal to

$$\int_{-R}^{R} \frac{e^{i\omega x}}{1+x^2}\, dx \tag{38}$$

and the integral around the semicircle, which we denote by (L'_R). Let us estimate this latter integral. From the definition of an integral it is easy to derive the following estimate:

$$\left|\int\limits_{(L)} f(z)\,dz\right| \leqslant \max_{(L)} |f(z)| \cdot \text{length } (L)$$

Therefore

$$\left|\int\limits_{(L'_R)} \frac{e^{i\omega z}}{1+z^2}\,dz\right| \leqslant \max_{(L'_R)} \frac{|e^{i\omega z}|}{|1+z^2|} \cdot \pi R \tag{39}$$

Representing $z = x + iy$ and noting that $y \geqslant 0$ on (L'_R), we find, on (L'_R), $|e^{i\omega z}| = |e^{i\omega(x+iy)}| = |e^{i\omega x}|\,|e^{-\omega y}| = e^{-\omega y} \leqslant 1$. On the other hand, on (L'_R) we have

$$|1+z^2| = \left|z^2\left(1+\frac{1}{z^2}\right)\right| = R^2\left|1+\frac{1}{z^2}\right| \approx R^2$$

(for large $R = |z|$). Hence the right member and, with it, the left member of (39) too tend to zero as $\frac{1}{R^2}R = \frac{1}{R}$ as $R \to \infty$.

Summarizing, we see that $I_R = \pi e^{-\omega}$ may be represented as the sum of the integral (38) (which in the limit tends to (36), that is, to I_1, as $R \to \infty$) and the integral around (L'_R) that tends to zero. Passing to the limit as $R \to \infty$, we get $I_1 = \pi e^{-\omega}$, or

$$\int\limits_{-\infty}^{\infty} \frac{\cos \omega x}{1+x^2}\,dx = \pi e^{-\omega} \tag{40}$$

This last formula is not obvious in the least and to obtain it without invoking complex numbers is extremely difficult and requires consummate skill. With the aid of the properties of integrals of complex functions it is possible to find many integrals of that kind of real-valued functions via standard procedures that only require care and accuracy.

In practice, the integral just considered is obtained in the problem of exciting oscillations in a system with a natural frequency ω under the action of a force varying with time by the law $f(t) = \frac{1}{1+t^2}\cdot$ This result gives the law of diminishing amplitude of the oscillations being excited as the natural frequency ω of a system is increased (see Sec. 7.5).

It is interesting to view formula (40) from the standpoint of the results of Sec. 3.4. This is precisely the case where the function $f(x) = \frac{1}{1+x^2}$ vanishes together with all its derivatives at the endpoints $x = \pm\infty$ of the interval of integration. Since all these derivatives are devoid of discontinuities, from Sec. 3.4 it follows that integral $I_1(\omega)$ tends to zero, as $\omega \to 0$, faster than any negative power of ω; however, the exact rate of this approach is only found through the use of residue theory. The measure of nonsmoothness of the function $f(x)$ that determines this rate is the distance from the pole of this function, which is continued into the complex plane, to the imaginary axis.

Of course, so brief a section as this could not, unfortunately, teach the techniques of integrating complex functions, but we hope that it has given the reader a feeling of the beauty of the theory of functions of a complex variable.

Exercises

1. Evaluate the integral $\oint\limits_{(L)} \frac{dz}{e^z - 1}$ around a circle (L) of radius 4 with centre at the point $3i$.
2. Compute the integral $\int\limits_{-\infty}^{\infty} \frac{dx}{x^6 + 1}$.

ANSWERS AND SOLUTIONS

Sec. 5.1

$\sqrt{2}\left(\cos\frac{7\pi}{4} + i\sin\frac{7\pi}{4}\right)$, $5(\cos\varphi + i\sin\varphi)$ $\left(\varphi = \operatorname{arc\,tan}\frac{4}{3}\right)$, $2\left(\cos\frac{3\pi}{2} + i\sin\frac{3\pi}{2}\right)$, $3(\cos\pi + i\sin\pi)$, $1(\cos 0 + i\sin 0)$, $0(\cos\varphi + i\sin\varphi)$ (φ arbitrary).

Sec. 5.2

1. If $\frac{z_1}{z_2} = w$, then $z_1 = z_2 w$, whence $z_1^* = (z_2 w)^* = z_2^* w^*$, i.e. $w^* = \frac{z_1^*}{z_2^*}$.
2. Since the second root is equal to $2 + i$, the left-hand member must be divisible by $[z - (2 - i)]\,[z - (2 + i)] = z^2 - 4z + 5$. Performing the division and solving the remaining quadratic equation, we get $z_{3,4} = 1 \pm \sqrt{3}$.
3. If $z = x + iy$, then $z^* = x - iy = x + i(-y)$ and $z^{**} = x - i(-y) = x + iy = z$.
4. $zz^* = x^2 + y^2 = |z|^2$. It is interesting that the real nature of the product zz^* may be proved on the basis of the properties of conjugate numbers without passing to the real and imaginary

parts: $(zz^*)^* = z^*z^{**} = z^*z = zz^*$. Now a number equal to its conjugate is of necessity real.

Sec. 5.3

1. (a) $1 + i = \sqrt{2}e^{i\frac{\pi}{4}}$, (b) $1 - i = \sqrt{2}e^{i\frac{7\pi}{4}} = \sqrt{2}e^{-i\frac{\pi}{4}}$, (c) $-1 = e^{i\pi}$, (d) $3i = 3e^{i\frac{\pi}{2}}$.
2. (a) First write the number $1 + i$ in exponential form to get $1 + i = \sqrt{2}e^{i\frac{\pi}{4}}$, whence $(1 + i)^{16} = (\sqrt{2}e^{i\frac{\pi}{4}})^{16} = (\sqrt{2})^{16} \cdot e^{i\frac{\pi}{4}\cdot 16} = 2^8 \cdot e^{i4\pi} = 2^8 = 256$, (b) -1.
3. $\cos 3\varphi = \cos^3 \varphi - 3 \cos \varphi \sin^2 \varphi$, $\sin 4\varphi = 4 \cos^3 \varphi \sin \varphi - 4 \cos \varphi \sin^3 \varphi$.
4. In Euler's formula $e^{i\varphi} = \cos \varphi + i \sin \varphi$, we replace φ by $-\varphi$ to get $e^{-i\varphi} = \cos \varphi - i \sin \varphi$. Adding these two formulas term by term, we get $e^{i\varphi} + e^{-i\varphi} = 2 \cos \varphi$, whence $\cos \varphi = \dfrac{e^{i\varphi} + e^{-i\varphi}}{2}$. The second formula is obtained by term-by-term subtraction.

Sec. 5.4

1. (a) $\ln(-1) = i(\pi + 2k\pi)$, (b) $\ln i = i\left(\frac{\pi}{2} + 2k\pi\right)$, (c) $\ln(-i) = i\left(\frac{3\pi}{2} + 2k\pi\right)$, (d) $\ln(1 + i) = \frac{1}{2}\ln 2 + i\left(\frac{\pi}{4} + 2k\pi\right)$.
2. Since $-1 = \cos \pi + i \sin \pi$, the general form of numbers x_k such that $x_k^3 = -1$ is

$$x_k = \cos \frac{\pi + 2k\pi}{3} + i \sin \frac{\pi + 2k\pi}{3} \quad (k = 0, 1, 2)$$

Therefore

$$x_0 = \cos \frac{\pi}{3} + i \sin \frac{\pi}{3} = \frac{1}{2} + i \frac{\sqrt{3}}{2}, \quad x_1 = \cos \pi + i \sin \pi = -1,$$

$$x_2 = \cos \frac{5\pi}{3} + i \sin \frac{5\pi}{3} = \frac{1}{2} - i \frac{\sqrt{3}}{2}.$$

3. $1, \frac{1}{2} + i\frac{\sqrt{3}}{2}, -\frac{1}{2} + i\frac{\sqrt{3}}{2}, -1, -\frac{1}{2} - i\frac{\sqrt{3}}{2}, \frac{1}{2} - i\frac{\sqrt{3}}{2}$.

4. On the basis of Exercise 4, Sec. 5.3, we arrive at the equation $\dfrac{e^{i\varphi} + e^{-i\varphi}}{2} = 2$, whence, after a few simple manipulations, we get

$$e^{2i\varphi} - 4e^{i\varphi} + 1 = 0$$

This is a quadratic equation in $e^{i\varphi}$; solving it, we get $e^{i\varphi} = 2 \pm \sqrt{3}$. From this we have $i\varphi = \ln(2 \pm \sqrt{3}) + 2k\pi i$ (k any integer).

Dividing by i and noting that $2 - \sqrt{3} = \frac{1}{2 + \sqrt{3}}$ and therefore $\ln(2 - \sqrt{3}) = -\ln(2 + \sqrt{3})$, we get the final answer: $\varphi = \pm \ln(2 + \sqrt{3})\, i + 2k\pi = \pm 1.317i + 2k\pi$. All these solutions are imaginary.

5. In similar fashion we obtain $\varphi = \pm \ln(2 + \sqrt{3})\, i + \frac{\pi}{2} + 2k\pi$.

Sec. 5.5

1. $J = \frac{\varphi_0 e^{i\beta}}{R + i\left(\omega L - \frac{1}{\omega C}\right)} e^{i\omega t}$. We thus get the same formulas as for the case of the RL circuit discussed in the text; but in place of ωL we have to substitute $\omega L - \frac{1}{\omega C}$. In the special case of $\omega^2 = \frac{1}{LC}$, we get $\omega L - \frac{1}{\omega C} = 0$, which is to say the inductance and capacitance cancel, as it were.

2. $J = \left(\frac{1}{R} + \frac{1}{i\omega L}\right) \varphi_0 e^{i\beta} e^{i\omega t}$, whence

$$j_0 = \sqrt{\frac{1}{R^2} + \frac{1}{\omega^2 L^2}}\, \varphi_0, \quad \alpha = \beta + \arg(R + i\omega L) - \frac{\pi}{2}$$

for the same meanings of the notation as in the text.

Sec. 5.6

$f(z) = (x + iy)^3 = x^3 + i3x^2y - 3xy^2 - iy^3$, whence

$u = x^3 - 3xy^2$, $v = 3x^2y - y^3$, $\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} = 3x^2 - 3y^2$,

$\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y} = 6xy$.

Sec. 5.7

1. From the condition $u = x + x^2 - y^2$ we get $\frac{\partial u}{\partial x} = 1 + 2x$. Using the first of the Cauchy-Riemann conditions, we get $\frac{\partial v}{\partial y} = 1 + 2x$. To find v from this condition, it is sufficient to integrate the equation with respect to y, holding x constant. This yields

$v(x, y) = y + 2xy + \varphi(x)$ *

* Since x is constant in the integration, the role of the constant of integration can be taken by any function φ that is dependent solely on the single variable x.

We now find $\frac{\partial v}{\partial x} = 2y + \varphi'(x)$. But according to the second of the Cauchy-Riemann conditions, $\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}$. Since $\frac{\partial u}{\partial y} = -2y$, we get $2y + \varphi'(x) = 2y$, whence $\varphi'(x) = 0$. That is, $\varphi(x) \equiv C$ where C is a constant. Hence $v(x, y) = y + 2xy + C$.
To determine C take advantage of the condition $f(0) = 0$. It means that $u = 0$, $v = 0$ for $x = 0$, $y = 0$. Thus, $v = 0$ for $x = 0$, $y = 0$, and so $C = 0$, $v(x, y) = y + 2xy$.

2. $u(x, y) = -x^2 + y^2 + 1$.

Sec. 5.8

(a) $\frac{2}{3}$, $\frac{2}{3}$, (b) $-\pi i$, πi, (c) $\frac{2}{3}(1 + i)$, $\frac{2}{3}(-1 + i)$ (we first write the result for the upper semicircle and then for the lower one). In the case of (a) the integrand is everywhere analytic and therefore the integral is independent of the path of integration. In case (b) the integrand is infinite at $z = 0$, in case (c) it has a branch point there.

Sec. 5.9

1. The integrand has two first-order poles inside (L): $z_1 = 0$, $z_2 = 2\pi i$. Therefore the integral is equal to

$$2\pi i\left(\operatorname{Res}_{z=0}\frac{1}{e^z - 1} + \operatorname{Res}_{z=2\pi i}\frac{1}{e^z - 1}\right) = 2\pi i\left(\frac{1}{e^0} + \frac{1}{e^{2\pi i}}\right) = 4\pi i$$

2. Like Example (34), the integral here is equal to the integral of the function $f(z) = \frac{1}{1 + z^6}$ over the contour (L_R) of Fig. 67 for large R. But within (L_R), $f(z)$ has three simple poles: $z_1 = \frac{\sqrt{3}}{2} + i\frac{1}{2}$, $z_2 = i$, $z_3 = -\frac{\sqrt{3}}{2} + i\frac{1}{2}$ and so the integral is equal to

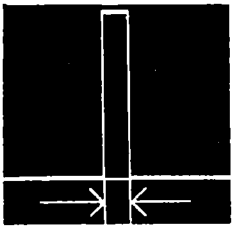
$$2\pi i\left(\frac{1}{6z_1^5} + \frac{1}{6z_2^5} + \frac{1}{6z_3^5}\right)$$

$$= \frac{2\pi i}{6}\left(\frac{z_1}{z_1^6} + \frac{z_2}{z_2^6} + \frac{z_3}{z_3^6}\right) = -\frac{\pi i}{3}(z_1 + z_2 + z_3) = \frac{2}{3}\pi$$

In this example, the indefinite integral can be expressed in terms of elementary functions, but the method of evaluation given here is much simpler.

Chapter 6

DIRAC'S DELTA FUNCTION[*]



6.1 Dirac's delta function $\delta(x)$

Take the function $y = \Phi_1(x)$ with a maximum at $x = 0$ and rapidly decreasing on both sides of $x = 0$; a function such that

$$\int_{-\infty}^{+\infty} \Phi_1(x)\,dx = 1$$

These conditions do not in any way determine the type of function $\Phi_1(x)$, because it is easy to think up a variety of functions that satisfy all the above requirements. For example,

$$\Phi_1(x) = \frac{1}{\pi}\,\frac{1}{1+x^2} \tag{1}$$

$$\Phi_1(x) = \frac{1}{\sqrt{\pi}}\,e^{-x^2}\,{}^{**} \tag{2}$$

The numerical factor ensures that the integral equals unity. The graphs of these functions are shown in Fig. 68. Now transform the curve $y = \Phi_1(x)$ as follows: increase the height m-fold and diminish the width the same number of times. It will be recalled (see, for example, HM, Sec. 1.7) that if we increase the height of a curve $y = \Phi_1(x)$ m times, then the equation of the curve takes the form $y = m\,\Phi_1(x)$ and if we diminish the width m times, the equation becomes $y = \Phi_1(mx)$. Thus, after both transformations, the equation of the curve becomes $y = \Phi_m(x) = m\Phi_1(mx)$. For example, from (1) we get $\Phi_m(x) = \dfrac{m}{\pi}\,\dfrac{1}{1+(mx)^2}$. It is clear that the area between the curve and the x-axis increases m-fold when stretched upwards and diminishes the same number of times when squeezed at the sides, which means that it remains unchanged. Incidentally, this can readily

[*] This chapter is closely related to the material of HM, Ch. 9.

[**] It can be demonstrated that $\int_{-\infty}^{+\infty} e^{-x^2}\,dx = \sqrt{\pi}$. See Sec. 4.7.

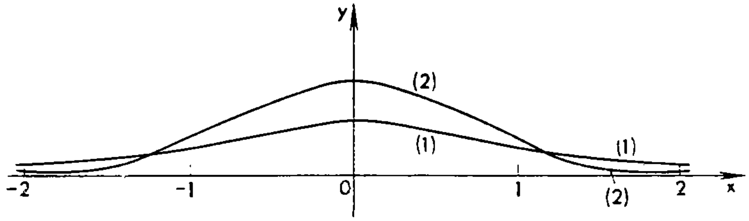


Fig. 68

be demonstrated also by integration after changing the variable of integration $mx = s$:

$$\int_{-\infty}^{\infty} \Phi_m(x)\, dx = \int_{-\infty}^{\infty} m\Phi_1(mx)\, dx = \int_{-\infty}^{\infty} \Phi_1(mx)\, d(mx)$$

$$= \int_{-\infty}^{\infty} \Phi_1(s)\, ds = \int_{-\infty}^{\infty} \Phi_1(x)\, dx$$

What form does the transformed function take for very large m or, to put it more precisely, in the limit, when m increases beyond all bounds? For any fixed $x \neq 0$, the quantity $y = m\Phi_1(mx)$ will approach zero without bound as m increases without bound because $\Phi_1(mx)$ decreases faster, with increasing m, than the growth of the factor m. For this it is necessary that, as $x \to \pm\infty$, $\Phi_1(x)$ approach zero faster than $\frac{1}{x}$ (this is what is meant by saying that the function is rapidly decreasing one).* For example, in the expression $\Phi_m(x) = \frac{m}{\pi} \frac{1}{1+(mx)^2}$ for a given $x \neq 0$ and for sufficiently great m it will be true that $(mx)^2 \gg 1$ and, hence, $\Phi_m(x) \approx \frac{m}{\pi} \frac{1}{m^2x^2} = \frac{1}{\pi m x^2}$, which is to say that it decreases indefinitely as m increases. The quantity $\Phi_m(x)$ obtained from formula (2) decreases faster than m increases. Indeed, in this case $\Phi_m(x) = \frac{m}{\sqrt{\pi}} e^{-(mx)^2}$, and we know that an exponential function with a negative exponent decreases faster than any power of m (see, for example, HM, Sec. 3.21).

Now let $x = 0$. Then $\Phi_1(mx) = \Phi_1(0)$ is constant for any m and, therefore, $\Phi_m(0) = m\Phi_1(0)$ increases without bound with increasing m.

* Such a rate of decrease automatically follows from the convergence of the integral (see Sec. 3.1).

Consequently, by increasing m without bound we obtain a function with the following properties:

(1) the function is zero for all $x < 0$ and for all $x > 0$;

(2) the function is infinite when $x = 0$;

(3) the integral of this function taken from $-\infty$ to $+\infty$ is equal to 1.

A function having these properties is termed the *Dirac delta function*, denoted by $\delta(x)$.* The delta function is extremely convenient and is broadly used today in physics.

We arrived at the concept of the delta function by considering ordinary familiar functions and transforming them in a special way. It is a remarkable fact, however, that to use the delta function it suffices to know the three properties that we listed above and one does not at all need to know from which function $\left(\frac{1}{\pi}\,\frac{1}{1+x^2}\right.$ or $\frac{1}{\sqrt{\pi}}e^{-x^2}$ or any other one) the delta function is derived. Crudely stated, the delta function is a function that assumes large values on a narrow interval, and these values are in agreement with the width of the interval in such a manner that condition (3) holds true. (From this there follows, for one thing, that the dimensions of $[\delta(x)] = 1/[x]$.)

From the properties of $\delta(x)$ follows the basic relation

$$I = \int_{-\infty}^{+\infty} \delta(x)\, f(x)\, dx = f(0) \tag{3}$$

Indeed, $\delta(x) = 0$ for all $x \neq 0$ and so

$$I = \int_{-\infty}^{+\infty} \delta(x) f(x)\, dx = \int_{-\varepsilon}^{+\varepsilon} \delta(x)\, f(x)\, dx$$

where ε is a small quantity. In the latter integral the interval of integration is small (it is of length 2ε), and so, there, $f(x) \approx f(0)$; consequently

$$I = \int_{-\varepsilon}^{+\varepsilon} \delta(x)\, f(x)\, dx = \int_{-\varepsilon}^{+\varepsilon} \delta(x)\, f(0)\, dx = f(0) \int_{-\varepsilon}^{+\varepsilon} \delta(x)\, dx$$

$$= f(0) \int_{-\infty}^{+\infty} \delta(x)\, dx = f(0)$$

* Paul Adrien Maurice Dirac, in honour of whom this function is named, is the celebrated English theoretical physicist who in 1929 predicted the existence of antiparticles: the positron, the antiproton, and others, which were later discovered experimentally.

To summarize, then, the formula (3) follows from the three properties of $\delta(x)$. The converse is valid as well: if we define $\delta(x)$ as a function for which the relation (3) holds for any $f(x)$, then from this follow all the three properties of the delta function. Without dwelling on a detailed proof of this fact, we will show, for example, that (3) implies the first property: $\delta(x) = 0$ for $x \neq 0$.

Indeed, from (3) it is clear that the value of the integral does not depend on the behaviour of the function $f(x)$ for $x \neq 0$, but depends only on $f(0)$. This means that $f(x)$ stands under the integral sign with a factor equal to zero for $x \neq 0$, i.e. $\delta(x) = 0$ for $x \neq 0$.

Note that $\delta(x - a)$ is different from zero (it is infinite) only when $x=a$. Reasoning as in the derivation of formula (3), we get the formula

$$\int_{-\infty}^{+\infty} \delta(x-a) f(x)\, dx = f(a) \tag{4}$$

It is worth noting certain other interesting formulas for the delta function. First,

$$\delta(ax) = \frac{1}{|a|}\delta(x) \quad (a = \text{constant} \neq 0)$$

Indeed, the function $|a|\,\delta(ax)$ satisfies all three properties that define a delta function. The slightly less obvious third property is verified via the substitution $ax = x_1$ thus:

$$a > 0, \quad \int_{-\infty}^{\infty} |a|\,\delta(ax)\, dx = \int_{-\infty}^{\infty} \delta(ax)\, a\, dx = \int_{-\infty}^{\infty} \delta(x_1)\, dx_1 = 1$$

$$a < 0, \quad \int_{-\infty}^{\infty} |a|\,\delta(ax) dx = -\int_{-\infty}^{\infty} \delta(ax)\, a\, dx$$

$$= -\int_{\infty}^{-\infty} \delta(x_1)\, dx_1 = \int_{-\infty}^{\infty} \delta(x_1)\, dx_1 = 1$$

And yet another property:

$$\delta(\varphi(x)) = \frac{1}{|\varphi'(x_0)|}\delta(x - x_0)$$

if $\varphi(x)$ vanishes only when $x = x_0$. This property follows from the preceding one since close to $x = x_0$, to within higher infinitesimals, we can write

$$\varphi(x) = \varphi(x_0) + \varphi'(x_0)\,(x - x_0) = \varphi'(x_0)\,(x - x_0)$$

Finally,

$$f(x)\,\delta(x-a) = f(a)\,\delta(x-a)$$

which follows immediately from the fact that the function $\delta(x - a)$ is equal to zero for $x \neq a$.

Many physical relations are very conveniently written with the aid of the delta function. Consider, for instance, a narrow rod with loads suspended at a variety of points. Let the dimensions of the loads be small compared to the length of the rod, and their masses of the same order as that of the rod. Then when solving problems (determining the total mass, the equilibrium position, and the like), one has to consider both the masses of the loads (called the localized masses) and the mass of the rod (distributed mass). Suppose the density of the rod is $\rho_{\text{dis}}(x)$.[*] Then the mass of the rod, one endpoint of which lies at the coordinate origin and the other end at the point $x = l$, is equal to $m_{\text{dis}} = \int_0^1 \rho_{\text{dis}}(x)\,dx$.

Suppose there is a weight m_a at the point $x = a$ of the rod. Then the total mass of the rod and the weight is

$$m = m_a + \int_0^l \rho_{\text{dis}}(x)\,dx$$

Using the delta function, we can represent the localized mass as a mass distributed with density $\rho_{\text{loc}}(x) = m_a\delta(x - a)$. Indeed, the last formula means that the density is different from zero only in the small neighbourhood of the point $x = a$, and

$$\int_0^l \rho_{\text{loc}}(x)\,dx = m_a \int_0^l \delta(x - a)\,dx = m_a$$

This means that by introducing the function $\rho_{\text{loc}}(x)$ we can write the localized mass in a form that in aspect coincides with the notation for a distributed mass.

Now set

$$\rho(x) = \rho_{\text{dis}}(x) + \rho_{\text{loc}}(x) = \rho_{\text{dis}}(x) + m_a\delta(x - a)$$

Then the total mass is

$$m = \int_0^l \rho(x)\,dx$$

In other words, the total mass need not be written as the sum of terms of different types, but, via the integral, in the same manner as the distributed mass. The different nature of the distributed and localized masses only affects the aspect of the function $\rho(x)$. Thanks to this, we can now write in a unified manner all the relations that we have already obtained for a distributed mass.

[*] The subscripts "dis" and "loc" stand for "distributed" and "localized", respectively.

For example, the mass between the points $x = b$ and $x = c$ of the rod is equal to $\int_b^c \rho(x)\,dx$. No stipulations need be made now that $b > a$ or that $b < a$, and so forth: the function $\rho(x)$ contains $\delta(x - a)$, and integration of the function $\delta(x - a)$ automatically adds to the mass of the rod the mass of the load, provided that the load is located between the points $x = b$ and $x = c$.

In the same way, the position of the centre of gravity of the rod is given by the formula $x_c = \dfrac{\int_0^l x\rho(x)\,dx}{\int_0^l \rho(x)dx}$, irrespective of whether there is a localized mass on the rod or not.

The case of several localized masses is considered in exactly the same way.

Let us take another example. In mechanics one considers forces smoothly varying with time and also sharp impacts, collisions of bodies. In the case of an impact, the body is acted on by a large force during a brief time interval. In most cases, any detailed study of the dependence of the force on the time during which the impact lasted is of no interest (see, for instance, HM, Sec. 6.5). It is enough to know the impulse i.e. $I_{\text{imp}} = \int F\,dt$. Then the force at impact can be written thus: $F(t) = I_{\text{imp}}\delta(t - t_{\text{imp}})$, where t_{imp} is the time of impact and I_{imp} is the impulse of the impact. This notation shows that the force is different from zero only at the instant of impact and the impulse is equal to I_{imp}.

Note in conclusion that the delta function is considered only for real values of the independent variable.

Exercises

1. Evaluate $\int_{-\infty}^{\infty} x^2\delta(x - 3)\,dx$.
2. Simplify the expressions:
   (a) $(x^2 + 3)\,\delta(x + 5)$; (*b*) $\delta(2x - 8)$;
   (c) $\delta(x^2 + x - 2)$.

6.2 Green's function

Let us first consider an example. Suppose a thin flexible string of length l is stretched along the x-axis by a constant force F. In the system depicted in Fig. 69, this tension is accomplished by a block

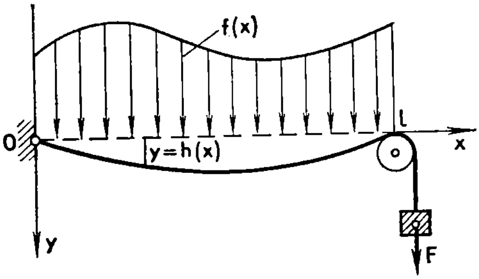


Fig. 69

and weight. Let the string be acted on perpendicular to the x-axis by a force distributed with density $f(x)$; that is, a force $f(x)\,dx$ operates over a small section of the string between the points x and $x + dx$, and the force $\int_0^l f(x)\,dx$ over the whole string. Let us find the shape of $y(x)$ that the string then takes up. Here the function $y(x)$ is the deflection of that point of the string which was at x of the x-axis in the original state.

We will assume that the tension F of the string is much greater than the total force acting on the string so the deflection is slight. Then we can take advantage of the law of linearity, according to which the corresponding deflections in the case of several superimposed weights are additive.

To begin with, let us suppose that the applied weight is of a special type, namely , that it is a unit localized weight applied to a certain *point of action* ξ of the x-axis; we denote by $y = G(x, \xi)$ the corresponding deflection at any *point of observation* x (Fig. 70). This function $G(x, \xi)$ is called the *influence function* or *Green's function* (after the English mathematician George Green (1793—1841)) of this problem. We will now show that if it is known, then it is also easy to find the deflection due to the action of an arbitrary weight with density $f(x)$.

Consider the load on the section of the axis from point ξ to point $\xi + d\xi$. It is equal to $f(\xi)\,d\xi$; and so the deflection due to it at the point x is equal to $G(x, \xi)\,f(\xi)\,d\xi$, since it follows from the law of linearity that if an external load is multiplied by a constant factor, then the deflection is likewise multiplied by that factor. Adding together such infinitesimal deflections due to all elements of the load from $\xi = 0$ to $\xi = l$, we get the overall deflection (see Fig. 69):

$$y = h(x) = \int_0^l G(x, \xi)\,f(\xi)\,d\xi \qquad (5)$$

In the example at hand it is easy to write down the function $G(x, \xi)$ in explicit form. Find the components of the tensile force of

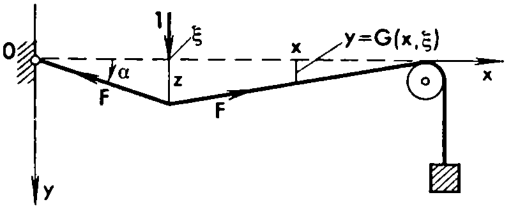


Fig. 70

the string along the y-axis. To the left of the point ξ it is equal (see Fig. 70) to

$$-F\sin\alpha = -F\frac{z}{\xi}$$

where z is the deflection of the point ξ which is not specified beforehand. Note that in this derivation we took advantage of the small nature of the deflections and for this reason we replaced the hypotenuse of the triangle by the larger leg when computing the sine. In a similar manner, we obtain the component of the tensile force to the right of ξ:

$$-F\frac{z}{l-\xi}$$

If under the given force the string is in equilibrium, this means that the sum of all forces acting on the string, that is, the sum of the forces of tension and the given force, is equal to zero. For this reason, the sum of the components of these forces along the y-axis is also zero. Since in our case the given force is equal to 1 and acts along the y-axis, we get, on the basis of the foregoing,

$$1-F\frac{z}{\xi}-F\frac{z}{l-\xi}=0$$

whence we find $z=\frac{\xi(l-\xi)}{Fl}$. If z is known, then the deflection of any point of the string can readily be found by using the fact that the string has the shape of a broken line. We obtain

$$y(x)=z\frac{x}{\xi} \qquad \text{if } x<\xi,$$

$$y(x)=z\frac{l-x}{l-\xi} \qquad \text{if } x>\xi$$

Substituting the value of z just found and recalling that the shape of the deflection under a unit localized load yields Green's function, we get

$$G(x,\xi)=\begin{cases}\frac{1}{Fl}\,x(l-\xi) & \text{if } x<\xi,\\ \frac{1}{Fl}\,\xi(l-x) & \text{if } x>\xi\end{cases}$$

This expression for Green's function may be substituted into (5) for the deflection due to an arbitrary load. Since $G(x, \xi)$ for $\xi < x$ and $\xi > x$ is written with the aid of different formulas, the integral is split into two integrals:

$$y = h(x) = \int_0^x G(x, \xi)\, f(\xi)\, d\xi + \int_x^l G(x, \xi)\, f(\xi)\, d\xi$$

$$= \frac{l-x}{Fl}\int_0^x \xi f(\xi)\, d\xi + \frac{x}{Fl}\int_x^l (l-\xi)\, f(\xi)\, d\xi$$

The same result may be obtained by writing the differential equation for the function $y(x)$ and solving it. However, the remarkable thing is that we succeeded in finding the solution without even writing down the differential equation. All we needed to know was that the law of linearity is applicable.

Now let us take up the general scheme for constructing the influence function. Let the external action exerted on an object be described by the function $f(x)$ ($a \leqslant x \leqslant b$; in the example above this was the function $f(x)$) and let the result of the action be given by the function $F(x)$ (this was $h(x)$ in the foregoing example). We can imagine that to every given function f there corresponds a new function F, that is to say, every function f is transformed via some specific law into a new function F. In mathematics, such a law of transformation of *preimage functions* into *image functions* is termed an *operator.* For example, the familiar differentiation operator D, which operates via the law $Df = f'$, i.e. $D(\sin x) = \cos x$, $D(x^3) = 3x^2$, and so on. Here, $\sin x$ is the preimage (inverse image) that is transformed by the operator D into the image $\cos x$. The notion of operator is similar to that of function, but whereas the function gives a law for the transformation of numbers (values of the independent variable) into numbers (values of the dependent variable), an operator transforms functions into functions.

Let us denote by L the transition operator from the function of external action $f(x)$ to the response function $F(x)$, so that $F = Lf$. We assume that the *law of linearity* (or the *principle of superposition*) holds: this means that when external actions are added, so are their results. This law holds true with sufficient accuracy when the external actions are not too great. It can be written as follows:

$$L(f_1 + f_2) = Lf_1 + Lf_2 \tag{6}$$

An operator having this property is called a linear operator. (To illustrate, verify that the differentiation operator is linear.) From this property we can conclude that the result obtained by multiplying an external action by a constant is also multiplied by that constant:

$$L(Cf) = CLf \quad (C \text{ constant}) \tag{7}$$

We do not give the proof here. (Try to justify it by first assuming C to be a positive integer, then set $C = \frac{1}{n}$ $(n = 2, 3, 4, \ldots)$ and $C = \frac{m}{n}$, where m and n are positive integers, then for $C = 0$, and finally when C is negative.)

In the example discussed at the beginning of this section we regarded $G(x, \xi)$ as the result of the action of a unit force localized in a certain point ξ, or, in other words (see Sec. 6.1), distributed with density $\delta(x - \xi)$. So in the general case as well we denote by $G(x, \xi)$ the result of an external action described by the delta function with a singularity at some fixed point ξ, that is, the function $\delta(x - \xi)$. Thus

$$G(x, \xi) = L[\delta(x - \xi)]$$

How can we, using Green's function $G(x, \xi)$, express the result of transforming any given function $f(x)$? To do this, represent f as a sum of "column" functions (Fig. 71), each of which is localized in the single point ξ and is zero outside an infinitesimal neighbourhood of this point. Such a column function is proportional to the delta function $\delta(x - \xi)$, and since the integral of a column function is equal to $f(\xi)\, d\xi$, it is simply equal to $[f(\xi)\, d\xi]\, \delta(x - \xi)$. We thus get the representation

$$f(x) = \sum [f(\xi)\, d\xi]\, \delta(x - \xi)$$

Strictly speaking, in the case of infinitesimal $d\xi$ we ought to write the integral sign instead of the summation symbol, so that actually this is formula (4) in different notation.
But the law of linearity for sums carries over, in the limit, to integrals as well.

By virtue of property (7), every column function is transformed to

$$L[[f(\xi)\, d\xi]\, \delta(x - \xi)] = [f(\xi)\, d\xi]\, L[\delta(x - \xi)] = f(\xi)\, G(x, \xi)\, d\xi$$

And so the sum of such functions, by property (6), is transformed to

$$L[\sum [f(\xi)\, d\xi]\, \delta(x - \xi)] = \sum L[[f(\xi)\, d\xi]\, \delta(x - \xi)] = \sum f(\xi)\, G(x, \xi)\, d\xi$$

This sum, for infinitesimal $d\xi$, is an integral, and we finally get

$$F(x) = L[f(x)] = \int_a^b G(x, \xi)\, f(\xi)\, d\xi \tag{8}$$

(compare with formula (5)).

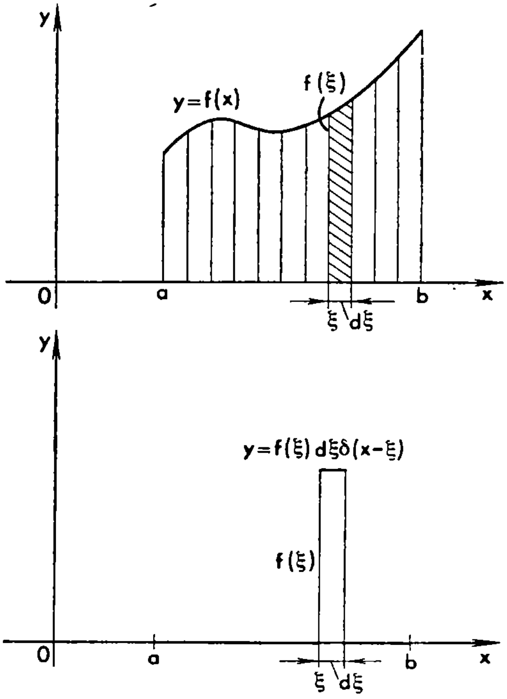


Fig. 71

The influence function can be computed theoretically in the simpler cases, as in the foregoing example, and experimentally in the more complicated problems by carrying out the necessary measurements (for example, measuring the deformation of a system under the action of a localized force). In such cases,the possibility of applying the superposition principle (or, as we say, the linearity of the system) follows either from general theoretical principles or can be verified experimentally. After Green's function has been found and the linearity of the system established, the solution of the problem is given by formula (8) for any external action f.

Thus, at times, even the most general conceptions of the properties of physical systems point the way to a solution of specific problems.

In conclusion, note that the image functions Lf need not necessarily be defined on the same interval as the preimage functions f; what is more, the independent variables x and ξ in (8) can have different physical meaning. The independent variable ξ can play the role of time; then the influence function describes the result of a "unit impulse" acting at time ξ.

Exercise

Write down the influence function for the operators:

(a) $Lf = 2f(x)$, (b) $Lf = \sin x \cdot f(x)$, (c) $Lf = f(x + 1)$,
(d) $Lf = f(x^2)$.

6.3 Functions related to the delta function

The delta function has been found useful in writing down certain other important functions. An important example is the *unit-step function* $e(x)$ (also denoted as $\theta(x)$):

$$e(x) = \int_{-\infty}^{x} \delta(x)\, dx \tag{9}$$

Clearly, for $x < 0$ we get $e(x) = 0$ and for $x > 0$ we get $e(x) = 1$. Thus, $e(x)$ is a discontinuous function that suffers a jump at $x = 0$. Its graph ("step") is shown in Fig. 72. It results from a sudden switching-in of some constant action, say, voltage in an electric circuit (the independent variable there is the time).

Equation (9) can also be obtained with the aid of approximate representations of the delta function. In Sec. 6.1 we saw that one such representation is the function $\frac{1}{\pi}\frac{m}{1 + (mx)^2}$ for m large. However

$$\int_{-\infty}^{x} \frac{1}{\pi}\frac{m}{1 + (mx)^2}\, dx = \frac{1}{\pi} \arctan mx \Big|_{-\infty}^{x} = \frac{1}{\pi} \arctan mx + \frac{1}{2}$$

The graph of this integral for $m = 1$ and $m = 5$ is shown in Fig. 73. In the limit, as $m \to \infty$, we obtain $e(x)$ from the integral, which means we arrive at (9).

The same result may be obtained with the aid of the column function, whose graph is shown in Fig. 74*a* (it too, in the limit as $N \to \infty$, yields the delta function). The graph of the integral of this function is shown in Fig. 74*b*. Here too we get (9) in the limit as $N \to \infty$.

Equation (9) can be demonstrated in the following example from physics. Consider the rectilinear motion of a mass m under the action of a variable force $F(t)$ directed along that same straight line. Writing down the expression for Newton's second law and carrying out the integration, we obtain the following equation (see, for example, HM, Sec. 6.4):

$$v(t) = \frac{1}{m} \int_{-\infty}^{t} F(t)\, dt$$

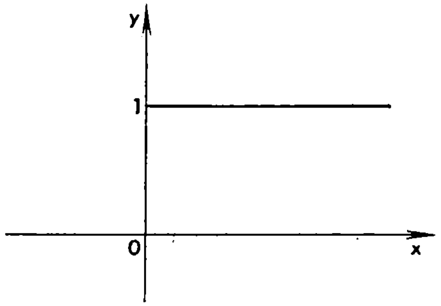


Fig. 72

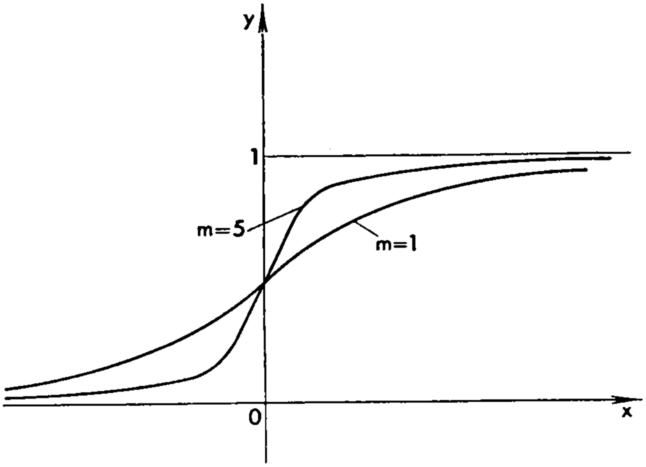


Fig. 73

(we assume that at $t = -\infty$ the initial velocity is zero). Let the force be in the nature of an impact: $F(t) = I_{\text{imp}}\delta(t - t_{\text{imp}})$ (see end of Sec. 6.1). Integrating, we get

$$v(t) = \frac{1}{m}\int_{-\infty}^{t} I_{\text{imp}}\delta(t - t_{\text{imp}})\,dt = \frac{I_{\text{imp}}}{m}e(t - t_{\text{imp}})$$

which is to say that the velocity v is equal to zero prior to the impact and is equal to $\frac{I_{\text{imp}}}{m}$ after the impact.

Now let us return to mathematics. If one differentiates (9), the result is

$$e'(x) = \delta(x) \tag{10}$$

This equation can also be shown in the examples we have just discussed. For an approximate representation of the function $e(x)$ we can take the function given in Fig. 74*b*, the derivative of which is shown in Fig. 74*a*; in the limit, as $N \to \infty$, we get (10).

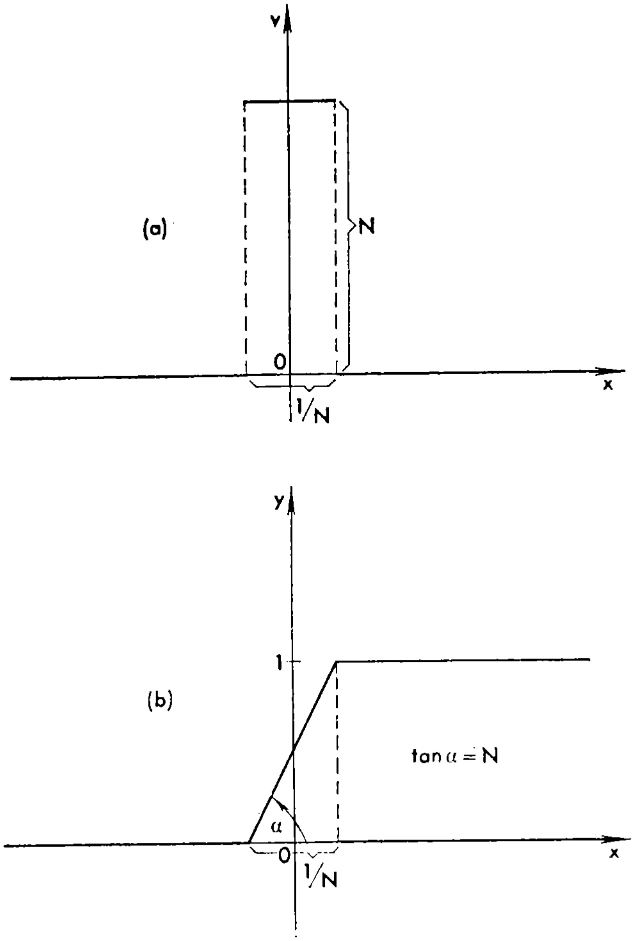


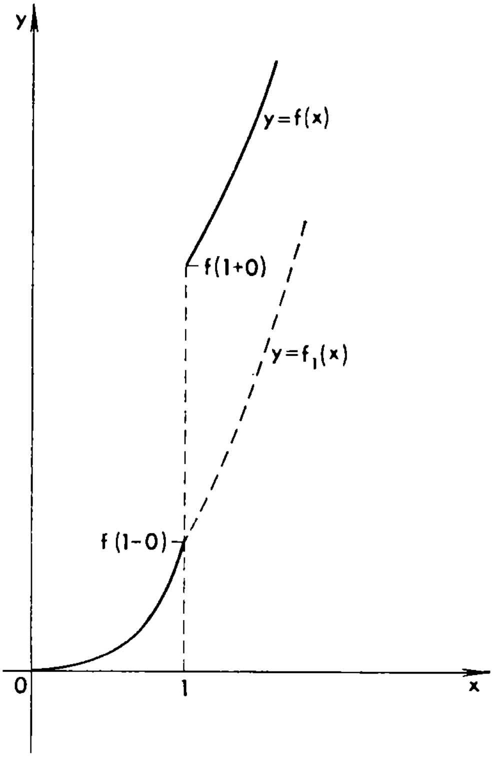
Fig. 74

Here, we proceeded from the delta function and arrived at the unit-step function with the aid of integration. We could go in the reverse direction: we could proceed from the function $e(x)$ and obtain the function $\delta(x)$ with the aid of differentiation. In other words, the delta function is obtained from the discontinuous function $e(x)$ via differentiation. In like manner, delta terms appear when differentiating any discontinuous function.

Consider, say, the function $f(x)$ specified by two formulas:

$$f(x) = \begin{cases} x^3 & \text{if } 0 < x < 1, \\ x^2 + 2 & \text{if } 1 < x < \infty \end{cases}$$

Fig. 75



The graph of this function, which has a discontinuity at $x = 1$, is shown by the heavy lines in Fig. 75. It would be a mistake to equate $f'(x)$ merely to the function $\varphi(x)$ obtained by differentiation of both formulas:

$$\varphi(x) = \begin{cases} 3x^2 & \text{if } 0 < x < 1, \\ 2x & \text{if } 1 < x < \infty \end{cases} \tag{11}$$

Actually, if we integrate the last function, say, from the value $x = 0$, then we get

for $0 < x < 1$: $\int\limits_0^x \varphi(x)\,dx = \int\limits_0^x 3x^2\,dx = x^3$

for $x > 1$: $\int\limits_0^x \varphi(x)\,dx = \int\limits_0^1 \varphi(x)\,dx$

$$+ \int\limits_1^x \varphi(x)\,dx = \int\limits_0^1 3x^2\,dx + \int\limits_1^x 2x\,dx = x^3\Big|_0^1 + x^2\Big|_1^x = x^2$$

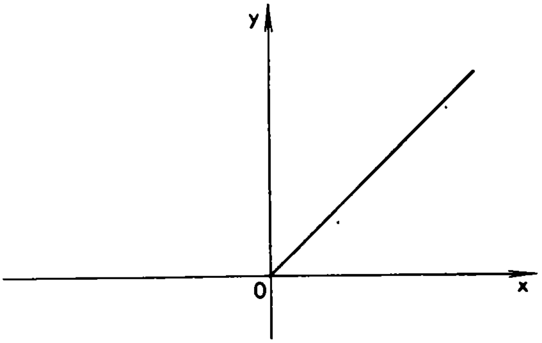


Fig. 76

Thus, we do not get $f(x)$ but rather a continuous function $f_1(x)$, whose graph for $x > 1$ is shown in Fig. 75 by the dashed line. In order to obtain $f(x)$ from $f_1(x)$, we have to add to the first function the "step" with discontinuity at $x = 1$ equal to the discontinuity of the function $f(x)$, that is to say, equal to

$$f(1+0) - f(1-0) = (1^2 + 2) - 1^3 = 2\,^*$$

And so $f(x) = f_1(x) + 2e(x-1)$ whence we finally get

$$f'(x) = f_1'(x) + 2e'(x-1) = \varphi(x) + 2\delta(x-1)$$

where $\varphi(x)$ is given by the formulas (11).

Closely related to the function $e(x)$ is the signum function, sgn x, which can be defined thus:

$$\operatorname{sgn} x = \frac{x}{|x|}$$

It is equal to -1 when $x < 0$ and to $+1$ when $x > 0$, which means it indicates the sign of the number x. The validity of the relation

$$\operatorname{sgn} x = 2e(x) - 1$$

is readily seen.

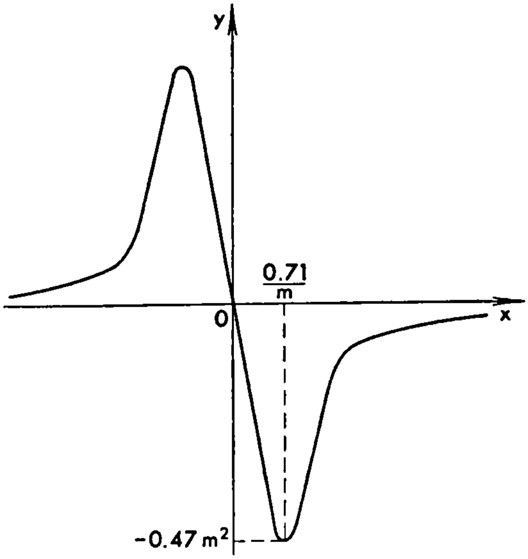
Integrating the function $e(x)$ we get a continuous function (the graph is shown in Fig. 76) since

$$\int_{-\infty}^{x} e(x)\,dx \begin{cases} \text{for } x < 0 \text{ is equal to } \displaystyle\int_{-\infty}^{x} 0\,dx = 0 \\ \text{for } x > 0 \text{ is equal to } \displaystyle\int_{-\infty}^{0} 0\,dx + \int_{0}^{x} 1\,dx = x \end{cases}$$

$\left(\text{Check to see that this function is equal to } \dfrac{x + |x|}{2}.\right)$

[*] The notation $f(1-0)$ is convenient for designating the limit of the value of $f(1-\varepsilon)$ as $\varepsilon \to 0$ $(\varepsilon > 0)$; the notation $f(1+0)$ is similarly defined.

Fig. 77



The delta function can be integrated and it can be differentiated; its derivative $\delta'(x)$ has a still "sharper" singularity than $\delta(x)$, and it assumes values of both signs. If we proceed from an approximate representation of the function $\delta(x)$ in the form $\frac{m}{\sqrt{\pi}} e^{-(mx)^2}$ for m large (see Sec. 6.1), then we get an approximate representation of $\delta'(x)$ in the form of the function

$$\frac{d}{dx}\left[\frac{m}{\sqrt{\pi}} e^{-(mx)^2}\right] = -\frac{2}{\sqrt{\pi}} m^3 x e^{-(mx)^2}$$

whose graph is shown in Fig. 77. This function assumes extremal values for $x = \pm \frac{1}{\sqrt{2}m} = \pm \frac{0.71}{m}$, which are equal, in absolute value, to $\sqrt{\frac{2}{\pi e}}\, m^2 = 0.47\, m^2$. (Check this.) These values are proportional to m^2 and not to m, as in the representation of the function $\delta(x)$.

We can start from an approximate representation of the function $\delta(x)$ in the form of a triangle, as shown in Fig. 78*a* for M large; then the function $\delta'(x)$ will be approximately represented by the graph shown in Fig. 78*b*.

If the delta function describes the density of a unit charge located at the coordinate origin (cf. Sec. 6.1), then $\delta'(x)$ describes the density of a dipole located at the origin. Such a dipole results if we place charges q and $-q$, respectively, at the points $x = 0$ and $x = l$, and then, while leaving $p = ql$ (the dipole moment) unchanged, let l approach zero and q infinity, so that in the limit we get two equal infinitely large

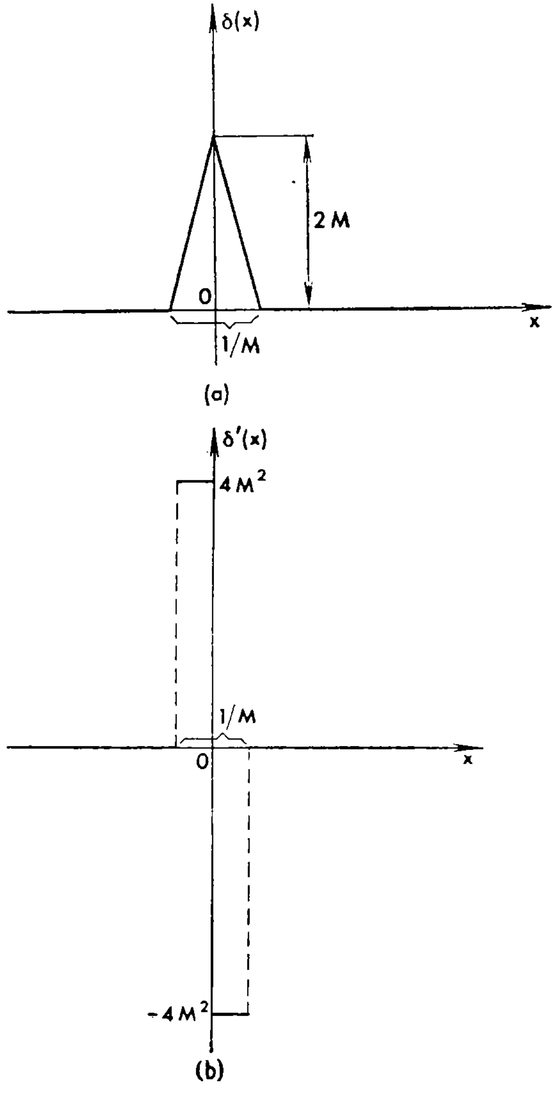


Fig. 78

charges of opposite sign at an infinitely close distance. Prior to passing to the limit, the charge density is of the form

$$q\delta(x) - q\delta(x - l) = p\,\frac{\delta(x - l) - \delta(x)}{-l}$$

For this reason, the charge density, after passing to the limit as $l \to 0$, is equal to $p\delta'(x)$.

Integrals involving $\delta'(x)$ are evaluated via integration by parts:

$$\int_{-\infty}^{\infty} f(x)\,\delta'(x-a)\,dx = f(x)\,\delta(x-a)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f'(x)\,\delta(x-a)\,dx = -f'(a) \qquad (12)$$

The delta function can also be considered in the plane and in space. For example, in space we are to understand the function $\delta(x, y, z)$ as a function equal to zero everywhere outside the coordinate origin (0, 0, 0) and equal to infinity at the origin; a function such that the integral of that function over the entire space is equal to unity. It is easy to verify that these conditions are fully satisfied, for example, by the function

$$\delta(x, y, z) = \delta(x)\,\delta(y)\,\delta(z)$$

Thus, a mass m localized in the point (a, b, c) may be regarded as a mass distributed in space with density

$$\rho(x, y, z) = m\delta(x-a)\,\delta(y-b)\,\delta(z-c)$$

Exercises

1. Find $|x|'$, $|x|''$.
2. Find $\left[\left(1 + e^{\frac{1}{x}}\right)^{-1}\right]'$.
3. Verify the truth of the formula (12) in straightforward fashion, taking advantage of the series expansion of the function $f(x)$ in powers of $x - a$ and use an approximate representation of the function $\delta'(x)$ in the form shown in Fig. 78*b*.

6.4 On the Stieltjes integral

The delta function is directly connected with a useful extension of the integral concept. Let's begin with an example first. Suppose we have a mass m located on a segment (l) of the x-axis with endpoints a and b and it is required to determine the force with which this mass attracts a unit point mass m_0 located at the point $x = c$ to the left of (l) on the same axis. The answer is very simple. Since by Newton's law the mass dm located at point x attracts m_0 with the force $dF = G\dfrac{m_0\,dm}{(x-c)^2}$ (the proportionality factor G here is the so-called gravitation constant), the total force is

$$F = \int_{(l)} G\,\frac{m_0\,dm}{(x-c)^2} = Gm_0 \int_{(l)} \frac{1}{(x-c)^2}\,dm \qquad (13)$$

If the mass m is distributed along (l) in a manner such that at each point x it has a finite density $\rho = \rho(x)$, then $dm = \rho(x)\,dx$ and from the integral (13) we can go over to the ordinary integral

$$Gm_0 \int_a^b \frac{\rho(x)}{(x-c)^2}\,dx$$

However, as was pointed out in Sec. 6.1, the mass m may contain a portion localized at separate points. Then the integral (13) can be understood to be the *integral with respect the measure m*. By this is meant that to every portion (Δl) of the line segment (l) (which also means to every point of this segment) there corresponds a "measure" (in our case it is the mass) $m(\Delta l)$, and the law of addition holds: the measure of the whole is equal to the sum of the measures of the parts. The integral with respect to measure (it is also called the *Stieltjes integral*) is, in the general case, of the form

$$\int_{(l)} f(x)\,d\mu \tag{14}$$

and, by definition, is equal to the limit

$$\lim \sum_{i=1}^{n} f(x_i)\,\mu(\Delta l_i)$$

which is constructed exactly by the same rule that the ordinary integral is (see, for instance, HM, Sec. 2.8), except that instead of the length of the subintervals (Δl_i) of the basic interval (l) we take their measure $\mu(\Delta l_i)$. If for the measure we take ordinary length, then we return to the ordinary definition of an integral, which means the Stieltjes integral is a generalization of this ordinary integral (which is then often called the *Riemann integral* in order to distinguish it from the Stieltjes integral).

If a given measure has a finite density $\rho = \frac{d\mu}{dx}$, then we can pass from the Stieltjes integral to the ordinary integral

$$\int_{(l)} f(x)\,d\mu = \int_a^b f(x)\,\rho(x)\,dx \tag{15}$$

But if there are points with nonzero measure, then, as we saw in Sec. 6.1, the density $\rho(x)$ will have delta-like terms. By assuming such terms we can make the transition (15) in this case as well.

Exercise

Frequently, the measure on the x-axis is given with the aid of an auxiliary function $g(x)$ by the rule: the measure of the interval $\alpha \leqslant x \leqslant \beta$ is equal to $g(\beta + 0) - g(\alpha - 0)$ (or simply $g(\beta) - g(\alpha)$ if the function $g(x)$ is continuous). Then instead of the integral (14) we write $\int\limits_{(l)} f(x)\, dg(x)$, and formula (15) assumes the form $\int\limits_{(l)} f(x)\, dg(x) = \int\limits_a^b f(x)\, g'(x)\, dx$. Find $\int\limits_0^1 x^3\, d(x^2)$, $\int\limits_{-1}^1 \sin x\, de(x)$, $\int\limits_{-1}^1 \cos x\, de(x)$.

ANSWERS AND SOLUTIONS

Sec. 6.1

1. $3^2 = 9$.

2. (a) $[(-5)^2 + 3]\,\delta(x+5) = 28\,\delta(x+5)$,

   (b) $\delta(2x-8) = \delta(2(x-4)) = \frac{1}{2}\,\delta(x-4)$,

   (c) the polynomial $P(x) = x^2 + x - 2$ has the zeros $x_1 = 1$, $x_2 = -2$ and $P'(x_1) = 3$, $P'(x_2) = -3$, whence $\delta(x^2 + x - 2) = \frac{1}{3}\,\delta(x-1) + \frac{1}{3}\,\delta(x+2)$.

Sec. 6.2

(a) $2\delta(x-\xi)$, (b) $\sin x \cdot \delta(x-\xi) = \sin \xi \cdot \delta(x-\xi)$,

(c) $\delta(x - \xi + 1)$,

(d) $$\delta(x^2 - \xi) = \begin{cases} 0 & (\xi < 0) \\ \dfrac{1}{2\sqrt{\xi}}[\delta(x-\sqrt{\xi}) + \delta(x+\sqrt{\xi})] & (\xi > 0) \end{cases}$$

Sec. 6.3

1. $\operatorname{sgn} x$, $2\delta(x)$.

2. $\left(1 + e^{\frac{1}{x}}\right)^{-2} e^{\frac{1}{x}}\, x^{-2} - \delta(x)$.

3. $$\int_{-\infty}^{\infty} f(x)\,\delta'(x-a)\,dx \approx \int_{a-\frac{1}{2M}}^{a} f(x)\,4M^2\,dx - \int_{a}^{a+\frac{1}{2M}} f(x)\,4M^2\,dx$$

$$\approx 4M^2\left\{\int_{a-\frac{1}{2M}}^{a} [f(a)+f'(a)(x-a)]\,dx - \int_{a}^{a+\frac{1}{2M}} [f(a)+f'(a)(x-a)]\,dx\right\}$$

$$= 4M^2\left\{\left[f(a)\frac{1}{2M} - \frac{f'(a)}{2}\,\frac{1}{4M^2}\right] - \left[f(a)\,\frac{1}{2M} + \frac{f'(a)}{2}\,\frac{1}{4M^2}\right]\right\} = -f'(a)$$

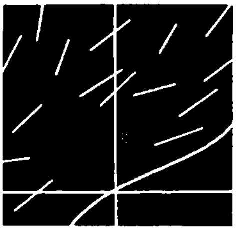
Passing to the limit as $M \to \infty$, we get the exact equation (12).

Sec. 6.4

$\frac{2}{5}$, 0, 1.

Chapter 7

DIFFERENTIAL EQUATIONS



The most elementary types of differential equations that arise when considering various physical processes, like water flowing out of a vessel, radioactive decay, the motion of a material particle, are handled by the fundamentals of integral calculus (see HM, Chs. 5, 6, 7). Here we introduce some general concepts referring to differential equations and will consider certain classes of equations that are not hard to solve. The subject will be discussed further in Ch. 8.

Differential equations involving a single independent variable are called *ordinary differential equations*. If there are two or more independent variables, then the partial derivatives with respect to these variables enter into the differential equation. Such equations are termed *partial differential equations.*

In this and the next chapters we will consider only ordinary differential equations. We will need the results of Secs. 6.1 and 6.2 and Euler's formula given in Sec. 5.3.

7.1 Geometric meaning of a first-order differential equation

A *differential equation of the first order* is a relation of the form

$$F\left(x, y, \frac{dy}{dx}\right) = 0$$

where y is an unknown function of x. From now on we will consider this equation to be solved for the derivative:

$$\frac{dy}{dx} = f(x, y) \tag{1}$$

It turns out that even without searching for an analytic solution of $y(x)$ in the form of a formula, it is possible to get an idea of the general pattern of these solutions on the basis of the geometric meaning of equation (1). That is precisely what we shall do in this section.

Let us recall the geometric meaning of the derivative $\frac{dy}{dx}$. In the xy-plane, the quantity $\frac{dy}{dx}$, for the curve $y = y(x)$, is equal to

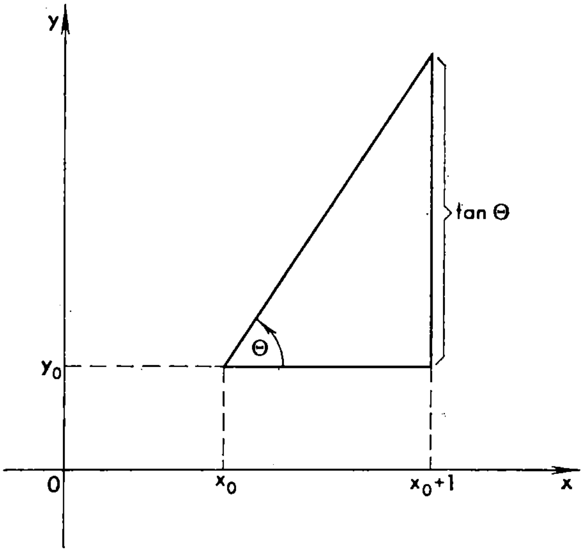


Fig. 79

the slope of the tangent line to the curve. Hence, knowing the dependence of $\frac{dy}{dx}$ on the variables x and y expressed by (1), we can find the direction of the tangent line to the curve that serves as the graph of the solution of (1), and this direction can be determined for any point of the plane. The graph of a solution of a differential equation is called an *integral curve* of the equation.

The direction of the tangent line can be shown by drawing through any point (x, y) a short straight line at the angle θ that satisfies the condition $\tan \Theta = f(x, y)$.*

For example, let $\frac{dy}{dx} = x^2 + y^2$, then $f(x, y) = x^2 + y^2$.

Set up the table

x	y	$\frac{dy}{dx} = \tan \Theta$	x	y	$\frac{dy}{dx} = \tan \Theta$
-1	-1	2	0	1	1
-1	0	1	1	-1	2
-1	1	2	1	0	1
0	-1	1	1	1	2
0	0	0			

* There is no need to find the angle Θ and construct it. The required direction can be found much faster by laying off on the x-axis a segment of unit length, and on the y-axis a segment of length $\frac{dy}{dx} = \tan \Theta$ (Fig. 79).

Fig. 80 shows the directions of the tangent lines at each of the nine points given in the table. If the number of points thus indicated is increased, one gets a picture of the collection of curves that satisfy the differential equation. (See Fig. 81 that corresponds to the equation $\frac{dy}{dx} = x^2 + y^2$.) It is clear that the equation has an infinity of integral curves and that one such curve passes through every point (x_0, y_0) of the plane. And so in order to isolate from all the solutions of equation (1) some one definite *particular solution*, we have to specify a supplementary condition:

$$\text{for some value } x = x_0 \text{ is given the value } y = y_0 \qquad (2)$$

This condition is termed the *initial condition* because if time is the independent variable, then the condition (2) signifies specifying the desired function at the initial instant of time.

Although two parameters, x_0 and y_0, are given in the initial condition (2), actually there is only one degree of freedom in choosing a particular solution of equation (1), for the point (x_0, y_0) can move along the integral curve determined by it, and the curve of course does not change because of this. In that motion there is one degree of freedom, which is thus extra in the choice of the integral curve; that is, in such a choice there is actually $2-1 = 1$ degree of freedom (see a similar discussion in Sec. 4.8). In order to indicate an essential parameter in this choice we can fix x_0 and draw the vertical line $x = x_0$; then various integral curves will intersect it at different heights. This means that the different curves correspond to distinct values of $y(x_0) = y_0$.

In order to draw a large number of line segments that yield the direction of a tangent line, it is convenient to take advantage of the following device. In the drawing, construct the lines $f(x, y) = C$ for several values of the constant C. By (1), the value of $\tan \theta$ at each point of this curve is constant and equal to C, which means that all the segments indicating the direction of the tangent line* at any point of the curve $f(x, y) = C$ are parallel.

The curves $f(x, y) = C$ are called *isoclines*. In particular, the curve $f(x, y) = 0$ is called the *isocline of zeros*. At each point of this curve the tangent to the integral curves of equation (1) is horizontal. The line at whose points the tangents are vertical is called the *isocline of infinities*. For example, the isocline of infinities for the equation $\frac{dy}{dx} = \frac{2x + y}{x - y - 1}$ is the straight line $x - y = 1$.

In Fig. 81 it is clearly seen that the integral curves do not intersect: at any rate, not at a nonzero angle. Indeed, (1) shows that for given

* Bear in mind that we are dealing with tangents to the integral curves of the differential equation $\frac{dy}{dx} = f(x, y)$ and not with the tangents to the curve $f(x, y) = C$ itself.

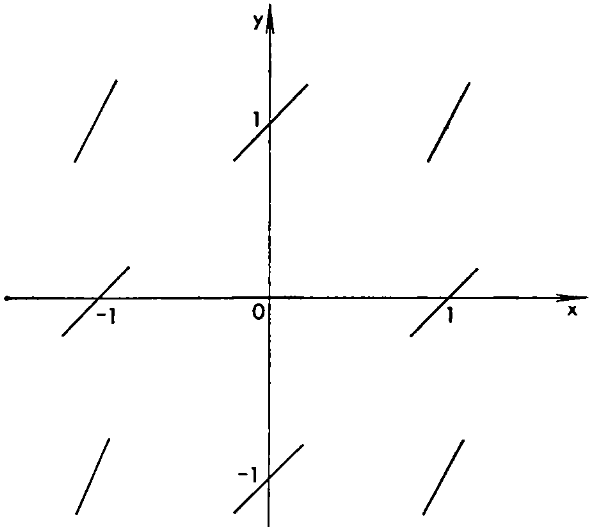


Fig. 80

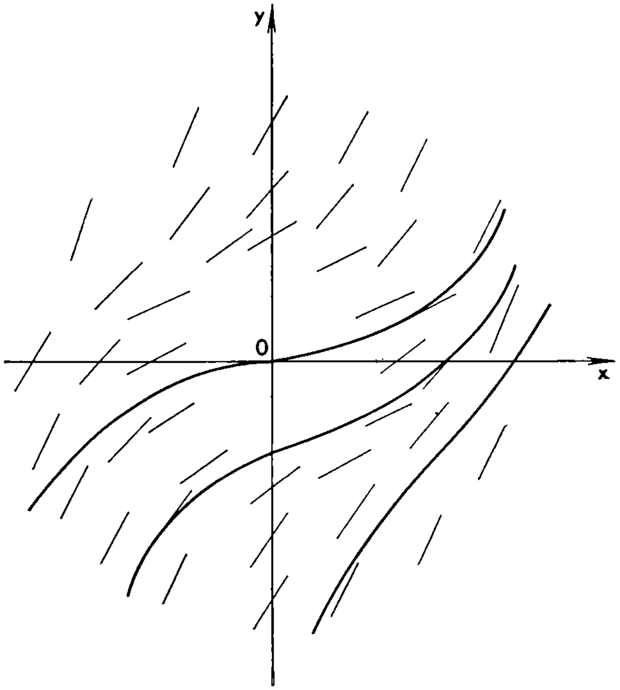


Fig. 81

x and y there is only one definite value of $\frac{dy}{dx}$; that is, a curve can pass through this point only with one specific slope. A more detailed investigation shows that distinct integral curves cannot be in contact at any point if at that point the right member of (1) and its partial derivative with respect to y assume finite values. Thus, the condition (2) does indeed define a unique solution to equation (1).

Exercises

1. Find the isoclines of the equation $\frac{dy}{dx} = x^2 + y^2$.
2. Find the equation of the locus of points of inflection of the integral curves of the general equation (1); of the equation $\frac{dy}{dx} = x^2 + y^2$.

7.2 Integrable types of first-order equations

We now consider several types of differential equations that can be solved without difficulty.

I. Equations with variables separable. These are equations of the form

$$\frac{dy}{dx} = \varphi(y) \cdot \Psi(x)^{*} \tag{3}$$

Here the right-hand member $\varphi(y) \cdot \Psi(x)$ is a product of two functions, one of which depends on y alone and the other on x alone.

We rewrite the equation thus:

$$\frac{dy}{\varphi(y)} = \Psi(x)\,dx$$

Integrating the right and left members, we get

$$\int \frac{dy}{\varphi(y)} = \int \Psi(x)\,dx + C \tag{4}$$

(we write only one arbitrary constant because both constants that appear when evaluating the integrals can be combined into one). From this general solution of equation (3) we get the particular solutions by assigning all possible values to C. We see that in the general solution of (1) there is one arbitrary constant, which is in accord with the existence of one degree of freedom in the choice of the particular solution (Sec. 7.1).

It is easy to find C if we have the extra initial condition (2). For this purpose we write (4), for brevity, thus:

$$\Phi(y) = \Psi(x) + C$$

[*] Equations of this type are found, for example, in the problem of radioactive decay and in the problem of water discharge from a vessel (see HM, Ch. 5).

Setting $x = x_0$, $y = y_0$, we get

$$\Phi(y_0) = \Psi(x_0) + C$$

whence

$$C = \Phi(y_0) - \Psi(x_0)$$

and, finally,

$$\Phi(y) = \Psi(x) + \Phi(y_0) - \Psi(x_0)$$

that is,

$$\Phi(y) - \Phi(y_0) = \Psi(x) - \Psi(x_0)$$

The particular solution thus found can also be written as

$$\int_{y_0}^{y} \frac{dy}{\varphi(y)} \int_{x_0}^{x} \Psi(x)\,dx$$

It is immediately apparent that this solution satisfies the condition (2). We obtain the desired solution by carrying out the integration.

II. Homogeneous linear equations. Suppose we have a homogeneous linear equation of the first order:

$$\frac{dy}{dx} = f(x)y \tag{5}$$

This equation is a particular case of equation (3), but we will examine it in more detail because of its great importance. Separating variables in (5) and integrating, we get

$$\frac{dy}{y} = f(x)\,dx, \quad \ln\, y = \int_a^x f(x)\,dx + \ln C$$

On the right we write the arbitrary constant in the form $\ln C$ for convenience of manipulation. From this we find y:

$$y = Ce^{\int_a^x f(x)\,dx} \tag{6}$$

where a is a fixed value of x. Formula (6) provides the general solution of (5).

For brevity put

$$y_1(x) = e^{\int_a^x f(x)\,dx} \tag{7}$$

Since this function is obtained from (6) for $C = 1$, it is a particular solution of (5). Formula (6) may be written as

$$y = Cy_1(x) \tag{8}$$

It is also easy to verify directly that if $y_1(x)$ is some particular solution of equation (5), then the function (8) for any constant C also satisfies (5):

$$\frac{dy}{dx}=\frac{d(Cy_1)}{dx}=C\frac{dy_1}{dx}=Cf(x)\,y_1=f(x)\,y$$

Thus, in order to obtain the general solution of equation (5), take any particular solution and multiply it by an arbitrary constant. Setting $C=0$, for instance, we see that one of the particular solutions of (5) is identically zero; this zero solution is of course not suitable for constructing the general solution.

III. Nonhomogeneous linear equations. Suppose we have the first-order nonhomogeneous linear equation

$$\frac{dy}{dx}=f(x)\,y+g(x)^{*} \qquad (9)$$

We seek the solution of (9) that vanishes for some value $x=x_0$. For a fixed function $f(x)$, this solution $y(x)$ is determined by the choice of the function $g(x)$, that is, $g(x)$ may be interpreted as a kind of external action and $y(x)$ as its result (in other words, the law by which the function $g(x)$ is associated with the solution $y(x)$ is an operator, see Sec. 6.2). It is easy to verify that the principle of superposition holds true here; which means that if the functions $g(x)$ can be added, so also can the results. Indeed, if

$$\frac{dy_1}{dx}=f(x)\,y_1+g_1(x),\qquad \frac{dy_2}{dx}=f(x)\,y_2+g_2(x)$$

and $y_1(x_0)=0$, $y_2(x_0)=0$, then the function $y=y_1(x)+y_2(x)$ satisfies the equation

$$\frac{dy}{dx}=f(x)\,y+[g_1(x)+g_2(x)]$$

and the condition $y(x_0)=0$ (why?).

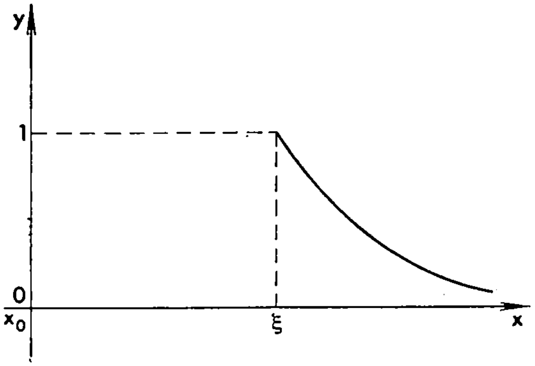
On the basis of Sec. 6.2, the solution of equation (9) may be obtained by constructing the appropriate influence function $G(x,\xi)$, which serves as a solution of the equation

$$\frac{dy}{dx}=f(x)\,y+\delta(x-\xi) \qquad (10)$$

for any fixed ξ. Let us assume that $\xi>x_0$; then for $x_0<x<\xi$ the equation (10) becomes (5), which means the solution is of the form (8), but since a solution is being sought for which $y(x_0)=0$,

[*] This kind of equation occurs for example in the problem of radioactive decay (see HM, Ch. 5). Here, the independent variable x is the time, and the function y is the amount of radioactive material in the system, so that the desired solution $y(x)$ describes the law of variation of quantity in time. The coefficient $f(x)$ is equal to a negative constant whose absolute value is the probability of decay of the given substance in unit time, and the free term $g(x)$ is equal to the rate at which this substance is introduced into the system at hand.

Fig. 82



it follows that $C = 0$, or $y(x) \equiv 0$. If we also integrate (10) from $x = \xi - 0$ to $x = \xi + 0$, then we get

$$y(\xi + 0) - y(\xi - 0) = \int_{\xi-0}^{\xi+0} f(x)\, y\, dx + \int_{\xi-0}^{\xi+0} \delta(x - \xi)\, dx = 0 + 1 = 1$$

(since the solution y is a finite solution, the integral of the first term on the right of (10) over an infinitesimal interval is infinitely small and can therefore be disregarded). But according to what has just been proved, $y(\xi - 0) = 0$, whence

$$y(\xi + 0) = 1 \tag{11}$$

However, for $x > \xi$ equation (10) also becomes (5) and therefore has the solution (8); the condition (11) yields

$$1 = Cy_1(\xi), \text{ or } C = \frac{1}{y_1(\xi)} \text{ and } y = \frac{1}{y_1(\xi)} y_1(x)$$

Thus, in the given problem Green's function is of the form

$$G(x, \xi) = \begin{cases} 0 & (x_0 < x < \xi) \\ \dfrac{y_1(x)}{y_1(\xi)} & (x > \xi) \end{cases}$$

The graph of the function $G(x, \xi)$ for fixed ξ is shown in Fig. 82 by the heavy line, which has a discontinuity at $x = \xi$.

We can now write the desired solution of (9) for any function $g(x)$ on the basis of the appropriately transformed general formula (8) of Ch. 6:

$$y(x) = \int_{x_0}^{\infty} G(x, \xi)\, g(\xi)\, d\xi = \int_{x_0}^{x} G(x, \xi)\, g(\xi)\, d\xi + \int_{x}^{\infty} G(x, \xi)\, g(\xi)\, d\xi$$

$$= \int_{x_0}^{x} \frac{y_1(x)}{y_1(\xi)}\, g(\xi)\, d\xi \tag{12}$$

where the function $y_1(x)$ is given by formula (7). In a similar manner it can be verified that the same final formula (12) holds true also for $x < x_0$. Note that in the right-hand member we can take $y_1(x)$ (but not $y_1(\xi)$!) out from under the integral sign.

Particularly frequent use is made of the last formula for $x_0 = -\infty$. Physically speaking, this is the most natural one, for the solution (12) has the property that if the function $g(x)$ is identically equal to zero up to some x, then the solution too is identically zero up to this x. Thus, in this case the formula (12) yields a sort of "pure" result of the action of the function $g(x)$.

Formula (12) gives a particular solution of equation (9): that solution which vanishes for $x = x_0$. To obtain the general solution of (9), note that the difference of any two solutions of this equation satisfies the appropriate homogeneous equation (5): if

$$\frac{dy_1}{dx} = f(x)\, y_1 + g(x), \quad \frac{dy_2}{dx} = f(x)\, y_2 + g(x)$$

then

$$\frac{d(y_1 - y_2)}{dx} = f(x)\,(y_1 - y_2)$$

(why?). Hence this difference must be of the form (8), where C is arbitrary. To summarize, the general solution of a nonhomogeneous linear equation is the sum of any particular solution and the general solution of the corresponding homogeneous equation. Choosing for this particular solution the solution (12) that we found, we obtain the general solution of (9) in the form

$$y = \int_{x_0}^{x} \frac{y_1(x)}{y_1(\xi)}\, g(\xi)\, d\xi + C y_1(x) \tag{13}$$

If we are interested in the particular solution that satisfies condition (2), then, substituting $x = x_0$ into (13), we get

$$y_0 = 0 + C y_1(x_0), \text{ or } C = \frac{y_0}{y_1(x_0)}$$

and finally we get the desired solution

$$y = \int_{x_0}^{x} \frac{y_1(x)}{y_1(\xi)}\, g(\xi)\, d\xi + y_0 \frac{y_1(x)}{y_1(x_0)}$$

The same results can be attained in a different way, with the aid of a faster but artificial device called the *variation of constants* (parameters). The celebrated French mathematician and mechanician Lagrange proposed seeking the solution of equation (9) in the form (by proceeding from (8)) of

$$y = u(x)\, y_1(x) \tag{14}$$

where $u(x)$ is some unknown function and $y_1(x)$ has the earlier sense of (7). Substituting (14) into (9), we get

$$uy_1' + u'y_1 = f(x)\, uy_1 + g(x)$$

But since y_1 is a solution of a homogeneous equation, the first terms on the left and right cancel out, and we get

$$u'y_1(x) = g(x), \text{ whence } u' = \frac{g(x)}{y_1(x)}, \quad u(x) = \int_{x_0}^{x} \frac{g(\xi)}{y_1(\xi)} d\xi + C$$

(in the last integral we changed the variable of integration to ξ in order to distinguish it from the upper limit). Substituting the result into (14), we arrive at formula (13).

IV. Simple cases of linear equations. There are cases where the solution of a linear equation appears to be particularly simple, as, for instance, if the coefficient $f(x)$ is constant:

$$f(x) \equiv p = \text{constant}$$

Then the homogeneous equation (5) is of the form

$$\frac{dy}{dx} = py$$

and the general solution is

$$y = Ce^{px} \tag{15}$$

This can be obtained from (7) and (8), assuming for the sake of simplicity $a = 0$, but it can also be obtained easily and directly by means of separation of variables or simple substitution.

The corresponding nonhomogeneous equation is particularly simple to solve if the free term, $g(x)$, is a constant or an exponential function. First consider the equation

$$\frac{dy}{dx} = py + A \quad (A = \text{constant}) \tag{16}$$

As we know, to find the general solution it is necessary to find some particular solution and then add to it the general solution (15) of the corresponding homogeneous equation. However, a particular solution of (16) can easily be found in the form $y = B = \text{constant}$. Substituting into (16), we get

$$0 = pB + A, \text{ or } B = -\frac{A}{p}$$

Thus, the general solution of equation (16) is of the form

$$y = -\frac{A}{p} + Ce^{px}$$

Now consider the equation

$$\frac{dy}{dx} = py + Ae^{kx} \tag{17}$$

The derivative of an exponential function is proportional to the function itself, and so it is natural to seek the particular solution of (17) in the form

$$y = Be^{kx} \tag{18}$$

because then, after substitution into (17), all terms will be similar and we can hope to attain equality by a proper choice of B. We then have

$$Bke^{kx} = pBe^{kx} + Ae^{kx}$$

whence it is easy to find $B = \frac{A}{k-p}$. Substituting into (18), we get a particular solution of equation (17) and, after adding the general solution of the corresponding homogeneous equation, we get the general solution of (17):

$$y = \frac{A}{k-p} e^{kx} + Ce^{px}$$

where C is an arbitrary constant.

This solution is clearly not suitable if $k = p$. In this special case we find the solution of (17) with the aid of the general formula (13), noting that in the case at hand the particular solution $y_1(x)$ of the homogeneous equation is equal to e^{px}. Taking $x_0 = 0$ for the sake of simplicity, we get

$$y = \int_0^x \frac{e^{px}}{e^{p\xi}} Ae^{p\xi} d\xi + Ce^{px} = Ae^{px}\int_0^x d\xi + Ce^{px} = Axe^{px} + Ce^{px} \tag{19}$$

Thus, for $k = p$, we have the supplementary factor x in the exponential in the particular solution.

The general nonhomogeneous equation

$$\frac{dy}{dx} = py + g(x)$$

is solved with the aid of Green's function, which in this case is particularly simple:

$$G(x, \xi) = \begin{cases} 0 & (x_0 < x < \xi) \\ \frac{e^{px}}{e^{p\xi}} = e^{p(x-\xi)} & (x > \xi) \end{cases}$$

The general solution, because of formula (13), is of the form

$$y = \int_{x_0}^x e^{p(x-\xi)} g(\xi)\, d\xi + Ce^{px}$$

The equations considered here do not exhaust the types of equations whose solutions can be written down in the form of exact formulas involving elementary functions and integrals of them. Several other such types can be found in texts on differential equations (they are fully covered in Kamke [10]).

Exercises

Find the solutions to the following differential equations that satisfy the indicated initial data.

1. $\frac{dy}{dx} = 2xy$, $y = 1$ for $x = 0$.

2. $\frac{dy}{dx} = \frac{y}{x}$, $y = 1$ for $x = 1$.

3. $\frac{dy}{dx} = e^{-y}$, $y = 1$ for $x = 0$.

4. $\frac{dy}{dx} = \frac{x}{y}$, $y = 1$ for $x = 0$.

5. $\frac{dy}{dx} = -y + e^{x}$, $y = \frac{1}{e}$ for $x = 1$.

6. $\frac{dy}{dx} = -2y + 4x$, $y = -2$ for $x = 0$.

7. $\frac{dy}{dx} + y = \cos x$, $y = \frac{3}{2}$ for $x = 0$.

7.3 Second-order homogeneous linear equations with constant coefficients

A *differential equation of the second order* contains the second derivative of the desired function and has the general aspect

$$F\left(x, y, \frac{dy}{dx}, \frac{d^2y}{dx^2}\right) = 0$$

If this equation depends linearly on the desired function and its derivatives (the dependence on x may be of any kind), it is called a *linear equation*. Thus, a second-order linear equation has the general form

$$a(x)\frac{d^2y}{dx^2} + b(x)\frac{dy}{dx} + c(x)\,y = f(x)$$

We will consider the most important special case where the coefficients a, b, c of the desired function and its derivatives are constant. This equation is encountered in problems involving mechanical vibrations and electric oscillations (see HM, Chs. 6 and 8); the independent variable in these problems is the time t. Let us examine the mechanical vibrations of the most elementary type of oscillator. The equation then has the form

$$m\frac{d^2y}{dt^2}+h\frac{dy}{dt}+ky=f(t) \tag{20}$$

where y is the deviation of an oscillating point from the position of equilibrium, m is its mass, h is the coefficient of friction (the friction is assumed to be proportional to the velocity), k is the coefficient of elasticity of the restoring force (which force is assumed to be proportional to the deviation), and $f(t)$ is an external force.*

We first consider the equation of free oscillations, which has the form

$$m\frac{d^2y}{dt^2}+h\frac{dy}{dt}+ky=0 \tag{21}$$

and is called a *homogeneous* linear equation. In many respects, its properties are similar to those of the homogeneous linear equation of the first order that was discussed in Sec. 7.2. For instance, it is easy to verify that if $y_1(t)$ is a particular solution of (21), then so also is $Cy_1(t)$, where C is any constant. In particular, for $C = 0$ we get the identically zero solution of equation (21). Besides, it is easy to verify that if $y_1(t)$ and $y_2(t)$ are two particular solutions of (21), then their sum, $y = y_1(t) + y_2(t)$, is also a solution of that equation (check this by substituting this sum into (21)).

From the foregoing it follows that if we have found two particular solutions $y_1(t)$ and $y_2(t)$ of equation (21), then their *linear combination*,

$$y = C_1y_1(t) + C_2y_2(t) \tag{22}$$

where C_1 and C_2 are arbitrary constants, is also a solution of that equation. But the general solution of a second-order differential equation is obtained via two integrations and for that reason contains two arbitrary constants. This means that the expression (22) serves as the general solution of equation (21). Of course, $y_1(t)$ and $y_2(t)$ should not be proportional here, since if $y_2(t) = ky_1(t)$, then

$$C_1y_1(t) + C_2y_2(t) = (C_1 + C_2k)\,y_1(t) = Cy_1(t)$$

which is to say that actually there is only one arbitrary constant here (the condition that the parameters be essential is not fulfilled, cf. Sec. 4.8).

How do we find two "independent" solutions of equation (21)? Euler, by proceeding from the property of an exponential being proportional to its derivatives, proposed seeking the particular solutions in the form

$$y = e^{pt} \tag{23}$$

* The reader will do well to review the results of HM, Ch. 6. Here the exposition is more complete and will be conducted from a broader base.

where p is a constant that has to be found. Since in such a choice $\frac{dy}{dt} = pe^{pt}$ and $\frac{d^2y}{dt^2} = p^2e^{pt}$, after substituting into (21) and cancelling out e^{pt}, we get, for a determination of p, the quadratic equation (the so-called *characteristic equation*)

$$mp^2 + hp + k = 0 \tag{24}$$

As we know from algebra, in solving a quadratic equation there may be different cases, depending on the sign of the discriminant:

$$D = h^2 - 4mk$$

If the friction is great, to be more exact, if $h^2 > 4mk$, then (24) has real roots:

$$p_{1,2} = \frac{-h \pm \sqrt{h^2 - 4mk}}{2m}$$

Denote them by $p_1 = -a$ and $p_2 = -b$, since they are both negative. Then, on the basis of (22) and (23), we get the general solution of (21) in the form

$$y = C_1e^{-at} + C_2e^{-bt} \tag{25}$$

where C_1 and C_2 are constants. Thus, in the case of considerable friction, the deviation of a point from the equilibrium position tends to zero exponentially with t without oscillations.

If the friction is small, to be more exact, if $h^2 < 4mk$, then (24) has imaginary conjugate roots:

$$p_{1,2} = -\frac{h}{2m} \pm i\sqrt{\frac{k}{m} - \frac{h^2}{4m^2}} = -\gamma \pm i\omega$$

where $\gamma = \frac{h}{2m}$, $\omega = \sqrt{\frac{k}{m} - \frac{h^2}{4m^2}}$. Having in view (23), we get the general solution (22):

$$y = C_1e^{-\gamma t + i\omega t} + C_2e^{-\gamma t - i\omega t} = e^{-\gamma t}(C_1e^{i\omega t} + C_2e^{-i\omega t}) \tag{26}$$

As was determined in Sec. 5.4, the multipliers $e^{i\omega t}$ and $e^{-i\omega t}$ are periodic functions with period $T = \frac{2\pi}{\omega}$. For this reason, ω is the circular frequency. The multiplier $e^{-\gamma t}$ characterizes the rate of decay of oscillations. The expression (26) will be a solution of equation (21) for any constants C_1 and C_2. To obtain a real solution, take any $C_1 = \frac{1}{2}re^{i\varphi}$ and put $C_2 = \frac{1}{2}re^{-i\varphi} = C_1^*$. Then

$$y = e^{-\gamma t}\left(\frac{1}{2}re^{i\varphi}e^{i\omega t} + \frac{1}{2}re^{-i\varphi}e^{-i\omega t}\right) = \frac{1}{2}re^{-\gamma t}[e^{i(\omega t + \varphi)} + e^{-i(\omega t + \varphi)}]$$

Taking advantage of Euler's formula, we get

$$y = \frac{1}{2} re^{-\gamma t}[\cos(\omega t + \varphi) + i \sin(\omega t + \varphi) + \cos(\omega t + \varphi)$$
$$- i \sin(\omega t + \varphi)] = re^{-\gamma t} \cos(\omega t + \varphi)$$

The solution $y = re^{-\gamma t} \cos(\omega t + \varphi)$ is real and contains two arbitrary constants, r and φ. This is sometimes written differently, the cosine of the sum being expanded:

$$\begin{aligned} y &= re^{-\gamma t}(\cos \omega t \cos \varphi - \sin \omega t \sin \varphi) \\ &= (r \cos \varphi) e^{-\gamma t} \cos \omega t + (-r \sin \varphi) e^{-\gamma t} \sin \omega t \\ &= C_1 e^{-\gamma t} \cos \omega t + C_2 e^{-\gamma t} \sin \omega t \end{aligned}$$

where C_1 and C_2 denote the appropriate multipliers in brackets (they are not equal to C_1 and C_2 in formula (26)!). Here the real independent solutions of equation (21) are quite apparent (see formula (22)).

Ordinarily, this transition to the real solution is carried out faster via the following reasoning. If into the left member of (21) we substitute a complex function of the real argument t and carry out all operations, then, by virtue of the properties of these functions given in Sec. 5.5, the same operations will be carried out on the real and the imaginary parts of this function. Therefore, if zero results from performing these operations on a complex function, then we get zero after performing these operations on the real part of the function (and on the imaginary part as well). Which means that if we have a complex solution (23) of equation (21), then the real and imaginary parts of the solution are also solutions of (21).

Note that the last assertion holds true for any homogeneous linear equation (i.e. an equation involving y and its derivatives linearly and such that it has no term not involving y or a derivative of y) with real coefficients. Now if y and its derivatives occur nonlinearly in the equation, this assertion is no longer valid. For this reason, to take an example, if a quadratic equation has two imaginary roots, then the real and imaginary parts of the roots are not, separately, roots of this equation.

Thus, when the friction is slight, the oscillations decay exponentially. Absence of friction offers a special case: $h = 0$. In this case, the characteristic equation (24) is of the form

$$mp^2 + k = 0$$

whence

$$p_{1,2} = \pm i \sqrt{\frac{k}{m}}$$

The solution of the differential equation has the form

$$y = C_1 e^{i\omega t} + C_2 e^{-i\omega t} = r \cos(\omega t + \varphi) \quad \text{where } \omega = \sqrt{\frac{k}{m}} \tag{27}$$

What this means is that in the system at hand we have nondecaying harmonic oscillations with arbitrary amplitude and arbitrary initial phase and a quite definite frequency $\omega = \sqrt{\frac{k}{m}}$.

It is interesting to follow the behaviour of the total energy of the system. It is easy to show (see HM, Sec. 6.10) that this energy has the expression

$$E = \frac{mv^2}{2} + \frac{ky^2}{2} = \frac{1}{2}\left[m\left(\frac{dy}{dt}\right)^2 + ky^2\right] \tag{28}$$

The first term on the right is equal to the kinetic energy, the second term, to the potential energy of the oscillator. Substituting (27) into (28) for a solution, we get

$$E = \frac{1}{2} m\omega^2 r^2 \sin^2(\omega t + \varphi) + \frac{1}{2} kr^2 \cos^2(\omega t + \varphi) \equiv \frac{1}{2} kr^2$$

Thus, in the case of $h = 0$ the total energy remains constant and all we have is a "pumping" of kinetic energy into potential energy and back again.

If there is friction, the total energy of the system diminishes, being dissipated (it passes into heat, which is not taken into consideration in the differential equation). Differentiating (28) and using the differential equation (21), we get

$$\begin{aligned}\frac{dE}{dt} &= \frac{1}{2}\left[2m\frac{dy}{dt}\,\frac{d^2y}{dt^2} + 2ky\frac{dy}{dt}\right] \\ &= \frac{dy}{dt}\left(-h\frac{dy}{dt} - ky\right) + ky\frac{dy}{dt} = -h\left(\frac{dy}{dt}\right)^2\end{aligned}$$

The derivative is negative and hence E decreases.

In practical problems we are not interested in the general solution but in a specific particular solution. Since there are two arbitrary constants in the general solution of a second-order differential equation (that is to say, two degrees of freedom), a particular solution can be chosen only if we specify two supplementary relations that allow us to find the values of these arbitrary constants. Such supplementary conditions are usually in the form of *initial conditions:* the values of the desired function and its derivative for a certain value of the independent variable are specified.

The initial conditions for (21) consist in specifying the values

$$y\Big|_{t=t_0} = y_0, \quad \frac{dy}{dt}\Big|_{t=t_0} = v_0 \tag{29}$$

that is, the initial deviation and the initial velocity of the oscillating point. From physical reasoning it is clear that these conditions fully determine the process of oscillation. This can readily be proved mathematically as well. Consider, say, the case of considerable friction when the general solution is of the form (25). Differentiating this solution and substituting $t = t_0$, we get, on the basis of (29),

$$C_1 e^{-at_0} + C_2 e^{-bt_0} = y_0,$$
$$-C_1 a e^{-at_0} - C_2 b e^{-bt_0} = v_0$$

whence for C_1 and C_2 we find the definite values

$$C_1 = \frac{By_0 + v_0}{b - a} e^{at_0}, \quad C_2 = \frac{ay_0 + v_0}{a - b} e^{bt_0}$$

Putting these values into (25), we obtain the sought-for particular solution that satisfies the conditions (29):

$$y = \frac{by_0 + v_0}{b - a} e^{at_0} e^{-at} + \frac{ay_0 + v_0}{a - b} e^{bt_0} e^{-bt}$$
$$= y_0 \frac{be^{-a(t-t_0)} - ae^{-b(t-t_0)}}{b - a} + v_0 \frac{e^{-a(t-t_0)} - e^{-b(t-t_0)}}{b - a} \qquad (30)$$

Formula (30) enables us to consider the intermediate case of $h^2 = 4mk$ that we have not yet examined; for this reason the characteristic equation (24) has the coincident roots $p_{1,2} = -a$. Then the solution (25) is not the general solution, since $C_1 e^{-at} + C_2 e^{-at} = (C_1 + C_2) e^{-at} = Ce^{-at}$, for, actually, we have here only one degree of freedom, not two, as required. If in (30) we pass to the limit as $b \to a$, then, taking advantage of l'Hospital's rule (see HM, Sec. 3.21), in the limit we get the solution

$$y = y_0[e^{-a(t-t_0)} + a(t - t_0) e^{-a(t-t_0)}] + v_0(t - t_0) e^{-a(t-t_0)}$$
$$= y_0 e^{-a(t-t_0)} + (ay_0 + v_0)(t - t_0) e^{-a(t-t_0)} \qquad (31)$$

We see that in the solution here there appears a term of the form te^{-at}. This term tends to zero as t increases, since the exponential function tends to zero faster than any power of t tends to infinity (see HM, 3.21). Hence, here again we have decay without oscillations.

If one notes that the initial conditions are arbitrary and regroups the terms in the right-hand member of (31), one can write down the general solution of equation (21) in the given intermediate case as

$$y = C_1 e^{-at} + C_2 t e^{-at}$$

Exercises

Find solutions to the following differential equations that satisfy the indicated initial data.

1. $y'' + y = 0, \quad y = 0, \quad y' = -2 \quad \text{for } t = \frac{\pi}{2}.$
2. $4y'' - 8y' + 5y = 0, \quad y = 0, \quad y' = \frac{1}{2} \quad \text{for } t = 0.$

3. $y'' - 3y' + 2y = 0,\quad y = 2,\quad y' = 3$ for $t = 0$.
4. $y'' - y = 0,\quad y = 4,\quad y' = -2$ for $t = 0$.
5. $y'' - 2y' + y = 0,\quad y = 0,\quad y' = e$ for $t = 1$.
6. $y'' + 4y' + 4y = 0,\quad y = 1,\quad y' = 3$ for $t = 0$.

7.4 A simple second-order nonhomogeneous linear equation

We now consider the equation of *forced oscillations*, equation (20). Note first of all that the reasoning given in Sec. 7.2 on the relationship of the solutions of a nonhomogeneous equation with the solutions of the corresponding homogeneous equation hold true: thus the general solution of (20) is of the form

$$y = Y(x) + C_1y_1(x) + C_2y_2(x)$$

where $Y(x)$ is a particular solution and $C_1y_1(x) + C_2y_2(x)$ is the general solution of the corresponding homogeneous equation (21).

As in Sec. 7.2, a particular solution to equation (20) may be constructed with the aid of the Green's function. First let us consider the simplest case of $h = k = 0$, that is, when (20) is of the form

$$m\frac{d^2y}{dt^2} = f(t) \tag{32}$$

This means that a body is being acted upon by a force whose dependence on time is specified. Let us find a solution that satisfies the zero initial conditions:

$$y\big|_{t=t_0} = 0, \quad v = \frac{dy}{dt}\bigg|_{t=t_0} = 0^* \tag{33}$$

As in Sec. 7.2 (see the beginning of Item III), we see that the solution $y(t)$ is fully determined by specifying the function $f(t)$, and the principle of superposition applies. Thus, on the basis of the general procedure of Sec. 6.2 we can construct the Green's function by solving the equation

$$m\frac{d^2y}{dt^2} = \delta(t - \tau) \tag{34}$$

for any fixed τ and for the initial conditions (33). For the sake of definiteness we assume that $t > t_0$. Then over the interval from t_0 to τ the right side of (34) is equal to zero, and so the solution is also zero for the zero conditions (33). Integrating the equation (34) from $\tau - 0$ to $\tau + 0$,** we get

$$m\left(\frac{dy}{dt}\bigg|_{\tau+0} - \frac{dy}{dt}\bigg|_{\tau-0}\right) = 1$$

[*] Here we can take $t_0 = -\infty$, which is physically natural for many problems (cf. Sec. 7.2). However, we leave the arbitrary t_0 for what follows, since for $t_0 = -\infty$ it is inconvenient to specify nonzero initial conditions.

[**] See the footnote on p. 218.

But by what has just been demonstrated $\left.\frac{dy}{dt}\right|_{\tau-0} = 0$, whence

$$\left.\frac{dy}{dt}\right|_{\tau+0} = \frac{1}{m} \tag{35}$$

We see that for $t = \tau$ the derivative $\frac{dy}{dt}$ has a finite discontinuity (a finite jump); and so the function itself $y(t)$ does not have a discontinuity at $t = \tau$, that is,

$$y|_{\tau+0} = y|_{\tau-0} = 0 \tag{36}$$

But for $t > \tau$ equation (34) is of the form $m\frac{d^2y}{dt^2} = 0$, or $\frac{d^2y}{dt^2} = 0$, and we have to find the solution of this equation for the initial conditions (35) and (36). It is not difficult to verify directly that all these requirements are satisfied by the function $y = \frac{1}{m}(t - \tau)$. We thus have the Green's function of the problem at hand:

$$G(t, \tau) = \begin{cases} 0 & (t_0 < t < \tau), \\ \frac{1}{m}(t - \tau) & (\tau < t < \infty) \end{cases} \tag{37}$$

(The fact that this continuous function with a "corner" at $t = \tau$ satisfies equation (34) also follows from the considerations of Sec. 6.3; see Fig. 76 which depicts a function whose second derivative is equal to $\delta(x)$. By virtue of the general formula (8) of Ch. 6 applied to the given case, we obtain the desired solution of equation (32) for the initial conditions (33):

$$\begin{aligned} y(t) &= \int_{t_0}^{\infty} G(t, \tau) f(\tau)\, d\tau \\ &= \int_{t_0}^{t} G(t, \tau) f(\tau)\, d\tau + \int_{t}^{\infty} G(t, \tau) f(\tau)\, d\tau = \frac{1}{m}\int_{t_0}^{t} (t - \tau) f(\tau)\, d\tau \end{aligned} \tag{38}$$

This completes the derivation of the formula for this solution, and the reader can go straight to Sec. 7.5 where we discuss the general equation of forced oscillations. However, we have a few more remarks to make with respect to the particular case we are now discussing.

The Green's function (37) and also formula (38) may be obtained directly by physical arguments. Equation (34), which defines the Green's function, signifies that a body which at the initial time was located at the origin and was not in motion was acted upon at time τ by an instantaneous force with unit impulse (see the end of

Sec. 6.1). But after the action of the brief force the body moves with a constant velocity equal to the ratio of the impulse to the mass of the body (see HM, Sec. 6.5), which in this case is 1/m. For this reason, $y(t)$, or the path traversed, is expressed just by formulas (37).

Formula (38) may be obtained without invoking the Green's function, although, actually, the method is the same. One imagines the force f over the interval t_0 to t as a sequence of short-term forces, each of which operates over a certain interval from τ to $\tau + d\tau$ and therefore has the impulse $f(\tau)\,d\tau$. If this short-term force were acting alone, then the body would attain a velocity $\frac{f(\tau)\,d\tau}{m}$ and by time t would traverse a path

$$\frac{f(\tau)\,d\tau}{m}(t - \tau) \tag{39}$$

But due to the linearity of equation (32), we have the principle of superposition (in other words, the principle of addition of motions, according to which the laws of motion are additive when several forces are superimposed on one another). Therefore, the results (39) must be added over all τ from t_0 to t; we thus arrive at formula (38).

Now let us derive the formula (38) in a different way, by a twofold integration of (32). The first integration yields

$$m[y'(t) - y'(t_0)] = \int_{t_0}^{t} f(t)\,dt$$

or, taking into account the second condition of (33),

$$my'(t) = \int_{t_0}^{t} f(t)\,dt$$

In this formula the variable of integration is denoted by the same letter t as the upper limit. Ordinarily, this does not cause any confusion, but it is more convenient here to use the more accurate notation

$$my'(t) = \int_{t_0}^{t} f(\tau)\,d\tau \tag{40}$$

in which we can strictly distinguish between the variable of integration τ and the upper limit t: the velocity of the body at time t depends on the values of the force at all preceding instants τ, which means it depends on $f(\tau)$, where τ assumes all values from t_0 to t. (This distinction is absolutely necessary in formula (38), where the difference $t - \tau$ appears in the integrand.)

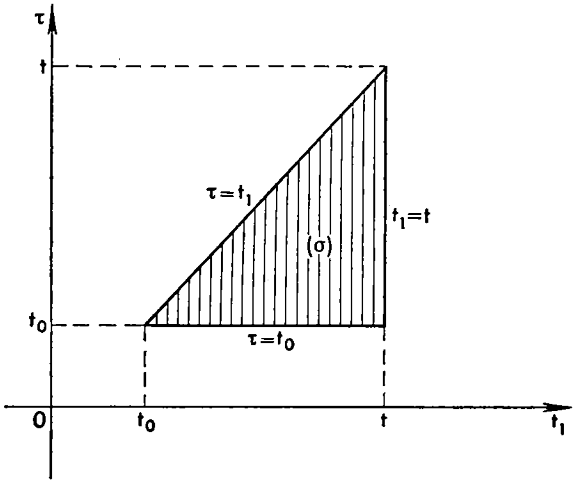


Fig. 83

Again integrating (40), we obtain, with account taken of the first condition of (33),

$$my(t) = \int_{t_0}^{t} my'(t_1)\,dt_1 = \int_{t_0}^{t}\left(\int_{t_0}^{t_1} f(\tau)\,d\tau\right)dt_1$$

We have a twofold iterated integral (cf. Sec. 4.7) in which the variable t_1 of the outer integration varies from t_0 to t, and for each t_1 the variable τ of inner integration varies from $\tau = t_0$ to $\tau = t_1$. Thus the integration is performed over the triangle (σ) shown in Fig. 83. But, as we know, a double integral can be evaluated in the reverse order, first with respect to t_1 and then with respect to τ. Then the integration with respect to t_1 will be carried out for fixed τ from $t_1 = \tau$ to $t_1 = t$ (see Fig. 83) and the outer integration with respect to τ from t_0 to t. We then get

$$my(t) = \iint_{(\sigma)} f(\tau)\,d\tau\,dt_1 = \int_{t_0}^{t} d\tau \int_{\tau}^{t} f(\tau)\,dt_1$$

But we can take $f(\tau)$ out from under the inner integral sign: it does not depend on the variable of integration t_1, so that

$$my(t) = \int_{t_0}^{t} d\tau f(\tau) \int_{\tau}^{t} dt_1 = \int_{t_0}^{t} d\tau\cdot f(\tau)\,(t-\tau)$$

and we arrive at formula (38).

Let us verify that (38) does indeed yield the solution of this problem. It will also be a useful exercise in differentiation. We will proceed from two formulas. The first,

$$\frac{d}{dt}\left(\int_a^t \varphi(s)\,ds\right) = \varphi(t) \tag{41}$$

(the derivative of the integral with respect to the upper limit) is familiar from integral calculus; the second formula,

$$\frac{d}{dt}\left(\int_a^b F(s, t)\,ds\right) = \int_a^b F'_t(s, t)\,ds \quad (a, b = \text{constant}) \tag{42}$$

(the derivative of the integral with respect to a parameter) was given in Sec. 3.6. But how do we find the derivative of the right-hand side of (38), where t appears both as a limit of integration and as the parameter under the integral sign? To do this we have to take the sum of two terms, one of which is obtained by differentiating the integral (38) for fixed t under the integral sign, and the other for fixed t in the upper limit (cf. HM, Sec. 3.4). The differentiation is performed by formulas (41) and (42):

$$\frac{dy}{dt} = \frac{1}{m}(t - \tau)f(\tau)\Big|_{\tau=t} + \frac{1}{m}\int_{t_0}^{t}\frac{\partial}{\partial t}[(t - \tau)f(\tau)]\,d\tau = \frac{1}{m}\int_{t_0}^{t} f(\tau)\,d\tau \tag{43}$$

We use (41) in the second differentiation:

$$\frac{d^2y}{dt^2} = \frac{1}{m}f(t)$$

This implies that equation (32) is satisfied. The conditions (33) hold if we put $t = t_0$ in (38) and (43).

By way of an illustration, let us find the law of motion of a body acted upon by a force proportional to the time elapsed from the start of motion. In this case, $f(t) = a(t- t_0)$ and so $f(\tau) = a(\tau-t_0)$. We get

$$y(t) = \frac{1}{m}\int_{t_0}^{t}(t - \tau)\,a(\tau - t_0)\,d\tau = \frac{a}{m}\int_{t_0}^{t}[(t - t_0) - (\tau - t_0)]\,(\tau - t_0)\,d\tau = \frac{a}{m}(t - t_0)\,\frac{(t - t_0)^2}{2} - \frac{a}{m}\,\frac{(t - t_0)^3}{3} = \frac{a(t - t_0)^3}{6m}$$

We solved the problem with zero initial data. This is the most important part of the job because if we can solve a problem with zero initial data, then the solution of the problem with arbitrary initial data, $y = y_0$, $v = v_0$, for $t = t_0$ offers no difficulties at all. Indeed, suppose

$$m\frac{d^2y}{dt^2} = f(t),\ y = y_0,\ v = v_0 \text{ for } t = t_0$$

Then

$$y(t) = y^{(1)}(t) + y^{(2)}(t)$$

where $y^{(1)}(t)$ is the solution of the problem with zero initial data.

$$m\frac{d^2y^{(1)}}{dt^2} = f(t), \quad y^{(1)}(t_0) = 0, \quad \left.\frac{dy^{(1)}}{dt}\right|_{t=t_0} = v^{(1)}(t_0) = 0$$

and $y^{(2)}(t)$ is the solution without a force:

$$m\frac{d^2y^{(2)}}{dt^2} = 0, \quad y^{(2)}(t_0) = y_0, \quad \left.\frac{dy^{(2)}}{dt}\right|_{t=t_0} = v^{(2)}(t_0) = v_0$$

(The reader can easily verify that $y = y^{(1)} + y^{(2)}$ is a solution of this problem.) It is easy to find the function $y^{(2)}(t)$: $y^{(2)}t = v_0(t - t_0) + y_0$ and so

$$y(t) = \frac{1}{m}\int_{t_0}^{t}(t - \tau)f(\tau)\,d\tau + v_0(t - t_0) + y_0$$

Let us investigate the solution (38) which corresponds to zero initial conditions; for simplicity we assume that $t_0 = -\infty$. We then write

$$y(t) = \frac{1}{m}\int_{-\infty}^{t}(t - \tau)f(\tau)\,d\tau = \frac{1}{m}\int_{-\infty}^{t}tf(\tau)\,d\tau - \frac{1}{m}\int_{-\infty}^{t}\tau f(\tau)\,d\tau$$

Since t is not the variable of integration, we can take it outside the integral sign, and so

$$y(t) = \frac{t}{m}\int_{-\infty}^{t}f(\tau)\,d\tau - \frac{1}{m}\int_{-\infty}^{t}\tau f(\tau)\,d\tau$$

On the basis of (43) we can rewrite this formula thus:

$$y(t) = tv(t) - \frac{1}{m}\int_{-\infty}^{t}\tau f(\tau)\,d\tau$$

Factor out $v(t)$ to get

$$y(t) = v(t)\cdot(t - \theta)$$

where

$$\theta = \frac{\frac{1}{m}\int_{-\infty}^{t}\tau f(\tau)\,d\tau}{v(t)} = \frac{\int_{-\infty}^{t}\tau f(\tau)\,d\tau}{\int_{-\infty}^{t}f(\tau)\,d\tau} \quad \text{and} \quad v(t) = \frac{1}{m}\int_{-\infty}^{t}f(\tau)\,d\tau$$

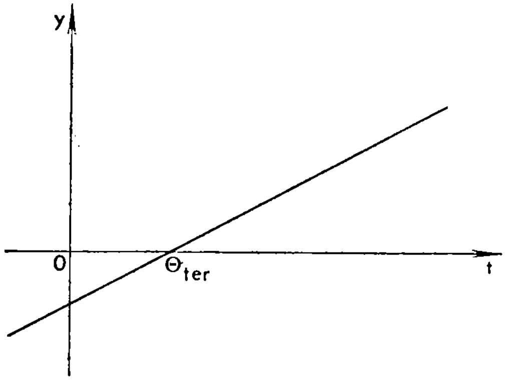


Fig. 84

The formulas written in this fashion are particularly convenient if the force ceases to act after a time interval. For times t after the force has ceased to act, the integrals $\int\limits_{-\infty}^{t} \tau f(\tau)\,d\tau$ and $\int\limits_{-\infty}^{t} f(\tau)\,d\tau$ no longer depend on t. An increase in t in these integrals only leads to an increase in that portion of the region of integration where the integrand is zero. After the force ceases to act, the body moves with a constant velocity $v = v_{\text{ter}}$ and the quantity $\Theta = \Theta_{\text{ter}}$ is also constant. Therefore, after the action has ceased, the graph of $y(t)$ is a straight line:

$$y = v_{\text{ter}} \cdot (t - \Theta_{\text{ter}})$$

The quantity Θ_{ter} is the abscissa of the point of intersection of this straight line with the t-axis (Fig. 84). The physical meaning of Θ_{ter} is this: if a body begins to move at time $t = \Theta_{\text{ter}}$ with velocity $v = v_{\text{ter}}$, then it will move via the same law as a body actually moves after the force ceases to act.

Exercises

Find solutions to the following differential equations that satisfy the indicated initial data.

1. $\dfrac{d^2x}{dt^2} = 0, \quad x(2) = 1, \quad x'(2) = -3.$

2. $\dfrac{d^2x}{dt^2} = 1, \quad x(0) = -2, \quad x'(0) = 0.$

3. $\dfrac{d^2x}{dt^2} = \sin t, \quad x(0) = 0, \quad x'(0) = 1.$

4. $\dfrac{d^2x}{dt^2} = e^t, \quad x(-\infty) = 0, \quad x'(-\infty) = 0.$

7.5 Second-order nonhomogeneous linear equations with constant coefficients

Green's function can be applied to the general equation (20) which describes the motion of an elastically attached body under the action of an external force dependent solely on the time in the presence of friction that is proportional to the velocity.

As in Sec. 7.4, we seek the solution for the zero initial conditions (33). To construct the Green's function, it is necessary, as in the case of Sec. 7.4 (see (34)), to solve the equation

$$m\frac{d^2y}{dt^2}+h\frac{dy}{dt}+ky=\delta(t-\tau) \tag{44}$$

under the initial conditions (33). Assuming $t>t_0$, we find that $y(t)\equiv 0$ for $t_0<t<\tau$; and integrating (44) from $t=\tau-0$ to $t=$ $=\tau+0$, we arrive at the same conditions (35) and (36), since the integrals of the finite second and third terms in (44) are zero. Thus, for $t>\tau$, it is required to solve the homogeneous equation (21) for the initial conditions (35) and (36). Proceeding from the general solution of equation (21),

$$y=C_1e^{p_1t}+C_2e^{p_2t}$$

where p_1 and p_2 are roots of the characteristic equation (24), and reasoning as we did at the end of Sec. 7.3, we obtain the desired solution

$$y=\frac{e^{-p_1\tau}}{m(p_1-p_2)}e^{p_1t}+\frac{e^{-p_2\tau}}{m(p_2-p_1)}e^{p_2t}=\frac{1}{m(p_1-p_2)}\left[e^{p_1(t-\tau)}-e^{p_2(t-\tau)}\right]$$

This solution is suitable both for a lesser and greater friction. We thus obtain the Green's function:

$$G(t,\tau)=\begin{cases}0 & (t_0<t<\tau),\\ \dfrac{1}{m(p_1-p_2)}\left[e^{p_1(t-\tau)}-e^{p_2(t-\tau)}\right] & (\tau<t<\infty)\end{cases}$$

(As under the conditions of Sec. 7.4, this function is continuous but has a salient point at $t=\tau$.) From this, like (38), we obtain the solution of equation (20) under the zero initial conditions (33):

$$y(t)=\frac{1}{m(p_1-p_2)}\int_{t_0}^{t}\left[e^{p_1(t-\tau)}-e^{p_2(t-\tau)}\right]f(\tau)\,d\tau \tag{45}$$

As in Sec. 7.2 (Item IV), equation (20) can be solved without any Green's function for an external load, particularly of a simple type. This occurs if $f(t)=$ constant, that is, if we consider the equation

$$m\frac{d^2y}{dt^2}+h\frac{dy}{dt}+ky=A\ \ (=\text{constant}) \tag{46}$$

It is easy to find a particular solution of the form $y = B =$ constant. Substituting into (46), we get

$$0 + 0 + kB = A, \text{ or } B = \frac{A}{k}$$

Taking note of the remark made at the beginning of Sec. 7.4, we get the general solution of equation (46):

$$y = \frac{A}{k} + C_1 e^{p_1 t} + C_2 e^{p_2 t} \tag{47}$$

where C_1 and C_2 are arbitrary constants determined from the initial conditions. In Sec. 7.3 we say that the solution of the homogeneous equation (21) tends to zero as t increases, since p_1 and p_2 are either negative real or imaginary with a negative real part; thus, from (47) we get, for t large,

$$y = \frac{A}{k} \tag{48}$$

Physically speaking, this result is clear. Given a constant external force and friction, the oscillations decay, and after the "transient process" determined by the initial conditions passes, the body will stop in a position where the elastic force ky (with sign reversed) will be equal to the external force A, whence we get (48). This stationary position no longer depends on the initial conditions.

The solution of the equation

$$m\frac{d^2y}{dt^2} + h\frac{dy}{dt} + ky = Ae^{qt} \tag{49}$$

is also simple. Here it is easy to find a particular solution of the form $y = Be^{qt}$. Substituting, we get $mBq^2e^{qt} + hBqe^{qt} + kBe^{qt} =$ $= Ae^{qt}$, or $B = \frac{A}{mq^2 + hq + k}$, whence the desired particular solution is of the form

$$y = \frac{Ae^{qt}}{mq^2 + hq + k} \tag{50}$$

This solution is unsuitable if q is a root of the characteristic equation (24), because then the denominator vanishes. As in Sec. 7.2, it can be shown, by proceeding from the general formula (45), that in this special case the equation (49) has a particular solution of the form Bte^{qt}.

Finally, let us consider the solution of the equation

$$m\frac{d^2y}{dt^2} + h\frac{dy}{dt} + ky = A\sin\omega t \tag{51}$$

Here we can take advantage of the fact that by virtue of Euler's formula the right member is the imaginary part of the function $Ae^{i\omega t}$. Hence, it suffices to solve the equation

$$m\frac{d^2y}{dt^2}+h\frac{dy}{dt}+ky=Ae^{i\omega t}$$

and to take the imaginary part (see similar reasoning in Sec. 5.5). On the basis of (50) we get the complex solution

$$y=\frac{Ae^{i\omega t}}{m(i\omega)^2+hi\omega+k}=\frac{A}{k-m\omega^2+ih\omega}e^{i\omega t}$$
$$=\frac{A[(k-m\omega^2)-ih\omega]}{(k-m\omega^2)^2+h^2\omega^2}(\cos\omega t+i\sin\omega t)$$

whence it is easy to isolate the imaginary part, that is, a particular solution of (51):

$$y=\frac{A}{(k-m\omega^2)^2+h^2\omega^2}[(k-m\omega^2)\sin\omega t-h\omega\cos\omega t] \tag{52}$$

In order to find the general solution of equation (51) we have to add to the particular solution (52) the general solution of the corresponding homogeneous equation. But since each of the solutions of the homogeneous equation tends to zero as t increases, it follows that after the transient process the body will begin to vibrate by the harmonic law (52) that does not depend on the initial conditions. This steady-state solution could have been found using the methods of Sec. 5.5.

Let us investigate the case where friction is absent: that is, equation (20) is replaced by

$$m\frac{d^2y}{dt^2}+ky=f(t) \tag{53}$$

Here the characteristic equation has the roots $p_{1,2}=\pm i\omega_0$, where

$$\omega_0=\sqrt{\frac{k}{m}}$$

is the frequency of natural oscillations of the system (that is, without an external force). Formula (45), transformed by Euler's formula, yields a solution for the zero initial conditions (33):

$$y=\frac{1}{m2i\omega_0}\int_{t_0}^{t}[e^{i\omega_0(t-\tau)}-e^{-i\omega_0(t-\tau)}]f(\tau)\,d\tau$$
$$=\frac{1}{m\omega_0}\int_{t_0}^{t}\sin\omega_0(t-\tau)\,f(\tau)\,d\tau \tag{54}$$

(Verify by means of differentiation that the right member does indeed satisfy (53) and the initial conditions (33).) Using the formula

$$\sin \omega_0(t - \tau) = \sin \omega_0 t \cdot \cos \omega_0 \tau - \sin \omega_0 \tau \cdot \cos \omega_0 t$$

we can rewrite the solution as

$$y(t) = \frac{1}{m\omega_0} \sin \omega_0 t \cdot \int_{t_0}^{t} f(\tau) \cos \omega_0 \tau \, d\tau - \frac{1}{m\omega_0} \cos \omega_0 t \cdot \int_{t_0}^{t} f(\tau) \sin \omega_0 \tau \, d\tau$$

If after the elapse of a time interval t_1 the action of the force $f(t)$ ceases, then the integrals in this formula will cease to depend on the time for $t > t_1$ and the solution becomes

$$y(t) = a \cos \omega_0 t + b \sin \omega_0 t \quad \text{if} \quad t > t_1$$

where

$$a = -\frac{1}{m\omega_0} \int_{t_0}^{t_1} f(\tau) \sin \omega_0 \tau \, d\tau, \quad b = \frac{1}{m\omega_0} \int_{t_0}^{t_1} f(\tau) \cos \omega_0 \tau \, d\tau$$

Thus, if originally a body is at rest and then for a time is acted upon by an external force $f(t)$, then when the force ceases to act the body will perform natural oscillations with a frequency ω_0 and an amplitude $\sqrt{a^2 + b^2}$.

Formula (54) can be written differently if we introduce a "complex deviation from the equilibrium position":

$$w(t) = \frac{1}{m\omega_0} \int_{t_0}^{t} e^{i\omega_0(t-\tau)} f(\tau) \, d\tau = \frac{1}{m\omega_0} \int_{t_0}^{t} e^{-i\omega_0\tau} f(\tau) \, d\tau \cdot e^{i\omega_0 t}$$

for which the real deviation $y(t)$ serves as the imaginary part. If t_0 is the initial time of action of the force, then the formula can be rewritten as

$$w(t) = \frac{1}{m\omega_0} \int_{-\infty}^{t} e^{-i\omega_0\tau} f(\tau) \, d\tau \cdot e^{i\omega_0 t}$$

because under this transformation a part equal to zero is added to the integral. If the force acts only up to a certain time t_1, then, from that time onwards, we can write

$$w(t) = \frac{1}{m\omega_0} \int_{-\infty}^{\infty} e^{-i\omega_0\tau} f(\tau) \, d\tau \cdot e^{i\omega_0 t} \tag{55}$$

(this integral is actually extended over the interval from $t = t_0$ to $t = t_1$). The result is a harmonic oscillation with frequency ω_0 and amplitude

$$\frac{1}{m\omega_0} \left| \int_{-\infty}^{\infty} e^{-i\omega_0\tau} f(\tau) \, d\tau \right|$$

Integrals like these (they are called Fourier integrals) will be considered in Ch. 14.

If the driving force is of the form $f(t) = A \sin \omega t$, then we can find a particular solution of the equation by formula (52), which is simplified in the absence of friction (i.e. for $h = 0$) and assumes the form

$$y = \frac{A}{k - m\omega^2} \sin \omega t = \frac{A}{m(\omega_0^2 - \omega^2)} \sin \omega t \qquad (56)$$

Superimposed on this oscillation, which occurs with the frequency ω of the driving force, is a natural oscillation with natural frequency ω_0, which depends on the initial conditions and does not decay in the absence of friction.

Of particular interest is the case where a sinusoidal external force acts on an oscillator without friction under zero initial conditions. The solution is then found from formula (54) and, assuming $t_0 = 0$ for the sake of simplicity, we get

$$\begin{aligned} y &= \frac{1}{m\omega_0} \int_0^t \sin \omega_0 (t - \tau) A \sin \omega\tau \, d\tau \\ &= \frac{A}{m\omega_0} \int_0^t \frac{1}{2} [\cos (\omega\tau - \omega_0(t - \tau)) - \cos (\omega\tau + \omega_0(t - \tau))] \, d\tau \\ &= \frac{A}{2m\omega_0} \left[\frac{\sin (\omega\tau - \omega_0(t - \tau))}{\omega + \omega_0} - \frac{\sin (\omega\tau + \omega_0(t - \tau))}{\omega - \omega_0} \right]_{\tau=0}^{t} \\ &= \frac{A}{2m\omega_0} \left[\frac{\sin \omega t}{\omega + \omega_0} - \frac{\sin \omega t}{\omega - \omega_0} - \frac{\sin (-\omega_0 t)}{\omega + \omega_0} + \frac{\sin \omega_0 t}{\omega - \omega_0} \right] \\ &= \frac{A}{m\omega_0(\omega + \omega_0)(\omega - \omega_0)} (\omega \sin \omega_0 t - \omega_0 \sin \omega t) \qquad (57) \end{aligned}$$

Let the frequency ω of the driving force be close to the natural frequency ω_0 of the oscillator. Transforming the right side of (57) by the formula

$$\begin{aligned} y &= \frac{A}{m\omega_0(\omega + \omega_0)(\omega - \omega_0)} [\omega_0 \sin \omega_0 t - \omega_0 \sin \omega t + (\omega - \omega_0) \sin \omega_0 t] \\ &= \frac{A}{m(\omega + \omega_0)(\omega - \omega_0)} (\sin \omega_0 t - \sin \omega t) + \frac{A}{m\omega_0(\omega + \omega_0)} \sin \omega_0 t \\ &= \frac{2A}{m(\omega + \omega_0)(\omega - \omega_0)} \sin \frac{\omega_0 - \omega}{2} t \cos \frac{\omega_0 + \omega}{2} t + \frac{A}{m\omega_0(\omega + \omega_0)} \sin \omega_0 t \end{aligned}$$

and replacing (approximately) $\omega_0 + \omega$ by $2\omega_0$, we get

$$y \approx \frac{-A}{m\omega_0(\omega - \omega_0)} \sin \frac{\omega - \omega_0}{2} t \cos \omega_0 t + \frac{A}{2m\omega_0^2} \sin \omega_0 t$$

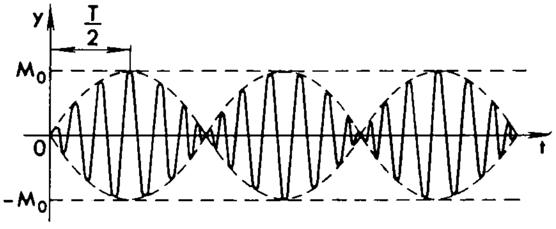


Fig. 85

The most interesting thing is the first, principal, term. It can be written in the form

$$M(t)\cos\omega_0 t, \text{ where } M(t) = \frac{-A}{m\omega_0(\omega-\omega_0)}\sin\frac{\omega-\omega_0}{2}t$$

and interpreted as a harmonic oscillation having frequency and a slowly varying amplitude. This amplitude varies from 0 to

$$M_0 = \max|M(t)| = \frac{A}{m\omega_0|\omega-\omega_0|} \tag{58}$$

with period

$$T = \frac{2\pi}{|\omega-\omega_0|}$$

Oscillations like these are termed *beats* (the graph is shown in Fig. 85). They result from the interference of a forced frequency (56) and a natural frequency due to their being close in amplitude and frequency (see formula (57)). We see that both the build-up time for beats and the amplitude of the beats are inversely proportional to $|\omega-\omega_0|$, which is the difference between the frequency of natural oscillations and that of the driving force.

If the oscillator has a low friction, then the oscillation process will also begin with beats, but after a sufficient time the natural oscillation decays and only the constrained oscillation (52) is left. Its amplitude is equal, for very small h, to

$$\frac{A}{(k-m\omega^2)^2+h^2\omega^2}\sqrt{(k-m\omega^2)^2+h^2\omega^2} = \frac{A}{\sqrt{(k-m\omega^2)^2+h^2\omega^2}}$$

$$\approx \frac{A}{|k-m\omega^2|} \approx \frac{A}{m|\omega_0^2-\omega^2|} = \frac{A}{m|\omega_0-\omega|(\omega_0+\omega)} \approx \frac{A}{2m\omega_0(\omega-\omega_0)}$$

Comparing this with formula (58), we see that the amplitude of the constrained oscillations is equal to half the amplitude of the beats. The oscillation curve thus has the form shown in Fig. 86. The time interval during which the beats turn into purely harmonic oscillations is the transient period, and the process itself is called a transient process.

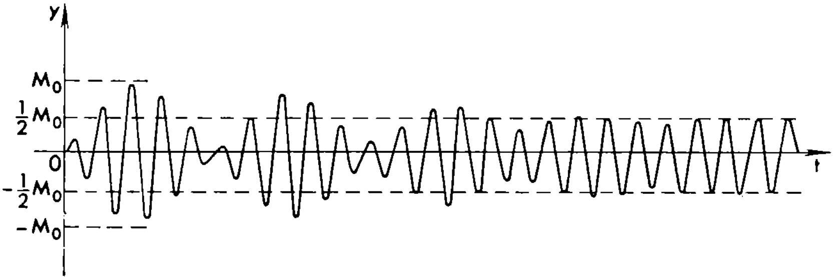


Fig. 86

If the oscillator is without friction, then in the special case of $\omega = \omega_0$, that is, the frequency of the driving force coincides with the natural frequency, formula (56) is inapplicable. Recall that this is precisely the case that we skipped at the end of Sec. 5.5. Let us take advantage of the general formula (54) and assume for the sake of simplicity $t_0 = 0$:

$$\begin{aligned} y &= \frac{1}{m}\int_0^t \sin \omega_0(t-\tau)\cdot A \sin \omega_0\tau \, d\tau \\ &= \frac{A}{2m}\int_0^t [\cos \omega_0(t-2\tau) - \cos \omega_0 t] \, d\tau \\ &= \frac{A}{2m}\left[\frac{\sin \omega_0(t-2\tau)}{-2\omega_0} - \tau \cos \omega_0 t\right]_{\tau=0}^{t} \\ &= \frac{A}{2m\omega_0}\sin \omega_0 t - \frac{A}{2m} t \cos \omega_0 t \end{aligned}$$

The first of the terms is a natural harmonic oscillation and is present solely to satisfy the zero boundary conditions. In contrast, the second term is an oscillation whose amplitude tends to infinity linearly with time. Therein lies the very important phenomenon of resonance, which results when the frequency of the driving force coincides with the natural frequency of the system.

Exercises

Find solutions of the following differential equations that satisfy the indicated initial data.

1. $y'' - y = 1, \quad y = 0, \quad y' = 0 \quad$ for $t = 0$.
2. $y'' + y = t, \quad y = 1, \quad y'' = 0 \quad$ for $t = 0$.

7.6 Stable and unstable solutions

We begin with the simplest equation

$$\frac{dy}{dt} = ay \quad (a = \text{constant}) \tag{59}$$

with the variable t interpreted as the time. It is easy to obtain the general solution:

$$y = Ce^{at} \tag{60}$$

where C is an arbitrary constant determined from the initial condition

$$y(t_0) = y_0$$

Substituting into (60), we get

$$y_0 = Ce^{at_0}, \text{ or } C = y_0 e^{-at_0}$$

and finally

$$y = y_0 e^{a(t-t_0)} \tag{61}$$

In particular, when $y_0 = 0$ we get the zero solution $y \equiv 0$. Now suppose that the initial value y_0, which we regarded as zero, is actually different from zero, albeit just slightly. Then how will the solution behave with time, that is, as t increases? Will such a *perturbed* solution approach the *unperturbed* zero solution or will it recede from it?

The answer to these questions depends essentially on the sign of the coefficient a. If $a < 0$, then (61) shows immediately that as t increases the solutions approach zero without limit, so that for large t they practically vanish. In such a situation, the unperturbed solution is said to be *asymptotically stable relative to the change* (perturbation) *of the initial condition* or *asymptotically stable in the sense of Lyapunov* (after the celebrated Russian mathematician A. M.Lyapunov, 1857—1918), who was the first to begin a systematic study of the concept of stability of processes.

The picture will be quite different for $a > 0$. Here, when $y_0 \neq 0$ and t is increasing, the solution increases in absolute value without bound, that is to say, it becomes significant even if y_0 was arbitrarily small. Here the unperturbed solution is said to be unstable. For $a > 0$ we have equation (59), for example, when considering the growth of bacteria in a nutrient medium with y denoting the mass of bacteria per unit volume and a, the intensity of growth. It is clear that if at the initial time there were no bacteria at all, then of course none will appear in the course of time. But this picture is unstable in the sense that a purposeful or accidental introduction of any arbitrarily small amount of bacteria into the medium will, in time, lead to extreme growth and pollution of the medium with bacteria.

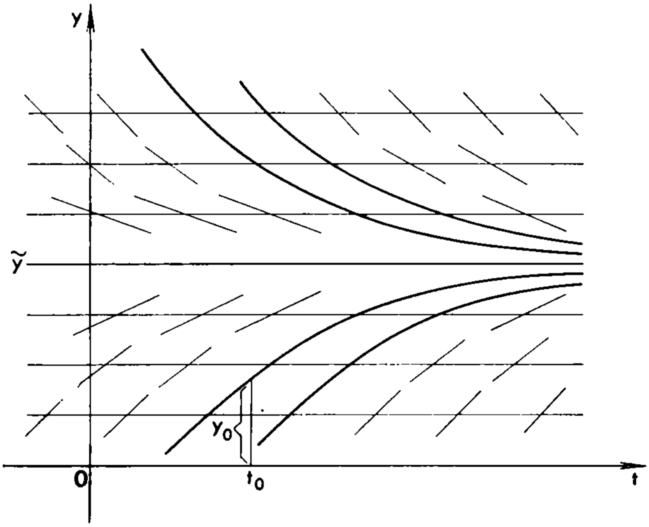


Fig. 87

The intermediate case $a = 0$ is interesting. Here the solutions will merely be constant and for this reason for a small initial deviation of the perturbed solution from the unperturbed solution the former will be close to the latter even when t increases, although the approach will not be asymptotic (as $t \to \infty$). Such a picture is called *nonasymptotic stability of an unperturbed solution.*

Now let us consider the more general equation

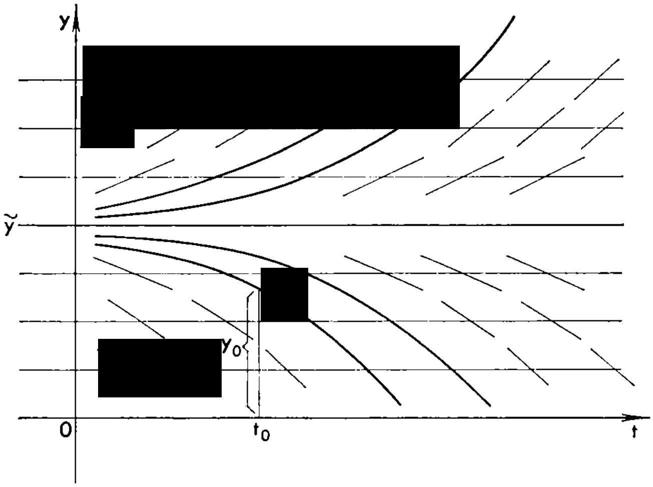
$$\frac{dy}{dt} = f(y) \tag{62}$$

It is easy to find all the stationary solutions, that is, solutions of the form $y =$ constant. To do this, in (62) set $y = \tilde{y} =$ constant to get

$$f(\tilde{y}) = 0 \tag{63}$$

Thus the stationary solutions of (62) are the zeros of the function $f(y)$ in the right-hand member. Let us examine one such solution $y = \tilde{y}$ and determine whether it is stable or not.

First assume that the function $f(y)$ is decreasing, at least in a certain neighbourhood of the value $y = \tilde{y}$; then if y passes through the value $\tilde{y}$, it follows that $f(y)$ passes from positive values to negative values. In this case the approximate picture of the direction field defined by equation (62) (cf. Sec. 7.1) is shown in Fig. 87. In constructing this field, bear in mind that the isoclines for (62) have the form $y =$ constant (why?), that is, they are straight lines parallel to the t-axis (they are also shown in Fig. 87). In Fig. 87 the heavy lines indicate the integral straight line $y = \tilde{y}$, which depicts an unperturbed stationary solution, and several other integral curves that result



[redacted] [redacted] [redacted]

Fig. 88

from changes in the initial condition. It is clear that if y_0 differs but slightly from $\tilde{y}$ (say, within the limits of the drawing), then the perturbed solution does not differ appreciably from the unperturbed one even when t increases, and when $t \to \infty$ it asymptotically approaches the unperturbed solution. Thus, in this case the unperturbed solution is asymptotically stable.

Now let $f(y)$ increase from negative values to positive ones when y passes through the value $\tilde{y}$. The appropriate picture is given in Fig. 88. It is clear that no matter how close y_0 is to $\tilde{y}$ (but not equal to $\tilde{y}$!), the appropriate solution $y(t)$ will recede from the unperturbed solution to a finite but substantial distance as t increases. This means that in the case at hand the unperturbed stationary solution is unstable. (Check to see that the earlier obtained criteria for stability and instability for equation (59) can be obtained as a consequence of general criteria indicated for equation (62).)

The criteria obtained here can be derived differently, without resorting to a geometric picture. Expand the right side of (62) in a power series about the value $y = \tilde{y}$; then by virtue of the condition (63) there will not be a constant term in the expansion and we get

$$\frac{dy}{dt} = f'(\tilde{y})\,(y - \tilde{y}) + \ldots$$

That is,

$$\frac{d(y - \tilde{y})}{dt} = f'(\tilde{y})\,(y - \tilde{y}) + \ldots \tag{64}$$

where the dots stand for higher-order terms. It must be stressed that when determining stability in the sense of Lyapunov, we study the behaviour of perturbed solutions that differ but slightly from the unperturbed solution, that is to say, we consider only small values of $y - \tilde{y}$. For this reason, the main role in the right-hand member of (64) is played by the linear term that is written out. Discarding higher-order terms, we obtain an equation of the form (59) in which $a = f'(\tilde{y})$. Applying the results obtained above for equation (59), we see that for $f'(\tilde{y}) < 0$ the solution $y - \tilde{y} = 0$, or $y = \tilde{y}$, is asymptotically stable; but if $f'(\tilde{y}) > 0$, then the solution $y = \tilde{y}$ is unstable. But in the first case the function $f(y)$ decreases (at least about the value $y = \tilde{y}$), while in the latter case, it increases, so that we arrive at the same conclusions that were obtained via geometrical reasoning. In a particular case where $f'(y) = a = 0$ for equation (59) we have nonasymptotic stability, that is, the solutions close to the unperturbed solution do not tend to it and do not recede from it; then in the complete equation (64) a substantial role is played by higher-order terms, which in one instance can direct the perturbed solutions to the unperturbed solution, while in another can carry them away to a substantial distance. We will not discuss this particular case.

To illustrate, let us examine the thermal regime in a certain volume where we have a chemical reaction associated with release of heat, which is carried off into the ambient space. Since the rate of the reaction depends on the temperature T in the given volume (we consider the mean temperature at a given time t), the rate Q of heat release in the reaction depends on T, $Q = Q(T)$. We assume the relationship to be as shown in Fig. 89. We also suppose the rate of heat dissipation into the ambient space to be equal to $a(T - T_0)$, where a is the proportionality constant and T_0 is the temperature of the ambient medium. Then, for a constant heat capacity c of the volume at hand, the differential equation of the process becomes

$$\frac{d(cT)}{dt} \equiv c\frac{dT}{dt} = Q(T) - a(T - T_0) \qquad (65)$$

By virtue of the foregoing, a stationary state (in which the temperature remains constant during the reaction) is possible for those T for which the right side vanishes, which means that the graph of $Q(T)$ crosses the graph of $a(T - T_0)$ (see Fig. 89). We see that if the ambient temperature T_0 is sufficiently great (when $T_0 = \overline{T}_0$ in Fig. 89), a stationary state is impossible: the supply of heat will all the time be greater than dissipation and the volume will constantly heat up. If the temperature is low (when $T_0 = \overline{\overline{T}}_0$ in Fig. 89), two stationary states having temperatures T_1 and T_2 are conceivable. Near the value T_1, the right member of (65) passes from positive values to negative values, which means it decreases. We have already seen that such a state is a stable state. This is evident from Fig. 89, for if the temperature T falls below T_1, then more heat will be released

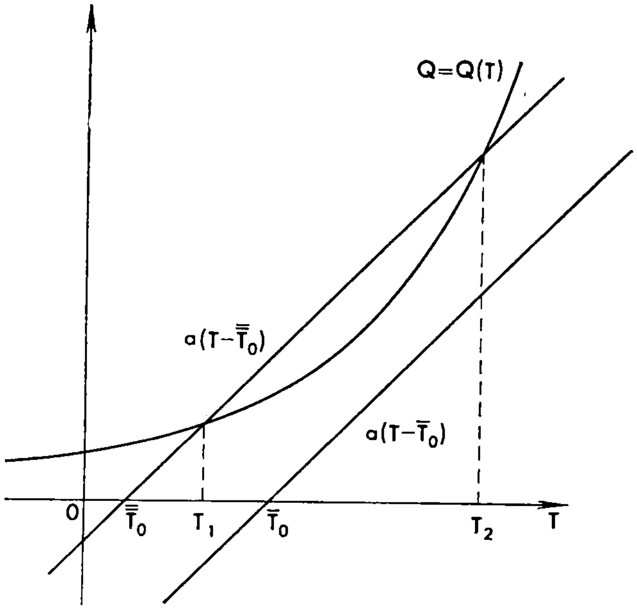


Fig. 89

in the reaction than is dissipated, which means the volume will heat up, and if T rises above T_1, then more heat will be dissipated than released and the volume will cool off. In similar fashion we can verify that the stationary temperature T_2 will be unstable. Thus, when $T_0 = \bar{\bar{T}}_0$ the development of the process depends on the initial temperature as follows: if it was less than T_2, then in time the temperature tends to the stationary value T_1; if the initial temperature was greater than T_2, then the temperature builds up catastrophically. Such was the reasoning that served as the basis for the theory of thermal explosion developed by Nobel Laureate N. N. Semenov in 1927.

Now let us take up the equation of free oscillations:

$$m\frac{d^2y}{dt^2} + h\frac{dy}{dt} + ky = 0 \qquad (m,\ h,\ k > 0) \tag{21}$$

with the general solution

$$y = C_1e^{p_1t} + C_2e^{p_2t} \tag{66}$$

where p_1 and p_2 are roots of the characteristic equation (24) and C_1 and C_2 are arbitrary constants determined from the initial conditions. This equation has the stationary solution $y \equiv 0$. In Sec. 7.3 we saw that all other solutions tend to zero (in an oscillatory or nonoscillatory manner) as t increases; thus the indicated stationary solution is asymptotically stable. In the absence of friction, that is when $h = 0$, we saw that the solutions are periodic. For this reason, the solution will be small even when t increases, given a small initial deviation and a small initial velocity, but it will not tend to zero. Hence in the absence of friction the stationary solution will be stable but not asymptotically stable.

Using specially chosen schemes, one can construct systems with one degree of freedom described by equation (21) (where y is the deviation of the system from the stationary state) for which $h < 0$ or $k < 0$. Such systems may be interpreted as systems with negative friction or with negative elasticity. (See, for example, the description of the operation of a tunnel diode given in HM, Sec. 8.16; in this mode of operation, the system may be interpreted as an oscillator with negative friction.) It can be readily verified that in all such systems the stationary solution $y \equiv 0$ is unstable. Indeed, from algebra we know the properties of the roots p_1 and p_2 of the quadratic equation (24):

$$p_1 + p_2 = -\frac{h}{m}, \quad p_1 p_2 = \frac{k}{m}$$

From the first equation it is seen that if $h < 0$, then either at least one root is positive and real or the roots are conjugate imaginaries with positive real part. From the second equation we see that if $k < 0$, then the roots are of different sign and for this reason there is one positive root. Thus in all cases there is at least one root that is positive and real or imaginary with a positive real part. Let p_1 be that root. Then the first term on the right of (66) is of the form

$$C_1 e^{p_1 t} (p_1 > 0) \text{ or } C_1 e^{(\gamma + i\omega)t} = C_1 e^{\gamma t} (\cos \omega t + i \sin \omega t) \ (\gamma > 0)$$

and therefore, for arbitrarily small C_1 (which is to say, for arbitrarily small initial data), it tends to infinity as t increases. This is what signifies the instability of a stationary solution.

Exercise

Find the stationary solutions of the equation $\frac{dy}{dt} = y^3 - 4y$ and determine their stability.

ANSWERS AND SOLUTIONS

Sec. 7.1

1. The circles $x^2 + y^2 = C$ with centre at the coordinate origin.
2. At points of inflection it must be true that $y'' = 0$. By virtue of equation (1) and the formula for the derivative of a composite function (Sec. 4.1), we get $y'' = (f(x, y))' = f'_x(x, y) + f'_y(x, y)\, y' = f'_x + f'_y f$. Hence the desired equation is of the form $f'_x + f'_y f = 0$. For the equation $\frac{dy}{dx} = x^2 + y^2$ we get $x + y\,(x^2 + y^2) = 0$.

Sec. 7.2

1. $y = e^{x^2}$.
2. $y = x$.
3. $y = \ln (x + e)$.
4. $y = \sqrt{1 + x^2}$.
5. $y = \frac{e^x}{2} - \left(\frac{e^2}{2} - 1\right) e^{-x}$.
6. $y = 2x - 1 - e^{-2x}$.
7. $y = e^{-x} + \frac{1}{2} (\sin x + \cos x)$.

Sec. 7.3

1. $y = 2 \cos t$.
2. $y = \frac{1}{4} e^{4t} \sin 2t$.
3. $y = e^{2t} + e^t$.
4. $y = e^t + 3e^{-t}$.
5. $y = (t - 1) e^t$.
6. $y = (5t + 1) e^{-2t}$.

Sec. 7.4

1. $x = -3(t - 2) + 1 =$
   $= - 3t + 7$.
2. $x = \frac{t^2}{2} - 2$.
3. $x = 2t - \sin t$.
4. $x = \int\limits_{-\infty}^{t} (t - \tau) e^{\tau} d\tau = e^t$.

Sec. 7.5

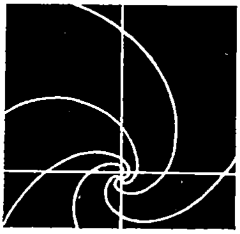
1. $y = \frac{1}{2} (e^t + e^{-t}) - 1$.
2. $y = t + \cos t - \sin t$.

Sec. 7.6

$y = \pm 2$ (unstable), $y = 0$ (stable).

Chapter 8

DIFFERENTIAL EQUATIONS CONTINUED



This chapter is a direct continuation of the preceding one, which serves as an essential foundation.

8.1 Singular points

In Sec. 7.1 we established that the integral curves of the equation $\frac{dy}{dx} = f(x, y)$ do not intersect. However there is an important exception to this rule. It may indeed happen that for certain values of x and y the function $f(x, y)$ does not have a definite value. For example, $f(x, y) = \frac{y}{x}$ does not have a definite value for $x = 0$, $y = 0$. The point in the plane at which $f(x, y)$ becomes meaningless is called a *singular point* of the differential equation $\frac{dy}{dx} = f(x, y)$. Several integral curves can pass through a singular point.

If $f(x, y)$ has the form of a ratio of two functions of a simple type (say, two polynomials), $f(x, y) = \frac{\varphi(x, y)}{\psi(x, y)}$, then the coordinates of the singular point are found from the system of equations

$$\left.\begin{aligned} \varphi(x, y) &= 0, \\ \psi(x, y) &= 0 \end{aligned}\right\}$$

Let us examine several examples of singular points that belong to the types most frequently encountered in applications.

1. $\frac{dy}{dx} = \frac{y}{x}$. The solution of this equation is the function $y = Cx$ for any constant C. The collection of integral curves are all straight lines passing through the origin (Fig. 90). Thus the integral curves intersect at the singular point $x = 0$, $y = 0$.

2. $\frac{dy}{dx} = \frac{2y}{x}$. The solution is $y = Cx^2$. The integral curves are parabolas with vertex at the origin. Here, all the integral curves are tangent at the singular point $x = 0$, $y = 0$ (Fig. 91).

Fig. 90

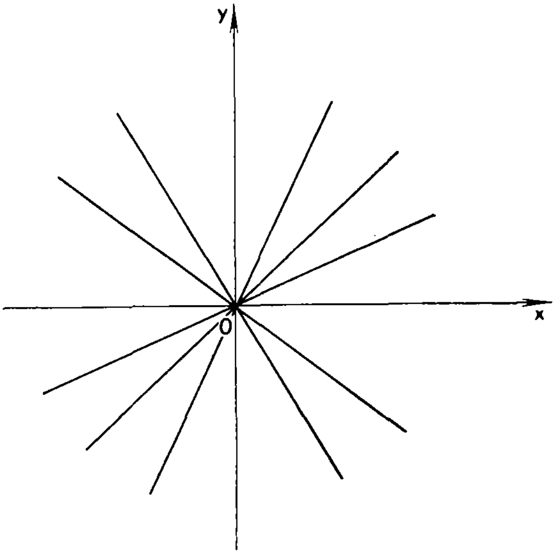
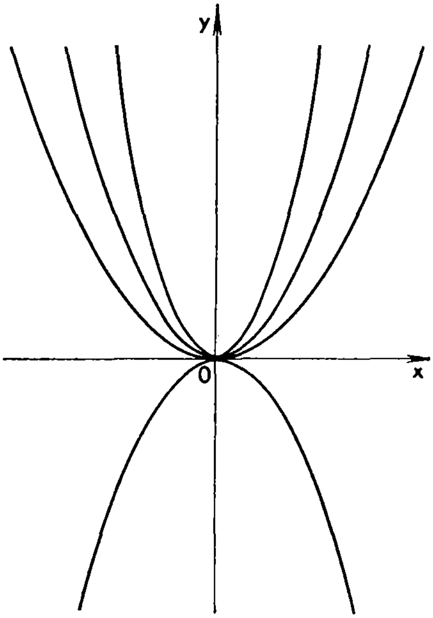


Fig. 91



In the foregoing examples, all the integral curves pass through the singular point and have a definite direction there. Such a singular point is called a *nodal point.*

3. There are also singular points near which integral curves behave differently. Let $\frac{dy}{dx} = -\frac{y}{x}$. The integral curves have the equation $xy = C$. For $C = 0$ we get $x = 0$ or $y = 0$, which are two straight lines passing through the origin. For $C \neq 0$, the integral curves are hyperbolas. Thus, two integral curves pass through the singular point $x = 0$, $y = 0$, and the others do not pass through it. A singular point of this type is called a *saddle point* (Fig. 92).

4. The integral curves of the equation $\frac{dy}{dx} = -\frac{x}{y}$ are the circles $x^2 + y^2 = C$. Here the integral curves enclose the singular point but not a single integral curve passes through it (Fig. 93). A singular point of this type is called a *vortex point.*

5. When integrating the equation $\frac{dy}{dx} = \frac{x+y}{x-y}$ with singular point at the origin, it is convenient to pass to polar coordinates ρ, φ via the formulas $x = \rho \cos \varphi$, $y = \rho \sin \varphi$, whence

$$dx = \cos \varphi \cdot d\rho - \rho \sin \varphi \, d\varphi, \quad dy = \sin \varphi \cdot d\rho + \rho \cos \varphi \, d\varphi$$

After substituting into the equation and collecting terms, we get $d\rho = \rho \, d\varphi$ (verify this), whence $\frac{d\rho}{\rho} = d\varphi$ and $\rho = Ce^{\varphi}$. Assigning all possible values to C, we get a family of spirals closing in on the origin (Fig. 94). A singular point of this type is called a *focal point.*

In the foregoing examples it is easy to establish the type of behaviour of integral curves near the singular point, since it is easy to solve the differential equations. In more complicated cases one can get a rough idea of the nature of the singular point by drawing isoclines. More effective procedures for investigating singular points lie outside the scope of this book. Application of these procedures shows, in particular, that if the condition $\varphi'_x\psi'_y \neq \varphi'_y\psi'_x$ holds, the singular point must definitely belong to one of the foregoing types.

Exercise

Determine the nature of the singular point for the equations

$$\frac{dy}{dx} = \frac{x}{y}, \quad \frac{dy}{dx} = \frac{x + 2y}{x}, \quad \frac{dy}{dx} = -\frac{2x + y}{x + 2y}.$$

8.2 Systems of differential equations

So far we have assumed that there is one differential equation from which it is necessary to find one desired function. But there may be more than one unknown function, say two: $y(x)$ and $z(x)$.

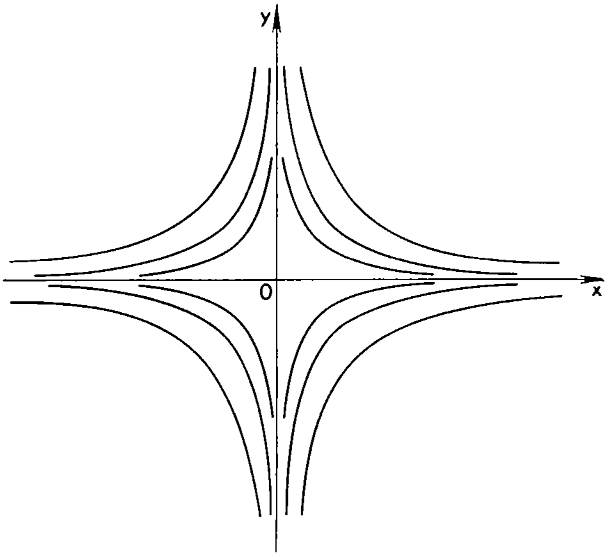


Fig. 92

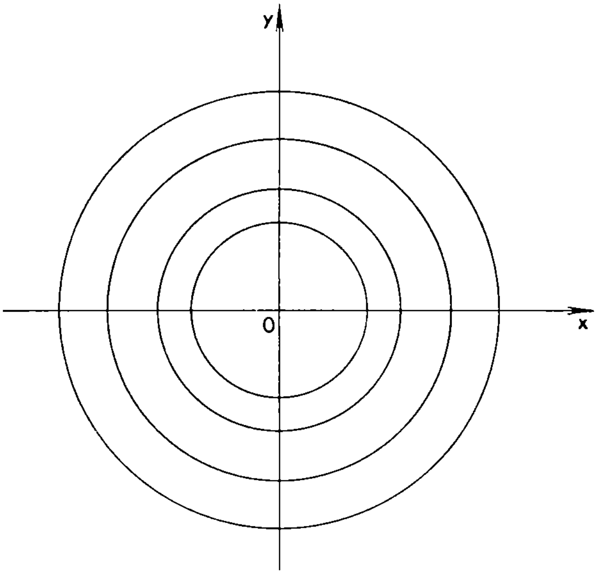


Fig. 93

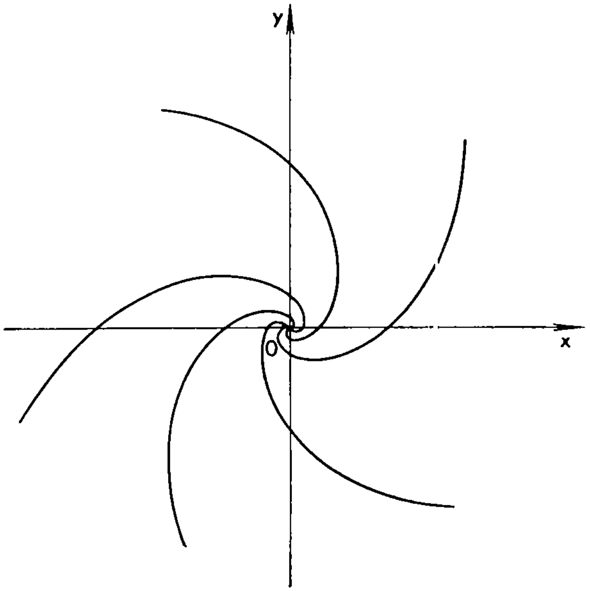


Fig. 94

Then there must be two differential equations as well. If these are of the first order (solved for the derivatives of the desired functions), then they have the general form

$$\left.\begin{aligned} \frac{dy}{dx} &= f(x, y, z), \\ \frac{dz}{dx} &= \varphi(x, y, z) \end{aligned}\right\} \qquad (1)$$

We thus obtain a system of first-order differential equations.

It is easy to pass from one equation of the second order,

$$\frac{d^2y}{dx^2} = F\left(x, y, \frac{dy}{dx}\right) \qquad (2)$$

in one unknown function to an equivalent system of two first-order equations with two unknown functions. For this we have to regard $\frac{dy}{dx}$ as the additional unknown function. Denoting it by z, we get

$$\frac{dy}{dx} = z, \quad \frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{dz}{dx}$$

And so instead of (2) we can write the equivalent system

$$\left.\begin{aligned} \frac{dy}{dx} &= z, \\ \frac{dz}{dx} &= F(x, y, z) \end{aligned}\right\}$$

Similarly, an equation of the third order,

$$\frac{d^3y}{dx^3} = \Phi\left(x, y, \frac{dy}{dx}, \frac{d^2y}{dx^2}\right)$$

can be replaced by an equivalent system of three equations of the first order. To do this, set

$$\frac{dy}{dx} = z, \ \frac{d^2y}{dx^2} = \frac{dz}{dx} = u$$

to get the system

$$\left.\begin{aligned} &\frac{dy}{dx} = z, \\ &\frac{dz}{dx} = u, \\ &\frac{du}{dx} = \Phi(x, y, z, u) \end{aligned}\right\}$$

In similar fashion it is possible to go from an equation of any order or even a system of equations of any order to a system of the first order. Conversely, it can be verified that it is possible to pass from a system of n equations of the first order with n unknown functions to one equation of order n with one unknown function. For this reason, the general solution of such a system is obtained as the result of n integrations and, thus, contains n arbitrary constants. The specification of n initial conditions (for some value of x we specify the values of all desired functions) is just sufficient to find the values of these constants and, thus, to obtain a particular solution.

For the sake of simplicity, let us consider a system of two first-order equations of type (1). It is sometimes possible, even without solving the system, to find a relation connecting the components (that is, $y(x)$ and $z(x)$) of any particular solution. Such a relation is of the form

$$H(x, y, z; C) = 0 \tag{3}$$

(here, C is a constant that varies from solution to solution), and is called the *first integral* of the system of equations. A knowledge of the first integral makes it possible to lower the number of equations in the system by one, that is, to pass to one equation of the first order with one unknown function. We can use (3) and express z in terms of the rest and substitute the result into the first equation of (1) to get one first-order equation with one desired function $y(x)$. If this equation is integrated and we find $y(x)$, then $z(x)$ may be found from (3) without integrations.

Similarly, a knowledge of two independent first integrals (that is, such that neither is a consequence of the other),

$$\left.\begin{aligned} H_1(x,\ y,\ z;\quad C_1) &= 0, \\ H_2(x,\ y,\ z;\quad C_2) &= 0 \end{aligned}\right\}$$

produces the general solution of system (1) written in implicit form. For a system of n first-order equations the general solution is obtained from n independent first integrals.

In certain cases the first integrals are suggested by physical reasoning, most often by conservation laws. For example, let us write down the equation of one-dimensional elastic linear oscillations without friction (see Sec. 7.3),

$$m\,\frac{d^2x}{dt^2} + kx = 0$$

in the form of a system of the first order:

$$\frac{dx}{dt} = v, \quad m\,\frac{dv}{dt} = -\,kx \tag{4}$$

In Sec. 7.3 we already mentioned the expression for the total energy of an oscillating point:

$$E = \frac{mv^2}{2} + \frac{kx^2}{2} \tag{5}$$

The energy should be conserved in the case of free oscillations without friction. Sure enough, by virtue of (4),

$$\frac{dE}{dt} = mv\,\frac{dv}{dt} + kx\,\frac{dx}{dt} = -\,kxv + kxv = 0$$

(this is a mathematical demonstration of the law of conservation of energy in the given instance). Thus, E = constant for any solution of system (4), which means that formula (5), in which E plays the part of the arbitrary constant C, serves as the first integral of the system (4).

As is evident from the foregoing, it is most natural to consider systems in which the number of equations is equal to the number of unknown functions; such systems are called *closed* systems. If there are fewer equations than desired functions, the system is said to be open (underdetermined); in most cases the openness of the system indicates merely that not all the necessary relations have been written down. If there are more equations than unknown functions, then the system is said to be overdetermined; overdetermination of a system ordinarily indicates that it is dependent (i.e. that some of the equations are consequences of others and therefore superfluous) or that a mistake has been made in setting up the system.

Exercise

Consider the system of equations

$$\frac{dy}{dx} = y + z, \quad \frac{dz}{dx} = -y + z$$

Multiply the first equation by y and the second by z and then add the results to find the first integral of the system. What conclusions can be drawn from this concerning the behaviour of particular solutions as $x \to \pm \infty$?

8.3 Determinants and the solution of linear systems with constant coefficients

First let us examine the notion of a determinant, for it will play an important role in the solution and investigation of systems of linear equations of various kinds. We begin with a system of two algebraic equations of the first degree in two unknowns:

$$\left.\begin{aligned} a_1x + b_1y &= d_1, \\ a_2x + b_2y &= d_2 \end{aligned}\right\} \tag{6}$$

Solving it (we leave this to the reader), we get

$$x = \frac{d_1b_2 - b_1d_2}{a_1b_2 - b_1a_2}, \quad y = \frac{a_1d_2 - d_1a_2}{a_1b_2 - b_1a_2} \tag{7}$$

The expression $a_1b_2 - b_1a_2$ is called a *determinant* of the second order and is written as

$$a_1b_2 - b_1a_2 = \begin{vmatrix} a_1 b_1 \\ a_2 b_2 \end{vmatrix} \tag{8}$$

where the vertical lines are the sign of the determinant. Using this notation, we can write (7) as

$$x = \frac{\begin{vmatrix} d_1 b_1 \\ d_2 b_2 \end{vmatrix}}{\begin{vmatrix} a_1 b_1 \\ a_2 b_2 \end{vmatrix}}, \quad y = \frac{\begin{vmatrix} a_1 d_1 \\ a_2 d_2 \end{vmatrix}}{\begin{vmatrix} a_1 b_1 \\ a_2 b_2 \end{vmatrix}} \tag{9}$$

To illustrate, let us evaluate a determinant:

$$\begin{vmatrix} 0 & -3 \\ 2 & 1 \end{vmatrix} = 0 \cdot 1 - (-3) \cdot 2 = 0 + 6 = 6$$

A similar solution of the system of equations

$$\left.\begin{aligned} a_1x + b_1y + c_1z &= d_1, \\ a_2x + b_2y + c_2z &= d_2, \\ a_3x + b_3y + c_3z &= d_3 \end{aligned}\right\} \tag{10}$$

yields the formulas

$$x = \frac{\begin{vmatrix} d_1 & b_1 & c_1 \\ d_2 & b_2 & c_2 \\ d_3 & b_3 & c_3 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}}, \quad y = \frac{\begin{vmatrix} a_1 & d_1 & c_1 \\ a_2 & d_2 & c_2 \\ a_3 & d_3 & c_3 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}}, \quad z = \frac{\begin{vmatrix} a_1 & b_1 & d_1 \\ a_2 & b_2 & d_2 \\ a_3 & b_3 & d_3 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}} \qquad (11)$$

where we have third-order determinants in the numerators and denominators; they are computed from the formula

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} = a_1b_2c_3 - a_1c_2b_3 - b_1a_2c_3 + b_1c_2a_3 + c_1a_2b_3 - c_1b_2a_3 \qquad (12)$$

Formulas (11) are completely analogous to formulas (9). In the denominator we have one and the same determinant made up of the coefficients of the unknowns (the so-called *determinant of the system*). In the numerator, for each of the unknowns we have a determinant obtained from the determinant of the system by substituting the column of constant terms for the column of coefficients of the given unknown.

A determinant of the third order can be expressed in terms of determinants of the second order. To do this, transform the expression (12) and take advantage of (8):

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} = a_1(b_2c_3 - c_2b_3) - b_1(a_2c_3 - c_2a_3) + c_1(a_2b_3 - b_2a_3)$$

$$= a_1\begin{vmatrix} b_2 & c_2 \\ b_3 & c_3 \end{vmatrix} - b_1\begin{vmatrix} a_2 & c_2 \\ a_3 & c_3 \end{vmatrix} + c_1\begin{vmatrix} a_2 & b_2 \\ a_3 & b_3 \end{vmatrix} \qquad (13)$$

This formula can be used to compute the value of the determinant. For example,

$$\begin{vmatrix} 1 & 0 & -2 \\ -1 & 1 & \frac{1}{2} \\ 3 & 1 & 2 \end{vmatrix} = 1\begin{vmatrix} 1 & \frac{1}{2} \\ 1 & 2 \end{vmatrix} - 0\begin{vmatrix} -1 & \frac{1}{2} \\ 3 & 2 \end{vmatrix} + (-2)\begin{vmatrix} -1 & 1 \\ 3 & 1 \end{vmatrix}$$

$$= 1\left(1\cdot 2 - \frac{1}{2}\cdot 1\right) - 2\,(-1\cdot 1 - 1\cdot 3)$$

$$= \frac{3}{2} + 8 = 9\,\frac{1}{2}$$

It turns out that formulas like (9) and (11) hold true for systems consisting of any number of equations of the first degree if the number of unknowns is equal to the number of equations. Determinants of the fourth order are computed by analogy with (13):

$$\begin{vmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \\ a_4 & b_4 & c_4 & d_4 \end{vmatrix} = a_1 \begin{vmatrix} b_2 & c_2 & d_2 \\ b_3 & c_3 & d_3 \\ b_4 & c_4 & d_4 \end{vmatrix} - b_1 \begin{vmatrix} a_2 & c_2 & d_2 \\ a_3 & c_3 & d_3 \\ a_4 & c_4 & d_4 \end{vmatrix}$$

$$+ c_1 \begin{vmatrix} a_2 & b_2 & d_2 \\ a_3 & b_3 & d_3 \\ a_4 & b_4 & d_4 \end{vmatrix} - d_1 \begin{vmatrix} a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \\ a_4 & b_4 & c_4 \end{vmatrix}$$

Pay special attention to the structure of the expression on the right. Determinants of order five are determined in terms of determinants of the fourth order, and so on.

In formulas (9) and (11) it is assumed that the determinant of the system in the denominators is not equal to zero. In this case, the system (6) itself (and respectively, (10)) has a unique solution, which is understood as the solution set, or the set of values of all the unknowns. Occasionally, one encounters systems with a determinant equal to zero; their properties are quite different.

For the sake of simplicity, consider system (6). If its determinant is zero,

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = a_1b_2 - b_1a_2 = 0$$

then

$$a_1b_2 = b_1a_2, \quad \frac{a_1}{a_2} = \frac{b_1}{b_2}$$

which means the left members of the system are proportional. For instance, a system can have the form

$$\left.\begin{aligned} 2x + 3y &= d_1, \\ 8x + 12y &= d_2 \end{aligned}\right\} \qquad (14)$$

It is clear that if the right-hand sides are chosen arbitrarily, the system is inconsistent, or contradictory (it does not have a single solution). And only if the right members are in the same proportion as the left members (in this example, $d_2 = 4d_1$), then one of the equations is a consequence of the other and so can be dropped. But we then get one equation in two unknowns of the type

$$2x + 3y = d_1$$

which has an infinity of solutions: we can assign any value to x and find the corresponding value of y.

It appears that this situation is typical. Namely, that if the determinant of a system is zero, then a definite relationship exists between the left members of the system. If that relation holds true for the right members, then the system has an infinity of solutions; otherwise there is not a single solution.

An important particular case is a system of n linear homogeneous (without constant terms, that is) equations in n unknowns. For example, for $n = 3$ the system is of the form

$$\left.\begin{aligned} a_1x + b_1y + c_1z &= 0, \\ a_2x + b_2y + c_2z &= 0, \\ a_3x + b_3y + c_3z &= 0 \end{aligned}\right\}$$

Such a system of course has the zero (or trivial, hence uninterestingr solution $x = y = z = 0$. It is often important to determine whethe-there are any other (nontrivial) solutions. The foregoing gives the ans, wer immediately. If the determinant of the system is not equal to zero) then there is only one solution and, hence, there are no nontrivial solutions. But if it is equal to zero, then the system has an infinity of nontrivial solutions, since it cannot be inconsistent in this case. To find these solutions, we discard one of the equations of the system, as was done with respect to system (14).

Now let us apply the results obtained to the solution of a system of homogeneous linear differential equations with constant coefficients.

Consider the system

$$\left.\begin{aligned} \frac{dy}{dx} &= a_1y + b_1z, \\ \frac{dz}{dx} &= a_2y + b_2z \end{aligned}\right\} \qquad (15)$$

in which all coefficients a_1, b_1, a_2, b_2 are constants. The particular solutions are sought in the form

$$y = \lambda e^{px}, \quad z = \mu e^{px} \qquad (16)$$

where λ, μ, p are as yet unknown constants. Substitution into (15) yields, after cancelling e^{px} and transposing all terms to one side,

$$\left.\begin{aligned} (a_1 - p)\,\lambda + b_1\mu &= 0, \\ a_2\lambda + (b_2 - p)\mu &= 0 \end{aligned}\right\} \qquad (17)$$

These equations may be regarded as a system of two first-degree algebraic homogeneous equations in two unknowns λ, μ. For it to have a nontrivial solution (and, by (16), only such a solution interests us), it is necessary and sufficient that the determinant of the system be zero:

$$\begin{vmatrix} a_1 - p & b_1 \\ a_2 & b_2 - p \end{vmatrix} = 0 \tag{18}$$

This is the *characteristic equation* of the system (15), from which we find possible values for p. It can be rewritten by expanding the determinant:

$$p^2 - (a_1 + b_2)\, p + a_1 b_2 - b_1 a_2 = 0$$

(verify this).

We see that equation (18) is of second degree in p and so has two roots p_1, p_2. If these roots are distinct, then either one (p_k) can be substituted into (17) to find a nontrivial solution λ_k, μ_k and, by (16), to obtain the corresponding solution $y(x)$, $z(x)$ of the system (15). Since the quantities λ_k, μ_k are determined to within a common arbitrary factor, we multiply by an arbitrary constant C_1 the solution corresponding to the root p_1, and by C_2 the solution corresponding to the root p_2, and then add the results. We thus obtain the general solution to the system (15):

$$\left.\begin{aligned} y &= C_1\lambda_1 e^{p_1 x} + C_2\lambda_2 e^{p_2 x} \\ z &= C_1\mu_1 e^{p_1 x} + C_2\mu_2 e^{p_2 x} \end{aligned}\right\} \tag{19}$$

Here the arbitrary constants C_1 and C_2 may be found, for example, if we also give the initial conditions for (15):

$$\text{for } x = x_0 \text{ are given } y = y_0 \text{ and } z = z_0$$

If the roots of equation (17) are imaginary, then the solution can either be left in the form (19) or be written in real form, as in Sec. 7.3. If $p_1 = p_2$, then y and z are obtained as a combination of the functions $e^{p_1 x}$ and $xe^{p_1 x}$ (cf. Sec. 7.3).

Exercises

1. Investigate the solvability of the system of algebraic equations

$$\left.\begin{aligned} x + 2y &= 3 \\ 3x + ay &= b \end{aligned}\right\} \text{ for distinct } a,\ b.$$

2. Find the general solution to the system $\frac{dx}{dt} = 4x - y$, $\frac{dy}{dt} = - 6x + 3y$.

8.4 Lyapunov stability of the equilibrium state

The concept of stability as the capability of an object, state or process to resist external actions not indicated beforehand appeared in antiquity and today occupies one of the central places in physics and engineering. There are a variety of specific realizations of this general notion, depending on the type of object under study, the nature of the external actions, and so forth. Here we consider one of the most important stability concepts: stability in the sense of Lyapunov, which was introduced for simple cases in Sec. 7.6.

Let the state of some object be described by a finite number of parameters; for the sake of simplicity, we take two parameters x, y, so that changes in this entity in time are specified by two functions $x = x(t)$, $y = y(t)$, where t is time. Let the law of this change have the form of a system of differential equations:

$$\left.\begin{aligned} \frac{dx}{dt} &= P(x, y), \\ \frac{dy}{dt} &= Q(x, y) \end{aligned}\right\} \qquad (20)$$

мith specified right members that do not involve the independent variable t explicitly .This last condition signifies that the differential law of development of the process does not change with time. Such processes are called *autonomous* processes. (Autonomous systems occur, for example, when the equations involve all bodies participating in the problem, since the laws of nature are invariable in time.)

Let the state of equilibrium of the entity at hand (when it does not vary in the course of time) be described by the constant values $x = x_0$, $y = y_0$; then this system of constants, regarded as functions of time, must also satisfy the system (20). From a direct substitution into (20) it follows that for this it is necessary and sufficient that, simultaneously, we have

$$P(x_0, y_0) = 0, \quad Q(x_0, y_0) = 0 \qquad (21)$$

Suppose at a certain time t_0 the entity is brought out of the equilibrium state (by some cause) and the parameters x, y become $x = x_0 + \Delta x_0$, $y = y_0 + \Delta y_0$. Then, in order to determine subsequent changes in the entity we have to solve the system of equations (20) for the initial conditions

$$x(t_0) = x_0 + \Delta x_0, \quad y(t_0) = y_0 + \Delta y_0 \qquad (22)$$

The state of equilibrium that is being studied is said to be *Lyapunov stable* (stable in the sense of Lyapunov) if after a slight deviation from this state the entity continues to remain near it all the time. In other words, to solve the system (20) under the initial conditions (22) for small Δx_0, Δy_0, the differences $\Delta x = x(t) - x_0$, $\Delta y = y(t) - y_0$ must be small for all $t > t_0$.

In order to determine whether stability will occur, substitute

$$x = x_0 + \Delta x, \quad y = y_0 + \Delta y$$

into (20) to get

$$\left.\begin{aligned}\frac{d(\Delta x)}{dt} &= P(x_0 + \Delta x,\ y_0 + \Delta y) = (P'_x)_0\,\Delta x + (P'_y)_0\,\Delta y + \ldots \\ \frac{d(\Delta y)}{dt} &= Q(x_0 + \Delta x,\ y_0 + \Delta y) = (Q'_x)_0\,\Delta x + (Q'_y)_0\,\Delta y + \ldots\end{aligned}\right\} \qquad (23)$$

where $(P'_x)_0 = P'_x(x_0, y_0)$, and so forth. In manipulating the right members we take advantage of Taylor's formula (Sec. 4.6) and formulas (21). The dots stand for terms higher than first order.

Since in determining stability we consider only small Δx, Δy, the main part in the right members of (23) is played by the linear terms that are written out (cf. similar reasoning in Sec. 7.6). For this reason, we replace the system (23) by an *abridged system* (*a system of first-order approximation*) by discarding higher-order terms:

$$\left.\begin{aligned}\frac{d(\Delta x)}{dt} &= (P'_x)_0\,\Delta x + (P'_y)_0\,\Delta y, \\ \frac{d(\Delta y)}{dt} &= (Q'_x)_0\,\Delta x + (Q'_y)_0\,\Delta y\end{aligned}\right\}. \qquad (24)$$

System (24) is a linear system with constant coefficients that is solved by the method of Sec. 8.3. By formula (19) (the notation, however, was different there) the solution of system (24) is obtained as a combination of functions of the form e^{pt}, where p satisfies the characteristic equation

$$\begin{vmatrix}(P'_x)_0 - p & (P'_y)_0 \\ (Q'_x)_0 & (Q'_y)_0 - p\end{vmatrix} = 0 \qquad (25)$$

Here, to small Δx_0, Δy_0 there correspond small values of the arbitrary constants C_1, C_2 and so the whole matter lies in the behaviour of the function e^{pt} as t increases. Since p can be imaginary as well, $p = r + is$, and then

$$e^{pt} = e^{rt}(\cos st + i \sin st) \qquad (26)$$

it follows that the increase or decrease of the perturbation is determined by the sign of p if p is real, and by the sign of r if p is imaginary: if this sign is plus, then the perturbation increases, and if it is minus, it decreases. We arrive at the following conclusions. If all roots of the characteristic equation (25) have a negative real part (in particular, they can be real and negative), then the equilibrium state (x_0, y_0) is stable in the sense of Lyapunov. Besides, for small Δx_0, Δy_0 it will then be true that $x(t) \to x_0$, $y(t) \to y_0$ as $t \to \infty$; as we pointed out in Sec. 7.6, this stability is said to be *asymptotic*. Now if at least one of the roots of equation (25) has a positive real part, then the equilibrium state under consideration is Lyapunov unstable (unstable in the sense of Lyapunov).

We derived these results from the system (24), but by the foregoing the same assertions hold for the complete system (23). Note that if equation (25) has a double root, this does not violate our assertions because even though t appears in our solution as a factor, the exponential e^{pt} tends to zero for $p < 0$ faster than t tends to infinity.

The two conclusions given above do not embrace the case where there are no roots of equation (25) with a positive real part, but there is at least one root with a zero real part. Then in the general solution of system (24) there appear functions of the form

$$e^{ist} = \cos st + i \sin st, \quad |e^{ist}| = 1 \quad \text{or} \quad e^{0t} = 1$$

The entity would appear to be oscillating (or remaining stationary) about the equilibrium state and not striving towards it. But then, due to the unlimitedness of time, the discarded higher-order terms begin to exert an influence and are capable of upsetting the stability. Thus, in this special case, one cannot judge the stability or instability of the equilibrium state on the basis of the roots of equation (25). This can only be done by invoking supplementary considerations, for instance, more terms in the expansion (23). We will not carry out this investigation and will merely note that small perturbations will increase or decay much more slowly since the variation time of such a perturbation will be a given number of times (say, e) inversely proportional to this perturbation or will even have a higher order.

Exercise

Find the equilibrium states for the system

$$\left.\begin{aligned} \frac{dx}{dt} &= -x - y, \\ \frac{dy}{dt} &= y + y^3 \end{aligned}\right\}$$

and investigate them for stability.

8.5 Constructing approximate formulas for a solution

Methods for constructing approximate formulas for the solution of a differential equation are largely analogous to those methods described in Sec. 1.4 for the solution of "finite" equations. For the sake of simplicity we consider first-order equations, although the same methods can naturally be extended to equations of any order and to systems of equations.

We begin with the method of iteration. Suppose we are considering a first-order differential equation with specified initial condition:

$$\left.\begin{aligned}\frac{dy}{dx} &= f(x, y),\\ y(x_0) &= y_0\end{aligned}\right\} \tag{27}$$

Taking the integrals of both sides of the equation, we get

$$\int_{x_0}^{x} \frac{dy}{dx}\, dx = y - y_0 = \int_{x_0}^{x} f(s, y(s))\, ds\text{*}$$

That is,

$$y(x) = y_0 + \int_{x_0}^{x} f(s, y(s))\, ds \tag{28}$$

Equation (28) is equivalent at once to both the equations of (27), since after its differentiation we get the first equation, and after the substitution $x = x_0$ the second equation. Equation (28) is an *integral equation* because the unknown function is under the integral sign. Since it includes not only the first but also the second equation of (27), it has a unique solution and not an infinity of solutions, as the differential equation does.

The aspect of (28) is convenient for using the iteration method (compare with equation (22) of Ch. 1), although the unknown here is a function, not a number. Choosing a function $y_0(x)$ for the zero-order approximation (it is desirable that it be as close as possible to the desired solution; if we know nothing about this solution, we can at least put $y_0(x) \equiv y_0$), we obtain the first-order approximation via the formula

$$y_1(x) = y_0 + \int_{x_0}^{x} f(s, y_0(s))\, ds$$

Substituting the result into the right side of (28), we get the second-order approximation, and so forth. Generally,

$$y_{n+1}(x) = y_0 + \int_{x_0}^{x} f(s, y_n(s))\, ds \qquad (n = 0, 1, 2, \ldots)$$

[*] We frequently use a notation like $\int_{x_0}^{x} \varphi(x)\, dx$, but here we must be more careful and distinguish between the upper limit and the variable of integration (cf. HM, Sec. 2.14).

As in Sec. 1.4, if the process of iteration converges, that is, if the sequence of approximations tends to a certain limit function with increasing n, then it satisfies the equation (28).

It is a remarkable fact that the iteration method for (28) converges at least for all x sufficiently close to x_0. This is due to the fact that when computing subsequent approximations it is necessary to integrate the preceding ones, and in successive integration the functions are as a whole "smoothed", any inaccuracies due to the choice of the zero-order approximation, roundoff errors, and the like being gradually eliminated.

(In contrast, in the case of successive differentiation, the functions, as a rule, deteriorate, the initial inaccuracies increase, and so an iteration method based on successive differentiation would not yield a convergence. Cf. similar reasoning in Sec. 2.2.)

By way of an illustration, let us consider the problem

$$\left.\begin{aligned} y' &= x^2 + y^2, \\ y(0) &= 1 \end{aligned}\right\} \tag{29}$$

Integration yields

$$y(x) = 1 + \frac{x^3}{3} + \int_0^x y^2(s)\,ds$$

For the zero-order approximation to the desired solution, of which we know nothing as yet, take the function $y_0(x) \equiv 1$, for it at least satisfies the initial condition. Then, by writing out powers up to x^4 inclusive, we get (verify this)

$$y_1(x) = 1 + x + \frac{x^3}{3},$$

$$y_2(x) = 1 + \frac{x^3}{3} + \int_0^x \left(1 + s + \frac{s^3}{3}\right)^2 ds = 1 + x + x^2 + \frac{2}{3}x^3 + \frac{1}{6}x^4 + \dots,$$

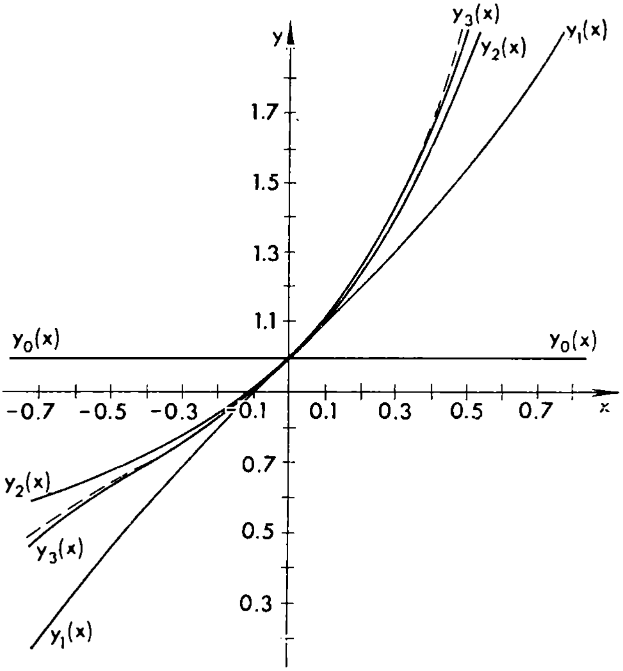
$$y_3(x) = 1 + x + x^2 + \frac{4}{3}x^3 + \frac{5}{6}x^4 + \dots,$$

$$y_4(x) = 1 + x + x^2 + \frac{4}{3}x^3 + \frac{7}{6}x^4 + \dots,$$

$$y_5(x) = 1 + x + x^2 + \frac{4}{3}x^3 + \frac{7}{6}x^4 + \dots$$

The graphs of the successive approximations are shown in Fig. 95, where the dashed line is the exact solution. It is clear that for small $|x|$ the process converges.

Fig. 95



The question of when to terminate the approximation is ordinarily settled by comparing adjacent approximations.

Another approximate method is based on the fact that one finds the values of $y'(x_0)$, $y''(x_0)$, etc. from the given conditions (27) with the aid of differentiation and then expands the solution in a Taylor series (see HM, Sec. 3.17). The requisite number of terms is determined in the form of a succession of calculations and their comparison with the desired degree of accuracy.

Let us consider the problem (29). Substitution into the right side of the equation yields $y'(0) = 0^2 + 1^2 = 1$. Differentiating both sides of the equation, we get $y'' = 2x + 2yy'$ and, substituting $x = 0$, we obtain $y''(0) = 2 \cdot 0 + 2 \cdot 1 \cdot 1 = 2$. In similar fashion we get

$$y''' = 2 + 2y'^2 + 2yy'', \qquad y'''(0) = 8,$$
$$y^{\mathrm{IV}} = 6y'y'' + 2yy''', \qquad y^{\mathrm{IV}}(0) = 28$$

and so on. Substituting this into the Taylor formula, we obtain

$$y = y(0) + \frac{y'(0)}{1!}x + \frac{y''(0)}{2!}x^2 + \dots$$
$$= 1 + x + x^2 + \frac{4}{3}x^3 + \frac{7}{6}x^4 + \dots \qquad (30)$$

We have the same formula as that obtained via the method of successive approximations. This formula can only be used for small $|x|$; for example, when $x = 1$ the series (30) (like the earlier described iteration method) diverges. It can be demonstrated that this is due to the essence of the problem at hand. Indeed, consider the solution $y^1(x)$ of the equation $\frac{dy}{dx} = y^2$ for the initial condition $y'(0) = 1$. Since $x^2 + y^2 > y^2$, the direction field in the xy-plane that determines the original solution $y(x)$ is located more steeply than the direction field that determines the solution $y^1(x)$. But $y(0) = y^1(0)$, and so for $x > 0$ the curve $y = y(x)$ passes above the curve $y = y^1(x)$. The solution $y^1(x)$ is easy to find by separating the variables, $y^1(x) = \frac{1}{1-x}$. Thus, $y(x) > \frac{1}{1-x}$ $(x > 0)$. The right side tends to infinity as $x \to 1 - 0$; hence the solution $y(x)$ also tends to infinity for a certain $x = x_1 \leqslant 1$ as x increases from 0. Computing the value of x_1 (which depends on the initial value $y(0)$) via the methods of Sec. 8.7 yields $x_1 = 0.959$. As $x \to x_1 - 0$ we will have $x^2 \ll y^2$ and so $y(x) \approx \approx 1/(x_1 - x)$.

Closely tied to this method is the method of using power series with undetermined coefficients. It consists in seeking the solution of an equation in the form of a series with unknown coefficients:

$$y = a + b(x - x_0) + c(x - x_0)^2 + d(x - x_0)^3 + \dots$$

which are found by substitution into the equation, subsequent equating of the coefficients of equal powers, and the use of the initial condition if it is given.

Apply the method of undetermined coefficients to the problem (29). Since $x_0 = 0$, we write

$$y = a + bx + cx^2 + dx^3 + ex^4 + \dots \tag{31}$$

Substituting $x = 0$, we get $a = 1$ by virtue of the initial condition. Before substituting the series (31) into equation (29), it is convenient to expand the right member of this equation in a series in powers of $y - 1$. (In the general case, the right member is expanded in a Taylor series in powers of $x - x_0$, $y - y_0$ by the formula

$$f(x, y) = f_0 + (f'_x)_0(x - x_0) + (f'_y)_0(y - y_0) + \dots$$

where the zero subscript denotes substitution of the values $x = x_0$, $y = y_0$.) We get

$$y' = x^2 + [(y - 1) + 1]^2 = x^2 + 1 + 2(y - 1) + (y - 1)^2$$

Substituting the series (31), we obtain

$$b + 2cx + 3\,dx^2 + 4ex^3 + \dots$$
$$= 1 + x^2 + 2(bx + cx^2 + dx^3 + ex^4 + \dots) + (bx + cx^2 + dx^3 + \dots)^2$$

Removing brackets on the right and collecting terms, and then equating the coefficients of equal powers of x, we arrive at the relations $b = 1$, $2c = 2b$, $3d = 1 + 2c + b^2$, $4e = 2d + 2bc$, ..., whence in succession we find $b = 1$, $c = 1$, $d = \frac{4}{3}$, $e = \frac{7}{6}$, ... Putting these values into (31), we again arrive at the series (30).

The perturbation method discussed in Sec. 1.4 is also used in solving differential equations. Here are some examples.

The problem

$$y' = \frac{x}{1 + 0.1xy}, \quad y(0) = 0 \tag{32}$$

does not contain any parameters. But we can consider the more general problem

$$y' = \frac{x}{1 + \alpha xy}, \quad y(0) = 0 \tag{33}$$

from which (32) is obtained when $\alpha = 0.1$. The problem (33) is readily solved for $\alpha = 0$: we then get $y = \frac{x^2}{2}$. Therefore we seek a solution by a series expansion in powers of α, i.e.

$$y = \frac{x^2}{2} + \alpha u + \alpha^2 v + \alpha^3 w + \dots \tag{34}$$

where $u = u(x)$, $v = v(x)$ and so on are the as yet unknown functions of x.

Substituting (34) into (33) yields, after multiplying by the denominator,

$$(x + \alpha u' + \alpha^2 v' + \alpha^3 w' + \dots)\left(1 + \frac{\alpha}{2} x^3 + \alpha^2 xu + \alpha^3 xv + \dots\right) = x, \tag{35}$$

$$\alpha u(0) + \alpha^2 v(0) + \dots = 0$$

or

$$u(0) = 0, \ v(0) = 0, \ w(0) = 0 \tag{36}$$

Opening brackets in (35) and equating to zero the coefficients of the powers of α, we successively get

$$u' + \frac{1}{2} x^4 = 0, \quad v' + \frac{x^3}{2} u' + x^2 u = 0,$$

$$w' + \frac{x^3}{2} v' + xuu' + x^2 v = 0 \quad \text{and so on}$$

whence, taking into account (36), we get (verify this!)

$$u = -\frac{x^5}{10}, \quad v = \frac{7}{160} x^8, \quad w = \frac{71}{1760} x^{11} \quad \text{and so on}$$

Therefore formula (34) yields

$$y = \frac{x^2}{2} - \frac{\alpha}{10} x^5 + \frac{7\alpha^2}{160} x^8 - \frac{71\alpha^3}{1760} x^{11} + \dots$$

In particular, for equation (32) we obtain

$$y = \frac{x^2}{2} - \frac{x^5}{100} + \frac{7x^8}{16\,000} - \frac{71x^{11}}{1\,760\,000} + \dots$$

This series converges excellently for $|x| \leqslant 1$ and rather decently for $1 < |x| < 2.$*

Now consider the problem

$$y' = \sin(xy), \quad y(0) = \alpha \tag{37}$$

Unlike the preceding example, the parameter here enters into the initial condition. For $\alpha = 0$ the problem (37) has the solution $y \equiv 0$. And so for small $|\alpha|$ we seek the solution in the form

$$y = \alpha u + \alpha^2 v + \alpha^3 w + \dots \quad (u = u(x), \quad v = v(x), \dots) \tag{38}$$

Substitution of the value $x = 0$ produces

$$u(0) = 1, \ v(0) = 0, \ w(0) = 0 \tag{39}$$

On the other hand, substituting (38) into the differential equation (37), we get (taking note of the Taylor series for the sine)

$$\alpha u' + \alpha^2 v' + \alpha^3 w' + \dots = \frac{(\alpha x u + \alpha^2 x v + \alpha^3 x w + \dots)}{1!} - \frac{(\alpha x u + \alpha^2 x v + \alpha^3 x w + \dots)^3}{3!} + \dots$$

Equating coefficients of like powers of α gives

$$u' = xu, \ v' = xv, \ w' = xw - \frac{x^3u^3}{3!}, \dots$$

Integrating these linear equations with account taken of the initial conditions (39), we find

$$u = e^{\frac{x^2}{2}}, \ v = 0, \ w = \frac{1}{12}(1 - x^2)\, e^{\frac{3}{2}x^2} - \frac{1}{12} e^{\frac{x^2}{2}}$$

(verify this!). Substitution of these expressions into (38) produces an expansion of the desired solution that is suitable for small $|\alpha|$.

* The substitution $s = -\alpha xy$, $v(s) = -\alpha x^3$, which we leave to the reader, carries (33) into an equation that does not involve a parameter. When $s = 1$ we have $|dy/dx| = \infty$. Numerical integration via the methods of Sec. 8.7 shows that $v(1) = 1.087$, and so the series (34) cannot converge for $x > \sqrt[3]{1.087/\alpha}$; for $\alpha = 0.1$ the right member is equal to 2.16.

Here, the greater $|x|$, the greater the coefficients, and for this reason the smaller the interval of values of α for which the series is applicable.

In more complicated cases of the use of the perturbation method it is often useful to find at least the first term of the expansion involving the parameter, since this gives us an idea of the behaviour of the solution for a small change in the parameter.

We conclude with the remark that in practice, particularly in the case of crude estimates, wide use is made of simplifying the original equation by dropping comparatively small terms, replacing slowly varying coefficients with constants, and the like. After such a simplification we might get an equation of one of the integrable types, which, when integrated, yields a function that can be regarded as an approximate solution of the original complete equation. At any rate, it frequently conveys the proper nature of the behaviour of the exact solution. With this "zero-order approximation" found, it is sometimes possible, using it, to introduce corrections that take into account the simplification and thus to find the "first-order approximation", and so on.

If the equation contains parameters (for example, masses, linear dimensions of the entities under study, and the like), the one has to bear in mind that for certain values of these parameters certain terms of the equation may be relatively small, and for other values, other terms will be small, so that the simplification is handled differently for different values of the parameters. Besides, it is sometimes necessary to split the range of the independent variable into parts, over each one of which the simplification procedure differs.

Such a simplification of the equation is particularly useful when— in the very derivation (writing) of the differential equation — essential simplifying assumptions are made or the accuracy to which the quantities in question are known is slight. For example, terms of the equation that are less than the admissible error in other terms must definitely be dropped.

To illustrate, consider the problem

$$y'' + \frac{1}{1 + 0.1x} y + 0.2y^3 = 0, \ y(0) = 1, \ y'(0) = 0, \ 0 \leqslant x \leqslant 2 \quad (40)$$

Since the coefficient of y varies slowly, we replace it by its mean value:

$$\frac{1}{1 + 0.1x} = k; \ x = 0, \ k = 1; \ x = 2, \ k = \frac{1}{1.2} = 0.83;$$

$$\bar{k} = \frac{1 + 0.83}{2} = 0.92$$

Also, we drop the relatively small third term and get the equation

$$y'' + 0.92\,y = 0$$

with the solution for the given initial conditions

$$y = \cos 0.96x \tag{41}$$

The aspect of this approximate solution confirms the justification for dropping the last term of the equation: the ratio of the third term to the second is of the order of $0.2\,y^2 < 0.2$ and therefore the sum of the first two terms is small in comparison with the second, i.e. the first term and the third should "almost cancel out".

We now introduce a correction for the last term by substituting into it the approximate solution (41) and leaving the averaged coefficient:

$$y'' + 0.92y = -0.2 \cos^3 0.96x$$

For the given initial conditions, the integration of this equation by the method of Sec. 7.5 yields

$$y = 0.99 \cos 0.96x - 0.08x \sin 0.96x + 0.01 \cos 2.88x$$

The difference as compared to the zero-order approximation (41) is slight, so that the conclusion concerning the significance of the separate terms in (40) remains valid; at the same time the third term of (40) made its contribution to the solution. (To take into account the variable nature of the coefficient k one could replace the second term in equation (40) by $[1.1 + 0.1(x-1)]^{-1}\, y = \frac{1}{1.1}[1 + 0.91(x-1)]^{-1}\, y \approx 0.91y - 0.08(x-1) \cos 0.96x$. Yet even this would not lead to a substantial change in the solution.)

Reasoning in this fashion is frequently rather nonrigorous (and at times leads to errors) but if it is coupled with common sense, it rather frequently produces solutions that are of practical use.

Exercises

1. Apply the method of successive approximation to the problem $\frac{dy}{dx} = y$, $y(0) = 1$.
2. By calculating derivatives, find the expansion, in powers of x, of the solution of the problem $\frac{dy}{dx} = e^{xy}$, $y(0) = 0$ up to x^5.
3. Find the first two terms of the expansion of the solution of the problem $\frac{dy}{dx} = y^2 + \alpha x$, $y(0) = 1$ in a series of powers of α.

8.6 Adiabatic variation of a solution

We now consider another important method of approximate solution of differential equations, this time using the example of the linear equation

$$\ddot{x} + \omega^2 x = 0 \quad (x = x(t), \quad \omega = \omega(t)) \tag{42}$$

where a dot indicates the derivative with respect to the time t, and the relationship of $\omega(t)$ is given. This is the equation of oscillations of an oscillator whose parameters vary in time; for example, this may be a pendulum the length of the suspension of which varies, and the like.

In the general case, equation (42) is not integrable by quadrature and its investigation is rather complicated. However, in an important particular case, namely when the coefficient $\omega(t) > 0$ varies slowly, such an investigation can be carried out. Here the concept of slow variation is made explicit as follows. To begin with, let $\omega =$ $=$ constant; then in Sec. 7.3 we saw that ω serves as the frequency of free oscillations of the oscillator, and so we have a natural measure of time equal to the period $\frac{2\pi}{\omega}$ of these oscillations. We agree to say that a quantity $p(t)$ *varies slowly* (we also say *adiabatically*) if its relative change during this period is small, that is, if

$$\left|p\,\frac{2\pi}{\omega}\right| \ll |p| \text{ or, what is the same thing, } |\dot{p}| \ll \omega\,|p|$$

We will use this definition for the case $\omega = \omega(t)$ as well; the adiabatic nature of the variation of ω means that $|\dot{\omega}| \ll \omega^2$

If $\omega =$ constant, then the general solution of equation (42) may be written in the form

$$x = C_1 \cos \omega t + C_2 \sin \omega t = A \sin (\omega t + \varphi_0)$$

where A and φ_0 are arbitrary constants. In short,

$$x = A \sin \varphi \tag{43}$$

where $\varphi = \omega t + \varphi_0$ and therefore

$$\frac{d\varphi}{dt} = \omega \tag{44}$$

Now if ω depends on t but varies slowly, then it is natural to assume that over each small time interval the oscillations of the oscillator are nearly harmonic with frequency equal to the current value of ω and to assume that the solution of equation (42) is still and all of the form (43) under the condition (44), where however we already have $A = A(t)$, $\omega = \omega(t)$. From (44) we get $\varphi = \int^{t} \omega(t)dt$, the constant lower limit of this integral being inessential.

Suppose that not only ω but also $\dot{\omega}$ varies slowly; then it is natural to assume that A and $\dot{A}$ also vary slowly. From (43) we get

$$\dot{x} = \dot{A} \sin \varphi + A \cos \varphi \cdot \omega$$

$$x = \ddot{A} \sin \varphi + 2\dot{A} \cos \varphi \cdot \omega - A \sin \varphi \cdot \omega^2 + A \cos \varphi \cdot \dot{\omega}$$

and substitution into (40) produces

$$\ddot{A}\sin\varphi + 2\dot{A}\cos\varphi\cdot\omega + A\cos\varphi\cdot\dot{\omega} = 0 \qquad (45)$$

Since, by assumption the first term is of higher order than the second, the second and third terms should cancel out. After cancelling out $\cos\varphi$, we get

$$2\dot{A}\omega + A\dot{\omega} = 0,\ \text{i.e.}\ 2\frac{dA}{dt}\omega + A\frac{d\omega}{dt} = 0,\ 2\frac{dA}{A} + \frac{d\omega}{\omega} = 0$$

and after integrating we obtain $2\ln A + \ln\omega = \ln C$, $A^2\omega = C$. Thus we see that, on our assumptions, the amplitude of oscillations varies in inverse proportion to the square root of the value of the natural frequency.

True, since in equation (45) we neglected the higher order term, the expression $A^2\omega$ is actually not constant. It varies with time, but the relative rate of its variation is of a higher order compared with the relative rate of variation of the natural frequency. We say that the quantity $A^2\omega$ is an *adiabatic invariant.*

A similar result is also obtained by energy reasoning. The energy of an oscillator is equal to (see formula (5))

$$E = \frac{mv^2}{2} + \frac{kx^2}{2} = \frac{m}{2}(\dot{x}^2 + \omega^2x^2)$$

whence, differentiating and using equation (42), we get, for $m =$ constant,

$$\dot{E} = \frac{m}{2}(2\ddot{x}\ddot{x} + 2\omega\dot{\omega}x^2 + 2\omega^2x\dot{x}) = m\omega\dot{\omega}x^2 \qquad (46)$$

If ω and ω vary slowly, then the coefficients of x^2 in the right members of the current period, that is, over a time interval of length $\frac{2\pi}{\omega}$, are almost constant and we can average over that interval. Taking into account $\sin^2(\omega t + \varphi_0) = \frac{1}{2} - \frac{1}{2}\cos 2(\omega t + \varphi_0)$, we get $\overline{x}^2 = \frac{1}{2}A^2$ and, similarly, $\bar{\dot{x}}^2 = \frac{1}{2}\omega^2A^2$. And so after averaging we have

$$E = \frac{1}{2}m\omega^2A^2,\ E = \frac{1}{2}m\omega\dot{\omega}A^2 \qquad (47)$$

(since E varies slowly, we can replace $\overline{E}$ again by E). From this we have $\frac{\dot{E}}{E} = \frac{\dot{\omega}}{\omega}$, $\ln E = \ln\omega + \ln C_1$, $E = C_1\omega$, which is to say that the energy of the oscillator is directly proportional to the instantaneous value of its natural frequency. Putting this result into the first equation of (47), we get $C_1\omega = \frac{1}{2} m\omega^2 A^2$, i.e. $\omega A^2 = \frac{2C_1}{m} =$ constant, as above.

It is interesting to note that the resulting proportionality, $E \propto \omega$, is readily grasped from the standpoint of quantum mechanics. The point is that the energy of one quantum is equal to $\hbar\omega$, where $\hbar \approx 10^{-27}$ g-cm/s^2 is Planck's constant; the energy of an oscillator at the nth level ($n = 1, 2, 3, \ldots$) is equal to $\hbar\omega n$. If the frequency ω varies slowly, the oscillator remains all the time at the same level, that is, $n =$ constant, whence the proportionality $E \propto \omega$. (See, for instance, P. Paradoksov, "How Quantum Mechanics Helps Us Understand Classical Mechanics." [12])

Now let us consider another important special case where ω and not $\dot{\omega}$ varies slowly; we take the example of equation (42) with $\omega = \omega_0 + \alpha \sin kt$, where $|\alpha| \ll \omega_0$ and the constant k is of the same order as ω_0. In this case, when averaging the right member of (46), it should first be transformed via the formula

$$m\omega_0 \alpha k \cos kt \cdot A^2\left[\frac{1}{2} - \frac{1}{2}\cos 2(\omega_0 t + \varphi_0)\right]$$
$$= \frac{1}{2} m\omega_0 \alpha k A^2 \cos kt - \frac{1}{4} m\omega_0 \alpha k A^2 \cos[(2\omega_0 + k)t + 2\varphi_0]$$
$$- \frac{1}{4} m\omega_0 \alpha k A^2 \cos[(2\omega_0 - k)t + 2\varphi_0]$$

There can be two cases now. If $k \neq 2\omega_0$, then the mean value of the right member, as the mean value of the sum of pure harmonics, is zero, $\dot{E} = 0$, with E the adiabatic invariant. But if $k = 2\omega_0$, then the last term in the right-hand member turns into a constant, whence, after averaging, we get

$$\dot{E} = -\frac{1}{4} m\omega_0 \alpha 2\omega_0 A^2 \cos(2\varphi_0) = -\alpha \cos(2\varphi_0)\, E, \text{ or}$$
$$E = Ce^{-\alpha\cos(2\varphi_0)\cdot t}.$$

To summarize, in the case at hand, the energy of the oscillator, and so also the amplitude of oscillations, are exponentially increasing or decreasing in time, depending on the sign of $\cos 2\varphi_0$. This phenomenon, which is similar to resonance (Sec. 7.5) and occurs due to the periodic variation of the parameters of the oscillator, is called *parametric resonance* (incidentally, parametric resonance is used to set an ordinary child's swing in motion).

8.7 Numerical solution of differential equations

It often happens that it is impossible to obtain an exact or sufficiently satisfactory approximate solution in the form of a formula. Then a numerical solution is found in which the desired particular solution (for concrete values of the parameters if they enter into the statement of the problem) is constructed in tabular form. The principle of a numerical solution of a differential equation is exceedingly simple and follows directly from the meaning of a derivative.

Suppose an equation is of the form $\frac{dy}{dx} = f(x, y)$ and the initial condition $y = y_0$ for $x = x_0$ is given. Then, putting the values x_0, y_0 into the function $f(x, y)$, we find the derivative at the point x_0:

$$\left.\frac{dy}{dx}\right|_{x=x_0} = f(x_0, y_0)$$

From this, assuming that Δx is a small quantity, we get

$$y(x_0 + \Delta x) = y(x_1) = y_1 = y_0 + \Delta y =$$

$$= y_0 + \left.\frac{dy}{dx}\right|_{x=x_0} \cdot \Delta x = y_0 + f(x_0, y_0) \cdot \Delta x$$

Writing $f(x_0, y_0) = f_0$ for brevity, we give this result as follows:

$$y_1 = y_0 + f_0 \cdot \Delta x \tag{48}$$

Now, taking the point (x_1, y_1) for the original one, we can obtain $y_2 = y(x_2)$, where $x_2 = x_1 + \Delta x$, in exactly the same manner. Thus, step by step, we can calculate various values of y for distinct values of x. This is *Euler's method.*

Of course this method provides approximate values of y, not exact values, for the derivative $\frac{dy}{dx}$ does not remain constant over the interval from $x = x_0$ to $x = x_1$. Therefore, by using formula (48) we err in determining y, and the error is the greater, the larger Δx is.

To be more exact, since the right side of (48) is a sum of the first two terms of the expansion of $y(x_0 + \Delta x)$ in powers of Δx, the error of (48) is of the order of $(\Delta x)^2$, i.e. it does not exceed $a(\Delta x)^2$, where the coefficient a depends on the type of function $f(x, y)$.

Suppose it is necessary, if we know the initial condition $y(x_0) = y_0$, to obtain the value of the solution for $x = x_0 + l$, where l is great, so that when using (48) and putting $\Delta x = l$ in it, we allow for an enormous error. To find $y(x_0 + l)$ we partition the interval between $x = x_0$ and $x = x_0 + l$ into n equal small subintervals: then the length of each subinterval is $\Delta x = \frac{1}{n}$. To obtain $y(x_0 + l)$ we will have to take n steps, finding in succession $y\left(x_0 + \frac{l}{n}\right)$, $y\left(x_0 + 2\frac{l}{n}\right)$, ..., $y(x_0 + l)$.

At each such step the error is of the order of $a\left(\frac{l}{n}\right)^2$, and for the n steps we have an error of the order of $a\left(\frac{l}{n}\right)^2 n = \frac{al^2}{n}$. Hence, the error that arises when using Euler's method is inversely proportional to the number of steps. If the accuracy ε is specified, then the necessary number of steps n is a quantity of the order of $\frac{al^2}{\varepsilon}$. Clearly, the larger the number of steps, the smaller the error and the more exact the quantity $y(x_0 + l)$ that we find. But the error decreases very slowly with increasing number of steps. For this reason, a great number of steps is needed in order to attain a specified accuracy.

We say that the approximate value of y given by formula (48) is the first approximation (y_{I}), so that $y_{\text{I}} = y_0 + f_0 \, \Delta x$. To obtain a more exact second-order approximation we will take the arithmetic mean of the derivative at the beginning and at the end of the interval, computing the derivative at the endpoint of the interval by means of the first approximation y_{I}. Thus

$$y_{\text{II}} = y_0 + \frac{1}{2}\left(\frac{dy}{dx}\bigg|_{\substack{x=x_0\\ y=y_0}} + \frac{dy}{dx}\bigg|_{\substack{x=x_0\ +\Delta x\\ y=y_{\text{I}}}}\right)\Delta x$$

or

$$y_{\text{II}} = y_0 + \frac{1}{2}\,[f(x_0, y_0) + f(x_0 + \Delta x, y_{\text{I}})]\,\Delta x$$

$$= y_0 + \frac{1}{2}[f_0 + f(x_0 + \Delta x, y_0 + f_0\,\Delta x)]\,\Delta x$$

It can be shown that y_{II} has an error of the order of $b(\Delta x)^3$, where b is a constant dependent on the type of $f(x, y)$. Therefore, the total error, over n steps, in determining $y(x_0 + l)$ will be $\varepsilon = b\left(\frac{l}{n}\right)^3 \cdot n = \frac{bl^3}{n^2}$ and the number of steps n needed to attain the given accuracy ε is a quantity of the order of $\sqrt{\frac{bl^3}{\varepsilon}}$. In this case the error is inversely proportional to the square of the number of steps, that is, as the number of steps increases it falls off much faster than when the first approximation y_{I} is used.

Note, however, that in finding y_{II} we have to compute $f(x, y)$ twice at each step, whereas in finding y_{I} we have to calculate only one value of $f(x, y)$ at each step. Indeed, using the Euler method, we begin the calculation with the point (x_0, y_0) and find $y_{\text{I}}(x_1) = y_0 + f_0 \cdot \Delta x$ and then compute $f(x_1, y_{\text{I}}(x_1))$ and pass to the next step.

If we want to find a second approximation, the scheme is as follows. First find $y_{\text{I}}(x_1) = y_0 + f_0 \cdot \Delta x$ and then determine $f(x_1, y_{\text{I}}(x_1))$ and

$$\bar{f}_0 = \frac{1}{2}[f(x_0, y_0) + f(x_1, y_{\text{I}}(x_1))]$$

We then find $y_{\text{II}}(x_1) = y_0 + \bar{f}_0 \cdot \Delta x$ and, finally, $f(x_1, y_{\text{II}}(x_1))$ Only then is everything ready for the next step. This calculation scheme is called the *recalculation method* because the quantity $f(x, y)$

is recalculated at each step and is replaced by the more reliable quantity $\bar{f}(x, y)$.

Computation of the values of $f(x, y)$ is, as a rule, the most cumbersome operation (the other operations — multiplying by Δx and addition — are done much faster), so that the amount of work done on n steps in this scheme involving recalculation is equivalent to the work needed for $2n$ steps in the scheme for calculating y_{I}. Despite this, however, if high accuracy is needed, that is, if ε is very small, then the recalculation scheme is better since $2\sqrt{\frac{bl^3}{\varepsilon}} \ll \frac{al^2}{\varepsilon}$ if ε is small.

This scheme has yet another advantage. In it there is a good check of the computations and of the choice of the quantity Δx, which is called a step: it is clear that a calculation is good only insofar as the values of $f(x, y)$ and $\bar{f}(x, y)$ differ but slightly.

Let us consider an example. Suppose y is a solution of the equation $y' = x^2 - y^2$ with the initial condition $y = 0$ when $x = -1$. Let us determine the solution for $x = 0$. We take advantage of the recalculation scheme by taking the step $\Delta x = 0.1$.

The calculations are given in Table 5. The intermediate results are given in the second and third columns of the table in parentheses,

Table 5

x	y	$f = y'$	$\bar{f}$	y_{exact}
−1.0	0.0000	1.0000		0.0000
	(0.1000)	(0.8000)	0.9000	
−0.9	0.0900	0.8019		0.0900
	(0.1702)	(0.6110)	0.7064	
−0.8	0.1606	0.6142		0.1607
	(0.2220)	(0.4407)	0.5274	
−0.7	0.2133	0.4444		0.2135
	(0.2577)	(0.2936)	0.3690	
−0.6	0.2502	0.2974		0.2504
	(0.2799)	(0.1717)	0.2345	
−0.5	0.2736	0.1752		0.2738
	(0.2911)	(0.0753)	0.1252	
−0.4	0.2861	0.0782		0.2862
	(0.2939)	(0.0036)	0.0409	
−0.3	0.2902	0.0058		0.2902
	(0.2908)	(−0.0446)	−0.0194	
−0.2	0.2883	−0.0699		0.2882
	(0.2840)	(−0.0706)	−0.0568	
−0.1	0.2826	−0.0699		0.2823
	(0.2756)	(−0.0760)	−0.0730	
0.0	0.2753	−0.0758		0.2749

and under them the results of the recalculation. In the last column are the values of y correct to four decimal places. Comparing them with the ones we obtained, we see that all the values obtained are correct to three decimal places.

The method of recalculation is amenable to further refinement, the result of which are the presently widely used *computational methods of Runge-Kutta* and *Milne* that may be found in texts dealing with numerical methods. The *method of Adams,* which is based on finite differences (Sec. 2.1) is also widely used. We now give this method in a simplified version (ordinarily it is carried to the third differences, but we will confine ourselves to second differences). We proceed from Newton's formula (Ch. 2, (5)) applied to the derivative of the desired solution, $y'(x)$, and in place of k we take $k - 1$:

$$y'(x) = y'_k + (-\delta y'_{k-\frac{1}{2}})\,\frac{h-s}{h} + \frac{\delta^2 y'_{k-1}}{2}\,\frac{h-s}{h}\left(\frac{h-s}{h} - 1\right),$$

$$s = x - x_{k-1}$$

Integrating this equation from $x = x_k$ to $x = x_{k+1}$, that is, from $s = h$ to $s = 2h$, we get (verify this)

$$\int\limits_{x_k}^{x_{k+1}} y'(x)\,dx = y_{k+1} - y_k = y'_k h\,.+ \delta y'_{k-\frac{1}{2}}\,\frac{h}{2} + \delta^2 y'_{k-1}\,\frac{5}{12}\,h$$

or

$$y_{k+1} = y_k + \left(y'_k + \frac{1}{2}\,\delta y'_{k-\frac{1}{2}} + \frac{5}{12}\,\delta^2 y'_{k-1}\right) h \tag{49}$$

This formula is used in the following manner. First, via some procedure (say, with the aid of Taylor's formula, Sec. 8.5, or by means of the recalculation method), we find the values $y_1 = y(x_0 + h)$ and $y_2 = y(x_0 + 2\,h)$. Then we calculate the corresponding values

$$y'_0 = f(x_0, y_0), \quad y'_1 = f(x_1, y_1) = f(x_0 + h, y_1), \quad y'_2 = f(x_2, y_2)$$

with the aid of which we get

$$\delta y'_{\frac{1}{2}} = y'_1 - y_0, \quad \delta y'_{1\frac{1}{2}} = y'_2 - y'_1, \quad \delta^2 y'_1 = \delta y'_{1\frac{1}{2}} - \delta y'_{\frac{1}{2}}$$

Furthermore, assuming in (49) $k = 2$, we compute y_3 and, with its aid,

$$y'_3 = f(x_3, y_3), \quad \delta y'_{2\frac{1}{2}} = y'_3 - y'_2, \quad \delta^2 y'_2 = \delta y'_{2\frac{1}{2}} - \delta y'_{1\frac{1}{2}}$$

Then, putting $k = 3$ in (49), we compute y_4, and with its aid, $y'_4 = f(x_4, y_4)$, and so on.

Particular care is needed in the numerical solution of differential equations when the desired function can increase indefinitely. For instance, suppose we have the equation $\frac{dy}{dx} = y^2$ and the initial condition $y = 1$ for $x = 1$.

It is easy to find the exact solution of this equation. Indeed, $\frac{dy}{y^2} = dx$, when $\int\limits_1^y \frac{dy}{y^2} = \int\limits_1^x dx$ or $1 - \frac{1}{y} = x - 1$, which yields $y = \frac{1}{2 - x}$. It is clear that y increases indefinitely as x approaches 2. If we were to solve this equation numerically, we would of course get a very definite though perhaps very large value of y for $x = 2$.

We found the value of the argument $x = 2$, during the approach to which the solution increases without bound, solely due to the fact that we could write the desired solution in explicit form. But many equations lack such a solution. How is it possible, when solving an equation numerically, to find that value of x during the approach to which the solution increases indefinitely?

What do we do, for example, with the *Riccati equation*

$$\frac{dy}{dx} = \varphi(x) \cdot y^2 + \psi(x) \tag{50}$$

which for arbitrary $\varphi(x)$ and $\psi(x)$ cannot be solved exactly?[*]

We introduce a new sought-for function $z = \frac{1}{y}$. Then $z = 0$ for the value of x we are interested in. Note that $\frac{dz}{dx} = -\frac{1}{y^2}\,\frac{dy}{dx}$. Divide (44) by $-y^2$ to get $-\frac{1}{y^2}\,\frac{dy}{dx} = -\varphi(x) - \frac{\psi(x)}{y^2}$, which is $\frac{dz}{dx} = -\varphi(x) - z^2 \cdot \varphi(x)$.

Solving the last equation numerically, we get the value of x at which $z = 0$.

The foregoing methods of numerical solution of differential equations are readily carried over to the case of a system of two or more equations. Let us illustrate this in the case of a system of two equations:

$$\left.\begin{aligned} \frac{dy}{dx} &= f(x, y, z), \\ \frac{dz}{dx} &= \varphi(x, y, z) \end{aligned}\right\}$$

[*] The equation $\frac{dy}{dx} = y^2$ just considered is obtained from (50) for $\varphi(x) \equiv 1$, $\psi(x) \equiv 0$.

Suppose we have the initial conditions $y = y_0$, $z = z_0$ for $x = x_0$. Set $f(x_0, y_0, z_0) = f_0$, $\varphi(x_0, y_0, z_0) = \varphi_0$. Then the first approximation is

$$y_{\mathrm{I}}(x_1) = y_{\mathrm{I}}(x_0 + \Delta x) = y_0 + f_0 \cdot \Delta x,$$

$$z_{\mathrm{I}}(x_1) = z_{\mathrm{I}}(x_0 + \Delta x) = z_0 + \varphi_0 \cdot \Delta x$$

To obtain a second approximation, we use the first approximation to find the values of the derivatives at the endpoint of the interval, i.e., for $x_1 = x_0 + \Delta x$. We get

$$\left.\frac{dy}{dx}\right|_{x=x_0+\Delta x} = f_{\mathrm{I}} = f(x_0 + \Delta x, \ y_{\mathrm{I}}, \ z_{\mathrm{I}}),$$

$$\left.\frac{dz}{dx}\right|_{x=x_0+\Delta x} = \varphi_{\mathrm{I}} = \varphi(x_0 + \Delta x, \ y_{\mathrm{I}}, \ z_{\mathrm{I}})$$

We then determine the mean values:

$$\bar{f} = \frac{1}{2}(f_0 + f_1), \quad \bar{\varphi} = \frac{1}{2}(\varphi_0 + \varphi_1)$$

The second approximation is

$$y_{\mathrm{II}} = y_0 + \bar{f} \cdot \Delta x, \quad z_{\mathrm{II}} = z_0 + \bar{\varphi} \cdot \Delta x$$

Finally, having obtained y_{II} and z_{II}, we recalculate the values of the derivatives at $x = x_0 + \Delta x$:

$$\left.\frac{dy}{dx}\right|_{x=x_0+\Delta x} = f_{\mathrm{II}} = f(x_0 + \Delta x, \ y_{\mathrm{II}}, z_{\mathrm{II}})$$

$$\left.\frac{dz}{dx}\right|_{x=x_0+\Delta x} = \varphi_{\mathrm{II}} = \varphi(x_0 + \Delta x, y_{\mathrm{II}}, z_{\mathrm{II}})$$

An equation of the second or higher order can, as was shown in Sec. 8.2, be reduced to a system of equations of the first order. Therefore the methods of numerical solution are applicable also to equations of higher than first order.

Note another special case. Suppose we have the equation

$$\frac{d^2x}{dt^2} = \varphi(x, t) \tag{51}$$

Unlike the general case of a second-order equation, the right side here does not contain $\frac{dx}{dt}$.*

Equation (51) admits an analytical solution in two cases.

* Such an equation describes, for example, the motion of a body under the action of a force that depends on the position of the body and on the time in the absence of friction.

1. If the right member $\varphi(x, t)$ does not depend on x. Then (45) assumes the form $\frac{d^2x}{dt^2} = \varphi(t)$, whence $\frac{dx}{dt} = \int \varphi(t)\,dt$. Integrating once again, we get $x(t)$.

2. If the right side does not depend on t, that is, (45) has the form $\frac{d^2x}{dt^2} = \varphi(x)$. In this case we set $v = \frac{dx}{dt}$ and then

$$\frac{d^2x}{dt^2} = \frac{dv}{dt} = \frac{dv}{dx} \cdot \frac{dx}{dt} = v\,\frac{dv}{dx} = \frac{1}{2}\,\frac{d(v^2)}{dx}$$

We can rewrite the equation thus: $\frac{1}{2}\,\frac{d(v^2)}{dx} = \varphi(x)$, whence

$$v^2 = 2\int \varphi(x)\,dx, \quad v = \frac{dx}{dt} = \sqrt{2\int \varphi(x)\,dx}$$

which is an equation with variables separable.

But when the right member depends both on x and on t, neither of these procedures can be used, so the equation has to be solved numerically. First put $\frac{dx}{dt} = v$ and then replace (51) by the first-order system of equations

$$\left.\begin{aligned} \frac{dx}{dt} &= v, \\ \frac{dv}{dt} &= \varphi(x, t) \end{aligned}\right\} \tag{52}$$

The system (52) has the following peculiarity: the right side of the first equation of the system does not involve x and the right side of the second equation does not contain v. This permits suggesting a very convenient computation scheme.

Suppose we have the initial conditions $x = x_0$, $\frac{dx}{dt} = v = v_0$ for $t = t_0$. We choose a step Δt, that is, we seek the values of the solutions for the following values of the arguments: $t_1 = t_0 + \Delta t$, $t_2 = t_0 + 2\,\Delta t$, $t_3 = t_0 + 3\Delta t$, and so forth.

We call the values of the arguments $t_0, t_1, t_2, t_3, \ldots$ integral and the values $t_{1/2}, t_{1^1/_2}, t_{2^1/_2} \ldots$ half-integral. We compute the values of $v = \frac{dx}{dt}$ for half-integral values of the arguments, and the values of x for integral values of the arguments. The sequence of operations is now as follows. Knowing x_0 and v_0, we find $v_{1/2}$ from the formula $v_{1/2} = v_0 + \varphi_0 \cdot \frac{\Delta t}{2}$, where $\varphi_0 = \varphi(x_0, t_0)$. Then we determine $x_1 = x_0 + v_{1/2} \cdot \Delta t$, $\varphi_1 = \varphi(x_1, t_1)$ and $v_{1^1/_2} = v_{1/2} + \varphi_1 \cdot \Delta t$. The process is then repeated: $x_2 = x_1 + v_{1^1/_2} \cdot \Delta t$, $\varphi_2 = \varphi(x_2, t_2)$, $v_{2^1/_2} = v_{1^1/_2} + \varphi_2 \cdot \Delta t$, and so on.

It is convenient to arrange the computations in a table.

Table 6

t	x	$\varphi(x, t)$	$v = \dfrac{dx}{dt}$
t_0	x_0	φ_0	v_0
$t_{1/2}$			$v_{1/2}$
t_1	x_1	$\varphi_1 = \varphi(x_1, t_1)$	
$t_{1^1/_2}$			$v_{1^1/_2}$
t_2	x_2	$\varphi_2 = \varphi(x_2, t_2)$	
$t_{2^1/_2}$			$v_{2^1/_2}$
t_3	x_3	φ_3	
$t_{3^1/_2}$			$v_{3^1/_2}$
t_4	x_4	φ_4	

Thus, when passing, say, from x_1 to x_2, we use the value of the derivative $v_{1^1/_2}$, which value corresponds to the midpoint of the interval. As a result, the accuracy of this procedure is of the same order as that of a second approximation in the ordinary method. (The error at each step is of the order of $(\Delta t)^3$.) The effort expended is the same as when using a first approximation, but we have higher accuracy due to a more reasonable construction of the computational scheme.

It is worth noting once again that this scheme is only possible because of the peculiarities of the system (52).

Exercises

1. Set up a table of the values of the function $y = e^x$ in a numerical solution of the equation $y' = y$ with initial condition $y = 1$ for $x = 0$.

   Obtain the values of e^x for $x = 0.1$, 0.2, 0.3, and so on at intervals of 0.1 up to $x = 1$. Carry the computations via the first approximation to four decimal places. Then take the step $\Delta x = 0.05$ and use the scheme involving recalculation. Compare the results with the tabulated figures.

2. Let $y(x)$ be a solution of the equation $\dfrac{dy}{dx} = x^2 + y^2$ with initial condition $y = 0.5$ for $x = 0$. Find $y(0.5)$.

   Work the problem in two ways:

   (a) using the first approximation with step $\Delta x = 0.025$;

   (b) using the second approximation with step $\Delta x = 0.05$.

   Compare the results.*

* The differential equations of Exercises (2) and (5) can not be solved analytically.

3. Let a body of mass m be in motion under the action of a force $f(t) = at(\theta - t)$ and let it experience a resistance of the medium proportional to the velocity (with proportionality factor k). Let the velocity at the initial time be 0. Set up a differential equation and solve it numerically for the case $m = 10$ g, $a = 5$ g-cm/s^4, $\theta = 20$ s, $k = 1$ g/cm. Determine the velocity of the body 1.5 s after the start of motion.
Hint. Take advantage of the recalculation scheme, taking $\Delta t = 0.05$. Carry the calculations to three places.
4. Set up a table of the values of the function $y = \sin x$, solving the equation $y'' + y = 0$ numerically with initial conditions $y = 0$, $\frac{dy}{dx} = 1$ at $x = 0$. Obtain the values for $x = 0.1$, 0.2, 0.3, etc. at 0.1 intervals up to $x = 1$. Use the recalculation scheme taking the step $\Delta x = 0.1$. Compare the results with those tabulated. Carry the computations to three decimal places.
5. Set up a table of values of the solution of the equation $\frac{d^2x}{dt^2} = t + x^2$ with initial conditions $x = 0, \frac{dx}{dt} = 1$ for $t = 0$. Form the table from $t = 0$ to $t = 0.5$. Take step length $\Delta t = 0.1$ and compute with a scheme involving half-integral values of the argument to three decimal places.

8.8 Boundary-value problems

The general solution of an nth order differential equation has n arbitrary constants, that is, it possesses n degrees of freedom (see Sec. 4.8). In order to pick a particular solution from the general solution we have, up to now, made use of the initial conditions, according to which the desired function and its derivatives are specified for a single value of the argument. This is quite natural if the independent variable is time, that is, if we are studying the development of some process: the initial conditions here simply serve as a mathematical notation for the initial state of the process. That is where the terms *initial conditions, initial-value problem* come from even when the independent variable has a quite different physical meaning. But there are also problems that are stated differently: for example, problems in which there are two "key" values of the independent variable having equal status for which the desired function is given. For example, when considering the deviation $y(x)$ of a string fixed at the endpoints $x = a$ and $x = b$, the conditions imposed on the desired function $y(x)$ are: $y(a) = 0$, $y(b) = 0$. There are also other methods for finding a particular solution from the general solution that are encountered in practical problems. Common to all these methods is that the number of supplementary equations imposed on the desired solution must be equal to the number of degrees of freedom in the general solution of the equation at hand, that is, to say, to the order of this equation.

We will consider a solution of the equation

$$y'' + p(x)y' + q(x)y = f(x) \ (a \leqslant x \leqslant b) \tag{53}$$

with the accessory conditions

$$y(a) = \alpha_1, \ y(b) = \alpha_2 \tag{54}$$

although all the general conclusions we obtained hold true for linear differential equations of any order n under linear accessory conditions of any kind. The conditions of type (54) that are imposed at the endpoints of the interval on which the solution is constructed are called *boundary conditions*, and a problem involving the solution of a differential equation with specified boundary conditions is said to be a *boundary-value problem.*

In Ch. 7 (Secs. 7.2, 7.3, 7.5) we saw that the general solution of the nonhomogeneous linear equation (53) has the following structure:

$$y = Y(x) + C_1y_1(x) + C_2y_2(x) \tag{55}$$

Here, $Y(x)$ is a particular solution of the equation (53), y_1 and y_2 are two independent solutions of the corresponding homogeneous equation, and C_1 and C_2 are arbitrary constants. Substituting (55) into the conditions (54), we get two relations for finding C_1 and C_2:

$$\left.\begin{aligned} C_1y_1(a) + C_2y_2(a) &= \alpha_1 - Y(a) \\ C_1y_1(b) + C_2y_2(b) &= \alpha_2 - Y(b) \end{aligned}\right\} \tag{56}$$

Two cases (see Sec. 8.3) can arise in the solution of this system of two algebraic equations of the first degree in two unknowns.

1. Basic case: the determinant of the system is different from zero. Here, the system (56) has a very definite solution, and therefore the equation (53) with conditions (54) has one and only one solution for any nonhomogeneous term of $f(x)$ and for any numbers α_1, α_2..

2. Particular case: the determinant of the system is zero. Here, the system (56) is, as a rule, inconsistent, but for certain right-hand members it has an infinitude of solutions. Which means that equation (53) for conditions (54) and for an arbitrary choice of the function $f(x)$ and the numbers α_1, α_2 also, does not as a rule have a single solution. However, for certain such choices the problem has an infinity of solutions. For example, it can be verified that if $f(x)$ and α_1 have already been chosen, then an infinitude of solutions results only for a single value of α_2 and there will not be a single solution for the remaining values.

As to which case occurs depends on the form of the left-hand members of equation (53) and conditions (54), this fact has to be stressed.

By Sec. 8.3, for the basic case to occur it is necessary and sufficient that the corresponding homogeneous problem (in which $f(x) \equiv 0$, $\alpha_1 = \alpha_2 = 0$) have only a trivial solution. In the particular case, the homogeneous problem has an infinity of solutions, and if the nonhomogeneous problem has at least one solution, then the general solution is obtained if to this particular solution we add the general solution of the corresponding homogeneous problem.

In solving an initial-value problem (that is, one with initial conditions), we always have to do with the basic case, since such a solution always exists and is unique. In solving a boundary-value problem we may also encounter the particular case. For example, consider the problem

$$y'' + y = 0 \ \left(0 \leqslant x \leqslant \frac{\pi}{2}\right), \quad y(0) = \alpha_1, \quad y\left(\frac{\pi}{2}\right) = \alpha_2$$

By virtue of Sec. 7.3, the general solution of the equation is of the form

$$y = C_1 \cos x + C_2 \sin x \tag{57}$$

Substituting the boundary conditions, we get

$$C_1 = \alpha_1, \quad C_2 = \alpha_2$$

Hence, for arbitrary α_1, α_2, we get a very definite solution

$$y = \alpha_1 \cos x + \alpha_2 \sin x$$

This is the basic case.

For the problem

$$y'' + y = 0 \ (0 \leqslant x \leqslant \pi), \quad y(0) = \alpha_1, \quad y(\pi) = \alpha_2 \quad \cdot \tag{58}$$

substitution of the boundary conditions into the same general solution (57) yields

$$C_1 = \alpha_1, \quad -C_1 = \alpha_2, \text{ that is, } C_1 = -\alpha_2$$

Thus, if $\alpha_1 \neq -\alpha_2$, then the problem (58) does not have a single solution. But if $\alpha_1 = -\alpha_2$, then the problem has the solution

$$y = \alpha_1 \cos x + C_2 \sin x$$

in which C_2 is quite arbitrary, which is to say that we have an infinitude of solutions. This is the particular case.

Finally, consider a problem with the parameter $\lambda =$ constant:

$$y'' + \lambda y = f(x) \ (0 \leqslant x \leqslant l), \quad y(0) = \alpha_1, \quad y(l) = \alpha_2 \tag{59}$$

To begin with we assume $\lambda > 0$. Then the independent solutions of the corresponding homogeneous differential equation are the functions $y_1(x) = \cos \sqrt{\lambda}\, x$, $y_2(x) = \sin \sqrt{\lambda}\, x$, and the determinant of system (56) is

$$\begin{vmatrix} y_1(0) & y_2(0) \\ y_1(l) & y_2(l) \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ \cos\sqrt{\lambda}l & \sin\sqrt{\lambda}l \end{vmatrix} = \sin\sqrt{\lambda}l$$

Equating it to zero, we get the values

$$\lambda = \left(\frac{\pi}{l}\right)^2, \quad \left(\frac{2\pi}{l}\right)^2, \left(\frac{3\pi}{l}\right)^2, \ldots \tag{60}$$

for which we have the particular case for problem (59), which means that either the existence or the uniqueness of solution is upset.

The set of values of the parameter involved in the statement of a problem, for which values the problem degenerates in one sense or another (which is to say that it loses certain very essential properties), is called the *spectrum* of the problem. We leave it to the reader to verify that for $\lambda \leqslant 0$ we always have the basic case for problem (59), and thus the set of values (60) constitutes the entire spectrum of the problem.

The spectrum (60) of problem (59) may also be obtained in a somewhat different but actually equivalent method. As has been pointed out, the particular case for a boundary-value problem is characterized by the fact that the corresponding homogeneous problem

$$y'' + \lambda y = 0, \quad y(0) = 0, \; y(l) = 0$$

can have a nontrivial solution. The general solution of this differential equation is of the form

$$y = C_1 \cos \sqrt{\lambda} x + C_2 \sin \sqrt{\lambda} x = C(D \cos \sqrt{\lambda} x + \sin \sqrt{\lambda} x)$$

where $C = C_2$, $D = \frac{C_1}{C_2}$. Using the boundary conditions yields

$$C(D \cos \sqrt{\lambda} 0 + \sin \sqrt{\lambda} 0) = 0, \; C(D \cos \sqrt{\lambda} l + \sin \sqrt{\lambda} l) = 0$$

We see that the constant C remains arbitrary, whereas there remain two equations to find the two constants D and λ:

$$D \cos \sqrt{\lambda} 0 + \sin \sqrt{\lambda} 0 = 0, \; D \cos \sqrt{\lambda} l + \sin \sqrt{\lambda} l = 0$$

From this we get

$$D = 0, \sin \sqrt{\lambda} l = 0, \quad \sqrt{\lambda} l = k\pi \quad (k = 1, 2, \ldots)$$

and we arrive at the same values (60) for λ.

The result just obtained has an interesting application to the investigation of the stability of an elastic rod in the case of compression. Let a homogeneous (which means being the same throughout its length) elastic rod be located along the x-axis and let it be compressed along this axis by a force P (both ends of the rod are held on the x-axis, but can be free to rotate about the points of attachment (Fig. 96*a*)). Now, when the force attains a certain critical value, P_{cr}, the rod bends and takes up the position shown in Fig. 96*b*. If we denote by y the transverse deviation of a point

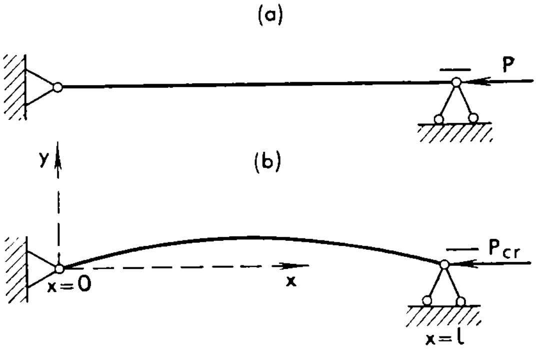


Fig. 96

of the rod from its original position, then, as is demonstrated in strength of materials courses, the function $y(x)$ satisfies, to a sufficient degree of accuracy, the differential equation and boundary conditions

$$y'' + \frac{P}{EJ}y = 0, \quad y(0) = y(l) = 0 \tag{61}$$

Here, E and J are the so-called *modulus of elongation* of the material of the rod and its *moment of inertia*, respectively.

As follows from (60), when

$$\frac{P}{EJ} < \left(\frac{\pi}{l}\right)^2 \tag{62}$$

then we have the basic case for the problem (62), that is, it has only a trivial solution: no bending occurs. As soon as, with the increase of P, the inequality (62) becomes an equality, the particular case sets in and problem (61) has, in addition to the trivial solution, a solution of the form $y = C \sin \frac{\pi}{l} x$, where C is an arbitrary constant. But then the rod cannot be held in the rectilinear state and small external forces* can lead to finite deviations from this state: the rod loses stability. The resulting expression for P_{cr},

$$P_{\text{cr}} = EJ\left(\frac{\pi}{l}\right)^2$$

was found by Euler in 1757. It might appear that for $P > P_{\text{cr}}$ the rod would straighten out, but this is not so. Equation (61) describes the deviation of the rod only in the limit for small deviations, whereas an analysis of the more exact equation that holds true for arbitrary deviations (it turns out to be nonlinear) shows that as

* We have in view external forces which tend to deflect the rod from the rectilinear state, for example, a small force directed perpendicularly to the rod.

P passes through P_{cr}, in addition to the unstable rectilinear form of equilibrium there appears a curved form of equilibrium, which is stable. As P increases, the curvature of this form rises rapidly and the rod is destroyed.

Green's function (Sec. 6.2) can be applied to the solution of a nonhomogeneous equation under homogeneous boundary conditions,

$$\left.\begin{aligned} y'' + p(x)y' + q(x)y &= f(x) \quad (a \leqslant x \leqslant b),\\ y(a) = 0, \quad y(b) &= 0 \end{aligned}\right\} \tag{63}$$

in the basic (nonparticular) case, since it is clear that when the right members are added, so are the solutions. In accordance with Sec. 6.2, if we denote by $G(x, \xi)$ the solution of problem (63), in which we take the delta function $\delta(x - \xi)$ instead of $f(x)$, then for an arbitrary function $f(x)$ the solution of (63) can be obtained from the formula

$$y(x) = \int_a^b f(\xi)\, G(x, \xi)\, d\xi \tag{64}$$

Here is a simple example. Suppose we have the problem

$$y'' = f(x) \quad (0 \leqslant x \leqslant l), \quad y(0) = y(l) = 0 \tag{65}$$

If instead of $f(x)$ we put $\delta(x - \xi)$, then for $0 \leqslant x < \xi$ (and $\xi < x \leqslant \leqslant l$ we simply get $y'' = 0$ or the solution

$$y = ax + b \;\; (0 \leqslant x < \xi), \quad y = cx + d \;\; (\xi < x \leqslant l)$$

where a, b, c, d are some kind of constants. Applying the boundary conditions shows that $b = 0$ and $cl + d = 0$, or

$$\begin{aligned} y &= ax \;\; (0 \leqslant x < \xi),\\ y &= c(x - l) \;\; (\xi < x \leqslant l) \end{aligned} \tag{66}$$

If the equation $y'' = \delta(x - \xi)$ is integrated from $x = \xi - 0$ to $x = \xi + 0$, then we find that $y'(\xi + 0) - y'(\xi - 0) = 1$. Incidentally, for the left-hand side of equation (63) we would have the same result since integration of a finite function over an interval of length zero yields zero. A second integration of the delta function yields a continuous function so that $y(\xi - 0) = y(\xi + 0)$ and from (66) we get $c - a = 1$, $a\xi = c(\xi - l)$, whence

$$a = -\frac{l - \xi}{l}, \quad c = \frac{\xi}{l}$$

Substituting into (66), we get Green's function for the problem (65):

$$G(x, \xi) = \begin{cases} -\dfrac{(l - \xi)\, x}{l} \;\; (0 \leqslant x < \xi),\\[2mm] -\dfrac{(l - x)\, \xi}{l} \;\; (\xi < x \leqslant l) \end{cases}$$

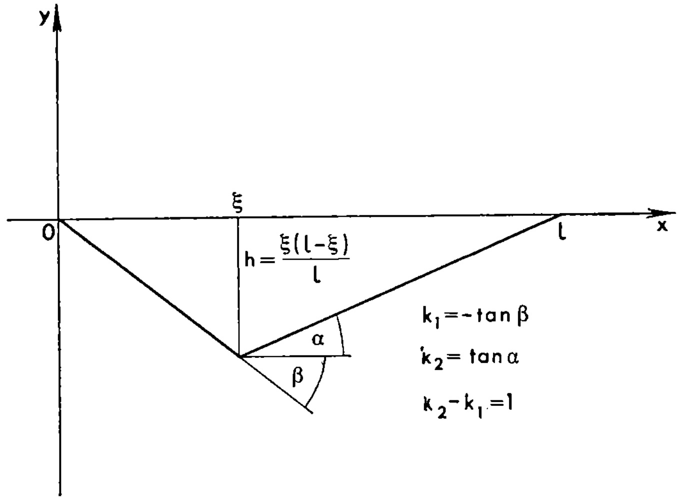


Fig. 97

This function is shown in Fig. 97. It will be seen that it differs from the function constructed in Sec. 6.2 only by the constant factor $-F$. By virtue of formula (64) we get the solution of problem (65) for any function $f(x)$:

$$y=\int_0^l G(x,\xi)\,f(\xi)\,d\xi=\int_0^x G(x,\xi)\,f(\xi)\,d\xi+\int_0^l G(x,\xi)\,f(\xi)\,d\xi=$$

$$=-\frac{l-x}{l}\int_0^x \xi f(\xi)\,d\xi-\frac{x}{l}\int_x^l (l-\xi)\,f(\xi)\,d\xi$$

Exercises

1. Find the spectrum of the boundary-value problem $y''+\lambda y=0$, $y(0)=0$, $y'(l)=0$.
2. Use Green's function to construct a solution of the problem
$$y''+y=f(x),\ y(0)=0,\ y(\pi/2)=0.$$

8.9 Boundary layer

It often happens that the differential equation being studied or a system of such equations contains one or several parameters that can take on a variety of constant values. For the sake of simplicity, let us consider the first-order equation

$$\frac{dy}{dx}=f(x,y;\lambda) \tag{67}$$

(where λ is a parameter) for definite initial conditions $x=x_0$, $y=y_0$.

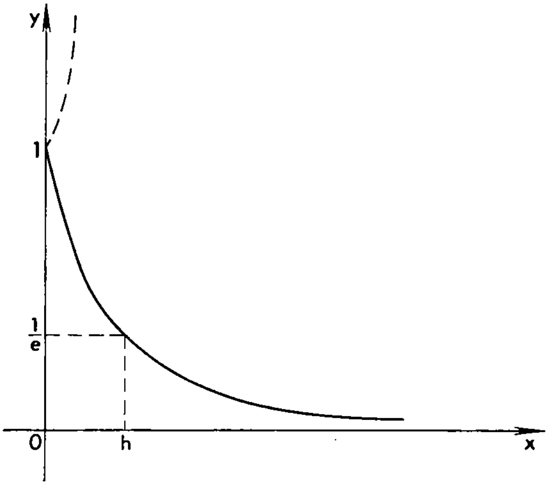


Fig. 98

We assume that the point (x_0, y_0) is not singular (Sec. 8.1), that is, for the given conditions there is a unique solution of the equation (67). Then from the geometric meaning of the equation (67) (Sec. 7.1) it follows that if the right side depends on λ in continuous fashion, then the direction field will vary but slightly for small changes in λ, and for this reason the solution $y(x, \lambda)$ will also depend on λ in continuous fashion.

However, it sometimes happens that the parameter occurs in the differential equation in a manner such that for certain values of the parameter the equation lowers its order (degenerates). Then new circumstances arise which we illustrate with an example.

Consider the problem

$$\lambda y' + y = 0, \quad y\,|_{x=0} = 1 \tag{68}$$

with the solution $y = e^{-x/\lambda}$. The equation degenerates when $\lambda = 0$ (why ?). Suppose the solution is considered for $x \geqslant 0$ and $\lambda \to +0$; this solution is shown in Fig. 98. The equation (68) passes into $y = 0$ in the limit, but we see that for small λ the solution is close to zero not from $x = 0$ at once but only from a certain $x = h$. The interval $0 < x < h$, which is called the *boundary layer*, serves as a transition from the unit initial value (68) to a value close to zero. The width of the boundary value is merely conventional, since theoretically the solution never becomes exactly equal to zero. If, say, for the width of the boundary layer we take the value $x = h$, at which the solution diminishes e-fold the original value, then we get

$$e^{-h/\lambda} = \frac{1}{e}, \quad h = \lambda$$

which means that for the problem (68) the width of the boundary layer is simply equal to λ.

This is the profile that the velocity of a fluid has in the sliding motion of a lamina in a fluid at rest. Here, x is the distance from the lamina, reckoned along the normal to it; the velocity of the fluid is laid off on the y-axis, and the parameter λ is proportional to the viscosity of the fluid. It turns out that for the equations of motion of a viscous fluid (this is a system of partial differential equations) the coefficient of viscosity serves as a factor of the highest derivative; in other words, for these equations we have the same situation as for the model problem (68). In the case of viscosity, the fluid adheres to the lamina, the layer of fluid entrained by the lamina being the narrower, the lower the viscosity. Incidentally, this is clear from physical reasoning. In the limit, when the viscosity is zero (this is called an ideal fluid), the lamina slides without entraining any fluid, and the velocity of the fluid is equal to zero right down to the very surface of the lamina.

If $\lambda \to -0$, then the resulting solution depicted in Fig. 98 by the dashed line tends to infinity for any $x > 0$. This case is of less interest.

Exercise

Consider the behaviour of the solution of the problem

$$\lambda y'' - y = 1, \quad y(-1) = y(1) = 0 \text{ as } \lambda \to +0.$$

8.10 Similarity of phenomena

Two or several phenomena are said to be *similar* (in physics, chemistry, engineering, sociology, etc.) if they differ in scale alone. For example, Fig. 99 shows similar processes of current variation in a circuit: in both cases the current builds up from zero over an interval of time, then remains constant, after which it drops suddenly to zero. Thus, if for the characteristic time t_{ch} (which in this process is the standard for comparison purposes) we take the time of current buildup, and for the characteristic current j_{ch}, the maximum current, and if we reckon the time from the onset of buildup and then denote

$$t = t_{\text{ch}}\tilde{t}, \quad j = j_{\text{ch}}\tilde{j}$$

where $\tilde{t}$ and $\tilde{j}$ are nondimensional time and current, then the relationship $\tilde{j}(\tilde{t})$ shown in Fig. 100 will be the same for both processes.

Note that the graphs in Fig. 99 are not similar in a geometric sense since the angles change when passing from one graph to the other and the lengths are not proportional. But if quantities with different dimensions are laid off on the axes, we are never interested in the angles and in the segments that are not parallel to the coordinate axes. For example, in Fig. 99 the distance of a point from the origin

Fig. 99

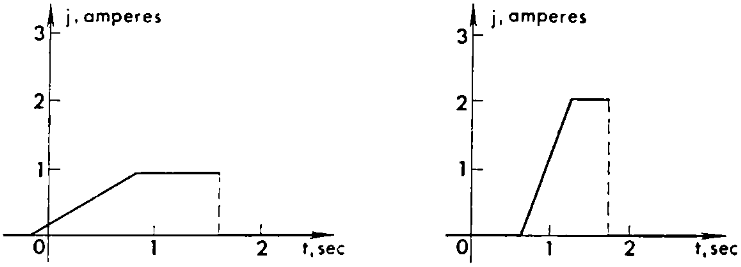
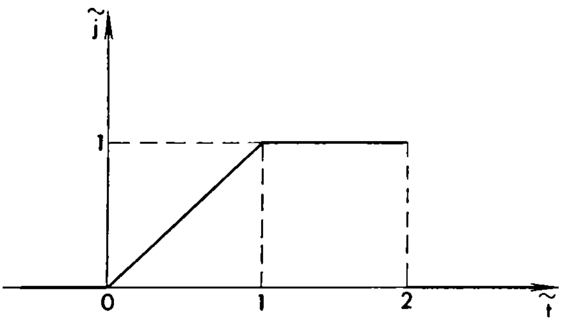


Fig. 100



is equal to $\sqrt{t^2 + j^2}$ by the rules of analytic geometry, but this expression is meaningless from the point of view of dimensions and for this reason will never be encountered.

If the relations $j(t)$ are similar, then so also are other relationships, such, say, as $\int^t j(t)\,dt$, which is the dependence of the magnitude of the charge flow on the time, or Rj^2 the dependence of the electric power on the time, and so on. But imagine that there is a device in the circuit that is turned on when the current reaches a definite value j_0; the operation of such a device is described by the function $e(j - j_0)$, where e is a unit function (Sec. 6.3). It is clear that, generally, there will be no similarity with respect to this device. Hence one speaks of similarity relative to certain characteristics and not in general.

How can we find out if two phenomena are similar or not? If the characteristics relative to which the similarity is considered are obtained from a solution of certain equations, then see if linear transformations can be performed with the quantities in these equations (i.e. substitutions of the type $x \to a_x\tilde{x} + b_x$, which signify changes in the scale and the reference point) so that the equations for both cases become the same. Of course if one is dealing with differential equations, then after the transformation the initial conditions (which are also needed for a solution) must likewise coincide. Here is an equivalent procedure: if the equations of two phenomena can be reduced via linear transformations to one and the same "standard" form, then the phenomena are similar.

Here is an example. Suppose we are considering forced oscillations of a linear oscillator without friction (Sec. 7.5) defined by the equation

$$m\frac{d^2x}{dt^2} + kx = A\sin\omega t \tag{69}$$

and by initial conditions

$$x\Big|_{t=0} = x_0, \quad \frac{dx}{dt}\Big|_{t=0} = 0 \tag{70}$$

There are five parameters in this problem: m, k, A, ω, x_0. To determine when the oscillations will be similar for various combinations of these parameters, take l/ω and $x_0 \neq 0$ for the characteristic time and length and denote $t = \frac{1}{\omega}\tilde{t}$, $x = x_0\tilde{x}$. After passing to the nondimensional variables $\tilde{t}$, $\tilde{x}$, and performing simple transformations, we get the differential equation and initial conditions in the form

$$\frac{d^2\tilde{x}}{d\tilde{t}^2} + \frac{k}{m\omega^2}\tilde{x} = \frac{A}{x_0 m\omega^2}\sin\tilde{t}, \quad \tilde{x}\Big|_{\tilde{t}=0} = 1, \quad \frac{d\tilde{x}}{d\tilde{t}}\Big|_{\tilde{t}=0} = 0 \tag{71}$$

Thus, if for one oscillator the values

$$I_1 = \frac{k}{m\omega^2}, \quad I_2 = \frac{A}{x_0 m\omega^2} \tag{72}$$

are the same as those of the other, then the standard problems (71) are the same for them, which means that these oscillations are similar. It is easy to verify that the quantities I_1 and I_2 are nondimensional. (By applying the results of Sec. 7.5 it is easy to obtain the formulas $I_1 = \omega_0^2/\omega^2$, $I_2 = x_{\text{free}}/x_0$, where ω_0 is the natural frequency of the oscillator and x_{free} is the amplitude of oscillations of the free mass m under the action of a force $A\sin\omega t$.) Such nondimensional quantities whose coincidence ensures the similarity of phenomena are called *similarity criteria*. In the problem at hand there are two criteria (72), of which only I_1 is determined by the parameters of the oscillator, whereas the coincidence of I_2 can be ensured through a choice of the initial conditions. (The case $x_0 = 0$ was not included in the foregoing reasoning. We leave it to the reader to see that in this case the sole similarity criterion will be I_1.)

The solution of the standard problem (71) has the form $\tilde{x} = \varphi(\tilde{t}; I_1, I_2)$, where the specific form of the function is readily obtainable by the method of Sec. 7.5. Returning to the original variables, we get the following oscillation law:

$$x = x_0\varphi\left(\omega t;\ \frac{k}{m\omega^2},\ \frac{A}{x_0 m\omega^2}\right)$$

In many cases the similarity criterion can be established directly from the dimensions. In the problem discussed above, we take for the basic dimensions the mass M^e, the time T and the length L. Then the dimensions of the parameters of the problem will be $[m] = M$, $[k] = MT^{-2}$, $[A] = MT^{-2}L$, $[\omega] = T^{-1}$, $[x_0] = L$ (verify this!). Using these parameters, we can form only one dimensionless combination not containing the initial data. This is I_1. (Of course, the quantities I_1^2, $I_1^{-1/2}$ and so on will also be nondimensional, but they do not yield any new and independent similarity criteria.) Another nondimensional combination that contains the initial data is I_2. It is easy to see — we leave this to the reader — that any nondimensional combination of the type $m^a k^b A^c \omega^d x_0^e$ can be represented in the ollowing manner: $I_1^b I_2^c$, that is, in this problem there are no similarity riteria independent of I_1 and I_2.

Let us consider another example. Suppose a ball of mass m is suspended by a string of length l and is oscillating with frequency ν and maximum angle α of deviation from the vertical; air resistance, the mass of the string, and other complicating factors are ignored. It is clear that there is one relation between the indicated parameters and the acceleration of gravity g: for example, we can arbitrarily specify m, l, g and α, and then ν is determined uniquely. The dimensions of the parameters are: $[m] = M$, $[l] = L$, $[\nu] = T^{-1}$,$[\alpha] = 1$, $[g] = LT^{-2}$. Here the complete system of nondimensional criteria is $S = l\nu^2 g^{-1}$ and α; and since there must be a relationship connecting hem, S must depend on α alone, whence

$$S^{1/2} + l^{1/2}\nu g^{-1/2} = f(\alpha), \text{ that is, } \nu = \sqrt{\frac{g}{l}}f(\alpha) \qquad (73)$$

The function $f(\alpha)$ may be obtained either numerically, by integrating the appropriate differential equation, or experimentally, by measuring ν for specified values of the remaining parameters. Since S is expressed in terms of α, it follows that α in this problem is the sole criterion of similarity.

We thus see that simple dimensional reasoning enabled us to obtain important information concerning oscillations. These are of course only the simplest examples of the application of the theory of similarity and dimensionality, which is today widely used in vaeious branches of physics. The theory and its applications are describcd very well in the following books: [2], [5], [14]. Beginning with a rertain level, the mathematical theory is closely tied up with physics and is being developed in various spheres of physics in different ways.

Simulation (*model-building*) is based on the similarity of phenomena. Here the behaviour of an object is studied on a model similar to the original in the sense described above. For example, to find out what the frequency will be of a pendulum on the moon for a given α, we can measure the frequency for that same α under terrestrial conditions, and then recalculate by the formula

$$\nu_1 : \nu_2 = \sqrt{\frac{g_1}{l_1}} : \sqrt{\frac{g_2}{l_2}}$$

that follows from the relation (73).

ANSWERS AND SOLUTIONS

Sec. 8.1

Saddle point, nodal point, vortex point.

Sec. 8.2

$yy' + zz' = y^2 + z^2$, $\frac{d(y^2+z^2)}{y^2+z^2} = 2\,dx$, whence $y^2 + z^2 = Ce^{2x}$. From this we see that for $x \to \infty$ all the particular solutions (except the trivial solution) become infinite, whereas they tend to zero as $x \to -\infty$.

Sec. 8.3

1. If $a \neq 6$, the system has exactly one solution. For $a = 6$, $b \neq 9$, the system is inconsistent. For $a = 6$, $b = 9$ the system has an infinity of solutions: $x = t$, $y = \frac{3-t}{2}$ (for arbitrary t).
2. $p_1 = 1$, $\lambda_1 = 1$, $\mu_1 = 3$, $p_2 = 6$, $\lambda_2 = 1$, $\mu_2 = -2$, $x = C_1e^t + C_2e^{6t}$, $y = 3C_1e^t - 2C_2e^{6t}$.

Sec. 8.4

At the point (0, 0) the equilibrium is stable; at the points (1, —1) and (—1, 1) it is unstable.

Sec. 8.5

1. $y_0(x) = 1$, $y_1(x) = 1 + x$, $y_2(x) = 1 + x + \frac{x^2}{2}$, $y_3(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{2\cdot 3}$, ... In the limit we obtain the exact solution $y = e^x$ expanded in a series of powers of x.
2. $y = x + \frac{x^3}{3} + \frac{x^5}{6} + \dots$.
3. $y = \frac{1}{1-x} + \frac{x^2(6 - 8x + 3x^2)}{12(1-x)^2}\alpha + \dots$.

Sec. 8.7

1. The following is a tabulation of the computations via the first approximation, the second approximation, and the exact values:

x	y_{first}	y_{second}	y_{exact}
0.0	1.0000	1.0000	1.0000
0.1	1.1000	1.1051	1.1052
0.2	1.2100	1.2212	1.2214
0.3	1.3310	1.3496	1.3499
0.4	1.4641	1.4915	1.4918
0.5	1.6105	1.6483	1.6487
0.6	1.7716	1.8216	1.8221
0.7	1.9488	2.0131	2.0138
0.8	2.1437	2.2248	2.2255
0.9	2.3581	2.4587	2.4596
1.0	2.5939	2.7172	2.7183

2. By the first method, $y(0.5) = 0.7081$; by the second, 0.7161.

3. The differential equation has the form $\frac{dv}{dt} = 0.5\, t\, (20 - t) - 0.2\, v$. The velocity 1.5 seconds after the start of motion is 9.682 cm/s.

4. The computed values and exact values are:

x	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
y_{approx}	0.000	0.100	0.199	0.296	0.390	0.480	0.565	0.645	0.718	0.784	0.842
y_{exact}	0.000	0.100	0.199	0.296	0.389	0.479	0.565	0.644	0.717	0.783	0.842

5. The following is a table of the values of the solution:

t	0.0	0.1	0.2	0.3	0.4	0.5
x	0.000	0.100	0.201	0.305	0.412	0.525

Sec. 8.8

1. Let $\lambda > 0$. Then $y_1(x) = \cos\sqrt{\lambda}\, x$, $y_2(x) = \sin\sqrt{\lambda}\, x$ and the determinant of the system similar to (56) is

$$\begin{vmatrix} y_1(0) & y_2(0) \\ y_1'(l) & y_2'(l) \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -\sqrt{\lambda}\sin\sqrt{\lambda}\, l & \sqrt{\lambda}\cos\sqrt{\lambda}\, l \end{vmatrix} = \sqrt{\lambda}\cos\sqrt{\lambda l}$$

Whence the spectrum is determined from the equation $\cos\sqrt{\lambda}\, l = 0$, that is, $\sqrt{\lambda} l = -\frac{\pi}{2} + k\pi$, $\lambda = \frac{(2k-1)^2\pi^2}{4l^2}$ $(k = 1, 2, \ldots)$.

If $\lambda < 0$, then $y_1(x) = e^{\sqrt{|\lambda|}\,x}$, $y_2(x) = \mathrm{e}^{-\sqrt{|\lambda|}\,x}$; the determinant is equal to $\begin{vmatrix} 1 & 1 \\ \sqrt{|\lambda|}\,e^{\sqrt{|\lambda|}\,l} & -\sqrt{|\lambda|}\,e^{-\sqrt{|\lambda|}\,l} \end{vmatrix} = -\sqrt{|\lambda|}\,(e^{\sqrt{|\lambda|}\,l} + e^{-\sqrt{|\lambda|}\,l}) <$ < 0, that is, it does not vanish. For $\lambda = 0$ we have $y_1 = 1$, $y_2 = x$; the determinant is equal to $\begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1$.

2. In this problem, $G(x, \xi) = \begin{cases} -\cos\xi \sin x & (0 \leqslant x \leqslant \xi), \\ -\sin\xi \cos x & \left(\xi \leqslant x \leqslant \frac{\pi}{2}\right) \end{cases}$ whence the solution of the problem is of the form

$$y = -\cos x \int\limits_0^x \sin\xi\, f(\xi)\, d\xi - \sin x \int\limits_x^{\frac{\pi}{2}} \cos\xi\, f(\xi)\, d\xi$$

Sec. 8.9

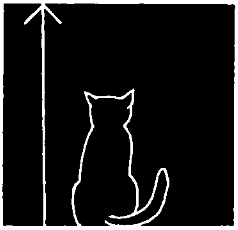
The solution has the form

$$y = -1 + \frac{e^{\frac{x}{\sqrt{\lambda}}} + e^{-\frac{x}{\sqrt{\lambda}}}}{e^{\frac{1}{\sqrt{\lambda}}} + e^{-\frac{1}{\sqrt{\lambda}}}}$$

As $\lambda \to +0$ it tends to the solution of the degenerate equation, that is, to $y = -1$ for all x between -1 and 1. Near both endpoints there arises a boundary layer whose width is asymptotically equal to $\sqrt{\lambda}$ as $\lambda \to 0$.

Chapter 9

VECTORS



In physics we often have to do with vectors, which are quantities endowed with numerical value and also direction. Examples are: a line segment connecting the origin of coordinates with a given point, the velocity of a moving material point, the force acting on a body.

If a body is in motion along a definite line, say a straight-line railway track, the position of the body can be determined by the distance from a specified point on the line measured along this line. Motion on such a line is possible only in two directions, which can be distinguished by affixing a plus sign to one direction and a minus sign to the other.

If a body is known to be moving in a plane (or in space), then we cannot indicate the position of the body at any time if only its distance is given from a specified point; we also have to specify the direction of the line connecting the body with that point (coordinate origin). In the same way, when specifying velocity we have to indicate the magnitude and the direction. Quantities endowed with direction are termed *vectors*. We will indicate them by boldface type or by an arrow over the letter. In contrast to vectors, quantities that do not have direction and are completely determined by their numerical value in a chosen system of units are called *scalars*. Examples are: the mass of a body, its energy, the temperature of a body at a given point. The term "scalar" was not needed as long as we got along without the word "vector".

Vectors can be treated in three-dimensional space or on a plane (in two-dimensional space).

If we disregard the direction of a vector quantity, we can deal with the absolute value (modulus) of the vector. The modulus is a positive scalar having the dimensions of the given quantity. For example, for a vector **F** of 5 newtons force having a definite direction, the modulus (denoted by $|\mathbf{F}|$ or F) is 5 newtons; $|\mathbf{F}| =$ 5 newtons.

Two vectors are taken to be *equal* if they have the same moduli, are parallel and in the same direction. This means that every vector can, without alteration, be *translated* parallel to itself to any spot, which means the origin of such a vector can have any location. To specify a vector means to specify its modulus and direction.

Geometrically, a vector is depicted by a line segment whose direction is indicated by an arrow. This of course requires a scale (for instance, a force of 1 newton can be represented by a line segment of 3 cm length, and so forth). Only a vector quantity having the dimensions of length can be represented without this condition, that is to say, on a 1-to-1 scale. For this reason, we can say, for example, that a vector of translation laid off from point A in space will reach point B, whereas for a vector of force such an assertion is meaningless.

9.1 Linear operations on vectors

Linear operations involving vectors include addition (and the associated subtraction) and multiplication of the vector by a scalar. (Adding a vector and a scalar is just as absurd as trying to add seconds and centimetres.) A quantity is characterized as a vector only if these operations are performed in accord with the rules given below.

Addition of two vectors obeys the familiar parallelogram rule of school physics (see Fig. 101). To find the sum here it is sufficient to construct only one of the two triangles shown in Fig. 101. From this it is easy to obtain a rule for the addition of several vectors: in Fig. 102 we have

$$\overrightarrow{OB} = \mathbf{a} + \mathbf{b},\ \overrightarrow{OC} = \overrightarrow{OB} + \mathbf{c} = \mathbf{a} + \mathbf{b} + \mathbf{c},$$

$$\overrightarrow{OD} = \overrightarrow{OC} + \mathbf{d} = \mathbf{a} + \mathbf{b} + \mathbf{c} + \mathbf{d}$$

(In the three-dimensional case the vectors **a**, **b**, **c**, **d** do not necessarily need to lie in a single plane.) Thus the sum of a number of vectors is represented by the line segment that closes the polygonal line whose segments are the vector summands; the direction of the closing vector is from the beginning of the first summand of the vector to the end of the last summand.

If, referring to Fig. 102, we reverse the direction of the vector $\overrightarrow{OD}$, the conclusion we come to is particularly interesting. If the vectors form a closed polygon (each vector is applied to the end of the preceding one and the end of the last one coincides with the origin of the first), then the sum of all these vectors is equal to the *zero (null) vector* **0**, which is a vector whose terminus coincides with its origin. The modulus of a null vector is zero and the direction is undefined.

Note that vectors cannot be connected by an inequality sign; in particular, there are no positive and negative vectors.

Multiplication of a vector by a scalar (say, by an ordinary number) is a natural generalization of addition and subtraction of

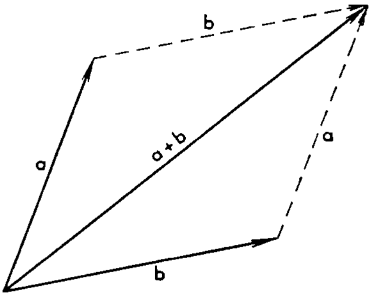


Fig. 101

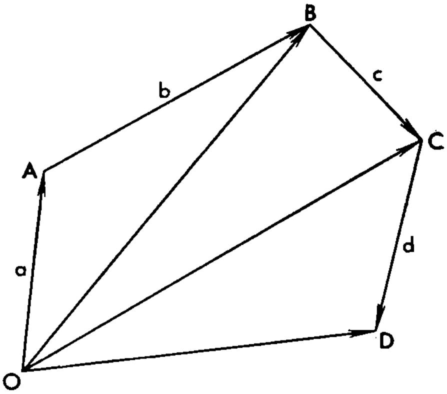


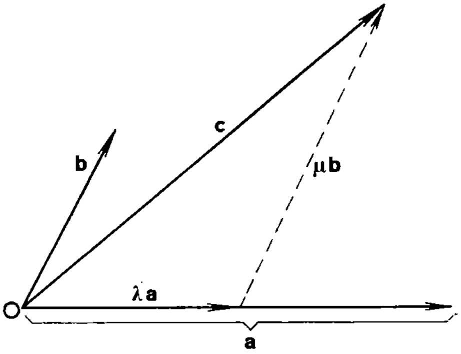
Fig. 102

vectors. The vector 3**F** stands for the sum **F** + **F** + **F**; from the construction it is clear that this vector had the direction of **F** but is three times as long. The vector (-1) **F** = −**F** is understood to be a vector, which, when added to **F**, yields **0**. It is clear that this vector is in the reverse direction to that of **F** and has the same modulus as **F**. Generalizing these definitions, we say that λ**F** (where λ is any scalar) is a vector whose modulus is equal to $|\lambda|\,|\mathbf{F}|$, and λ**F** is parallel to **F** and in the same direction if $\lambda > 0$ and in the opposite direction if $\lambda < 0$. (For $\lambda < 0$ we say that the vectors λ**F** and **F** are *antiparallel.*)

Linear operations involving vectors abide by all the ordinary rules of elementary mathematics. For instance, a term **a** can be transposed to the other side of an equation to become − **a**, both sides of a vector equation can be multiplied or divided by one and the same scalar by the ordinary rules, etc.

Any expression of the form $\lambda\mathbf{a} + \mu\mathbf{b} + \ldots + \xi\mathbf{d}$, where $\lambda, \mu, \ldots, \xi$ are scalars, is called a *linear combination* of the vectors **a**, **b**, ..., **d**. The vectors thus specified are said to be *linearly dependent* if any one of them is a linear combination of the others; otherwise these vectors are said to be *linearly independent* (among themselves).

Fig. 103



The linear dependence of two vectors means that they are parallel (think this over!). If any two nonparallel vectors **a** and **b** have been chosen in a plane, then any third vector **c** in that plane can be "resolved into the vectors **a** and **b**", which is to say, it can be represented in the form of a linear combination (Fig. 103):

$$\mathbf{c} = \lambda\mathbf{a} + \mu\mathbf{b} \tag{1}$$

It is therefore possible, in a plane, to indicate two linearly independent vectors, but any three vectors are then linearly dependent. The expansion (1) is frequently encountered in mechanics and other fields (the resolution of a force along two directions, and the like), each one of the terms $\lambda\mathbf{a}$ and $\mu\mathbf{b}$ being called a *component* of the vector **c**, and the two vectors **a**, **b** being termed the *basis.*

In the same way, in space we can indicate three linearly independent vectors (any three vectors not parallel to a single plane). They can be taken as a basis, which means that any fourth vector can be resolved into these three, and so any four vectors in space are linearly dependent. The difference lies in the fact that a plane is two-dimensional and ordinary space is three-dimensional. If we introduce the concept of a vector into n-dimensional space (Sec. 4.8), then the basis there will consist of n vectors (see Sec. 9.6).

The most widely used bases are those consisting of *unit* vectors (i.e. with nondimensional modulus equal to 1) that are mutually perpendicular. Such bases are termed *Cartesian* or *Euclidean.* Vectors used to form a Cartesian basis are ordinarily denoted by **i**, **j** (in the plane) and **i**, **j**, **k** (in space). Thus, by analogy with (1), we can write the resolution of any vector:

$$\mathbf{a} = a_x\mathbf{i} + a_y\mathbf{j}$$

in the plane and

$$\mathbf{a} = a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k}$$

in space. Here, a_x, a_y, a_z are the coefficients of the expansion.

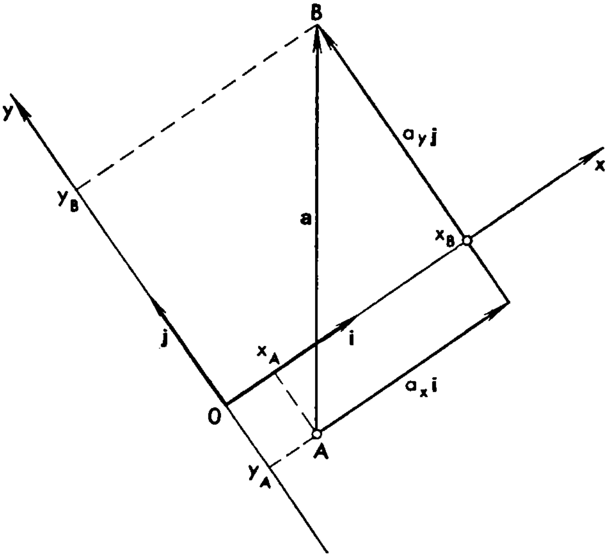


Fig. 104

It is easy to determine the geometric significance of these coefficients which are called the *Cartesian coefficients of the vector a*; for the sake of definiteness we will speak of vectors in the plane, since the results are quite analogous for space. In the plane choose a point O, called the *origin of coordinates,* and through it draw axes parallel to the vectors of the chosen basis. Denote these axes by the letters x and y (Fig. 104). We thus obtain a *Cartesian system of coordinates* in which the position of each point A in the plane is defined by the coordinates (x_A, y_A). From Fig. 104 we see that for the vector $\mathbf{a} = \overrightarrow{AB}$ we have

$$a_x = x_B - x_A, \quad a_y = y_B - y_A$$

The difference $x_B - x_A$ is also called the *projection* of the vector $\overrightarrow{AB}$ on the x-axis. Generally, the projection $\text{pr}_l\, \mathbf{a}$ of vector $\mathbf{a} = \overrightarrow{AB}$ on some axis l (that is, on a straight line on which is indicated which of the two directions is taken to be positive) is called the *modulus* (absolute value) of the vector $\overrightarrow{A'B'}$ defined by the feet of the perpendiculars (Fig. 105) dropped from points A and B to the axis l; this modulus is taken with the plus sign or the minus sign depending on whether the vector $\overrightarrow{A'B'}$ goes along the positive or negative l-axis. The projection of one vector on another is defined in similar fashion; in this case, perpendiculars are dropped to the other vector or to its prolongation.

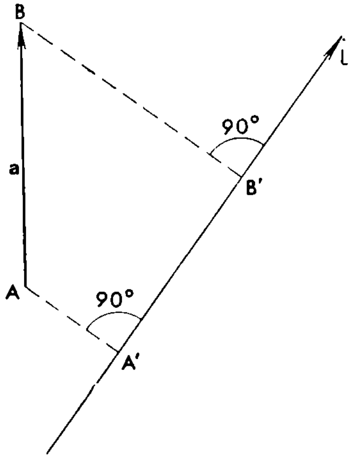


Fig. 105

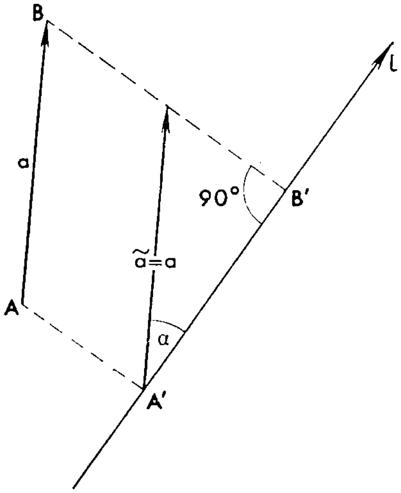


Fig. 106

Thus, the Cartesian coordinates of a vector are its projections on the basis vectors:

$$a_x = \text{pr}_x\mathbf{a} = \text{pr}_\mathbf{i}\mathbf{a}, \quad a_y = \text{pr}_y\mathbf{a} = \text{pr}_\mathbf{j}\mathbf{a}, \quad a_z = \text{pr}_z\mathbf{a} = \text{pr}_\mathbf{k}\mathbf{a}$$

We stress the fact that the projection of a vector is a scalar. Its physical dimensions are those of the vector being projected. From Fig. 106 there follows immediately a simple formula for computing projections:

$$\text{pr}_l\,\mathbf{a} = |\,\mathbf{a}\,|\cos\alpha = |\,\mathbf{a}\,|\cos(\widehat{\mathbf{a}, l}) \qquad (2)$$

From this follows, in particular, the widely used formula for the Cartesian resolution of the unit vector **e**:

$$\mathbf{e} = e_x\mathbf{i} + e_y\mathbf{j} + e_z\mathbf{k} = \cos(\widehat{\mathbf{e}, x})\,\mathbf{i} + \cos(\widehat{\mathbf{e}, y})\,\mathbf{j} + \cos(\widehat{\mathbf{e}, z})\,\mathbf{k}$$

The use of Cartesian projections makes possible the use of formulas and computations instead of geometric constructions, and ordinarily this turns out to be simpler. By way of an illustration, we obtain the condition of the parallelism of two vectors specified by their resolutions:

$$\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}, \quad \mathbf{G} = G_x\mathbf{i} + G_y\mathbf{j} + G_z\mathbf{k}$$

This condition can be written as a vector equation:

$$\mathbf{G} = \lambda\mathbf{F}$$

where λ is a scalar, or in terms of the projections onto the coordinate axes:

$$G_x = \lambda F_x, \quad G_y = \lambda F_y, \quad G_z = \lambda F_z$$

Solving for λ in each equation and then equating the results, we get the required condition:

$$\frac{G_x}{F_x} = \frac{G_y}{F_y} = \frac{G_z}{F_z}$$

This condition consists of two equations, but since a vector is determined by three, then in choosing a vector parallel to the given one, there still remains one degree of freedom, the modulus. (Derive these equations geometrically cn the basis of the properties of similarity of triangles.)

To conclude, let it be emphasized that the vector **a** is completely described by its Cartesian projections a_x, a_y, a_z only when the Cartesian basis **i**, **j**, **k** is fixed. If this basis is chosen in a different manner, that is, if the coordinate axes are rotated, then the same vector will have other projections (coordinates). We will discuss this in more detail in Sec. 9.5.

Exercises

1. Let $\overrightarrow{OA} = \mathbf{a}$, $\overrightarrow{OB} = \mathbf{b}$. Find the vector $\overrightarrow{OC}$, where C is the midpoint of the line segment AB.
2. Under the same conditions as in Exercise 1, find the vector $\overrightarrow{OD}$, where D bisects AB in the ratio $\lambda: 1$.
3. A vector has origin at the point $A(2, 3)$ and terminus at the point $B(-1, 4)$. Resolve this vector in terms of the unit vectors of the coordinate axes.

9.2 The scalar product of vectors

The *scalar product* (also sometimes called *dot product* or *inner product*) of two vectors is a scalar equal to the product of the modulus of one vector by the projection onto it of the other vector. The scalar product of the vectors **a** and **b** is denoted by $\mathbf{a}\cdot\mathbf{b}$ or $(\mathbf{a}, \mathbf{b})$. Thus, $\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}|\,\mathrm{pr}_{\mathbf{a}}\mathbf{b}$ and if we take advantage of formula (2) for computing the projection, we get

$$\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}|\,\mathrm{pr}_{\mathbf{a}}\mathbf{b} = |\mathbf{a}|\,|\mathbf{b}|\cos(\widehat{\mathbf{a}, \mathbf{b}}) \tag{3}$$

From this it is immediately apparent that a scalar product does not depend on the order of the factors: $\mathbf{a}\cdot\mathbf{b} = \mathbf{b}\cdot\mathbf{a}$. Besides, it is evident that for nonzero vectors **a** and **b** the scalar product $\mathbf{a}\cdot\mathbf{b}$ is positive, negative or zero according as the angle between the vectors **a** and **b** is acute, obtuse, or a right angle. A particularly important equation to remember is

$$\mathbf{a}\cdot\mathbf{b} = 0$$

as the necessary and sufficient condition for two vectors **a** and **b** to be perpendicular. A special case is the scalar product of a vector into itself (the *scalar square* of the vector):

$$\mathbf{a}\cdot\mathbf{a} = |\mathbf{a}|\,|\mathbf{a}|\cos 0 = |\mathbf{a}|^2$$

An important example of a scalar product is the work performed by a force **F** over a rectilinear path **S**. If the body is moving in a straight line and the force is directed along this line, the work is equal to the product of the force by the path length. If the direction of the force does not coincide with the direction of motion, then the work is equal to the product of the path into the component of the force acting in the direction of motion, that is,

$$A = |\mathbf{S}|\,F_S = |\mathbf{S}|\,|\mathbf{F}|\cos 0 = \mathbf{S}\cdot\mathbf{F}$$

where θ is the angle between the directions of force and motion. Thus, the work is equal to the scalar product of the force by the path length. The special case of a force acting in the direction of motion is embraced by this formula. Here, if the direction of the force coincides with that of the motion, then $\theta = 0$, $\cos\theta = 1$, $A = |\mathbf{F}|\,|\mathbf{S}|$ and the work is positive. But if the force acts in the reverse direction to the motion, then $\theta = \pi$, $\cos\theta = -1$, $A = -|\mathbf{F}|\,|\mathbf{S}|$, and the work is negative.

Also note the following properties of a scalar product:

$$(\lambda\mathbf{a})\cdot\mathbf{b} = \lambda(\mathbf{a}\cdot\mathbf{b}), \quad (\mathbf{a}+\mathbf{b})\cdot\mathbf{c} = \mathbf{a}\cdot\mathbf{c} + \mathbf{b}\cdot\mathbf{c}$$

We will not dwell on the simple proof of these properties, which stems directly from the definition of a scalar product. The reader can carry out the proof himself. These properties permit using the ordinary rules of algebra when dealing with scalar products, but one should remember that only two vectors can be multiplied together to form a scalar product.

Use is also made of the following formulas, which the reader can verify for himself: if **e**, **e**$'$, **e**$''$ are unit vectors, then

$$\mathbf{a}\cdot\mathbf{e} = \text{pr}_{\mathbf{e}}\mathbf{a}, \quad \mathbf{e}'\cdot\mathbf{e}'' = \cos(\widehat{\mathbf{e}', \mathbf{e}''})$$

Let us now derive an important formula that permits computing a scalar product if we know the projections of the vectors in some Cartesian basis **i**, **j**, **k** (Sec. 9.1):

$$\mathbf{a} = a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k}, \quad \mathbf{b} = b_x\mathbf{i} + b_y\mathbf{j} + b_z\mathbf{k} \qquad (4$$

If one takes note of the fact that

$$\mathbf{i}\cdot\mathbf{i} = \mathbf{j}\cdot\mathbf{j} = \mathbf{k}\cdot\mathbf{k} = 1, \quad \mathbf{i}\cdot\mathbf{j} = \mathbf{j}\cdot\mathbf{k} = \mathbf{k}\cdot\mathbf{i} = 0$$

(why is this?), then, substituting for **a** and **b** their expansions (4), we get

$$\mathbf{a}\cdot\mathbf{b} = a_xb_x + a_yb_y + a_zb_z \qquad (5)$$

(verify this!). As was pointed out at the end of Sec. 9.1, the projec) tions of two vectors **a** and **b** on the coordinate axes generally changes under a rotation of the axes; however, the right part of (5) remains unchanged (invariant), since it is equal to the left side; the definition of the scalar product was given irrespective of the location of the axes.

If in formula (5) we set $\mathbf{b} = \mathbf{a}$, then we have an expression for the square of the modulus of the vector in terms of its Cartesian coordinates:

$$\mathbf{a}\cdot\mathbf{a} = |\mathbf{a}|^2 = a_x^2 + a_y^2 + a_z^2$$

For a vector in a plane we get $|a|^2 = a_x^2 + a_y^2$. This formula is equivalent to the Pythagorean theorem, and the preceding one is equivalent to an analogue of the Pythagorean theorem for space (the square of a diagonal of a rectangular parallelepiped is equal to the sum of the squares of three edges).

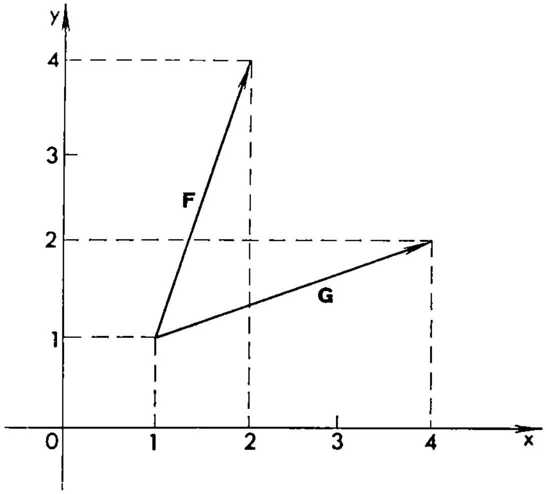
Exercises

1. Two vectors of unit length form an angle $\varphi = 30°$. Find their scalar product.
2. Find the scalar product of the vectors depicted in Fig. 107.
3. Compute the angle θ between the vectors

$$\mathbf{F} = \sqrt{3}\mathbf{i} + \mathbf{j} \text{ and } \mathbf{G} = -\sqrt{3}\mathbf{i} + \mathbf{j}$$

4. Prove that vectors having their origin at the point $A(-1, 1)$ and their termini at the points $B(1, 2)$ and $C(0, -1)$, respectively, are perpendicular.
5. A parallelogram $ABCD$ is constructed on the vectors $\overrightarrow{AB} = \mathbf{F}$ and $\overrightarrow{AD} = \mathbf{G}$. Express the diagonals $\overrightarrow{AC}$ and $\overrightarrow{DB}$ in terms of **F**

Fig. 107



and **G**, consider $\overrightarrow{AC} \cdot \overrightarrow{AC}$ and $\overrightarrow{DB} \cdot \overrightarrow{DB}$, and then prove the theorem: the sum of the squares of the diagonals of the parallelogram is equal to the sum of the squares of all its sides.

6. In a cube of side a find the length of the (inner!) diagonals; the angles between the diagonals; the projections of the sides on the diagonals.
7. A regular tetrahedron with side a is located so that one of its vertices lies at the coordinate origin, another on the positive x-axis, and a third in the first quadrant of the xy-plane. Find the coordinates of all vertices and of the centre of the tetrahedron and also the angle between the straight lines issuing from the centre to the vertices.

9.3. The derivative of a vector

Let us find the derivative of a vector that is dependent on some variable, say the time t, with respect to this variable.

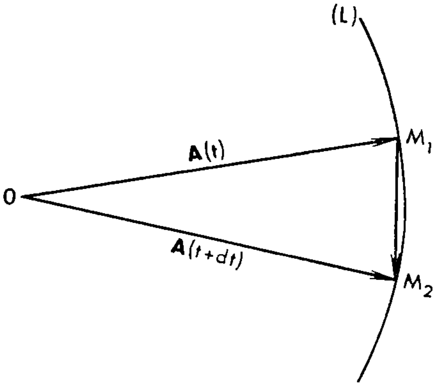
Lay off the vector $\mathbf{A}(t)$ from some point O. Then the terminus of the vector will, as t varies, trace out a line (L) (Fig. 108).

Take a very small dt and form the ratio

$$\frac{\mathbf{A}(t + dt) - \mathbf{A}(t)}{dt}$$

This ratio is also a vector. (The vector $\mathbf{A}(t + dt) - \mathbf{A}(t)$ is shown in Fig. 108 by the line segment M_1M_2.) It is the *derivative* of the vector

Fig. 108



$\mathbf{A}(t)$ with respect to the variable t and for this reason is denoted by $\frac{d\mathbf{A}}{dt}$ so that

$$\frac{d\mathbf{A}}{dt} = \frac{\mathbf{A}(t + dt) - \mathbf{A}(t)}{dt} \tag{6}$$

(To be more exact, we should put the limit sign in the right member as $dt \to 0$.)

Formula (6) can also be rewritten thus:

$$\mathbf{A}(t + dt) = \mathbf{A}(t) + \frac{d\mathbf{A}}{dt}\, dt \tag{7}$$

to within infinitesimals of higher order. As in the case of ordinary functions, we can write the Taylor series

$$\mathbf{A}(t) = \mathbf{A}(t_0) + \mathbf{B}(t_0)\,(t - t_0) + \mathbf{D}(t_0)\,\frac{1}{2}\,(t - t_0)^2 + \ldots$$

where $\mathbf{B}(t_0) = \left.\frac{d\mathbf{A}}{dt}\right|_{t-t_0}$, $\mathbf{D}(t_0) = \left.\frac{d^2\mathbf{A}}{dt^2}\right|_{t-t_0}$ and so forth.

It is clear that for very small dt the points M_2 and M_1 of (L) are very close together, so that the vector $\frac{d\mathbf{A}}{dt}$ is directed along the tangent to it.

If $\mathbf{C}$ is a constant vector, then $\mathbf{C}(t + dt) - \mathbf{C}(t) = \mathbf{0}$, so that in this case $\frac{d\mathbf{C}}{dt} = \mathbf{0}$.

Using the definition of a derivative, it is easy to prove the following two formulas:

1. $\frac{d}{dt}[a_1\mathbf{A}_1(t) + a_2\mathbf{A}_2(t)] = a_1\frac{d\mathbf{A}_1(t)}{dt} + a_2\frac{d\mathbf{A}_2(t)}{dt}$, where a_1, a_2 are constant scalars.

2. $\frac{d}{dt}[f(t)\,\mathbf{A}(t)] = \frac{df}{dt}\mathbf{A}(t) + f(t)\frac{d\mathbf{A}}{dt}$. In particular, $\frac{d}{dt}[f(t)\,\mathbf{C}] = \frac{df}{dt}\mathbf{C}$ if the vector $\mathbf{C}$ is a constant. Thus we see that the derivative of a vector of type $f(t)\mathbf{C}$ is parallel to the vector itself (whereas it is evident from Fig. 108 that this is not so in the general case).

Let us find the derivative of a scalar product. Suppose $\mathbf{A}$ and $\mathbf{B}$ are two variable vectors. By the definition of a derivative,

$$\frac{d}{dt}(\mathbf{A}\cdot\mathbf{B}) = \frac{\mathbf{A}(t+dt)\cdot\mathbf{B}(t+dt) - \mathbf{A}\cdot\mathbf{B}}{dt}$$

Using (7), we get

$$\frac{d}{dt}(\mathbf{A}\cdot\mathbf{B}) = \frac{\left(\mathbf{A}(t) + \frac{d\mathbf{A}}{dt}dt\right)\cdot\left(\mathbf{B}(t) + \frac{d\mathbf{B}}{dt}dt\right) - \mathbf{A}(t)\cdot\mathbf{B}(t)}{dt}$$

or

$$\frac{d}{dt}(\mathbf{A}\cdot\mathbf{B}) = \frac{\mathbf{A}\cdot\mathbf{B} + dt\left(\mathbf{A}\cdot\frac{d\mathbf{B}}{dt}\right) + dt\left(\frac{d\mathbf{A}}{dt}\cdot\mathbf{B}\right) + (dt)^2\left(\frac{d\mathbf{A}}{dt}\cdot\frac{d\mathbf{B}}{dt}\right) - \mathbf{A}\cdot\mathbf{B}}{dt}$$

Neglecting the term containing dt^2, we finally get

$$\frac{d}{dt}(\mathbf{A}\cdot\mathbf{B}) = \mathbf{A}\cdot\frac{d\mathbf{B}}{dt} + \frac{d\mathbf{A}}{dt}\cdot\mathbf{B} \tag{8}$$

Thus, the formula has the same aspect as the formula for the derivative of a product of two scalar functions.

In particular, putting $\mathbf{A} = \mathbf{B}$ in (8), we get

$$\frac{d}{dt}(\mathbf{A}\cdot\mathbf{A}) = \mathbf{A}\cdot\frac{d\mathbf{A}}{dt} + \frac{d\mathbf{A}}{dt}\cdot\mathbf{A} = 2\mathbf{A}\cdot\frac{d\mathbf{A}}{dt} \tag{9}$$

But $\mathbf{A}\cdot\mathbf{A} = |\mathbf{A}|^2$ and so $\frac{d}{dt}(\mathbf{A}\cdot\mathbf{A}) = 2|\mathbf{A}|\frac{d|\mathbf{A}|}{dt}$. Equating these results, we obtain

$$|\mathbf{A}|\frac{d|\mathbf{A}|}{dt} = \mathbf{A}\cdot\frac{d\mathbf{A}}{dt}, \quad \text{that is,} \quad \frac{d|\mathbf{A}|}{dt} = \frac{\mathbf{A}\cdot\frac{d\mathbf{A}}{dt}}{|\mathbf{A}|} \tag{10}$$

From this it is easy to see, in particular, that if the vector $\mathbf{A}(t)$ has a constant modulus and only the direction changes, then the vector $\frac{d\mathbf{A}}{dt}$ is perpendicular to the vector $\mathbf{A}$. Indeed, then $|\mathbf{A}| =$ constant, or $\frac{d|\mathbf{A}|}{dt} = 0$, and by virtue of this last formula, $\mathbf{A}\cdot\frac{d\mathbf{A}}{dt} = 0$, and the fact that the scalar product is equal to zero means that $\mathbf{A}$ and $\frac{d\mathbf{A}}{dt}$ are perpendicular.

Exercise

Find the angle that the helical curve $x = R\cos\omega t$, $y = R\sin\omega t$, $z = vt$ forms with its axis (z-axis).

9.4 The motion of a material point

When a material point (particle) is in motion in space, the relationship between the positions of the point, its velocity and acceleration are relationships between vectors. Suppose the position of the point is described by the vector $\mathbf{r}$ drawn from the origin to this point, the velocity by the vector $\mathbf{u}$, and the acceleration by the vector $\mathbf{a}$. Then from the definition of a derivative we immediately have

$$\mathbf{u} = \frac{d\mathbf{r}}{dt}, \quad \mathbf{a} = \frac{d\mathbf{u}}{dt} = \frac{d^2\mathbf{r}}{dt^2}$$

Let us write down the vector $\mathbf{r}$ in terms of the unit vectors of the coordinate axes:

$$\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$$

Here, x, y, z are the projections, which vary with time as the point moves, and the vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ are constants.

And so

$$\mathbf{u} = \frac{d\mathbf{r}}{dt} = \frac{d}{dt}(x\mathbf{i} + y\mathbf{j} + z\mathbf{k}) = \frac{dx}{dt}\mathbf{i} + \frac{dy}{dt}\mathbf{j} + \frac{dz}{dt}\mathbf{k}$$

Hence, the projections of the velocity vector $\mathbf{u}$ on the coordinate axes are equal to the derivatives of the corresponding projections of the vector $\mathbf{r}$:

$$u_x = \frac{dx}{dt}, \quad u_y = \frac{dy}{dt}, \quad u_z = \frac{dz}{dt}$$

In the same way, for the acceleration we get

$$a_x = \frac{du_x}{dt} = \frac{d^2x}{dt^2}, \quad a_y = \frac{du_y}{dt} = \frac{d^2y}{dt^2}, \quad a_z = \frac{du_z}{dt} = \frac{d^2z}{dt^2}$$

By virtue of (10), the rate of change of the distance $|\mathbf{r}|$ of the point from the origin is

$$\frac{d|\mathbf{r}|}{dt} = \frac{\mathbf{r}\cdot\frac{d\mathbf{r}}{dt}}{|\mathbf{r}|} = \frac{\mathbf{r}\cdot\mathbf{u}}{|\mathbf{r}|}$$

Using this formula it is easy to find, for example, the condition under which the distance of the point from the origin remains unaltered, i.e. $|\mathbf{r}| \equiv$ constant. This is precisely what happens in the following two cases. Firstly, if $\mathbf{u} = \mathbf{0}$ (the point is motionless here), and, secondly, if at each instant of time the velocity $\mathbf{u}$ is perpendicular to the vector

$\mathbf{r}$ (in which case the body is in motion over a sphere of radius $|\mathbf{r}|$ with centre at the origin).

Similarly, from (9) we get

$$\frac{d}{dt}|\mathbf{u}|^2 = 2(\mathbf{a}\cdot\mathbf{u}) \tag{11}$$

From (11) it follows that the velocity is constant in two cases: if the acceleration is zero or if the acceleration is not zero but is perpendicular to the velocity.

The equation of motion of a material point has the form $m\mathbf{a} = \mathbf{F}$, where m is the mass of the point and $\mathbf{F}$ is the force acting on the point. (Newton's second law). Multiply both sides of this equation by the velocity vector $\mathbf{u}$ to get a scalar product:

$$m(\mathbf{a}\cdot\mathbf{u}) = \mathbf{F}\cdot\mathbf{u}$$

Using formula (11), we get

$$\frac{m}{2}\,\frac{d}{dt}|\mathbf{u}|^2 = \mathbf{F}\cdot\mathbf{u}$$

or

$$\frac{d}{dt}\left(\frac{m}{2}|\mathbf{u}|^2\right) = \mathbf{F}\cdot\mathbf{u} \tag{12}$$

We will show that the quantity $\mathbf{F}\cdot\mathbf{u}$ is the power of the force $\mathbf{F}$. Indeed, during the time dt the point moves by the amount $d\mathbf{r}$ and in doing so performs work $\mathbf{F}\cdot d\mathbf{r}$. The ratio of this work to the time dt, equal to

$$\frac{\mathrm{F}\cdot d\mathbf{r}}{dt} = \mathbf{F}\cdot\frac{d\mathbf{r}}{dt} = \mathbf{F}\cdot\mathbf{u}$$

is the work performed in unit time, which is the power.

If we set $\frac{m}{2}|\mathbf{u}|^2 = T$, equation (12) takes the form $\frac{dT}{dt} = \mathbf{F}\cdot\mathbf{u}$, whence $T_2 - T_1 = \int\limits_{t_1}^{t_2}\mathbf{F}\cdot\mathbf{u}\,dt$ (on the right we have the work done by the force $\mathbf{F}$).

Thus, from Newton's second law it follows that there is a definite quantity T expressed in terms of the mass and velocity of the moving point and it is such that the increase in T in the process of motion is exactly equal to the work of the external force $\mathbf{F}$. This is what justifies calling T the *kinetic energy* of a moving material point.

Sometimes, for the parameter defining the position of a point moving along a path one takes the length s of the arc reckoned along the path from some reference point, i.e., we set $\mathbf{r} = \mathbf{r}(s)$. Since the ratio of the length of an infinitesimal arc to the chord subtending the arc tends to unity (by virtue of the fact that the arc hardly has time to change its direction, or to "bend"), it follows that, as $\Delta s \to 0$,

$$\frac{|\Delta \mathbf{r}|}{\Delta s} \to 1, \text{ that is, } \left|\frac{d\mathbf{r}}{ds}\right| = \lim \left|\frac{\Delta \mathbf{r}}{\Delta s}\right| = 1 \qquad (13)$$

Therefore the derivative $d\mathbf{r}/ds$ is a unit vector directed along the tangent to the trajectory. This vector is customarily denoted by $\boldsymbol{\tau}$, whence

$$\mathbf{u} = \frac{d\mathbf{r}}{dt} = \frac{d\mathbf{r}}{ds} \cdot \frac{ds}{dt} = \boldsymbol{\tau} u = u \boldsymbol{\tau} \qquad (14)$$

where $u = |\mathbf{u}|$.

Also from (13) follows an expression for the differential of the arc in Cartesian coordinates:

$$ds = |d\mathbf{r}| = |d(x\mathbf{i} + y\mathbf{j} + z\mathbf{k})| = |dx\mathbf{i} + dy\mathbf{j} + dz\mathbf{k}|$$
$$= \sqrt{dx^2 + dy^2 + dz^2}$$

Since $|\boldsymbol{\tau}(s)| = 1 =$ constant, it follows (see Sec. 9.3) that $\frac{d\boldsymbol{\tau}}{ds} \perp \boldsymbol{\tau}$. Thus the straight line *pp* (Fig. 109) drawn through the running point M of the trajectory (L) parallel to $d\boldsymbol{\tau}/ds$ will serve as the normal to (L), that is, as a perpendicular to the tangent line *ll* at the point M. To distinguish it from the other normals (at every point of the line in space it is possible to draw an infinity of normals that will fill the entire "normal plane"), it is called the *principal normal* to the curve (L) at the point M. The length of the vector $d\boldsymbol{\tau}/ds$ is called the *curvature* of the curve (L) at the point M and is denoted *by* k; that is,

$$\left|\frac{d\boldsymbol{\tau}}{ds}\right| = k, \quad \frac{d\boldsymbol{\tau}}{ds} = k\mathbf{n}$$

where $\mathbf{n}$ is the unit vector of the principal normal (see Fig. 109).

The geometrical meaning of the curvature is seen in Fig. 110:

$$k = \left|\frac{d\boldsymbol{\tau}}{ds}\right| = \lim \left|\frac{\Delta \boldsymbol{\tau}}{\Delta s}\right| = \lim \frac{BC}{\Delta s} = \lim \frac{\alpha}{\Delta s}$$

where α is the angle between the tangents to (L) at close points B and C; in the last passage to the limit we replaced the line segment BC by the arc of a circle of unit radius. Thus, the curvature is the rate of rotation of a tangent per unit length of the path traversed. From this it is evident, for example, that at all points of a circle of radius R the curvature $k = 1/R$ (this is discussed in HM, Sec. 4.5, where the curvature of a plane curve is analyzed). It is also apparent here that

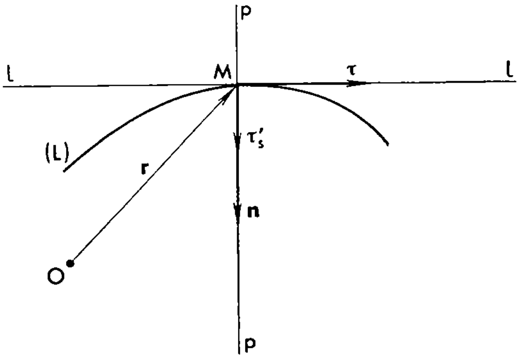


Fig. 109

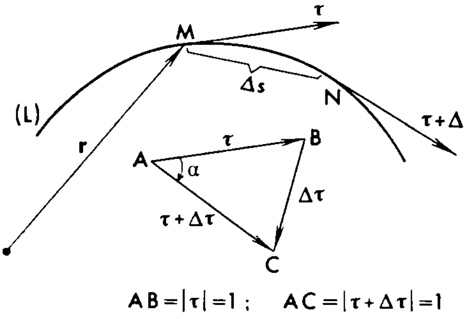


Fig. 110

the vector $\boldsymbol{\tau}'_s = \frac{d\boldsymbol{\tau}}{ds}$ and, with it, **n** as well are in the direction of the bending curve.

Differentiating (14) with respect to *t*, we get

$$\mathbf{a} = \frac{d^2\mathbf{r}}{dt^2} = \frac{du}{dt}\boldsymbol{\tau} + u\frac{d\boldsymbol{\tau}}{dt} = \frac{du}{dt}\boldsymbol{\tau} + u\frac{d\boldsymbol{\tau}}{ds}\,\frac{ds}{dt} = \frac{du}{dt}\boldsymbol{\tau} + u^2k\mathbf{n} \qquad (15)$$

This formula is widely used in physics, since it gives a resolution of the acceleration vector into a "tangential" component (i.e. one along the tangent) and a normal component; the latter, as we see, is directed along the principal normal.

Thus, since the vector **a** is laid off from point *M*, it must necessarily lie in the plane passing through the tangent line and the principal normal drawn at this point. This plane is termed the *osculating plane* to the curve (*L*) at the point *M*. It can be proved that this is nothing but a plane passing through three points of the curve (*L*) that are located infinitely close to *M*. (Just like a tangent is a straight line drawn through two infinitely close points of a curve.)

From formula (15) there follow very important conclusions on the forces acting on a point in the course of motion. Put the expression for **a** taken from (15) into the formula of Newton's second law: $\mathbf{F} = m\mathbf{a}$. We see that the acting force has a tangential component

$$\mathbf{F}_\tau = F_\tau \boldsymbol{\tau} = m \frac{du}{dt} \boldsymbol{\tau} \tag{16}$$

and a normal component

$$\mathbf{F}_n = F_{\mathrm{n}} \mathbf{n} = mu^2 k \mathbf{n} \tag{17}$$

directed along the principal normal to the trajectory. The osculating plane to the trajectory at some point is the plane of the vectors **u** and **F** at this point.

From formula (16) we get

$$uF_\tau = um \frac{du}{dt} = \frac{d}{dt}\left(\frac{mu^2}{2}\right) = \frac{dT}{dt}$$

whence the increment in the kinetic energy is equal to

$$T_2 - T_1 = \int_{t_1}^{t_2} uF_\tau \, dt = \int_{s_1}^{s_2} F_\tau \, ds$$

We see that the work is performed only by the tangential component of the force. The normal component (which is also called the centripetal force) does not alter the velocity of the point but generates a curving of the path, the curvature, by (17), being equal to

$$k = \frac{F_n}{mu^2}$$

Recall that the dimensions of curvature (cm^{-1}) are inverse to the dimensions of length.

Exercise

Find the curvature of the helical curve given in the Exercise to Sec. 9.3.

9.5 Basic facts about tensors

We have already mentioned, at the end of Sec. 9.1, that if we describe a vector by the triad of its Cartesian projections, then we have to bear in mind that this triad is essentially dependent on the choice of the Cartesian basis. Let us now investigate this in more detail. We will denote the vectors of the basis by $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ and the projections of the vector **a** by the symbols a_1, a_2, a_3, so that

$$\mathbf{a} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + a_3\mathbf{e}_3 \tag{18}$$

Now suppose we have chosen a new Cartesian basis $\mathbf{e}'_1$, $\mathbf{e}'_2$, $\mathbf{e}'_3$ that can be expressed in terms of the old basis by the formulas

$$\mathbf{e}'_i = \alpha_{i1}\mathbf{e}_1 + \alpha_{i2}\mathbf{e}_2 + \alpha_{i3}\mathbf{e}_3 = \sum_{j=1}^{3} \alpha_{ij}\mathbf{e}_j \quad (i = 1, 2, 3)$$

In tensor algebra this is customarily written compactly as follows:

$$\mathbf{e}'_i = \alpha_{ij}\mathbf{e}_j \tag{19}$$

with the convention that the summation is over the repeated index in accord with the dimensionality of the space; for three-dimensional space, we have $j = 1, 2, 3$. This is a *dummy* index, which means it can be denoted arbitrarily: $\alpha_{ij}\mathbf{e}_j = \alpha_{ik}\mathbf{e}_k = \alpha_{ii'}\mathbf{e}_{i'}$ and so on.

Forming the scalar product of both sides of (19) by $\mathbf{e}_j$, we get $\alpha_{ij} = \mathbf{e}'_i \cdot \mathbf{e}_j$. Similarly, from the formulas $\mathbf{e}_i = \beta_{ij}\mathbf{e}'_j$, which express the old basis in terms of the new basis, we find that $\beta_{ij} = \mathbf{e}_i \cdot \mathbf{e}'_j$. But this means that $\beta_{ij} = \alpha_{ji}$; the old basis is expressed in terms of the new basis by the formulas

$$\mathbf{e}_i = \alpha_{ji}\mathbf{e}'_j \tag{20}$$

(This equation can be rewritten as $\mathbf{e}_j = \alpha_{ij}\mathbf{e}'_i$ so that in the equation (19) connecting the Cartesian bases we can simply "transpose" the factor α_{ij} from one side to the other.)

Substituting the expressions (20) into (18) and denoting by a'_i the projections of the vector $\mathbf{a}$ in the basis $\mathbf{e}'_i$, we get

$$\mathbf{a} = \sum_i a_i\mathbf{e}_i = \sum_i a_i \sum_j \alpha_{ji}\mathbf{e}'_j = \sum_{i,j} \alpha_{ji}a_i\mathbf{e}'_j$$

Changing the notations $i \leftrightarrow j$, we get

$$\mathbf{a} = \sum_{i,j} \alpha_{ij}a_j\mathbf{e}'_i = \sum_i \left(\sum_j \alpha_{ij}a_j\right)\mathbf{e}'_i = \sum_i a'_i\mathbf{e}'_i$$

whence

$$a'_i = \alpha_{ij}a_j \tag{21}$$

Comparing this with (19), we see that the projections of any vector transform by the same formulas as the basis vectors.

In the foregoing, $\mathbf{a}$ need not be regarded as a "geometric" vector, that is, a directed line segment in a plane or in space. It may be a force vector or a velocity vector, and so forth; in all cases its projections transform via the formulas (21). It is also easy to verify that, conversely, any triad of quantities can be interpreted as a triple of coordinates of some vector, i.e. they can be regarded as defining a vector, only if they acquire specific values after indicating a Cartesian basis and if they transform via the formulas (21) under the change of basis (19). (On the other hand, it is hardly of any use to interpret just *any* triad of quantities as a vector. For example, in the study of a gas flux we can regard the triple of temperature θ, pressure p and density ρ, as characterizing the state of the gas at a point of space. But it is not wise to interpret this triad as a vector, since rotations of the basis in space do not affect the values of the quantities that make up the triad. Three scalar quantities do not constitute a vector).

There are quantities that transform via a more complicated law than (21) under a change of basis (19). An important instance of such quantities is obtained when considering a *linear mapping* of space into itself. We say that there is a mapping T if to every vector $\mathbf{a}$ is associated a vector $T\mathbf{a}$ (of the same space), and to the sum of the inverse images corresponds the sum of the images, or $T(\mathbf{a} + \mathbf{b}) = T\mathbf{a} + T\mathbf{b}$, $T(\lambda\mathbf{a}) = \lambda T\mathbf{a}$. Instances of linear mappings are: a rotation of space about a point, the uniform contraction of space to a plane, a straight line, or a point, and so forth (think this through!)

To describe such a mapping by numbers, choose a basis $\mathbf{e}_i$ in space and expand each of the vectors $T\mathbf{e}_i$ in terms of this basis:

$$T\mathbf{e}_i = p_{ji}\mathbf{e}_j \tag{22}$$

(note the order of the indices). A knowledge of the nine coefficients p_{ij} permits finding the image $\mathbf{b} = T\mathbf{a}$ of any vector (18). Indeed,

$$\mathbf{b} = \sum_i b_i\mathbf{e}_i = T(\sum_i a_i\mathbf{e}_i) = \sum_{i,\,j} a_i p_{ji}\mathbf{e}_j = \sum_{i,\,j} a_j p_{ij}\mathbf{e}_i$$

whence $b_i = p_{ij}a_j$.

The set of coefficients p_{ij} depending on two indices is frequently written in the form of a *matrix*;

$$(p_{ij}) = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}$$

Now let us see how the coefficients p_{ij} change under a transition to a new basis $\mathbf{e}'_i$. Since by virtue of (22) $p_{ij} = T\mathbf{e}_j \cdot \mathbf{e}_i$, we get, taking into account formula (19),

$$p'_{ij} = T\mathbf{e}'_j \cdot \mathbf{e}'_i = \alpha_{jl}T\mathbf{e}_l \cdot \alpha_{ik}\mathbf{e}_k = \alpha_{ik}\alpha_{jl}p_{kl} \tag{23}$$

From this it can be shown that the sum of p_{ii} remains invariant under such a change, which means it is a scalar. Indeed,

$$p'_{ii} = \alpha_{ik}\alpha_{il}p_{kl} = \mathbf{e}'_i \cdot \mathbf{e}_k\alpha_{il}p_{kl} = \mathbf{e}_k \cdot (\mathbf{e}'_i\alpha_{il})\,p_{kl} = \mathbf{e}_k \cdot \mathbf{e}_l p_{kl} = p_{ii}$$

In the example given here the quantities p_{ij} were nondimensional. Conversely, it can be verified that any set of nondimensional quantities p_{ij} that transform via the formulas (23) under the change of Cartesian basis (19) can be interpreted as the matrix of a linear mapping of space into itself. In Sec. 11.5 we will give examples of dimensional quantities with a different physical meaning, but such that also transform via the formulas (23) under a change of basis.

Let us rewrite formulas (21) and (23) and change the indices:

$$a'_{i'} = \alpha_{i'i}a_i, \quad p'_{i'j'} = \alpha_{i'i}\alpha_{j'j}p_{ij}$$

We see that these formulas have a common structure. Generalizing, we arrive at a law of transformation of quantities $q_{ij\ldots l}$, that depend on m indices $i, j, \ldots, l$, each of which takes on the values 1, 2, 3 (for three-dimensional space) or 1, 2 (for a plane, or two-dimensional space):

$$q_{i'j'\ldots l'} = \alpha_{i'i}\alpha_{j'j}\ldots, \alpha_{l'l}q_{ij\ldots l} \tag{24}$$

If the quantities $q_{ij\ldots l}$ assume definite values only after indication of the Cartesian basis and if they transform via the formulas (24) (on the right we have an m-fold sum!) under the change of basis (19), then we say that the quantities $q_{ij\ldots l}$ constitute a *tensor of rank m*. Thus, the set of projections of a vector form a tensor of rank one, the set of coefficients of a linear mapping, a tensor of rank two (in every basis this set of coefficients is specific, but the set defines one and the same mapping irrespective of the choice of basis). Note that from the viewpoint of this definition, a tensor of rank zero is to be regarded as a quantity that takes on a numerical value in the chosen units of measurement and that is invariant under a change of basis; in other words, it is a scalar.

We can also approach the tensor concept differently. For the sake of definiteness, consider a second-rank tensor, that is, a set of quantities p_{ij} that transform via formulas (23) under a change of basis (19). With this set let us associate the following formal expression:

$$\mathbf{P} = \sum_{i,j=1}^{3} p_{ij}\mathbf{e}_i\mathbf{e}_j \tag{25}$$

Here, the *tensor product* $\mathbf{e}_i\mathbf{e}_j$ (not to be confused with the scalar product $\mathbf{e}_i \cdot \mathbf{e}_j$!) does not reduce to any other simpler object like a vector or a scalar. The factors in a tensor product cannot be interchanged: $\mathbf{e}_i\mathbf{e}_j \neq \mathbf{e}_j\mathbf{e}_i$, but in a tensor product of sums the brackets can be removed by the ordinary rules, so long as we keep to the order of the factors, for example:

$$(2\mathbf{e}_1 + \mathbf{e}_2 - \mathbf{e}_3)\cdot(\mathbf{e}_1 + 3\mathbf{e}_3) = 2\mathbf{e}_1\mathbf{e}_1 + 6\mathbf{e}_1\mathbf{e}_3 + \mathbf{e}_2\mathbf{e}_1 + 3\mathbf{e}_2\mathbf{e}_3 - \mathbf{e}_3\mathbf{e}_1 - 3\mathbf{e}_3\mathbf{e}_3$$

In the new basis $\mathbf{e}'_i$ we have to write

$$\mathbf{P}' = \sum_{i,j} p'_{ij}\mathbf{e}'_i\mathbf{e}'_j$$

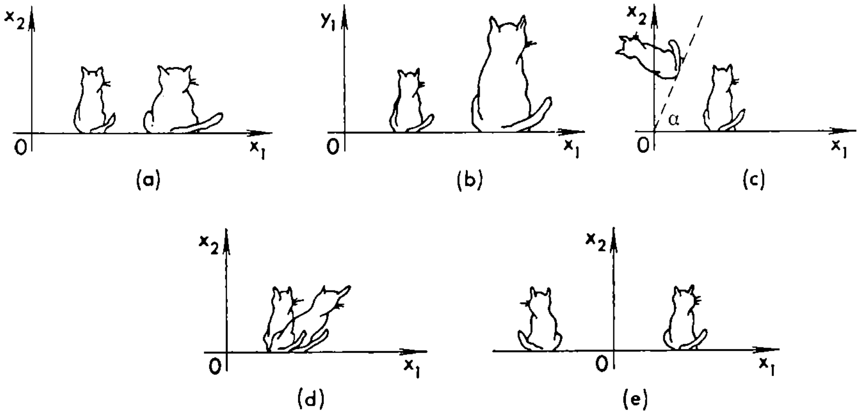


Fig. 111

instead of (25). However, taking into account the formulas (23) and (20), we get

$$\mathbf{P}' = \sum_{i,j,k,l} \alpha_{ik}\alpha_{jl}p_{kl}\mathbf{e}'_i\mathbf{e}'_j = \sum_{kl} p_{kl}\left(\sum_i \alpha_{ik}\mathbf{e}'_i\right)\left(\sum_j \alpha_{jl}\mathbf{e}'_k\right) = \sum_{k,l} p_{kl}\mathbf{e}_k\mathbf{e}_l = \mathbf{P}$$

Thus, expression (25), which is called a tensor, is invariant under a change of the Cartesian basis. In similar fashion we introduce a tensor of any rank, with the first-rank tensor $\sum_i a_i\mathbf{e}_i$ being a vector.

Tensors, tensor algebra, and tensor analysis (which handles the rules for differentiating tensors) play a very important role in modern physics. The basic laws of the theory of elasticity, electrodynamics, optics of an anisotropic medium, and so forth are all stated in tensorial terms. We are not forced to associate the consideration of a problem with any one artificially chosen system of coordinates that is not justified by the essence of the investigation. It is to be observed that in many cases one cannot limit oneself to Cartesian bases, in which case it becomes necessary to alter the definition of a tensor so that it remains invariant under a change to any basis (not necessarily a Cartesian basis). We will not dwell on this altered definition here.

Albert Einstein applied tensor analysis to electrodynamics and the theory of gravitation, thus attracting the attention of physicists and mathematicians to this field. In fact, the best brief exposition of the theory of tensors is given in Einstein's book *The Meaning of Relativity* [4].

Exercise

Write the matrices of the linear mappings of a plane into itself given in Fig. 111. Indicate into what the square is carried for different directions of its sides.

9.6 Multidimensional vector space

In Sec. 4.8 we saw that the concept of a multidimensional space can be approached in two ways: either by proceeding from a numerical scheme or by considering a system with many degrees of freedom. Analogous approaches are possible to the idea of a multidimensional vector space, which is distinguished from general spaces by the possibility of performing linear operations. Let us consider the first approach: for the sake of definiteness we will discuss a four-dimensional space (the general case is considered analogously).

As we have seen, a point in four-dimensional space may be any set of four numbers (x_1, x_2, x_3, x_4). The unit vector $\mathbf{e}_1$ of the first coordinate axis (the x_1-axis) is to be pictured as a line segment issuing from the coordinate origin (0,0,0,0) and with terminus at the point (1,0,0,0). Incidentally, because of the possibility of translating any vector parallel to itself, $\mathbf{e}_1$ may be regarded as a line segment with its origin at any point (x_1, x_2, x_3, x_4) and its terminus at the point $(x_1 + 1, x_2, x_3, x_4)$. The unit vectors $\mathbf{e}_2$, $\mathbf{e}_3$, $\mathbf{e}_4$ of the other three coordinate axes are introduced similarly. By considering a linear combination of these four vectors,

$$\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + x_3\mathbf{e}_3 + x_4\mathbf{e}_4 \tag{26}$$

we obtain a vector that issues from the coordinate origin and extends to the point (x_1, x_2, x_3, x_4), that is, the radius vector of that point. The vector $\mathbf{x}$ may be laid off from any point other than the origin of coordinates. In any case, the coefficients x_1, x_2, x_3, x_4 in formula (26) are the projections of $\mathbf{x}$ on the coordinate axes.

The operations involving four-dimensional vectors are performed by the same formal rules that govern two-dimensional and three-dimensional vectors. The simplest thing is to consider vectors resolved along the coordinate axes, that is, as represented in the form (26). If besides the vector (26) we have a vector $\mathbf{y} = y_1\mathbf{e}_1 + y_2\mathbf{e}_2 + y_3\mathbf{e}_3 + y_4\mathbf{e}_4$, then

$$\begin{gathered}\mathbf{x} + \mathbf{y} = (x_1 + y_1)\,\mathbf{e}_1 + (x_2 + y_2)\,\mathbf{e}_2 + (x_3 + y_3)\,\mathbf{e}_3 + (x_4 + y_4)\,\mathbf{e}_4,\\ \lambda\mathbf{x} = \lambda x_1\mathbf{e}_1 + \lambda x_2\mathbf{e}_2 + \lambda x_3\mathbf{e}_3 + \lambda x_4\mathbf{e}_4 \quad (\lambda \text{ a scalar}),\\ \mathbf{x}\cdot\mathbf{y} = x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4\end{gathered} \tag{27}$$

It can be verified that all the basic properties described above for two-dimensional and three-dimensional vectors hold true also for four-dimensional vectors. However, the properties associated with the idea of linear dependence (Sec. 9.1) require special attention. The reason is that four-dimensional space has planes of two kinds: two-dimensional and three-dimensional planes. The two-dimensional plane results if we take two linearly independent (i.e. nonparallel) vectors and then lay off all possible linear combinations of them from some fixed point, say the coordinate origin. If this is done with three linearly independent vectors, then we get a three-dimensional plane (also spoken of as a "three-dimensional hyperplane") in four-dimensional space. Now if we take four linearly independent vectors (that is, such as are not parallel to a single three-dimensional plane), then their linear combinations will fill all the space, which means that any fifth vector can be resolved in terms of these four. Which means that in four-dimensional space, any four linearly independent vectors can be taken as a basis (cf. Sec.9.1).

We now consider an alternative approach to the concept of a multidimensional vector space. We proceed from some collection R of arbitrary entities, upon which we can perform linear operations without going outside R (linear operations, it will be recalled, involve adding and multiplying by a scalar). Besides this it is also required that the very same properties indicated in Sec. 9.1 for linear operations on vectors in a plane and in space hold true. Then R is called a *linear* (or *vector*) *space*, and the component entities are termed *(generalized) vectors*.

The basic numerical characteristic of a specified vector space is the maximum possible number of linearly independent vectors, which is also called the *dimensionality* of the space. For example, if R is four-dimensional, this means that in it we can indicate four linearly independent vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, $\mathbf{a}_4$ such that any other vector $\mathbf{x}$ in R can be represented as

$$\mathbf{x} = \lambda_1\mathbf{a}_1 + \lambda_2\mathbf{a}_2 + \lambda_3\mathbf{a}_3 + \lambda_4\mathbf{a}_4 \qquad (28)$$

We can thus take these vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, $\mathbf{a}_4$ in R for a basis. Since the coefficients of the expansion in (28) can assume all possible values for distinct vectors $\mathbf{x}$, it follows that in choosing $\mathbf{x}$ in R there are four degrees of freedom, which means the space R is four-dimensional also in the sense of Sec. 4.8.

It may happen that in a linear space it is possible to indicate an arbitrarily large number of linearly independent vectors. Such a space is said to be *infinite-dimensional*. An example of an infinite-dimensional space will be examined in Sec. 14.7.

If in a finite-dimensional linear space we introduce a scalar product that satisfies the properties described in Sec. 9.2 for an ordinary scalar product, then this space is said to be *Euclidean*. In Euclidean space it is natural to introduce the following notions: the modulus of a vector (via the formula $|\mathbf{x}| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$), the unit vector, and the orthogonality of vectors (if $\mathbf{x} \cdot \mathbf{y} = 0$). It is customary, in Euclidean space, to choose an orthogonal basis and not just any basis. Thus, for example, for the four-dimensional case we have a collection of four mutually orthogonal vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, $\mathbf{a}_4$. In this case, the coefficients of the expansion (28) of any vector $\mathbf{x}$ are readily found in the following manner: form the scalar product of both sides of (28) by the vectors $\mathbf{a}_j$ to get

$$\mathbf{x} \cdot \mathbf{a}_j = \lambda_j(\mathbf{a}_j \cdot \mathbf{a}_j)$$

The remaining terms on the right side drop out because of the condition of orthogonality. And so

$$\lambda_j = \frac{\mathbf{x} \cdot \mathbf{a}_j}{\mathbf{a}_j \cdot \mathbf{a}_j} \quad (j = 1, 2, 3, 4) \tag{29}$$

The existence of a scalar product makes possible rotations in linear space, which then acquires the property of *isotropy* (the properties of the space are the same in all directions). The vector space thus becomes more valuable. Recall the characteristic, mentioned in Sec. 9.5, of the state of a gas at a point given by the quantities θ, p, ρ. Increments of these quantities can be subjected to linear operations, which means triples of such increments form a three-dimensional linear space. But this space is devoid of a scalar product and of rotations, which means it is not Euclidean and, hence, deficient in a sense.*

There are also what are known as *pseudo-Euclidean* (or quasi-Euclidean) spaces in which the square of the modulus of the vector (the scalar product of the vector into itself) can be positive, negative, or zero. Such is (by virtue of the theory of relativity) the four-dimensional space x, y, z, t of Cartesian coordinates and time. Rotations are possible in pseudo-Euclidean space, although not all directions in it are of equal status.

Up to now we assumed scalars to be arbitrary real numbers, in which case the linear space is called a *real linear space.* But we can also consider *complex linear spaces* in which the scalars are arbitrary complex numbers. Then a new factor appears in defining a Euclidean space: it is required that

$$(\mathbf{y}, \mathbf{x}) = (\mathbf{x}, \mathbf{y})^* \tag{30}$$

That is, when the factors in a scalar product are interchanged, the product is replaced by the conjugate complex number. However, if in (30) we set $\mathbf{y} = \mathbf{x}$, then we see that the scalar square of any vector is a real number (and, as may be verified, it is even positive, so that the modulus of the vector may be found from the same formula as before, and this modulus turns out to be positive). Besides, when taking the scalar factor outside the sign of the scalar product, we make use of the following formulas:

$$(\lambda\mathbf{x}, \mathbf{y}) = \lambda(\mathbf{x}, \mathbf{y}), \quad (\mathbf{x}, \lambda\mathbf{y}) = \lambda^*(\mathbf{x}, \mathbf{y})$$

The simplest instance of a four-dimensional complex Euclidean space is the space considered at the beginning of this section, but in this instance the numbers $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$ must be complex numbers, and instead of formula (27) we have to use the formula

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1^* + x_2 y_2^* + x_3 y_3^* + x_4 y_4^*$$

* Do not confuse this "non-Euclidicity" with the so-called non-Euclidean geometry, which has nothing to do with the theory of linear spaces.

Exercises

1. Give the formula for the length of the vector (26).
2. Verify the fact that the radius vectors of the points $(1, 1, 1, -1)$, $(1, 1, -1, 1)$ and $(0, 0, 1, 1)$ are mutually perpendicular. Construct the fourth (nonzero) vector perpendicular to these three.

ANSWERS AND SOLUTIONS

Sec. 9.1

1. $\frac{\mathbf{a}+\mathbf{b}}{2}$. 2. $\frac{\mathbf{a}+\lambda\mathbf{b}}{1+\lambda}$ 3. $\overrightarrow{AB} = -3\mathbf{i} + \mathbf{k}$.

Sec. 9.2

1. $\frac{\sqrt{3}}{2}$.
2. From the figure it is evident that $F_x = 1, F_y = 3, G_x = 3, G_y = 1$. For this reason $\mathbf{F}\cdot\mathbf{G} = 6$.
3. $\cos\theta = -1/2$, $\theta = 120°$.
4. $\overrightarrow{AB} = 2\mathbf{i} + \mathbf{j}$, $\overrightarrow{AC} = \mathbf{i} - 2\mathbf{j}$ and so $\overrightarrow{AB}\cdot\overrightarrow{AC} = 2 - 2 = 0$.
5. $\overrightarrow{AC} = \mathbf{F} + \mathbf{G}$, $\overrightarrow{DB} = \mathbf{F} - \mathbf{G}$, whence

$$AC^2 + DB^2 = (\mathbf{F}+\mathbf{G})\cdot(\mathbf{F}+\mathbf{G}) + (\mathbf{F}-\mathbf{G})\cdot(\mathbf{F}-\mathbf{G})$$
$$= 2\mathbf{F}\cdot\mathbf{F} + 2\mathbf{G}\cdot\mathbf{G} = AB^2 + BC^2 + CD^2 + DA^2$$

6. Locate the axes as in Fig. 112. Then $\overrightarrow{OA} = a\mathbf{i}, \overrightarrow{OB} = a\mathbf{j}, \overrightarrow{OC} = a\mathbf{k}$. The lengths of all diagonals are the same; for instance $\overrightarrow{OD} = a\mathbf{i} + a\mathbf{j} + a\mathbf{k}$, $OD = \sqrt{a^2 + a^2 + a^2} = 1.73a$. The angles between diagonals are also the same; for example, the angle between $\overrightarrow{OD}$ and $\overrightarrow{CE} = a\mathbf{i} + a\mathbf{j} - a\mathbf{k}$ is computed from the formula

$$\cos\Theta = \frac{\overrightarrow{OD}\cdot\overrightarrow{CE}}{OD\cdot CE} = \frac{a^2}{\sqrt{3}a\cdot\sqrt{3}a} = \frac{1}{3}, \text{ that is, } \Theta = 70°32'.$$

And the projections of the sides on the diagonals are the same; for example, the projection of side OA on the diagonal OD is $\frac{\overrightarrow{OA}\cdot\overrightarrow{OD}}{OD} = \frac{a^2}{\sqrt{3}a} = \frac{a}{\sqrt{3}} = 0.58a$.

7. Label the points given in the statement of the problem as in Fig. 113. It is clear that O has the coordinates $(0, 0, 0)$ and $A(a, 0, 0)$. Let B have the coordinates $(x_B, y_B, 0)$ and $C(x_C, y_C, z_C)$. Then $OB = a$, $\overrightarrow{OB}\cdot\overrightarrow{OA} = a\cdot a\cdot\cos 60° = \frac{a^2}{2}$, that is, $x_B^2 + y_B^2 = a^2$, $x_B a = \frac{a^2}{2}$, whence $x_B = \frac{a}{2}, y_B = \frac{a\sqrt{3}}{2}$. Further-

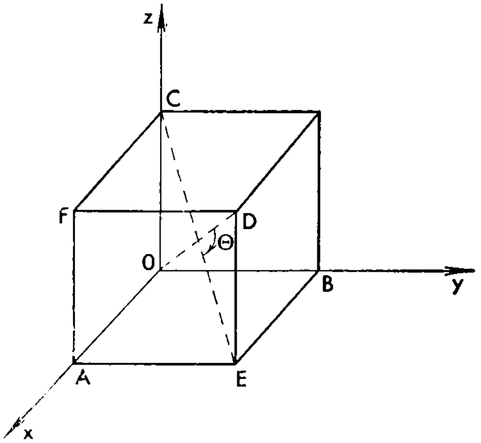


Fig. 112

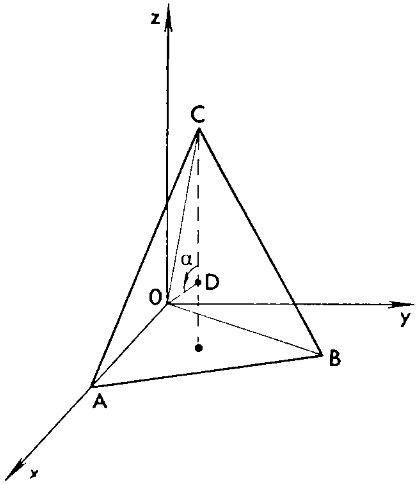


Fig. 113

more, $OC = a$, $\overrightarrow{OC}\cdot\overrightarrow{OA} = \frac{a^2}{2}$, $\overrightarrow{OC}\cdot\overrightarrow{OB} = \frac{a^2}{2}$, that is, $x_C^2+y_C^2+ z_C^2 = a^2$, $x_C a = \frac{a^2}{2}$, $x_C\frac{a}{2}+y_C\frac{a\sqrt{3}}{2}=\frac{a^2}{2}$, whence $x_C=\frac{a}{2}$, $y_C = \frac{a}{2\sqrt{3}}$, $z_C=\sqrt{\frac{2}{3}}\,a$.

Since point D, by symmetry, is located exactly under C, we denote its coordinates by $\left(\frac{a}{2}, \frac{a}{2\sqrt{3}}, z_D\right)$. Equating the lengths of the vectors $\overrightarrow{DC}$ and $\overrightarrow{DO}$, we get $\left(\sqrt{\frac{2}{3}}\,a - z_D\right)^2 = \left(\frac{a}{2}\right)^2 +$

$+\left(\frac{a}{2\sqrt{3}}\right)^2 + z_D^2$, whence $z_D = \frac{a}{2\sqrt{6}}\left(= \frac{z_C}{4}\right)$. Finally, since $\overrightarrow{DC} \parallel \mathbf{k}$ and $\overrightarrow{DO} = -\frac{a}{2}\mathbf{i} - \frac{a}{2\sqrt{3}}\mathbf{j} - \frac{a}{2\sqrt{6}}\mathbf{k}$, the desired angle may be found from the formula

$$\cos\alpha = \frac{\mathbf{k}\cdot\overrightarrow{DO}}{|\mathbf{k}|\cdot DO} = \frac{-\frac{a}{2\sqrt{6}}}{\sqrt{\frac{a^2}{4}+\frac{a^2}{12}+\frac{a^2}{24}}} = -\frac{1}{3}$$

whence $\alpha = 109°28'$.

Sec. 9.3

The radius vector of the point of the helical curve is $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k} = R\cos\omega t\,\mathbf{i} + R\sin\omega t\,\mathbf{j} + vt\mathbf{k}$. From this we have

$$\frac{d\mathbf{r}}{dt} = -R\omega\sin\omega t\,\mathbf{i} + R\omega\cos\omega t\,\mathbf{j} + v\mathbf{k} \qquad (31)$$

Since this vector is directed along the tangent line, it remains to find its angle θ with the vector $\mathbf{k}$:

$$\cos\theta = \frac{\frac{d\mathbf{r}}{dt}\cdot\mathbf{k}}{\left|\frac{d\mathbf{r}}{dt}\right|\cdot|\mathbf{k}|} = \frac{v}{\sqrt{R^2\omega^2+v^2}}$$

Sec. 9.4

By virtue of (31), $ds = |d\mathbf{r}| = |-R\omega\sin\omega t\,\mathbf{i} + R\omega\cos\omega t\,\mathbf{j} + v\mathbf{k}|\,dt = \sqrt{R^2\omega^2+v^2}\,dt$. And so

$$\boldsymbol{\tau} = \frac{d\mathbf{r}}{ds} = \frac{-R\omega\sin\omega t\,\mathbf{i} + R\omega\cos\omega t\,\mathbf{j} + v\mathbf{k}}{\sqrt{R^2\omega^2+v^2}},$$

$$k = \left|\frac{d\boldsymbol{\tau}}{ds}\right| = \left|\frac{d\boldsymbol{\tau}}{dt}\,\frac{dt}{ds}\right| = \frac{R\omega^2}{R^2\omega^2+v^2}$$

Sec. 9.5

(a) $\begin{pmatrix} k & 0 \\ 0 & 1 \end{pmatrix}$, (b) $\begin{pmatrix} k & 0 \\ 0 & k \end{pmatrix}$, (c) $\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$,

(d) $\begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}$, (e) $\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$.

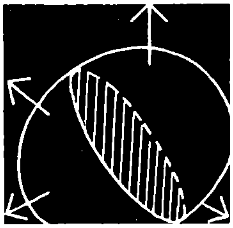
In examples (b), (c) and (e) every square goes into a square. In examples (a) and (d) the square generally yields an oblique-angled parallelogram. If the sides of the square are parallel to the coordinate axes in the case (a) or have the slope $\frac{k}{2} \pm \sqrt{\frac{k^2}{4}+1}$ in the case (d), then we get a rectangle. If the square is rotated from the indicated position through 45°, the result is a rhombus.

Sec. 9.6

1. $\sqrt{x_1^2+x_2^2+x_3^2+x_4^2}$.
2. Perpendicularity is verified by means of the formula (27) since the scalar product of any pair of vectors is equal to zero. If the fourth vector has the projections x_1, x_2, x_3, x_4, then from the perpendicularity condition we get $x_1+x_2+x_3-x_4=0$, $x_1+x_2-x_3+$ $+x_4=0$, $x_3+x_4=0$, whence it is easy to derive that $x_3=$ $=x_4=0$, $x_1+x_2=0$. Thus, we can take the projections of the fourth vector to be equal to 1, -1, 0, 0.

Chapter 10

FIELD THEORY [*]



10.1 Introduction

We say that a *field* of some quantity has been specified in space if the value of the quantity is stated at every point in the space (or in some *region* of it). For example, when studying a gas flow, one has to investigate the temperature field (at every point the temperature has a definite value), the field of densities, the field of pressures, the field of velocities, and so forth. We have a *scalar* field or a *vector* field depending on the character of the quantity: for instance, fields of temperatures, pressures or densities are scalar fields, whereas fields of velocities and forces are vector fields. A field is *stationary* (or *steady-state*) if it does not vary with time at every point of the space, or it is *nonstationary* (*nonsteady-state*) if it does vary.

For the sake of definiteness let us take a scalar field and denote it by u; also, let us introduce a Cartesian system of coordinates x, y, z. Specifying these coordinates determines a point in space and thus the corresponding value of $u = u(x, y, z)$. (If the field is nonstationary, then $u = u(x, y, z, t)$, where t is the time. Here however we do not regard time as a fourth coordinate of equal status, but rather as a sort of accessory parameter so that subsequent constructions will refer to any fixed instant of time.) Thus, from a purely formal point of view, a stationary field is merely a function of three variables x, y, z. We must bear in mind, however, that the coordinates can be introduced in space in different ways, and this will cause the expression $u(x, y, z)$ to change. But in any given point M the value of u will of course be independent of any choice of the coordinate system. For this reason, we often say that u is a *point function*, $u = u(M)$, since specifying M fully determines the appropriate value of u, that is, the value of u at the point M. When considering a field, a point function is primary relative to a function of the coordinates for the reason that a field is meaningful and can be investigated irrespective of any system of coordinates.

[*] This chapter is a direct continuation of the preceding one and draws on the material of Secs. 9.1 and 9.2. Use will also be made of the concept of a multiple integral (Sec. 4.7) and Green's function (Sec. 6.2).

If the quantity under study is specified in a plane, then the corresponding field is said to be a *plane field*; such fields occur in the study of heat processes in a plate (lamina) whose width is disregarded.

If the field of u is a spatial field, but proves to be independent of z in the Cartesian coordinate system x, y, z, the field is said to be *plane-parallel*. It is then often possible to disregard the z-coordinate by considering the field in the xy-plane (i.e. to consider a plane field instead of a plane-parallel one) and bearing in mind that the field in all parallel planes has exactly the same form, that is, all quantities describing the field are constant on all perpendiculars to these planes.

Exercise

Let the quantity u have the expression $u = x^2 + y^2 + 2yz + z^2$ in a Cartesian system of coordinates. Prove that this quantity forms a plane-parallel field.

Hint. Turn the system of coordinates through 45° about the x-axis.

10.2 Scalar field and gradient

For the sake of simplicity, we will assume an xyz-coordinate system (Cartesian) in space and will consider the stationary scalar field $u = u(x, y, z)$. Let there be a point M from which it is possible to issue in all directions; Fig. 114 shows one such direction l. The derivative of u in the direction l is the rate of change of the field in that direction per unit length:

$$\frac{du}{dl} = \lim_{N\to M} \frac{u(N) - u(M)}{MN} = \lim_{\Delta s\to 0} \frac{\Delta u}{\Delta s}\Big|_l$$

Reasoning as we did in the derivation of formula (5), Ch. 4, we get

$$\frac{du}{dl} = \frac{\partial u}{\partial x}\frac{dx}{dl} + \frac{\partial u}{\partial y}\frac{dy}{dl} + \frac{\partial u}{\partial z}\frac{dz}{dl}$$

The right-hand side is conveniently represented as a scalar product of two vectors (see formula (5) of Ch. 9):

$$\frac{du}{dl} = \left(\frac{\partial u}{\partial x}\mathbf{i} + \frac{\partial u}{\partial y}\mathbf{j} + \frac{\partial u}{\partial z}\mathbf{k}\right)\cdot\left(\frac{dx}{dl}\mathbf{i} + \frac{dy}{dl}\mathbf{j} + \frac{dz}{dl}\mathbf{k}\right)$$

The first one is called the *gradient* of the field u and is denoted by

$$\operatorname{grad} u = \frac{\partial u}{\partial x}\mathbf{i} + \frac{\partial u}{\partial y}\mathbf{j} + \frac{\partial u}{\partial z}\mathbf{k} \tag{1}$$

Its physical meaning will be explained later on. The second vector

$$\frac{dx}{dl}\mathbf{i} + \frac{dy}{dl}\mathbf{j} + \frac{dz}{dl}\mathbf{k} = \frac{d(x\mathbf{i} + y\mathbf{j} + z\mathbf{k})}{dl} = \frac{d\mathbf{r}}{dl} = \boldsymbol{\tau}$$

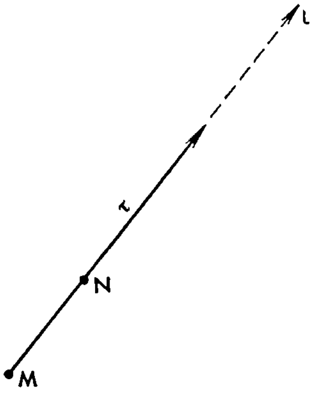


Fig. 114

is the unit vector of the direction l (see Sec. 9.4). Thus

$$\frac{du}{dl} = \text{grad } u \cdot \tau \tag{2}$$

The first factor on the right, for a given field u, depends only on the choice of the point M. The second factor depends only on the direction l.

Since the scalar product of a vector by the unit vector is merely equal to the projection of the first vector on the second, formula (2) can be rewritten thus:

$$\frac{du}{dl} = \text{grad}_l\, u \tag{3}$$

(this is the notation for the projection of a gradient on the direction l).

Suppose we have a field u and a point M. We ask: Along what direction l will the derivative du/dl be greatest? According to (3), this question reduces to the following one: On what direction is the projection of the vector grad u the largest? It is clear that any vector projected on different directions yields the larger projection, equal to its length, when projected on its own direction. Thus, the vector grad u at the point M indicates the direction of fastest buildup in the field u, and this fastest rate, referred to unit length, is equal to $|\text{grad } u|$; the faster the field varies, the longer grad u is. Fig. 115 shows the vectors of the temperature gradient at different points of a heat conductor heated from within (hatched zone) and cooled from without. The temperature gradient is directed towards the "heater".

The resulting physical meaning of the gradient also shows that the gradient is invariantly related to the field at hand, that is, it remains unchanged (invariant) under a change of the Cartesian axes (this was not apparent from the definition (1) that was given in "non-variant" form, i.e. associated with a specific system of coordinates).

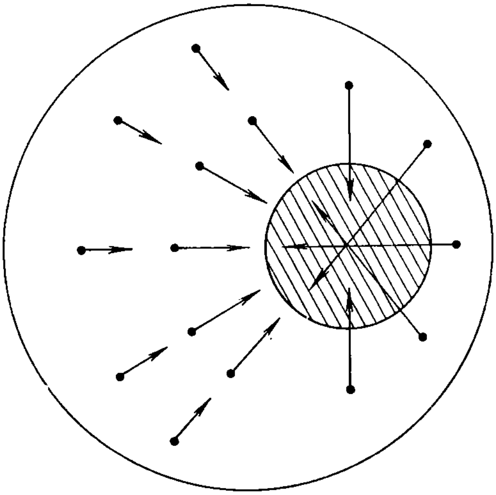


Fig. 115

What is more, if the field u is given, then at every point of space it is possible to find the direction and rate of fastest buildup in the field u; thus we can find the vector grad u without resorting to coordinates and to specifying $u(x, y, z)$. To summarize, the gradient of a scalar field forms a very definite vector field.

The same requirement of invariance is demanded of all the other basic notions in field theory. As we have already explained in Sec. 9.5, when the axes of a Cartesian coordinate system are changed, the vectors remain unchanged (invariant) but their projections change. Thus, if some concept in the theory of a vector field is formulated with the aid of coordinates and projections of that field, this concept must satisfy the requirement of invariance relative to any change of the coordinates and their projections under a rotation of the coordinate axes. Such will be the case, in particular, if the statement is expressed in terms of scalars, vectors or tensors (Sec. 9.5).

Note that the derivatives u'_x, u'_y, u'_z are also directional derivatives: for example, u'_x is a derivative in the direction of the x-axis.

The gradient of the field $u(x, y, z)$ is closely related to the *level surfaces* of the field, that is, the surfaces on which the field has a constant value, $u(x, y, z) =$ constant. Depending on the physical meaning of the field they are called isothermic, isobaric, etc. surfaces. Namely, at every point M the gradient of the field is normal (that is, perpendicular to the tangent plane) to the level surface passing through M. Indeed (Fig. 116), if ΔC is small, then the surfaces $u = C$ and $u = C + \Delta C$ near M may be regarded as nearly plane and

$$\frac{du}{dl} \approx \frac{\Delta u}{\Delta l} = \frac{\Delta C}{\Delta l} \tag{4}$$

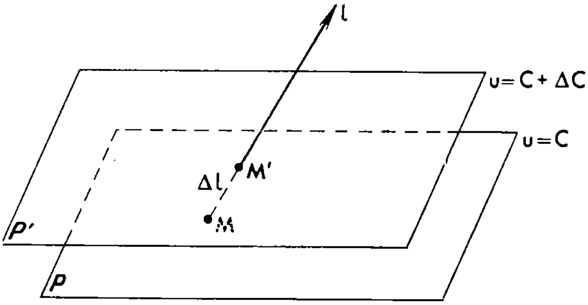


Fig. 116

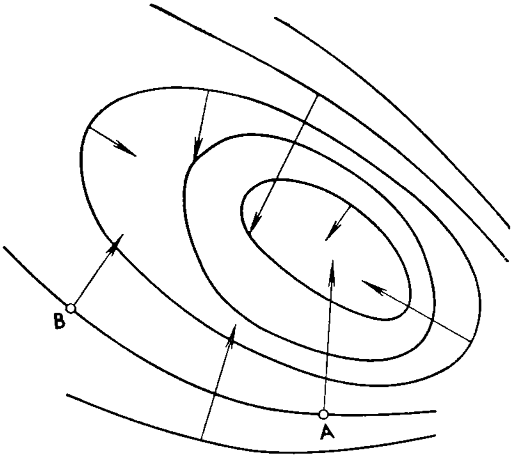


Fig. 117

where Δl is the distance between the surfaces in the direction l. But it is clear that Δl will be smallest, and therefore the derivative $\frac{du}{dl}$ largest, if l is directed along the normal to the surface. Whence follows our assertion.

All the concepts introduced for a spatial field carry over with appropriate simplifications to flat fields (see Sec. 10.1). For instance, the gradient of the field $u(x, y)$ computed from the formula grad $u = \frac{\partial u}{\partial x}\mathbf{i} + \frac{\partial u}{\partial y}\mathbf{j}$ is a vector in the xy-plane. At every point the gradient of the field is normal to the *level line* of the field, i.e. the line $u(x, y) =$ constant passing through this point (Fig. 117). Here it is evident from formula (4) that the modulus of the gradient is inversely proportional to the distance between the level lines. The gradient increases where the level lines come close together (compare, for example, the gradient of the field at points A and B in Fig. 117).

As an example, let us compute the gradient of the central-symmetric field $u = f(r)$, where $r = \sqrt{x^2 + y^2 + z^2}$. Here the value of u depends solely on the distance of the point from the origin and for this reason the level surfaces are spheres centred at the origin of coordinates. If we take two spheres whose radii differ by dr, then the values of the function f on them will differ by df. Therefore, the rate of change of the field across the level surfaces (that is, along the radii) is equal to $\frac{df}{dr}$ and so

$$\text{grad}\, u(r) = \frac{df}{dr}\frac{\mathbf{r}}{r} = \frac{df}{dr}\mathbf{r}^0 \tag{5}$$

Here, $\mathbf{r}$ is the radius vector drawn from the origin to any running point (x, y, z); $r = |\mathbf{r}|$; and $\mathbf{r}^0 = \frac{\mathbf{r}}{r}$ is a vector of unit length along $\mathbf{r}$.

Let us obtain formula (5) in a different way, relying on the expression of the gradient in coordinate form (1). We have

$$\frac{\partial u}{\partial x} = \frac{df}{dr}\frac{\partial r}{\partial x} = \frac{df}{dr}\frac{x}{\sqrt{x^2 + y^2 + z^2}} = \frac{df}{dr}\frac{x}{r}$$

Similarly, $\frac{\partial u}{\partial y} = \frac{df}{dr}\frac{y}{r}$, $\frac{\partial u}{\partial z} = \frac{df}{dr}\frac{z}{r}$, whence

$$\text{grad}\, u = \frac{df}{dr}\frac{x}{r}\mathbf{i} + \frac{df}{dr}\frac{y}{r}\mathbf{j} + \frac{df}{dr}\frac{z}{r}\mathbf{k} = \frac{df}{dr}\cdot\frac{1}{r}(x\mathbf{i} + y\mathbf{j} + z\mathbf{k}) = \frac{df}{dr}\frac{1}{r}\mathbf{r}$$

Since $\frac{1}{r}\frac{df}{dr}$ is a scalar, we see that for $\frac{df}{dr} > 0$ the vectors grad u everywhere continue the corresponding vectors $\mathbf{r}$ (like the spines of a rolled up hedgehog), and if $\frac{df}{dr} < 0$, then the vectors grad u are all directed to the origin.

Exercises

1. Find the derivative of the field $u = xy - z^2$ at the point $M(2, 1, -3)$ along the direction of the vector $a = \mathbf{i} + 3\mathbf{k}$.
2. Find the gradient of the field $u = \frac{k_1}{r_1} + \frac{k_2}{r_2}$, where k_1, k_2 = constant, and r_1, r_2 are the distances from certain fixed points M_1, M_2.

10.3 Potential energy and force

Suppose a body (one of very small size, often called merely a material point or particle) in arbitrary motion in space is acted upon by a force $\mathbf{F}$ depending solely on the position of the body. In other words, at every point M of space there is defined an appropriate force vector $\mathbf{F} = \mathbf{F}(M)$, which thus forms a vector *field of force*. As the body moves, this force performs work that is used to change the kinetic energy of the body or to overcome the resistance of certain external forces.

It is easy to calculate the work performed by the force **F** over a specified path traversed by the body. For an infinitesimal translation $d\mathbf{r}$, the force may be considered constant and therefore the corresponding work (see Sec. 9.2) is equal to the scalar product

$$dA = \mathbf{F} \cdot d\mathbf{r}$$

Summing these elementary pieces of work, we get the total work carried out by the force **F** when the body covers a certain path (L),

$$A = \int_{(L)} \mathbf{F} \cdot d\mathbf{r} \qquad (6)$$

Such an integral taken along the line (L) is called a *line integral.* To evaluate it, we must know not only the field **F** and the line (L), but also the direction in which this line is traversed; we say that the line (L) must be *oriented.* If (L) is open, all we need to know is which of the two endpoints is regarded as the initial point and which the terminal point, i.e. we have to indicate the limits of integration. This orientation is essential, for if we traverse the curve in the same field in the opposite direction, then $d\mathbf{r}$ becomes $-d\mathbf{r}$ and A becomes $-A$ (the work changes sign).

The expression (6) of a line integral can be written in other useful forms. If we recall that $d\mathbf{r} = \boldsymbol{\tau}\, ds$, where $\boldsymbol{\tau}$ is the unit vector of the tangent to (L) and ds is the differential of arc length, then we get

$$A = \int_{(L)} \mathbf{F} \cdot \boldsymbol{\tau}\, ds = \int_{(L)} F_{\tau}\, ds$$

where F_{τ} denotes the projection of the vector **F** on $\boldsymbol{\tau}$, that is, the "tangential" projection of the vector **F**. On the other hand, if we write, in Cartesian coordinates x, y, z,

$$\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}, \quad \mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$$

whence $d\mathbf{r} = dx\mathbf{i} + dy\mathbf{j} + dz\mathbf{k}$ and if we recall the expression (5), Ch. 9, for a scalar product in Cartesian projections, then we have

$$A = \int_{(L)} (F_x\, dx + F_y\, dy + F_z\, dz)$$

The work of a force can also be expressed as an integral of the time; if the motion of a body is given, that is, if we have the function $\mathbf{r}(t)$, then the force acting on the body is also a complex function of the time:

$$\mathbf{F} = \mathbf{F}(\mathbf{r}) = \mathbf{F}[\mathbf{r}(t)] = \mathbf{F}(t)$$

The work of a force during time dt can be expressed in terms of the velocity $\mathbf{v}$, where $\mathbf{v} = \frac{d\mathbf{r}}{dt}$; since $d\mathbf{r} = \mathbf{v}\,dt$, from formula (6) we get

$$A = \int_{t_{\text{init}}}^{t_{\text{ter}}} \mathbf{F}(t) \cdot \mathbf{v}(t)\, dt$$

where t_{init} and t_{ter} denote the initial and terminal instants of motion.

Thus, the quantity $\mathbf{F} \cdot \mathbf{v}$ is a scalar product of the force into the velocity, it is equal to the work per unit time and, as we know, is termed the power.

In this section we consider a force dependent solely on the position of the body. The dependence on time of the force acting on the body is a result of the translation of the body. In this case we can say that the work A during translation over a given path is actually independent of the velocity of the body, does not depend on the form of the functions $\mathbf{v}(t)$ and $\mathbf{r}(t)$, but depends only on the trajectory (path) over which the body moves, despite the fact that $\mathbf{v}(t)$ enters explicitly as a factor in the integrand for A. The point is that $\mathbf{r}(t)$ also varies when the function $\mathbf{v}(t)$ varies. Hence for a given $\mathbf{F} = \mathbf{F}(\mathbf{r}) = \mathbf{F}(x, y, z)$ the aspect of $\mathbf{F}(t)$ likewise changes. So also do the limits of integration t_{init} and t_{ter} that correspond to the given initial and terminal points of the path:

$$\mathbf{r}(t_{\text{init}}) = \mathbf{r}_{\text{init}}, \quad \mathbf{r}(t_{\text{ter}}) = \mathbf{r}_{\text{ter}}$$

For a large number of physical fields of force $\mathbf{F}$ it turns out that if the body traverses a closed contour, the overall work performed by the force $\mathbf{F}$ is equal to zero; in other words, if the force performs a certain amount of work on one section of the contour, then on the remaining portion it moves against the action of the force and returns the accumulated energy to the field. Such is the behaviour of a gravitational field, an electrostatic field and many other fields. Mathematically, this means that

$$\oint_{(L)} \mathbf{F} \cdot d\mathbf{r} = 0 \tag{7}$$

for any closed curve (L), where $\oint$ is to be understood as a line integral around a closed curve. The term *circulation* is also used for such an integral. Thus, we assume that the circulation of the force $\mathbf{F}$ around any closed path is equal to zero.

The assumption (7) can be stated equivalently thus: the work (6) of a force $\mathbf{F}$ along any open curve (L) depends only on the position of the beginning and end of the curve but does not depend on the shape of the curve. Indeed, suppose condition (7) is fulfilled, and the curves (L_1) and (L_2) have a common beginning M and end N, and suppose the work of the force $\mathbf{F}$ along (L_1) is equal to A_1 and along (L_2) is equal to A_2. We form a closed contour, going from M to N along (L_1) and then from N to M along (L_2). Then, by the foregoing, the work of the force $\mathbf{F}$ along such a contour is equal to $A_1 - A_2$. On the other hand, by virtue of (7) we get $A_1 - A_2 = 0$, whence $A_1 = A_2$. We leave it to the reader to prove in similar fashion the converse: that from the last equation follows condition (7).

We can introduce into the assumption (7) the concept of potential energy of a body; that is, the energy that depends on the position of the body (this notion is discussed in detail for one-dimensional motion in HM, Ch. 6). Namely, by definition the value of the potential energy U at any point M is equal to the work performed by the force $\mathbf{F}$ when the body is moved from M to a fixed point M_0. (In concrete problems, the point M_0 is frequently chosen at infinity, so that one speaks of the work performed when a body is carried to infinity; for this purpose, the work must be finite and must not depend on the mode of such a recession to infinity.) In other words

$$U(M) = \int\limits_{\frown{MM_0}} \mathbf{F} \cdot d\mathbf{r} \tag{8}$$

Here the choice of a particular path from M to M_0 is inessential.

Using the potential energy, it is easy to express the work performed by $\mathbf{F}$ when the body is moved from any point M_1 to any other point M_2. Precisely, for if in such a movement we pass through M_0 (this is possible since the choice of path has no effect on the work) and if we consider the work on the sections from M_1 to M_0 and from M_0 to M_2, we find that the work is equal to $U(M_1) - U(M_2)$. Thus, the work performed by a field is equal to the reduction in potential energy of the body.

Choice of the point M_0, i.e. the point at which the potential energy is set equal to zero, in determining the potential energy is rather arbitrary. Let us see what happens if M_0 is replaced by another point $\widetilde{M}_0$. Since the work $\widetilde{U}(M)$ performed in moving from M to $\widetilde{M}_0$ is equal to the work $U(M)$ performed in moving from M to M_0 plus the work performed in moving from M_0 to $\widetilde{M}_0$, it follows that $\widetilde{U}(M)$ differs from $U(M)$ by an additive constant. Thus, the potential energy is determined to within an additive constant. However, this arbitrariness does not affect the difference of the values of potential energy at the two points, that is, it disappears when we compute the work performed in moving from one point to the other.

We have proved the existence of potential energy in a field of force (that is, the potentiality of this field) by proceeding from condition (7) that the work performed around a closed path is equal to zero. It is easy to verify that, conversely, if the potential energy exists, then in calculating the work performed by a field during a translation of a body from point M to point N, only the position of the points is essential but not the path, which plays no role at all since this work is equal to $U(M) - U(N)$. But we have seen that this is equivalent to the condition (7).

Formula (8) defines the potential energy if a field of force is given. Let us derive the inverse formula that expresses the field in terms of the potential energy. Assume that the body performs an infinitesimal translation $d\mathbf{r} = dx\,\mathbf{i} + dy\,\mathbf{j} + dz\,\mathbf{k}$. Then the force $\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}$ performs the elementary work

$$dA = \mathbf{F} \cdot d\mathbf{r} = F_x\,dx + F_y\,dy + F_z\,dz$$

On the other hand, this work is equal to

$$dA = -dU = -\left(\frac{\partial U}{\partial x}dx + \frac{\partial U}{\partial y}dy + \frac{\partial U}{\partial z}dz\right)$$

Comparing both formulas, we see that

$$F_x = -\frac{\partial U}{\partial x}, \quad F_y = -\frac{\partial U}{\partial y}, \quad F_z = -\frac{\partial U}{\partial z}$$

whence

$$\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k} = -\frac{\partial U}{\partial x}\mathbf{i} - \frac{\partial U}{\partial y}\mathbf{j} - \frac{\partial U}{\partial z}\mathbf{k} = -\operatorname{grad} U \tag{9}$$

(see Sec. 10.2).

Let a particle of mass m be in motion under the action of a given force field $\mathbf{F} = -\operatorname{grad} U$. Then, by Sec. 9.4, the increment in kinetic energy is

$$\frac{mv_2^2}{2} - \frac{mv_1^2}{2} = \int_{s_1}^{s_2} F_\tau\,ds = -\int_{s_1}^{s_2} \frac{dU}{ds}\,ds = U_1 - U_2$$

and so

$$\frac{mv_2^2}{2} + U_2 = \frac{mv_1^2}{2} + U_1$$

Thus, the *total energy* of the particle,

$$E = \frac{mv^2}{2} + U$$

remains constant throughout the motion. We see that the total energy is the sum of the kinetic and the potential energy. (For one-dimensional motion a similar discussion is given in HM, Sec. 6.8.)

To illustrate, let us take the case of attraction via Newton's law, that is, the case where a body is acted upon by a force directed to a certain fixed point O (which we take for the coordinate origin) and is inversely proportional to the square of the distance from this point. Since the vector $-\frac{\mathbf{r}}{r}$ is directed towards the origin and has unit length, the force field is of the form

$$\mathbf{F} = -\frac{k}{r^2}\frac{\mathbf{r}}{r} = -k\frac{\mathbf{r}}{r^3} \tag{10}$$

where k is a constant of proportionality. By virtue of central symmetry of the force, it is easy to see* that the potential energy is also of the form $U = f(r)$. But then, by formulas (9) and (5), we get

$$\mathbf{F} = -\operatorname{grad} U = -\frac{df}{dr}\frac{\mathbf{r}}{r}$$

Comparing this with (10), we get

$$\frac{df}{dr} = \frac{k}{r^2}, \text{ whence } U = f(r) = -\frac{k}{r} \tag{11}$$

to within an arbitrary additive constant. Thus a Newtonian field of force is a potential field, and its *potential* is determined from formula (11). This formula gives the potential normalized (*normalization* is defined as a choice of one of a number of equivalent entities) by the condition of vanishing at infinity.

A problem sometimes arises of constructing *vector lines* of a force field **F** or, as we say for short, the *lines of force* of a field **F**. These are lines which at each point are in the direction of the field, that is to say, such that touch the vector of the field **F** at every point. The problem of constructing such lines readily reduces to the problem of integrating a system of differential equations: first introduce the Cartesian coordinates x, y, z and then write the condition of parallelism of the vectors:

$$d\mathbf{r} = dx\,\mathbf{i} + dy\,\mathbf{j} + dz\,\mathbf{k}$$

and

$$\mathbf{F} = F_x(x, y, z)\,\mathbf{i} + F_y(x, y, z)\,\mathbf{j} + F_z(x, y, z)\,\mathbf{k}$$

(see end of Sec. 9.1):

$$\frac{dx}{F_x(x, y, z)} = \frac{dy}{F_y(x, y, z)} = \frac{dz}{F_z(x, y, z)} \tag{12}$$

Here we use the fact that the vector $d\mathbf{r}$ lies precisely along the tangent to the curve (see Sec. 9.4). Equations (12) form a system of differential equations of the vector lines of the field **F** written in "symmetric form", in which all three coordinates are of equal status. From this it is easy to express $\frac{dy}{dx}$ and $\frac{dz}{dx}$ and go over to a system of two

* Consider the translation of a body over a sphere centred at O when r = constant. The work of the force is then equal to zero.

differential equations in $y(x)$ and $z(x)$ written in the ordinary form (Sec. 8.2).

As is clear from the geometric meaning (and as has been discussed in detail in the theory of differential equations), exactly one line of force passes through each point. Thus the entire space (or the portion of it in which the field of force is specified) is filled with lines of force.

For a Newtonian field, the lines of force are rays emanating from the centre of attraction, i.e. from point O. They do not intersect anywhere except at O itself, which is a singular point of the system of differential equations (cf. end of Sec. 10.2).

In Sec. 10.2 we saw that at every point the vector grad U is normal to the surface $U =$ constant passing through that point. It therefore follows from (9) that the lines of force at every point are normal to the equipotential surfaces. This, incidentally, is clear from the fact that the greatest work is performed per unit length of path if the body is moving normally to the surfaces of equal potential, and, on the other hand, the same occurs if the body is in motion along the lines of force.

Exercise

Let a flat field of force have a potential $U = \alpha xy$ ($\alpha =$ constant). Find the field itself and its lines of force.

10.4 Velocity field and flux

Consider the stationary flow of a fluid. At any point M the velocity of a particle of the fluid has a definite value $\mathbf{v} = \mathbf{v}(M)$; we have a vector velocity field.

Consider the trajectory (L) of a particle (this is called a *flow line*). It is known (see, for instance, Sec. 9.4) that at each point of (L) the velocity vector $\mathbf{v}$ is tangent to (L). Hence, for the velocity field the flow lines serve as vector lines (Sec. 10.3). As in (12), the differential equations for them can readily be obtained directly if we express dt in terms of the equations

$$\frac{dx}{dt} = v_x, \quad \frac{dy}{dt} = v_y, \quad \frac{dz}{dt} = v_z$$

and equate the results.

Let us now investigate something that will be of importance later on: the concept of the flux of a vector through a surface. Suppose we have chosen in space a closed or open "oriented" surface (σ). A surface is said to be *oriented* if we indicate which side is the inside and which the outside. Such an orientation can be done in two ways (Fig. 118).

Now let us calculate the volume of fluid carried outwards through (σ) in unit time. First find the volume carried during a small time dt

Fig. 118

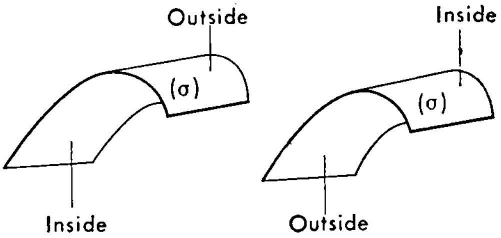
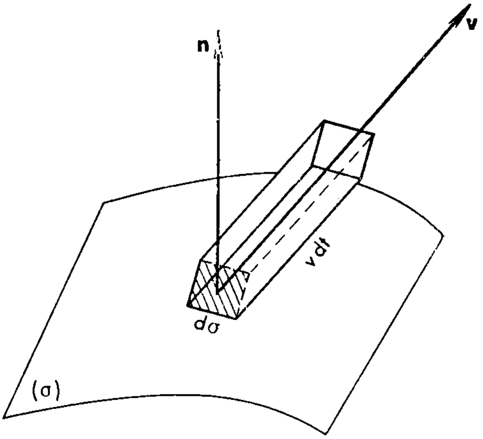


Fig. 119



through an element of surface $(d\sigma)$ (Fig. 119). This is the volume of an oblique cylinder with base $d\sigma$ and generatrix $v\,dt$. The altitude of the cylinder is equal to the projection of the generatrix on the perpendicular to the base, i. e. it is equal to $v_n\,dt$, where v_n is the projection of the velocity vector $\mathbf{v}$ on the unit vector $\mathbf{n}$ of the outer (directed outwards) normal to the surface. Thus, the volume of the cylinder is equal to $v_n\,dt\,d\sigma$.

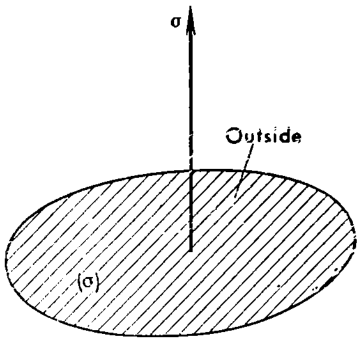
If in time dt there passes through the surface element $(d\sigma)$ a volume $v_n\,dt\,d\sigma$, then in unit time a volume $v_n\,d\sigma$ will pass through the same element. Summing these volumes for all elements, we find that in unit time the following volume will pass outward through the entire surface (σ):

$$Q = \int\limits_{(\sigma)} v_n\,d\sigma \qquad (13)$$

Such an integral of the normal projection is called the *flux of the vector* $\mathbf{v}$ through the surface (σ).*

* The integral (13) is a special case of the surface integral discussed in Sec. 4.7.

Fig. 120



In hydrodynamics, besides the velocity field **v** one frequently makes use of the field of "mass velocity" $\rho\mathbf{v}$, where ρ is the density of the gas at a given point. Reasoning exactly as before, we can easily prove that the flux

$$\int\limits_{(\sigma)} \rho v_n \, d\sigma$$

of this vector through any oriented surface (σ) is equal to the mass (and not the volume, as earlier!) of the gas carried outwards in unit time through (σ).

Clearly, the flux is a scalar quantity. It is essentially dependent on the orientation of the surface (σ): if the orientation is reversed, this reverses the sign of v_n and also of Q in formula (13). (Incidentally, this is also clear from the meaning of the integral (13) that we gave.) If the surface (σ) is located so that the outgoing flow lines intersect it everywhere, then $Q > 0$, and if the flow lines are inward, then $Q < 0$. Now if the flow lines partially intersect (σ) outwardly and partially inwardly, then the flux is equal to the sum of the positive and negative quantities (which ones?) and may be positive, negative, or equal to zero. The flux of the velocity vector through a surface completely filled with flow lines (that is, through every point of which a flow line passes that lies entirely on this surface) is always zero.*

The flux (13) is sometimes written in a different form. If an oriented plane area (σ) is given, then it is customary to associate with it a vector σ directed perpendicularly to (σ) from the inside to the outside (Fig. 120), the modulus of the vector being taken equal to the area (σ). This is done when the only essential things are the area (σ) and its direction in space while the specific form of (σ) (i.e. whether it is a circle or a rectangle, etc.) is immaterial. In this notation,

* More pictorially, this is a surface along which a gas slides.

the vector $d\boldsymbol{\sigma}$ will lie along $\mathbf{n}$ (see Fig. 119) and we can write $v_n\,d\sigma = \mathbf{v}\cdot d\boldsymbol{\sigma}$ (see formula (3) of Ch. 9). Thus

$$Q = \int\limits_{(\sigma)} v_n\,d\sigma = \int\limits_{(\sigma)} \mathbf{v}\cdot d\boldsymbol{\sigma} \tag{14}$$

The flux of vector $\mathbf{v}$ through the surface (σ) is also called the number of outward vector lines (flow lines) cutting (σ). This is a mere convention because the indicated number has dimensions and as a rule is fractional, but it is widely used because of its pictorial nature. Bear in mind that this "number" is to be understood algebraically, so that if a portion of (σ) is cut by outward lines and another portion is cut by inward lines, the result can have any sign (and even be zero) depending on which portion is cut by the larger number of lines.

Here are some simple examples of computing flux. First let the field $\mathbf{v}$ be constant (homogeneous, i.e. the same at all points) and let the surface (σ) be flat. Then from (14) it follows that

$$Q = v_n\sigma = \mathbf{v}\cdot\boldsymbol{\sigma} \tag{15}$$

Let the field $\mathbf{v}$ be proportional to the radius vector $\mathbf{r}$, that is, $\mathbf{v} = k\mathbf{r}$ where k is a constant of proportionality, and let (σ) be a sphere with centre at the origin and oriented in a natural fashion (the inside facing the origin). Then $v_n = v_r = kr$, and since $r =$ constant on (σ), it follows that

$$Q = \int\limits_{(\sigma)} kr\,d\sigma = kr\int\limits_{(\sigma)} d\sigma = kr\sigma = 4\pi kr^3$$

Finally, consider a general central-symmetric field in space defined by

$$\mathbf{v} = f(r)\,\mathbf{r}^0 \quad \left(\mathbf{r}^0 = \frac{\mathbf{r}}{r}\right) \tag{16}$$

where $\mathbf{r}$ is the radius vector of the running point and r is its length, so that $\mathbf{r}^0$ is a unit vector along the radius vector indicating the direction of the field. Then the flux of the field through a sphere of radius r with centre at the origin is equal to

$$Q = Q(r) = \int v_n\,d\sigma = \int f(r)\,d\sigma = f(r)\,4\pi r^2 \tag{17}$$

An interesting corollary follows from (15). Consider a closed polyhedron (Fig. 121); orient its faces so that one of them ((σ) in Fig. 121) is the closing face relative to the collection of other faces. Then it is clear that if this polyhedron is imagined to be in a homogeneous field of velocity $\mathbf{v}$, the flux through (σ) is equal to the algebraic sum of the fluxes through the other faces: $Q = Q_1 + Q_2 + Q_3$, whence

$$\mathbf{v}\cdot\boldsymbol{\sigma} = \mathbf{v}\cdot\boldsymbol{\sigma}_1 + \mathbf{v}\cdot\boldsymbol{\sigma}_2 + \mathbf{v}\cdot\boldsymbol{\sigma}_3 = \mathbf{v}\cdot(\boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_2 + \boldsymbol{\sigma}_3)$$

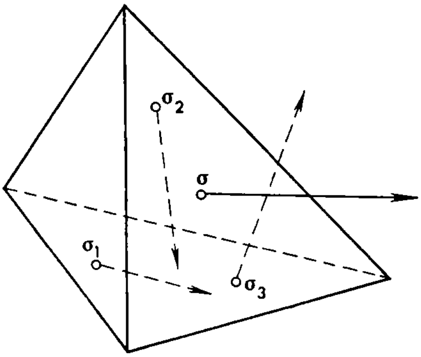


Fig. 121

and, because of the arbitrary nature of $\mathbf{v}$,

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_2 + \boldsymbol{\sigma}_3 \tag{18}$$

To summarise, then, the vector of the closing area is equal to the sum of the vectors of the "component" areas, which is to say that the areas can be added like vectors. We made essential use of the fact that the areas are represented by vectors precisely as shown in Fig. 120, which is a strong argument in favour of this representation.

This question can be approached from another angle. Orient all the faces of the polyhedron in Fig. 121 in natural fashion, so that the vectors $\boldsymbol{\sigma}_i$ and $\boldsymbol{\sigma}$ point outwards. Then in (18) substitute $-\boldsymbol{\sigma}_i$ for $\boldsymbol{\sigma}_i$ and transpose all terms to the left-hand side to get $\boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_2 + \boldsymbol{\sigma}_3 + \boldsymbol{\sigma} = \mathbf{0}$. This relation is fully analogous to the familiar (from vector algebra) rule on the sum of the vectors forming a closed polygon being equal to zero (see Sec. 9.1). This also justifies representing the areas by vectors.

In the limit, we can obtain from a polygon any closed curve (L) and from a polyhedron, any closed surface (σ); we thus arrive at the equations $\oint\limits_{(L)} d\mathbf{r} = \mathbf{0}$, $\oint\limits_{(\sigma)} d\boldsymbol{\sigma} = \mathbf{0}$ (the first one is obvious), where the circle on the integral sign (the circle can be dropped) stresses the fact that the integral is taken around a closed curve or surface.

Exercise

Consider the general axial-symmetric field in space defined by $\mathbf{v} = f(|\boldsymbol{\rho}|, z)\,\boldsymbol{\rho}^0 \left(\boldsymbol{\rho}^0 = \frac{\boldsymbol{\rho}}{|\boldsymbol{\rho}|}\right)$, where $\boldsymbol{\rho}$ is a vector from the point $(0, 0, z)$ to the point (x, y, z). What is the flux of this field through the surface of a right circular cylinder of radius $|\boldsymbol{\rho}|$ with the z-axis? What is the flux in the special case of a plane-parallel field?

10.5 Electrostatic field, its potential and flux

It is a well-known fact that a charge placed in space generates an electric field that can be analyzed on the basis of Coulomb's law. For the sake of simplicity, we consider space empty (a vacuum) and first examine the field of a charge q placed at the origin of coordinates O. To investigate the field, we can place at an arbitrary point M a trial unit charge and see with what force the field acts on it. This force $\mathbf{E}$ is termed the intensity of the electric field. Since at different points M the intensity E is different, we have a vector field $\mathbf{E} = \mathbf{E}(M)$, which is actually a special case of the force field considered in Sec. 10.3 (if all the time we have in view the action on a trial unit charge). The mathematical statement of Coulomb's law is similar to that of Newton's law considered in Sec. 10.3. For a point charge q it is of the form

$$\mathbf{E} = \frac{kq}{r^2}\mathbf{r}^0 = \frac{kq}{r^3}\mathbf{r}$$

$\left(\mathbf{r}^0 = \frac{\mathbf{r}}{r}\right.$ is a unit vector along the radius vector$\left.\right)$. This formula is somewhat different from (10) since the force must be directly proportional to q and, besides, for $q > 0$ a trial unit positive charge is not attracted but repulsed. The coefficient k in the last formula depends on the system of units chosen, and for simplicity we set it equal to unity.* Thus we write

$$\mathbf{E} = \frac{q}{r^2}\mathbf{r}^0 = \frac{q}{r^3}\mathbf{r} \tag{19}$$

In Sec. 10.3 we saw that such a field has a *potential*

$$\varphi(M) = \frac{q}{r} = \frac{q}{OM}$$

which is connected with the field $\mathbf{E}$ by the relation

$$\mathbf{E} = -\operatorname{grad}\varphi \tag{20}$$

and is normalized by the condition that it vanishes at infinity. By virtue of the homogeneity of space, the same kind of charge q placed at the point $M_0(x_0, y_0, z_0)$ generates a potential (we write $\varphi(\mathbf{r})$ instead of $\varphi(M)$, regarding $\mathbf{r}$ as the radius vector of the point M):

$$\varphi(\mathbf{r}) = \frac{q}{|\mathbf{r} - \mathbf{r}_0|} \tag{21}$$

where $\mathbf{r}_0$ is the radius vector of the point M_0.**

Now if we have a system of charges in space, the electric fields $\mathbf{E}$ associated with these charges combine.(This important fact, which

* This means that $\mathbf{E}$ and q are expressed in the electrostatic system of units CGSE.

** Do not confuse the radius vector $\mathbf{r}_0$ with the unit vector $\mathbf{r}^0$.

signifies the independence of the separate charges, has been established experimentally.) But from formula (20) it is easy to verify that the corresponding potentials are also additive. Indeed, if $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$, where

$$\mathbf{E}_l = -\operatorname{grad} \varphi_l = -\left(\frac{\partial \varphi_l}{\partial x}\mathbf{i} + \frac{\partial \varphi_l}{\partial y}\mathbf{j} + \frac{\partial \varphi_l}{\partial z}\mathbf{k}\right) \quad (l = 1, 2)$$

then, adding like projections, we get

$$\mathbf{E} = -\left[\frac{\partial(\varphi_1 + \varphi_2)}{\partial x}\mathbf{i} + \frac{\partial(\varphi_1 + \varphi_2)}{\partial y}\mathbf{j} + \frac{\partial(\varphi_1 + \varphi_2)}{\partial z}\mathbf{k}\right] = -\operatorname{grad}(\varphi_1 + \varphi_2)$$

which is to say that $\varphi = \varphi_1 + \varphi_2$ serves as the potential of the field $\mathbf{E}$.

We see that an electric field formed by any system of charges at rest possesses a potential and, besides, it is found that in studying the dependence of the potential on the field the superposition principle applies (the law of linearity, see Sec. 6.2). This makes it possible to construct, via a general method, the potential of an electric field resulting from charges being distributed in space with a certain variable density ρ. If the density has the aspect of a spatial delta function (see Sec. 6.3); that is, if

$$\rho(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{r}_0) \tag{22}$$

then the corresponding charge is a unit charge concentrated at a point with radius vector $\mathbf{r}_0$. For this reason, the potential corresponding to the density (22) — it is the Green's function in the given problem — is computed from formula (21) with $q = 1$; thus,

$$G(\mathbf{r}, \mathbf{r}_0) = \frac{1}{|\mathbf{r} - \mathbf{r}_0|}$$

Here, $\mathbf{r}_0$ determines the point of action, and $\mathbf{r}$ the point of observation.

Knowing the Green's function, it is easy, by the general method of Sec. 6.2, to obtain an expression for the potential in the case of an arbitrary density $\rho(\mathbf{r})$:

$$\varphi(\mathbf{r}) = \int\limits_{(\Omega_0)} G(\mathbf{r}, \mathbf{r}_0)\, \rho(\mathbf{r}_0)\, d\Omega_0 = \int\limits_{(\Omega_0)} \frac{\rho(\mathbf{r}_0)}{|\mathbf{r} - \mathbf{r}_0|}\, d\Omega_0$$

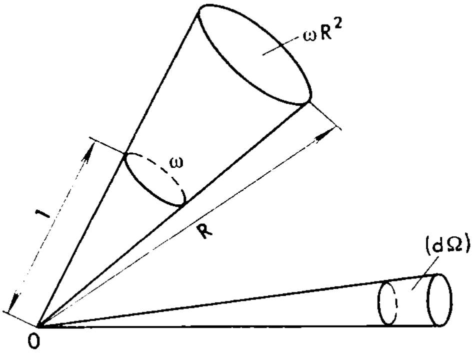
where the integration is over the entire region (Ω_0) occupied by the charge. In coordinate form this is written thus:

$$\varphi(x, y, z) = \iiint\limits_{(\Omega_0)} \frac{\rho(x_0, y_0, z_0)}{\sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2}}\, dx_0\, dy_0\, dz_0 \tag{23}$$

Examples in computing potential will be considered below in Sec. 10.6.

Now let us consider the flux of vector $\mathbf{E}$ through a closed surface (σ) oriented in "natural fashion": so that the inside faces a finite region bounded by (σ) and the outside faces infinity. In Sec. 10.4 we

Fig. 122



said that such a flux is also called the number of vector lines (i.e., for a field **E**, the "electric lines of force") cutting (σ) in the outward direction. Since this number is taken in the algebraic meaning, that is, the difference between the number of emanating lines and the number of entering lines, it can be understood as the number of lines originating inside (σ).

In the simplest case, where **E** is generated by a point charge q located at the coordinate origin and (σ) is a sphere with centre at the origin with radius R, the flux can readily be found: on (σ)

$$\mathbf{E} = \frac{q}{R^2}\mathbf{r}^0 = \frac{q}{R^2}\mathbf{n}$$

since $\mathbf{r}^0$ on a sphere coincides with the unit vector of the outer normal **n**. But then, on the sphere, $E_n = \frac{q}{R^2} = \text{constant}$ and, by definition (cf. formula (13)), the flux is

$$\int\limits_{(\sigma)} E_n\,d\sigma = \int\limits_{(\sigma)} \frac{q}{R^2}\,d\sigma = \frac{q}{R^2}\int\limits_{(\sigma)} d\sigma = \frac{q}{R^2}\,4\pi R^2 = 4\pi q \qquad (24)$$

Thus, the flux at hand does not depend on the radius R and is directly proportional to the charge.

As in the case of (24), proof can be given that the flux of the vector **E** through a portion of a sphere with centre in a point charge (Fig. 122) is equal to ωq, where ω is the corresponding solid angle (the corresponding area on a sphere of unit radius). But from this it follows immediately that for a field generated by a point charge the electric lines of force cannot begin or terminate in any volume that does not contain the charge (i. e. they begin or terminate only on that charge). Indeed, choosing a small volume ($d\Omega$) as shown in Fig. 122, it is easy to see that just as many lines emerge through its spherical walls as enter, but through the conical wall the lines do not pass at all.

Thus, for any surface (σ) containing a point charge q within it, the number of lines of force originating inside (σ) (i.e., the flux of the vector $\mathbf{E}$ through (σ)) is equal to $4\pi q$. In other words, for the field at hand, the lines of force, for $q > 0$, originate on the charge (their quantity is proportional to the charge) and go off to infinity. If the charge is negative, the lines of force originate at infinity and terminate on the charge.

Now suppose that there are several charges, say q_1, q_2 and q_3, inside the closed surface (σ). Then the associated fields and, for this reason, the fluxes are additive, which means the overall flux through (σ) is equal to $4\pi q_1 + 4\pi q_2 + 4\pi q_3 = 4\pi(q_1 + q_2 + q_3) = 4\pi q$, where $q = q_1 + q_2 + q_3$ is the total charge located inside (σ).

To summarize: the number of electric lines of force originating in any volume of space is proportional (with proportionality constant 4π) to the total charge contained in the volume. This assertion is known as *Gauss' theorem*.

10.6 Examples

The field formed by a point charge is the simplest type of electric field. More complicated fields result from a combination of fields of point charges located at distinct points.

In many important cases, a field can be constructed by direct application of the Gauss theorem (see end of Sec. 10.5). Let us consider, say, a field created by a sphere of radius R_0 charged with a constant volume density ρ.

If the centre of the sphere is taken as the origin, then by virtue of the spherical symmetry of the problem it is clear that the intensity $\mathbf{E}$ of the field has the form

$$\mathbf{E} = F(r)\,\mathbf{r}^0 \tag{25}$$

where $F(r)$ is a certain scalar function. If by Gauss' theorem we equate the vector flux computed from formula (17), through a sphere of radius r to the charge, inside the sphere, multiplied by 4π, then we get

$$F(r)\cdot 4\pi r^2 = \begin{cases} 4\pi\dfrac{4}{3}\pi r^3\rho & (r < R_0), \\ 4\pi\dfrac{4}{3}\pi R_0^3\rho = 4\pi q & (r > R_0) \end{cases}$$

where q is the total charge of the sphere. From this, finding $F(r)$ and substituting into (25), we get

$$\mathbf{E} = \begin{cases} 4\pi\dfrac{\rho}{3}r\mathbf{r}^0 = \dfrac{4}{3}\pi\rho\mathbf{r} & \text{(inside of sphere)}, \\ \dfrac{q}{r^2}\mathbf{r}^0 & \text{(outside of sphere)} \end{cases}$$

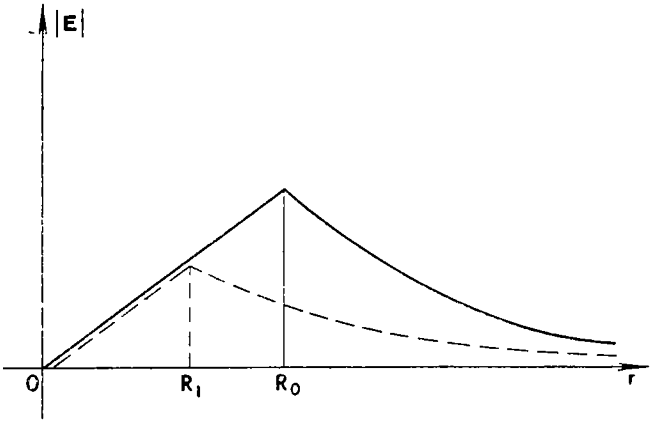


Fig. 123

In particular (cf. formula (19)), outside the sphere the field intensity is exactly as if the entire charge were concentrated at the centre of the sphere. The graph of the modulus of the vector **E** as a function of the radius is shown in Fig. 123 by the solid line.

By symmetry, the corresponding potential is of the form $U = U(r)$. Recalling formula (5), we find that

$$\frac{dU}{dr} = \begin{cases} -\frac{4}{3}\pi\rho r & (0 \leqslant r \leqslant R_0), \\ -\frac{q}{r^2} & (R_0 \leqslant r < \infty) \end{cases}$$

whence

$$U = \begin{cases} -\frac{2}{3}\pi\rho r^2 + C_1 & (0 \leqslant r \leqslant R_0), \\ \frac{q}{r} + C_2 & (R_0 \leqslant r < \infty) \end{cases}$$

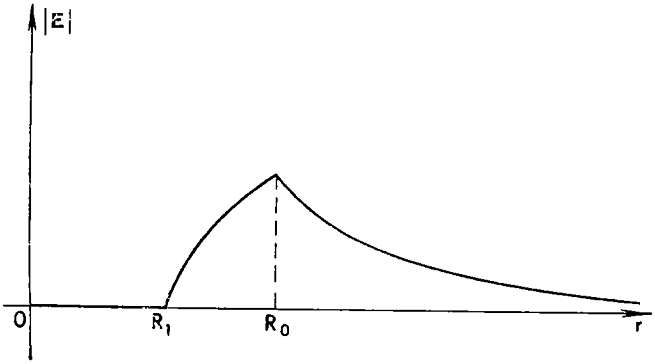
where C_1, C_2 are constants. Normalizing the potential by the condition $U(\infty) = 0$, we get $C_2 = 0$; and for $U(r)$ to be continuous for $r = R_0$, it must be true that

$$-\frac{2}{3}\pi\rho R_0^2 + C_1 = \frac{q}{R_0}, \quad \text{that is} \quad C_1 = \frac{q}{R_0} + \frac{2}{3}\pi\rho R_0^2$$

And so, finally,

$$U = \begin{cases} -\frac{2}{3}\pi\rho(R_0^2 - r^2) + \frac{q}{R_0} & (0 \leqslant r \leqslant R_0), \\ \frac{q}{r} & (R_0 \leqslant r < \infty) \end{cases}$$

Fig. 124



Now let us consider the field of a charge distributed with constant density σ between two concentric spheres of radii R_1 and R_0. By virtue of the principle of superposition, this field may be obtained as the difference of fields obtained in a distribution of charge of density ρ in a sphere of radius R_0 and in a sphere of radius R_1. The graphs of both fields are shown in Fig. 123. We see that inside the cavity (i.e., inside the sphere of radius R_1) the difference is zero, which means there is no electric field; in other words, the potential is constant. Outside the sphere of radius R_0 the field is of the form

$$\frac{q_0}{4\pi r^2}\mathbf{r}^0 - \frac{q_1}{4\pi r^2}\mathbf{r}^0 = \frac{q_0 - q_1}{4\pi r^2}\mathbf{r}^0 = \frac{q}{4\pi r^2}\mathbf{r}^0$$

Again, the field is as if the entire charge were concentrated in the centre. The graph of a field for a hollow charge is shown in Fig. 124. These same conclusions hold true for infinitely close R_1 and R_0, which is to say for uniform distribution of charge on the sphere.

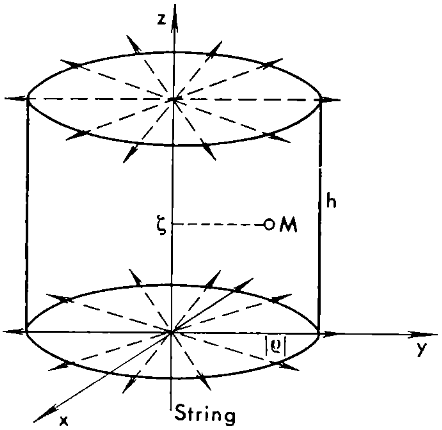
For the next example, we consider a charge distributed with constant linear density μ along a finite rectilinear string which will be taken for the z-axis. This is a plane-parallel field (Sec. 10.1) and it can be obtained by summing (integrating) the fields due to small charges located along the string. But it is simpler to use the Gauss theorem. Note that by virtue of the axial symmetry of the problem, the field should have the form

$$\mathbf{E} = \varphi(|\boldsymbol{\rho}|)\,\boldsymbol{\rho}^0$$

where $\boldsymbol{\rho}$ is a vector perpendicular to the string and directed away from it to the running point (if the string is along the z-axis, then $\boldsymbol{\rho} = x\mathbf{i} + y\mathbf{j}$); $\boldsymbol{\rho}^0 = \boldsymbol{\rho}/|\boldsymbol{\rho}|$ is the unit vector (do not confuse the vector $\boldsymbol{\rho}$ with the volume density of charge) and $\varphi(|\boldsymbol{\rho}|)$ is a certain scalar function. Applying the Gauss theorem to a cylinder of altitude h and radius $|\boldsymbol{\rho}|$ shown in Fig. 125, we get

$$\varphi(|\boldsymbol{\rho}|)\,2\pi|\boldsymbol{\rho}|\,h = 4\pi\mu h$$

Fig. 125



(here it must be taken into account that the vector flux **E** through the bases of the cylinder is equal to zero, since the electric lines of force lie along these bases). From this we have $\varphi(|\boldsymbol{\rho}|) = \frac{2\mu}{|\boldsymbol{\rho}|}$, or

$$\mathbf{E} = \frac{2\mu}{|\boldsymbol{\rho}|}\boldsymbol{\rho}^0 \tag{26}$$

In Fig. 125 are shown the corresponding lines of force, which (in the case $\mu > 0$) originate on the charged string and diverge to infinity.

An interesting difficulty arises in computing the potential of the field under consideration. We might attempt to sum the potentials of the elementary charges of the string. By (21) this would yield, at some point $M(x, y, z)$,

$$\varphi = \int_{-\infty}^{\infty} \frac{\mu\, d\zeta}{|x\mathbf{i} + y\mathbf{j} + z\mathbf{k} - \zeta\mathbf{k}|} = \int_{-\infty}^{\infty} \frac{\mu\, d\zeta}{\sqrt{x^2 + y^2 + (z - \zeta)^2}} \tag{27}$$

But for large $|\zeta|$ the denominator behaves like $\sqrt{\zeta^2} = |\zeta|$ and for this reason the integral (27) diverges to infinity (see Sec. 3.1), which means it has an infinite value! Clearly this is no way to find the potential.

But recall that we are always interested not in the value itself of the potential but in the derivatives of it (for finding the field **E**) or in the difference of its values for two points. Therefore the situation is exactly as described in Sec. 3.6 for a divergent integral depending on a parameter (here we have three parameters: x, y, z). Also note that the string is always finite, so instead of (26) we have to consider the integral

$$\varphi_N = \int_{-N}^{N} \frac{\mu\, d\zeta}{\sqrt{x^2 + y^2 + (z - \zeta)^2}}$$

for very large N. This integral can conveniently be split into two: $\int_{-N}^{z} + \int_{z}^{N}$; then in the first substitute ζ for $\zeta - z$, and in the second integral, η for $z - \zeta$. Then we get (verify this!)

$$\begin{aligned} \varphi_N &= -\mu \ln(x^2 + y^2) + \mu \ln[N - z + \sqrt{x^2 + y^2 + (N - z)^2}] \\ &\quad + \mu \ln[N + z + \sqrt{x^2 + y^2 + (N + z)^2}] \\ &= -\mu \ln(x^2 + y^2) + P_N(x, y, z) \qquad (28) \end{aligned}$$

where P_N denotes the sum of the "long logarithms". Hence, the potential at the fixed point $Q(x, y, z)$, which potential depends on the length of the string, becomes infinite when the length of the string tends to infinity. In nature we never have to do with infinitely long strings with constant charge density on them. In computing the potential, it is essential that when the string is of finite length, the quantity N, which characterizes the length, enters into the answer even when $N \to \infty$.

It is a remarkable fact that despite this the field $\mathbf{E}$ near the string is independent of the length of the string.*

The exact mode of operations is as follows: fix some large but finite N and find the field by the general rule (20):

$$\begin{aligned} E_x &= 2\mu \frac{x}{x^2 + y^2} + \frac{\partial P_N}{\partial x} \\ &= 2\mu \frac{x}{x^2 + y^2} + \frac{\mu x}{(N - z + \sqrt{x^2 + y^2 + (N - z)^2})\sqrt{x^2 + y^2 + (N - z)^2}} \\ &\qquad + \text{similar term with } N + z \end{aligned}$$

The expression E_y is obtained by substituting y for x:

$$E_z = \frac{\partial P_N}{\partial z} = \frac{-1}{\sqrt{x^2 + y^2 + (N - z)^2}} + \text{similar term with } N + z$$

After differentiating, let N go to infinity, then the terms $\frac{\partial P_N}{\partial x}$, $\frac{\partial P_N}{\partial y}$, $\frac{\partial P_N}{\partial z} \to 0$ and we obtain a simple answer that is not dependent on N:

$$E_x = 2\mu \frac{x}{x^2 + y^2}, \quad E_y = 2\mu \frac{y}{x^2 + y^2}, \quad E_z = 0$$

* For this assertion to hold, the following inequalities must be fulfilled: $\frac{|x|}{N}, \frac{|y|}{N}, \frac{|z|}{N} \ll 1$. Verify that if at least one of these quantities satisfies the reverse inequality, that is, $\frac{|x|}{N} \gg 1$, or $\frac{|y|}{N} \gg 1$, or $\frac{|z|}{N} \gg 1$, then the potential and the field are close to the field of a finite charge equal to $2\mu N$ and located at the coordinate origin.

We get along without the detailed procedure of obtaining φ for a finite N, of finding $\mathbf{E}$ and then passing to the limit as $N \to \infty$. We dispense with normalization of the potential φ by the condition $\varphi = 0$ at infinity. For example, we take for zero the potential at the point $x = 1$, $y = 0$, $z = 0$ and denote by Φ the thus normalized potential. It results from adding the constant $C = C_N$ to (28):

$$\Phi = -\mu \ln(x^2 + y^2) + P_N(x, y, z) + C$$

whence

$$C = -P_N(1, 0, 0) = -2\mu \ln(N + \sqrt{1 + N^2})$$

It is easy to see that in the limit, as $N \to \infty$, we have

$$P_N(x, y, z) + C = P_N(x, y, z) - P_N(1, 0, 0) \to 0$$

whence

$$\Phi = -\mu \ln(x^2 + y^2)$$

Computing $\mathbf{E} = -\operatorname{grad} \Phi$, we get the preceding result, which coincides with what was obtained via the Gauss theorem (cf. formula (26)). Unlike the case of the potential, when computing the field via the Gauss theorem, we can (and must!) consider at once an infinite string.

The field can also be found by integrating: by adding the elementary fields $d\mathbf{E}$ of the charges $dq = \mu d\zeta$ located on an element of the string. Since $\mathbf{E}$ is a vector, we have to add separately the components

$$dE_x = \frac{x\,dq}{[x^2 + y^2 + (z - \zeta)^2]^{3/2}} = \frac{\mu x\,d\zeta}{[x^2 + y^2 + (z - \zeta)^2]^{3/2}}$$

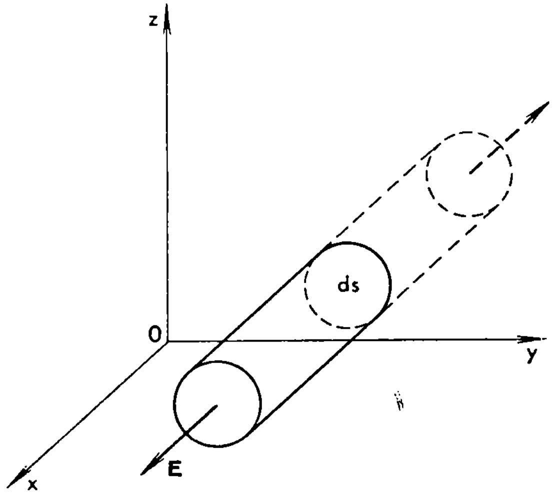
$$dE_y = \frac{\mu y\,d\zeta}{[x^2 + y^2 + (z - \zeta)^2]^{3/2}}, \quad dE_z = \frac{\mu(z - \zeta)\,d\zeta}{[x^2 + y^2 + (z - \zeta)^2]^{3/2}}$$

The corresponding integrals converge, since for large ζ the integrand diminishes as $|\zeta|^{-3}$ or $|\zeta|^{-2}$. For this reason we can at once take the integrals from $\zeta = -\infty$ to $\zeta = \infty$ (see Exercise 1).

Finally, let us consider a field generated by a charge that is uniformly distributed over, say, the yz-plane with constant surface density ν. Here again it is convenient to use the Gauss theorem. Note that because of symmetry the lines of force must be parallel to the x-axis. Let us take a thin cylinder (Fig. 126) located symmetrically about the yz-plane. The flux through its lateral surface is equal to zero, whereas through the bases it is equal to $2E ds$, again because of symmetry. On the other hand, by Gauss' theorem this flux is equal to the product of 4π by the charge enclosed in the cylinder, i.e. by $\nu\,ds$. Equating the results, we get $2E\,ds = 4\pi\nu\,ds$, whence $E = 2\pi\nu$. Thus, in the given example,

$$\mathbf{E} = 2\pi\nu\mathbf{i} \ (x > 0), \quad \mathbf{E} = -2\pi\nu\mathbf{i} \ (x < 0) \tag{29}$$

Fig. 126



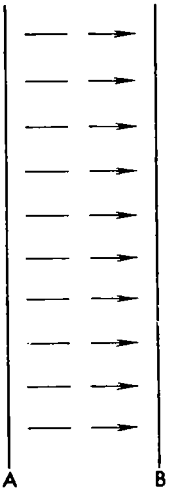
We see that the intensity on both sides of the charged plane is constant and experiences a jump of $4\pi\nu\mathbf{i}$ in a passage through this plane.

What will happen if, as is usual in practical situations, not the infinite plane but a finite portion of the plane is charged? Then near this portion (not near the boundary but inside) the finiteness of the piece has a weak effect and we have nearly a homogeneous field that is computed from formulas (29). Near the boundary of the piece and also at some distance from it, the homogeneity of the field is substantially disrupted. At a sufficient distance from the piece (as in the case of any finite system of charges) it is possible to replace it with a point-charge field equal to the total charge on the indicated piece, which means we can take advantage of formula (19).

An important case is that shown schematically in Fig. 127. Here we have a plane capacitor: two parallel plates with equal and opposite charges on them (ν units per cm^2 on one plate and $-\nu$ units per cm^2 on the other plate). Check to see that to the left of A and to the right of B the fields of the two planes are in opposite directions and together equal zero. Inside the capacitor, between A and B, the field is equal to $E = 4\pi\nu$ and directed from the positive plate to the negative plate. Bear in mind, however, that this field is half due to the charges of plate A and half due to the charges of plate B. In particular, the force acting on plate A is equal to $2\pi S\nu^2 = \frac{1}{2}qE$, since A is acted upon solely by the field of charges B (cf. HM, Sec. 8.4).

The coefficient $\frac{1}{2}$ can also be obtained in this way: the charges on A are located on the boundary between the region where $E = 0$

Fig. 127



(to the left of A) and where $E = 4\pi\nu$ (to the right of A). The mean field is therefore equal to $\frac{1}{2}(4\pi\nu + 0) = 2\pi\nu$. But this averaging is precisely the method for considering only that portion of the field that is created by B charges and is the same to the left and to the right of A. The field of charges A themselves is of opposite sign to the left and right of A and therefore vanishes in an averaging.

Now let us take up the case where the field is obtained by a direct superposition of the fields of the point charges.

We consider a combination of two charges, $+q$ and $-q$, at the points $\left(\frac{h}{2}, 0, 0\right)$ and $\left(-\frac{h}{2}, 0, 0\right)$. Denoting by $\mathbf{r}_+$ and $\mathbf{r}_-$ the radius vectors of these points, we get, by virtue of (21), the total potential:

$$\begin{aligned}\varphi(\mathbf{r}) &= \frac{q}{|\mathbf{r} - \mathbf{r}_+|} - \frac{q}{|\mathbf{r} - \mathbf{r}_-|} \\ &= q\left(\frac{1}{\sqrt{\left(x - \frac{h}{2}\right)^2 + y^2 + z^2}} - \frac{1}{\sqrt{\left(x + \frac{h}{2}\right)^2 + y^2 + z^2}}\right) \qquad (30)\end{aligned}$$

Since the field is symmetric about the x-axis, it suffices to imagine it in a plane passing through this axis, say in the xz-plane. The solid lines in Fig. 128 are the traces of the intersection by the xz-plane of the equipotential surfaces that result if we equate the right side of (30) to a constant (these lines are closed oval surfaces of the eighth order). The dashed lines are lines of force. They originate on the positive charge (the "source" of the lines of force having "strength" $4\pi q$; see

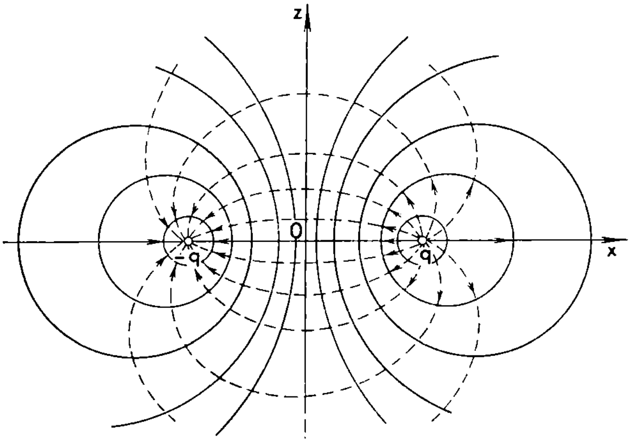


Fig. 128

Sec. 10.7) and terminate on the negative charge (this is the "sink"; see Sec. 10.7). By virtue of the relation (20), these lines are normal to the equipotential surfaces (they intersect them at right angles).

If h is infinitely small, that is, if the source and the sink are located infinitely close together, then the system of charges is called a *dipole*. However, if in this situation the charges themselves remain finite, then their fields merely cancel out to yield a zero field. For this reason, the only interest is in the case where the charges are infinitely large and such that the product $m = qh$ (this product is called the *dipole moment)* remains finite. In space, a dipole is characterized not only by its location and moment but also by its direction; to obtain the direction, draw the axis through the charges in the direction from the negative to the positive charge.

To derive the field of a dipole, consider Fig. 129, where l is the axis of the dipole. For a sufficiently small h, at any point M we will have

$$\mathbf{E} = \frac{q}{r_1^3}\mathbf{r}_1 - \frac{q}{r^3}\mathbf{r} = qh\,\frac{\dfrac{\mathbf{r}_1}{r_1^3} - \dfrac{\mathbf{r}}{r^3}}{h} = m\frac{d}{dl}\left(\frac{\mathbf{r}}{r^3}\right) \tag{31}$$

where $\frac{d}{dl}$ denotes the derivative obtained under a change in the position of the charge (it differs by the factor -1 from the derivative obtained under a change in the position of the observation point M). Simplifying the right side of (31), we get (verify this!)

$$\mathbf{E} = m\left(\frac{1}{r^3}\frac{d\mathbf{r}}{dl} - \frac{3\mathbf{r}}{r^4}\frac{dr}{dl}\right) = \frac{m}{r^3}\left(\frac{3\mathbf{r}}{r}\cos\alpha - \mathbf{l}^0\right)$$

where $\mathbf{l}^0$ is the unit vector along the l-axis and α is the angle between the vectors $\mathbf{l}^0$ and $\mathbf{r}$.

Fig. 129

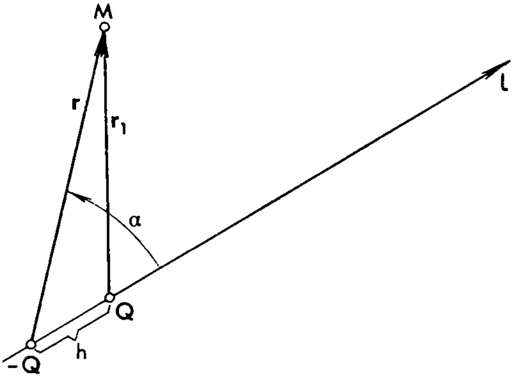
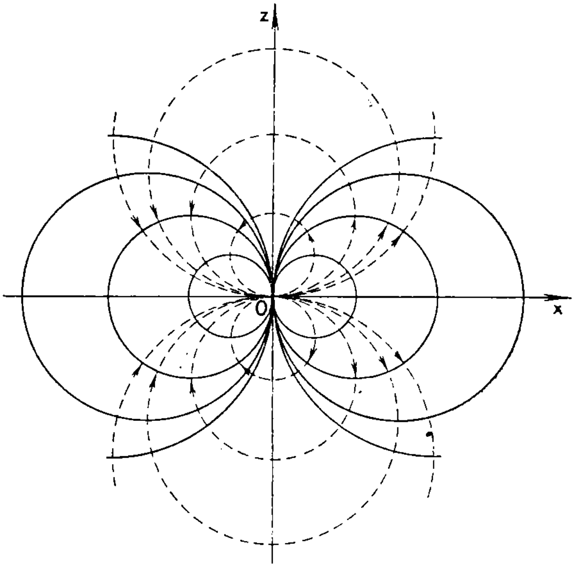


Fig. 130



A similar consideration of the potential of the dipole yields the expression

$$\varphi = m \frac{d}{dl}\left(\frac{1}{r}\right) = m \frac{\cos\alpha}{r^2}$$

If we assume the axis of the dipole to coincide with the x-axis and, like Fig. 128, consider the picture in the xz-plane, then we get $\cos\alpha = x/r$, or $\varphi = m\,\frac{x}{r^3} = m\,\frac{x}{(x^2+z^2)^{3/2}}$. Fig. 130 shows the traces of the equipotential surfaces (which are oval surfaces of order six) and the electric lines of force, which, as we see, originate on the dipole and terminate there. (Think about how to obtain Fig. 130 from Fig. 128).

Exercises

1. Using integration, compute the field of an infinite homogeneous charged rectilinear string.
2. Using integration, compute the field of an infinite homogeneous charged plane.
3. Compute the field and the potential of a "plane dipole" which is the result of superimposing two oppositely charged homogeneous rectilinear infinitely close strings.

10.7 General vector field and its divergence

It is now easy to formulate the foregoing concepts in general form for a certain stationary *vector field* $\mathbf{A} = \mathbf{A}(M)$, irrespective of the physical meaning. It is natural to generalize the concept of *vector lines* as lines along the field at every point. As in Sec. 10.4, we introduce the concept of the *flux of a vector* **A** through an oriented surface (σ) by the formula $Q = \int\limits_{(\sigma)} A_n\, d\sigma = \int\limits_{(\sigma)} \mathbf{A}\cdot d\boldsymbol{\sigma}$. This flux is otherwise spoken of as the *number of outward vector lines* intersecting (σ).

As in Sec. 10.5, we can regard the vector flux **A** through a closed surface (σ),

$$Q = \oint\limits_{(\sigma)} \mathbf{A}\cdot d\boldsymbol{\sigma} \tag{32}$$

(the circle on the integral sign is not obligatory, since it merely serves to emphasize that the integral is taken round a closed surface). Such a flux has the following simple property: if a certain body (Ω) is partitioned with the aid of certain surfaces into parts (Ω_1), (Ω_2), ..., then the outward flux of the field through the surface of the body (Ω) is equal to the sum of analogous fluxes taken for each of the bodies (Ω_1), (Ω_2), ... For example, in Fig. 131 we have a partition of (Ω) into the two parts (Ω_1) and (Ω_2). The flux of the vector A through the surface of the body (Ω_1) can be represented as a sum of two integrals, over (σ_1) and over (σ_3); now the flux through the surface of the body (Ω_2) is equal to the sum of the integrals over (σ_2) and over (σ_3). If these fluxes are combined, the integrals over (σ_3) cancel out (why?), and the integrals over (σ_1) and (σ_2) together yield the flux of vector **A** through the surface of the body (Ω).

This property enables us to interpret the integral (32) as the number of vector lines originating inside the volume (Ω) bounded by the surface (σ). If $Q > 0$, we say that there is a *source of vector lines* in (Ω), and Q is called the *strength* of this source. If $Q < 0$, we say that there is a *sink* (or, what is the same thing, a source of nega-

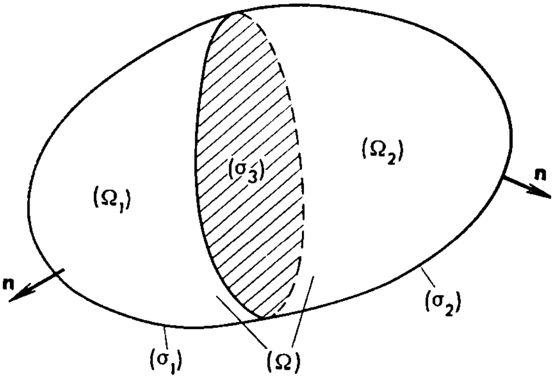


Fig. 131

tive strength) in (Ω). For the sake of simplicity we will always regard a sink as a special case of the source. If $Q = 0$, then either there are no sources or sinks in (Ω) or they cancel out; incidentally, even when $Q \neq 0$, there may be either sources or sinks in (Ω), which, however, are only partially compensated for in this case.

Suppose the vector lines arise throughout the space. We can speak of the mean density of the source Q/Ω in any volume (Ω) (by Ω is meant the numerical value of the volume (Ω)) and we can speak of the density of the source at any point M of the space. This density is equal to

$$\lim_{(\Delta\Omega)\to M} \frac{\Delta Q}{\Delta\Omega} = \lim_{(\Delta\Omega)\to M} \frac{\oint\limits_{(\Delta\sigma)} \mathbf{A}\cdot d\boldsymbol{\sigma}}{\Delta\Omega} \tag{33}$$

where $(\Delta\Omega)$ is a small volume containing the point M and $(\Delta\sigma)$ is the surface of the small volume. This density of the source is also known as the *divergence* of the vector field $\mathbf{A}$ and is denoted by div $\mathbf{A}$. We can thus say that the divergence of a vector field is the number of vector lines originating in an infinitely small volume (or, what is the same thing, the flux of field $\mathbf{A}$ through the surface of this volume) referred to a unit of this volume. Note that the divergence of a vector field is a scalar quantity or, to be more precise, it forms a scalar field, since it assumes a value at every point of space.

Formula (33) may be rewritten as

$$\operatorname{div} \mathbf{A} = \frac{dQ}{d\Omega}, \quad \text{i.e.} \quad dQ = \operatorname{div} \mathbf{A}\, d\Omega \tag{34}$$

The result is an expression for the number of vector lines originating in an elementary volume $(d\Omega)$. Summing (Sec. 4.7), we get an exprestion for the number of vector lines originating in a finite volume (Ω), i.e. for the flux of vector $\mathbf{A}$,

$$\oint_{(\sigma)} \mathbf{A}\cdot d\boldsymbol{\sigma} = \int_{(\Omega)} \operatorname{div} \mathbf{A}\, d\Omega \tag{35}$$

where (Ω) is a finite volume and (σ) is its surface. This important formula is called the *Ostrogradsky formula* (also known as the *divergence theorem*).

It is sometimes convenient to compute the divergence directly on the basis of its definition (33). Consider, for example, a central-symmetric field in space defined by the formula

$$\mathbf{A} = f(r)\, \mathbf{r}^0 \tag{36}$$

As was shown in Sec. 10.4 (formulas (16) and (17)), the field flux through a sphere of radius r with centre at the origin is equal to $Q(r) = 4\pi r^2 f(r)$, and so the number of vector lines originating in a thin layer between two such spheres is equal to $dQ = 4\pi d[r^2 f(r)] = 4\pi[2rf(r) + r^2 f'(r)]dr$.

Hence, per unit volume of this layer we have

$$\operatorname{div} \mathbf{A} = \frac{dQ}{4\pi r^2\, dr} = \frac{2}{r} f(r) + f'(r) \tag{37}$$

In particular, a central-symmetric field without sources and outside the coordinate origin is characterized by the fact that

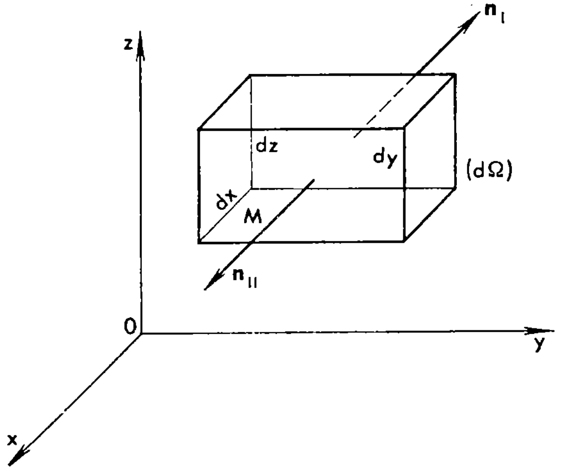
$$\frac{2}{r} f(r) + f'(r) = 0, \text{ whence } f(r) = \frac{C}{r^2} \ (C \text{ constant})$$

We have arrived at the Newtonian law (10). By virtue of formula (17), $Cr^{-2}\cdot 4\pi r^2 = 4\pi C$ vector lines originate at the very origin of coordinates. But a point source of strength $4\pi C$ at the origin has a density of $4\pi C\delta(\mathbf{r})$. Thus, by (36),

$$\operatorname{div}\left(\frac{C}{r^2}\, \mathbf{r}^0\right) = 4\pi C\delta(\mathbf{r}), \text{ i.e. } \operatorname{div}\left(\frac{\mathbf{r}}{r^3}\right) = 4\pi\delta(\mathbf{r}) \tag{38}$$

In this important instance, forgetting the delta function would have resulted in a gross error! Here we have a three-dimensional delta function (see Sec. 6.3): $\delta(\mathbf{r}) = \delta(x)\delta(y)\delta(z)$. Using the delta function is much more convenient than long verbal statements like "the divergence is everywhere zero except at a singular point of the field, where it is infinite", and the like. The origin of coordinates is a singular point since the velocity at this point is infinite and does not have a definite direction. In this example, all vector lines go to infinity. Think over where the vector lines originate and where they terminate (or where they go to) if $f(r)$ increases more slowly or faster than r^{-2} as $r\to 0$.

Fig. 132



The definition (33) of divergence was given in invariant form, which does not depend on the choice of coordinate system. It is also of interest to derive the formula for computing the divergence with the aid of an xyz-coordinate system. To do this, we take advantage of the fact that in formula (34) the form of the elementary volume $(d\Omega)$ is inessential, and for this volume we choose an infinitesimal rectangular parallelepiped with edges parallel to the coordinate axes (Fig. 132). Then the flux dQ of the vector $\mathbf{A}$ through the surface of the parallelepiped (that is, the numerator of the fraction on the right of (34)) can be represented as a sum of six terms corresponding to the six faces of the parallelepiped. Let us consider the sum of two of these terms that correspond to the back and front faces, we denote them by I and II respectively. Then $(A_n)_{II} = (A_x)_{II}$, $(A_n)_I = -(A_x)_I$ (why?), and for this reason the indicated sum can be written as

$$\left(\int A_n\,d\sigma\right)_I + \left(\int A_n\,d\sigma\right)_{II} = -\int (A_x)_I\,d\sigma + \int (A_x)_{II}\,d\sigma$$
$$= \int [(A_x)_{II} - (A_x)_I]\,d\sigma$$

To within higher order infinitesimals, the integrand is equal to $\partial_x A_x = \frac{\partial A_x}{\partial x}\,dx$. This is the "partial differential" of A_x with respect to x resulting from the fact that the points of the front face differ from the corresponding points of the back face by the value of the x-coordinate. For this reason the entire integral is, up to higher order terms, equal to

$$\frac{\partial A_x}{\partial x}\,dx \int d\sigma = \frac{\partial A_x}{\partial x}\,dx\,dy\,dz$$

Carrying out similar computations for the other two pairs of faces, we get the entire elementary flux:

$$dQ = \left(\frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}\right) dx\, dy\, dz$$

But since $dV = dx\, dy\, dz$, by the first formula of (34) we finally get

$$\operatorname{div} \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} \qquad (39)$$

In conclusion, note the following. From formula (33) it is readily apparent that the vector flux through a small closed surface is of the order not of the area of the surface (as one might expect from the surface integral) but of the volume bounded by it, that is, it is an infinitesimal of third order and not of second order compared with the linear dimensions. This can be explained as follows. The vector flux through an open surface is indeed of the second order. However, the flux of a constant vector through a closed surface is always equal to zero! (This property follows easily from the preceding material; prove it, for instance, with the aid of the last formula of Sec. 10.4 or of formula (35).) For this reason, if the field within a small closed surface is expanded by Taylor's formula (Sec. 6.6), then the integral of the first constant term yields zero, while the integral of first-order terms will be of the third order of smallness.

This remark does not refer to the case where, inside the closed surface, there are singularities of the field at which the volume density of the source becomes infinite. If there is a source of vector lines distributed over the area within a surface, the flux is of second order; if the source is distributed along a line, then the flux is of first order (proportional to the length of the line), finally, the flux through a surface inside which there is a point source (i.e. the divergence is of the nature of the delta function) is not an infinitesimal at all. It is precisely the case that we encountered above in the extremely important instance of a Coulomb field (see formula (38)).

Exercises

1. Use the Ostrogradsky formula to find the outward flux through the sphere $x^2 + y^2 + z^2 = R^2$ of the fields $\mathbf{A} = 2x\mathbf{i} + y\mathbf{k}$, $\mathbf{B} = y^2\mathbf{i} - z^2\mathbf{j}$.
2. Derive the formula (37) with the aid of (39).
3. Find the divergence of the plane-parallel axial symmetric field $\mathbf{A} = f(|\boldsymbol{\rho}|)\,\boldsymbol{\rho}^0$. In what case is this field devoid of sources outside the z-axis?

10.8 The divergence of a velocity field and the continuity equation

Let us now return to a consideration of the stationary flow of a gas (see Sec. 10.4); we assume that the mass of the gas does not change in the process. In Sec. 10.4 we established the fact that the flux of the velocity vector $\mathbf{v}$ through an oriented surface (σ) is equal to the volume of the gas carried outward through (σ) in unit time. Hence if the surface (σ) is closed, then this flux is equal to the difference between the volume emerging in unit time from (σ) over that portion of the surface where $\mathbf{v}$ is directed outwards, and the volume entering through (σ) over the same portion of surface where $\mathbf{v}$ is directed inwards.

Why can the vector flux $\mathbf{v}$ through a closed surface be different from zero? If the gas is taken to be incompressible during the flow process (such is the case for flow velocities considerably less than the speed of sound and also for liquids), then the incoming volume is equal to the outflow and so the indicated flux is zero. However, if (σ) is in a zone where the gas expands in its flow, then the outgoing volume is greater than the incoming volume and therefore the overall flux will be positive, similarly, in the compression zone the overall flux is negative.

It is now clear what the divergence of the velocity vector means. By virtue of Sec. 10.7, to obtain div $\mathbf{v}$ it is necessary to divide the flux of vector $\mathbf{v}$ through the surface of an infinitesimal volume $(d\Omega)$ by the numerical value $d\Omega$ of this volume. We thus have a ratio of the volume "originating" in $(d\Omega)$ in unit time to the volume $d\Omega$. It is natural to call this ratio the "rate of relative increase in volume". From this it is clear that the equation $\operatorname{div} \mathbf{v} = 0$ is the condition for incompressibility of a gas, whereas the relations $\operatorname{div} \mathbf{v} > 0$ and $\operatorname{div} \mathbf{v} < 0$ hold, respectively, in the zone of expansion and the zone of compression.

Consider the field of "mass velocity" $\rho\mathbf{v}$. In Sec. 10.4 we showed that the flux of such a vector through an oriented surface (σ) is equal to the mass of the gas carried outwards in unit time through (σ). But since the mass of the gas remains unchanged, the total flux of the vector $\rho\mathbf{v}$ through any closed surface is necessarily equal to zero, since the influx of mass is the same as the efflux. This also holds for an infinitesimal volume and so, by Sec. 10.7, we arrive at the equation

$$\operatorname{div} (\rho\mathbf{v}) = 0 \tag{40}$$

for stationary flow in which the mass of gas in every element of volume remains unchanged with time. This is the *continuity equation*, which is one of the basic relations of hydrodynamics that expresses the law of conservation of mass.

Let us write out this expression in coordinates. The vector $\rho\mathbf{v}$ clearly has the components ρv_x, ρv_y, ρv_z, and so

$$\begin{aligned}\operatorname{div}(\rho\mathbf{v}) &= \frac{\partial}{\partial x}(\rho v_x) + \frac{\partial}{\partial y}(\rho v_y) + \frac{\partial}{\partial z}(\rho v_z)\\ &= \rho\left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}\right) + v_x\frac{\partial\rho}{\partial x} + v_y\frac{\partial\rho}{\partial y} + v_z\frac{\partial\rho}{\partial z}\\ &= \rho\operatorname{div}\mathbf{v} + \mathbf{v}\cdot\operatorname{grad}\rho\end{aligned}$$

(this will be derived differently in Sec. 11.8). Hence, the condition for conservation of mass leads to the relation

$$\operatorname{div}\mathbf{v} = -\frac{1}{\rho}\mathbf{v}\cdot\operatorname{grad}\rho$$

For an incompressible fluid $\rho =$ constant, grad $\rho = 0$, div $\mathbf{v} = 0$. What is the quantity $\mathbf{v}\cdot\operatorname{grad}\rho$ in the stationary flow of a compressible fluid [such precisely is the flux we are considering, and for it, div $(\rho\mathbf{v}) = 0$]? Let us find the variation of density of a particle. In a stationary flux $\rho = \rho(x, y, z,)$ the density at every point does not depend on t, i.e. $\partial\rho/\partial t = 0$. But when we consider a *given particle*, it is necessary to take into account that its coordinates depend on the time. The velocity $\mathbf{v}$ is precisely what characterizes this relationship:

$$\frac{dx}{dt} = v_x,\quad \frac{dy}{dt} = v_y,\quad \frac{dz}{dt} = v_z$$

By the general rule (cf. formula (5) of Ch.4) and considering $\rho = \rho[x(t), y(t), z(t)] = \rho(t)$, we find

$$\frac{d\rho}{dt} = \frac{\partial\rho}{\partial x}\frac{dx}{dt} + \frac{\partial\rho}{\partial y}\frac{dy}{dt} + \frac{\partial\rho}{\partial z}\frac{dz}{dt} = \mathbf{v}\cdot\operatorname{grad}\rho$$

Thus

$$\operatorname{div}\mathbf{v} = -\frac{1}{\rho}\frac{d\rho}{dt}$$

This is the quantitative expression of the earlier statements concerning div $\mathbf{v}$ in a region where the gas is compressed $(d\rho/dt > 0)$ or expands $(d\rho/dt < 0)$ for $\frac{\partial\rho}{\partial t} = 0$.

Now let us consider a nonstationary gas flow when the velocity field and other fields vary with time. All the earlier introduced notions will have to be considered (as in the case of a nonstationary field with arbitrary physical meaning) for any fixed instant of time. Then the flow lines (i.e. lines which are constructed for any fixed time and at that time lie along the velocity vector at every point) no longer coincide with the particle paths of the liquid (give some thought to why this is so!).

If in the nonstationary case we consider a volume $(d\Omega)$, then the vector flux $\rho\mathbf{v}$ through the surface is equal, by (34), to $\operatorname{div}(\rho\mathbf{v})\, d\Omega$ and yields the rate of decrease in mass of gas in that volume, that is, the mass of gas emanating from this volume in unit time. However, the mass of gas contained in an infinitesimal volume $(d\Omega)$ is equal to $\rho\, d\Omega$ and for this reason increases* at the rate of

$$\frac{\partial}{\partial t}(\rho\, d\Omega) = \frac{\partial \rho}{\partial t}\, d\Omega$$

We thus get

$$\operatorname{div}(\rho\mathbf{v})\, d\Omega = -\frac{\partial \rho}{\partial t}\, d\Omega$$

whence

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho\mathbf{v}) = 0 \qquad (41)$$

This is the continuity equation, which for the nonstationary case as well signifies that in the process of motion the gas is conserved (does not appear or disappear). In the stationary case, equation (41) clearly becomes equation (40).

In the nonstationary case, the rate of change $\frac{d\rho}{dt}$ of density in a particle (in contrast to the rate of change $\frac{\partial \rho}{\partial t}$ of density in a fixed point) is equal to

$$\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t} + \mathbf{v} \cdot \operatorname{grad} \rho$$

From this we find that in the nonstationary case the formula

$$\operatorname{div} \mathbf{v} = -\frac{1}{\rho}\frac{d\rho}{dt}$$

remains unchanged.

Exercise

Consider a flux of "neutron gas" in a reactor, the neutrons entering the flux throughout the whole volume due to fission of uranium nuclei in the volume. How will the continuity equation change?

10.9 The divergence of an electric field and the Poisson equation

Let a charge with a certain density ρ be distributed in space and let it generate an electric field $\mathbf{E}$. In Sec. 10.5 we saw that the flux Q of the vector $\mathbf{E}$ through a surface of any volume (Ω), i.e.

[*] The word "increases" is to be understood in the algebraic sense: if $\partial\rho/\partial t > 0$, the mass inside $d\Omega$ increases, and if $\partial\rho/\partial t < 0$, the mass decreases.

the number of electric lines of force originating in (Ω), is equal to the product of 4π by the charge q contained in (Ω). From this, if the volume is infinitesimal, we get

$$\operatorname{div} \mathbf{E} = \frac{dQ}{d\Omega} = \frac{d(4\pi q)}{d\Omega} = 4\pi \frac{dq}{d\Omega}$$

or, finally,

$$\operatorname{div} \mathbf{E} = 4\pi\rho \tag{42}$$

Thus, the divergence of the intensity of an electric field, i.e. the density of origination of electric lines of force, is directly proportional to the density of distributed charges.

In Sec. 10.5 we saw that an electric field has a potential φ connected with the vector $\mathbf{E}$ by the relation (20). Substituting (20) into (42), we get the equation of the potential:

$$\operatorname{div} \operatorname{grad} \varphi = -4\pi\rho$$

The combination div grad is called the *Laplacian operator* and is denoted by Δ; we can therefore rewrite the equation for the potential as

$$\Delta\varphi = -4\pi\rho \tag{43}$$

This is the *Poisson equation*, its solution (23), equal to zero at infinity, was found earlier. In the portion of space free from charges, equation (43) becomes

$$\Delta\varphi = 0 \tag{44}$$

and is called the *Laplace equation*.

It is easy to obtain an expression for the Laplacian operator in Cartesian coordinates. From formulas (1) and (39) we get

$$\operatorname{div} \operatorname{grad} \varphi = \operatorname{div}\left(\frac{\partial\varphi}{\partial x}\mathbf{i} + \frac{\partial\varphi}{\partial y}\mathbf{j} + \frac{\partial\varphi}{\partial z}\mathbf{k}\right)$$
$$= \frac{\partial}{\partial x}\left(\frac{\partial\varphi}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial\varphi}{\partial y}\right) + \frac{\partial}{\partial z}\left(\frac{\partial\varphi}{\partial z}\right) = \frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} + \frac{\partial^2\varphi}{\partial z^2}$$

So equation (44) is the same Laplace equation we encountered in Sec. 5.7 but written for the three-dimensional case.

Sometimes it is more convenient to compute the Laplacian operator in other coordinate systems. For example, for a central-symmetric field $\varphi(r)$, we get, by formulas (5) and (36)-(37),

$$\Delta\varphi(r) = \operatorname{div} \operatorname{grad} \varphi(r)$$
$$= \operatorname{div}\left(\frac{d\varphi}{dr}\,\frac{\mathbf{r}}{r}\right) = \frac{2}{r}\left(\frac{d\varphi}{dr}\right) + \left(\frac{d\varphi}{dr}\right)'_r = \frac{d^2\varphi}{dr^2} + \frac{2}{r}\,\frac{d\varphi}{dr}$$

The Laplacian operator is connected by a simple formula with the mean value of the function on a small sphere. To derive this relationship, set $\mathbf{A} = \text{grad}\ \varphi$ in the Ostrogradsky formula (35). We then have

$$\oint\limits_{(\sigma)} \text{grad}_n\, \varphi\, d\sigma = \int\limits_{(\Omega)} \text{div grad}\, \varphi\, d\Omega, \text{ or } \oint\limits_{(\sigma)} \frac{d\varphi}{dn}\, d\sigma = \oint\limits_{(\Omega)} \Delta\varphi\, d\Omega$$

Suppose that $(\sigma) = (\sigma)_r$ is a sphere of radius r with centre at some point O. Then $d\sigma = r^2\, d\omega$, where $d\omega$ is an element of the solid angle with vertex O (Sec. 10.5), whence

$$\oint\limits_{(\sigma)_r} \frac{d\varphi}{dr}\, r^2\, d\omega = \oint\limits_{(\Omega)_r} \Delta\varphi\, d\Omega, \quad \text{or} \quad \frac{d}{dr}\oint \varphi\, d\omega = \frac{1}{r^2}\int\limits_{(\Omega)_r} \Delta\varphi\, d\Omega$$

(We passed to integration with respect to ω so that the range of integration should not be dependent on the parameter r and so that it would be possible to apply the rule of Sec. 3.6 for differentiating an integral with respect to a parameter.) Integrating this last equation with respect to r, beginning with $r = 0$, we get

$$\oint \varphi\, d\omega - \oint \varphi\, d\omega \Big|_{r=0} = \int_0^r \frac{ds}{s^2} \int\limits_{(\Omega)_s} \Delta\varphi\, d\Omega$$

But $\varphi\,|_{r=0} = \varphi(O) = \text{constant}$, $\oint d\omega = 4\pi$ (why?), whence, dividing by 4π, we get

$$\varphi(O) = \frac{1}{4\pi r^2}\int\limits_{(\sigma)_r} \varphi\, d\sigma - \int_0^r \frac{ds}{4\pi s^2}\int\limits_{(\Omega)_s} \Delta\varphi\, d\Omega \qquad (45)$$

We will derive certain consequences from this formula. First of all, suppose that $\Delta\varphi \equiv 0$ in some region including the sphere $(\Omega)_r$; then the function φ is called a *harmonic function* (cf. Sec. 5.7) and by (44)

$$\varphi(O) = \frac{1}{4\pi r^2}\int\limits_{(\sigma)_r} \varphi\, d\sigma \qquad (46)$$

Thus, the value of the harmonic function at the centre of a sphere is equal to the mean value of this function on the sphere.

From this it follows in particular that a harmonic function cannot have any maxima or minima inside the domain of its definition. Indeed, if, say, there were a minimum at O and r is sufficiently small, then the mean value of the function φ on $(\sigma)_r$ would be greater than $\varphi(O)$ (why?), but this contradicts (46). This property is demonstrated differently in HM, Sec. 6.3, where a corollary to it was also noted: a charge in an electrostatic field in a region free of other charges cannot have positions of stable equilibrium.

Another consequence of formula (45) is obtained for an arbitrary function φ if r is taken to be small. Then $(\Delta\varphi)_{(\Omega)_s} \approx \Delta\varphi\,|_O$, whence

$$\int_0^r \frac{ds}{4\pi s^2}\int_{(\Omega)_s}\Delta\varphi\,d\Omega \approx \int_0^r \frac{ds}{4\pi s^2}\left(\Delta\varphi\,|_O\cdot\frac{4}{3}\,\pi s^3\right) = \frac{1}{6}\,r^2\,\Delta\varphi\,|_O$$

For this reason, to within quantities that are small for small r,

$$\Delta\varphi\,|_O = \frac{6}{r^2}\left[\frac{1}{4\pi r^2}\int_{(\sigma)_r}\varphi\,d\sigma - \varphi(O)\right] = 6\,\frac{\overline{\varphi}^{(\sigma)_r} - \varphi(O)}{r^2} \qquad (47)$$

where $\overline{\varphi}^{(\sigma)_r}$ signifies the mean over the sphere $(\sigma)_r$. This then shows that, in particular, if $\overline{\varphi}^{(\sigma)_r} \equiv \varphi(O)$, that is to say, if the above described property of the mean over spheres holds, then $\Delta\varphi = 0$, which means the function φ is harmonic.

Formula (47) is very reminiscent of the formula

$$\varphi''(x) = \frac{2}{h^2}\left[\frac{\varphi(x+h)+\varphi(x-h)}{2} - \varphi(x)\right]$$

which is readily verified with the aid of the Taylor formula for small h for the function $\varphi(x)$ of one variable (do this!). In many cases, formula (47) makes it possible to get a deeper understanding of the meaning of the relations that involve the Laplacian operator. For example, in the theory of heat conduction the equation $\frac{\partial\nu}{\partial t} =$ $=$ constant $\cdot\,\Delta\nu$ is derived for the propagation of heat in a homogeneous medium, where ν is the temperature. By (47) this means that if the temperature around the point O is, on the average, higher than at O, then the temperature at O must increase.

Exercises

1. What is $\Delta\varphi(|\boldsymbol{\rho}|)$ equal to, where $\boldsymbol{\rho} = x\mathbf{i} + y\mathbf{j}$ (Sec. 10.6)?
2. Prove that the value of a harmonic function at the centre of a sphere of any radius is equal to the mean value of the function on the sphere.
3. Prove formula (47) by expanding the function φ in a Taylor series.

10.10 An area vector and pressure

In Sec. 10.4 we said that an oriented plane area is customarily depicted by a vector. This approach is particularly natural when investigating forces of pressure in a fluid. Indeed, suppose the pressure is p at some point. This is of course a scalar quantity.* Suppose

* The fluid obeys the Pascal law; we do not consider either strength or viscosity, which lead to distinct forces acting on areas that have different directions.

an oriented element of surface $(d\sigma)$ is located at this point (see Fig. 120). Then the force exerted on this element by a liquid located on the inside of the surface is directed outward along the normal to $(d\sigma)$ and is numerically equal to $p\,d\sigma$; this force is thus equal to

$$d\mathbf{F} = p\,d\boldsymbol{\sigma} \tag{48}$$

The vector of the area is thus interpreted in a straightforward manner.

It is now easy to compute the overall force which the liquid exerts on an arbitrarily chosen volume (Ω) with surface (σ). To do this, we have to sum the elementary forces (48) but with the sign reversed on the right side, since the surface (σ) will, as in the preceding sections, be considered to have the inner side facing (Ω) and not facing the pressing liquid. We then get

$$\mathbf{F} = -\int\limits_{(\sigma)} p\,d\boldsymbol{\sigma} \tag{49}$$

When computing the integral (49) it is often convenient to pass to integration over (Ω). Note for this purpose that

$$p\,d\boldsymbol{\sigma} = p\mathbf{n}\,d\sigma = p[\cos(\widehat{\mathbf{n}, x})\,\mathbf{i} + \cos(\widehat{\mathbf{n}, y})\,\mathbf{j} + \cos(\widehat{\mathbf{n}, z})\mathbf{k}]\,d\sigma \tag{50}$$

(see end of Sec. 9.1). On the other hand, $p\cos(\widehat{\mathbf{n}, x}) = (p\mathbf{i})_n$, where the subscript n denotes the projection on the outer normal. For this reason, by virtue of the Ostrogradsky formula (35) and formula (39) for computing the divergence, we get

$$\oint\limits_{(\sigma)} p\cos(\widehat{\mathbf{n}, x})\,d\sigma = \oint\limits_{(\sigma)} (p\mathbf{i})_n\,d\sigma = \int\limits_{(\Omega)} \operatorname{div}(p\mathbf{i})\,d\Omega = \int\limits_{(\Omega)} \frac{\partial p}{\partial x}\,d\Omega$$

and the integral of the first term on the right of (50) turns out equal to $\mathbf{i}\int\limits_{(\Omega)} \frac{\partial p}{\partial x}\,d\Omega$. The integrals of the other two terms are transformed in similar fashion and as a result we get

$$\int\limits_{(\sigma)} p\,d\boldsymbol{\sigma} = \int\limits_{(\Omega)} \left(\frac{\partial p}{\partial x}\mathbf{i} + \frac{\partial p}{\partial y}\mathbf{j} + \frac{\partial p}{\partial z}\mathbf{k}\right) d\Omega = \int\limits_{(\Omega)} (\operatorname{grad} p)\,d\Omega \tag{51}$$

(see formula (1)). The force of the pressure is

$$\mathbf{F} = -\int\limits_{(\Omega)} (\operatorname{grad} p)\,d\Omega \tag{52}$$

Let us consider some examples. First let the pressure be constant, $p = \text{constant}$. Then $\operatorname{grad} p = \mathbf{0}$ and formula (52) shows that $\mathbf{F} = \mathbf{0}$ as well.

The conclusion is that a constant pressure applied to all sides of a body of arbitrarily complex shape yields a resultant force of zero. It is precisely for this reason that the atmospheric pressure, which is very strong but acts with constant force on the complicated surface of our body, does not push us out of the atmosphere. Actually, this example was examined at the end of Sec. 10.4; in particular, we see that if the polyhedron shown in Fig. 121 is put in a liquid, then

$$p(-\boldsymbol{\sigma}) + p(\boldsymbol{\sigma}_1) + p(\boldsymbol{\sigma}_2) + p(\boldsymbol{\sigma}_3) = 0$$

Cancelling out the p, we arrive at formula (18).

Let us now consider the pressure in a liquid located in a field of gravitation. If the z-axis is directed upwards from the free surface, then it is a well-known fact that the pressure in the liquid depends on z in the following manner:

$$p = p_0 - g\rho z$$

where p_0 is the pressure on the free surface,* ρ is the density of the liquid, and g is the acceleration of gravity. From this, by (52), we get the buoyant force

$$\mathbf{F} = \int\limits_{(\Omega)} g\rho\mathbf{k}\, d\Omega = g\rho\Omega\mathbf{k}$$

Since the product $g\rho\Omega$ is equal to the weight of the liquid in the volume (Ω) we arrive at the famous Archimedean law.

Let us return to the general case and find the resultant of forces of pressure applied to an infinitesimal volume $(d\Omega)$. Since in the case of a constant pressure this resultant is, as we have just demonstrated, equal to zero, then by reasoning as we did at the end of Sec. 10.7 we can derive that although the forces of pressure are applied to the *surface*, the resultant is proportional not to the surface area but to the *volume*. This is also evident from formula (52), whence it follows that for an infinitesimal volume of integration

$$d\mathbf{F} = -(\operatorname{grad} p)\, d\Omega \qquad (53)$$

To the volume $(d\Omega)$ there can be applied certain volumetric forces (say, forces of gravitation, centrifugal forces, and the like) distributed with density **f**. The resultant of such forces is

$$d\mathbf{F}_1 = \mathbf{f}\, d\Omega$$

Both of these forces impart to the mass $\rho\, d\Omega$ an acceleration of $\frac{d\mathbf{v}}{dt}$, or

$$\rho\, d\Omega \frac{dv}{dt} = -(\operatorname{grad} p)\, d\Omega + \mathbf{f}\, d\Omega$$

* Note that in the liquid we obviously have $z < 0$, so that the pressure p in the formula is greater than p_0, as it should be.

whence, cancelling $d\Omega$, we get the equation of motion:

$$\rho \frac{d\mathbf{v}}{dt} = -\operatorname{grad} p + \mathbf{f} \tag{54}$$

In the left-hand member we have the rate of change of the vector $\mathbf{v}$ along the path of the particle. This rate is written down in the same way as was done in Sec. 4.1 for the case of a scalar field (that is, it consists of a local rate and a convective rate). Its coordinate representation will be obtained in Sec. 11.8.

Let us consider the stationary flow of an incompressible (i.e. for ρ = constant) liquid in a gravitational field. Directing the z-axis vertically upwards, we find that $\mathbf{f}$ (i.e. the force acting on unit volume) equals $-\rho g\mathbf{k}$. Forming the scalar product of both sides of equation (54) by $d\mathbf{r} = \mathbf{v}\, dt$, we get on the left

$$\rho\, d\mathbf{v} \cdot \mathbf{v} = \frac{\rho}{2} d(\mathbf{v} \cdot \mathbf{v}) = \frac{\rho}{2} d(v^2) \quad (\text{here, } v = |\mathbf{v}|)$$

On the right side, the first term after multiplying yields

$$-\operatorname{grad} p \cdot d\mathbf{r} = -\left(\frac{\partial p}{\partial x} dx + \frac{\partial p}{\partial y} dy + \frac{\partial p}{\partial z} dz\right) = -dp$$

and the second term, $-\rho g\mathbf{k} \cdot d\mathbf{r} = -\rho g\, dz$. Thus,

$$\frac{\rho}{2} d(v^2) = -dp - \rho g\, dz$$

It is to be stressed that this relation holds true along flow lines; we made use of this fact when we wrote $d\mathbf{r} = \mathbf{v}\, dt$, since this equation determines the displacement of a moving particle along the trajectory (see Sec. 9.4). Integrating along such a curve, we get the relation

$$\frac{\rho v^2}{2} + p + \rho g z = \text{constant} \tag{55}$$

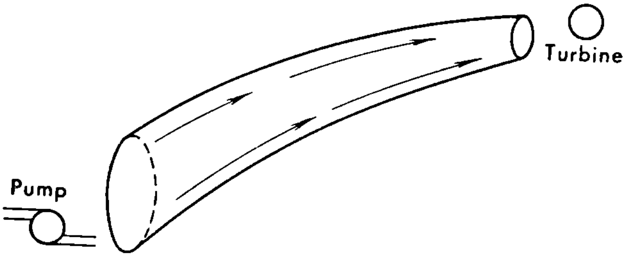
which should hold true in a stationary flux along every flow line. Here the constant of integration on the right assumes distinct values on distinct flow lines. The relation (55) is called the *Bernoulli integral* and expresses the law of conservation of energy in the motion of a particle of liquid. This is particularly clear if we multiply both sides by $d\Omega$ and take the increment Δ (do not confuse this with the Laplacian operator!) of both sides of the equation along a certain segment of the flow line. We then have

$$\Delta \frac{\rho v^2}{2} d\Omega + \Delta p\, d\Omega + \Delta \rho g z\, d\Omega = 0$$

or

$$\Delta\left(\frac{dm \cdot v^2}{2} + dm \cdot gz\right) = -\Delta p \cdot d\Omega$$

Fig. 133



Since in the brackets we have the sum of the kinetic energy and the potential energy of the moving particle, the left-hand side is equal to the increment in total energy of that particle. The right side may be represented in the form of an integral along the flow line:

$$-\int \frac{dp}{ds}\, ds \cdot d\Omega = -\int (\text{grad } p)_\tau\, ds \cdot d\Omega = \int F_\tau\, ds$$

(see formula (53)). Hence, this right-hand side is equal to the work of the forces of pressure during the motion of the particle. Thus, formula (55) means that the increment in the total energy of a particle is equal to the work of the forces acting on it.

In a stationary flux, the law of conservation of energy can be applied in yet another way. Consider a *stream tube,* which is a tube through the lateral surface of which the flow lines neither enter nor leave. Such a tube is shown schematically in Fig. 133. The walls may be solid since no work is performed on them and the fluid does not pass through them. We compare the power (i.e. the energy per unit time) of a pump required to deliver fluid to the tube with the power of a turbine at the output of the tube. These powers must be equal if we disregard losses to friction. In unit time, v_1S_1 units of volume of fluid enter through the entry opening into the tube, where S is the cross-sectional area of the tube and the subscript 1 indicates input. This fluid has the kinetic energy

$$\rho v_1 S_1 \frac{v_1^2}{2}$$

Besides, the pressure p_1 is created in the volume v_1S_1 by the power $p_1v_1S_1$ of the pump, the power $\rho v_1S_1gz_1$ is used to raise the liquid from some standard altitude $z = 0$ to an altitude z_1. Setting up similar expressions for the output cross-section and equating the sums, we get the following notation for the law of conservation of energy:

$$\rho v_1 S_1 \frac{v_1^2}{2} + p_1 v_1 S_1 + \rho v_1 S_1 g z_1 = \rho v_2 S_2 \frac{v_2^2}{2} + p_2 v_2 S_2 + \rho v_2 S_2 g z_2 \quad (56)$$

But because of conservation of mass and by virtue of the incompressibility of the fluid, we have $v_1S_1 = v_2S_2$. Cancelling this quantity out of (56), we get

$$\frac{\rho v_1^2}{2} + p_1 + \rho g z_1 = \frac{\rho v_2^2}{2} + p_2 + \rho g z_2$$

This equation is equivalent to the Bernoulli integral (55).

From the Bernoulli integral it is quite apparent that for a constant or slightly variable altitude, the pressure in the stream is substantially dependent on the rate of flow: the higher the rate, the smaller the pressure. Which is quite natural: if a particle passes from a region with low pressure into a region with higher pressure, then the resultant of the forces of pressure acting on it is directed counter to the velocity, which means the motion is hampered. Now if the pressure falls with the motion, then the pressure "behind" a particle is greater than that "in front", and the particle will be accelerated. By (55), when a definite velocity has been attained, the pressure should become negative; actually this does not occur, the liquid loses its continuity and cavities appear (what is called "cavitation" sets in).

Exercise

Using formula (51) derive an invariant (unrelated to any choice of coordinate system) definition for gradient that is similar to the definition (33) of divergence.

ANSWERS AND SOLUTIONS

Sec. 10.1

It is easy to verify that the relationship between the old coordinates x, y, z and the new ones x', y', z' is $x = x'$, $y = \frac{y' - z'}{\sqrt{2}}$, $z = \frac{y' + z'}{\sqrt{2}}$. Substituting these formulas into the expression for u, we get $u = x'^2 + 2y'^2$. Since z' does not appear, the field is plane-parallel.

Sec. 10.2

1. Since grad $u = y\mathbf{i} + x\mathbf{j} - 2z\mathbf{k}$, it follows that $(\text{grad } u)_M = \mathbf{i} + 2\mathbf{j} + 6\mathbf{k}$. The desired derivative is equal to $\frac{\mathbf{a}}{|\mathbf{a}|} \cdot \text{grad}\, u = \frac{19}{10} = 6.0$.

2. $\text{grad } u(M) = -\frac{\overrightarrow{M_1M}}{M_1M} - \frac{\overrightarrow{M_2M}}{M_2M}$.

Sec. 10.3

$\mathbf{F} = -\text{grad } U = -\alpha y\mathbf{i} - \alpha x\mathbf{j}$. The lines of force are determined by the equation $\frac{dx}{y} = \frac{dy}{x}$, whence $x\,dx = y\,dy$, $x^2 = y^2 + C$. These are hyperbolas with centre at the origin (Fig. 33).

Sec. 10.4

$Q = 2\pi |\boldsymbol{\rho}| \int\limits_a^b f(|\boldsymbol{\rho}|, z)\, dz$, where $z = a$ and $z = b$ are the planes of the bases of the cylinder. In the case of a plane-parallel field, f does not depend on z and $Q = 2\pi |\boldsymbol{\rho}| f(|\boldsymbol{\rho}|)\, h$, where h is the height of the cylinder.

Sec. 10.6

1. Integrating the expression for dE_x indicated in the text, we get (via the substitution $\zeta - z = \sqrt{x^2 + y^2} \tan \alpha$)

$$E_x = \int\limits_{-\infty}^{\infty} \frac{\mu x\, d\zeta}{[x^2 + y^2 + (z - \zeta)^2]^{3/2}}$$

$$= \int\limits_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{\mu x \sqrt{x^2 + y^2}\, \dfrac{d\alpha}{\cos^2\alpha}}{(x^2 + y^2)^{3/2}\, \dfrac{1}{\cos^3\alpha}} = \frac{\mu x}{x^2 + y^2} \int\limits_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos \alpha\, d\alpha = \frac{2\mu x}{x^2 + y^2}$$

Similarly we find $E_y = \dfrac{2\mu y}{x^2 + y^2}$, $E_z = 0$, whence

$$\mathbf{E} = \frac{2\mu x}{x^2 + y^2} \mathbf{i} + \frac{2\mu y}{x^2 + y^2} \mathbf{j} = \frac{2\mu \boldsymbol{\rho}}{|\boldsymbol{\rho}|^2} = \frac{2\mu}{|\boldsymbol{\rho}|} \boldsymbol{\rho}^0$$

The result coincides with (26).

2. Taking the charged plane for the yz-plane, we get $E_y = E_z = 0$ at the point $(x, 0, 0)$ for $x > 0$, whereas

$$E_x = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} \frac{\nu x}{(x^2 + \eta^2 + \zeta^2)^{3/2}}\, d\eta\, d\zeta$$

Passing to the polar coordinates ρ, φ in the η, ζ plane, as at the end of Sec. 4.7, we get

$$E_x = \int\limits_0^{2\pi} d\varphi \int\limits_0^{\infty} \frac{\nu x \rho\, d\rho}{(x^2 + \rho^2)^{3/2}} = 2\pi \nu x \int\limits_0^{\infty} \frac{\rho\, d\rho}{(x^2 + \rho^2)^{3/2}}$$

$$= 2\pi\, \nu x \left. \frac{-1}{(x^2 + \rho^2)^{1/2}} \right|_{\rho=0}^{\infty} = 2\pi\nu$$

From this $\mathbf{E} = 2\pi\nu\mathbf{i}$, which coincides with (29).

3. Arguing as at the end of Sec. 10.6, we obtain (for the natural meaning of the letters)

$$\mathbf{E} = 2m\frac{d}{dl}\left(\frac{\boldsymbol{\rho}}{|\boldsymbol{\rho}|^2}\right) = \frac{2m}{|\boldsymbol{\rho}|^2}\left(\frac{2\boldsymbol{\rho}}{|\boldsymbol{\rho}|}\cos\alpha - \mathbf{l}^0\right),$$

$$\varphi = 2m\frac{d}{dl}(-\ln|\boldsymbol{\rho}|) = 2m\frac{\cos\alpha}{|\boldsymbol{\rho}|}$$

It can be verified that the traces of cylindrical equipotential surfaces and also the electric lines of force are located as in Fig. 130, but here both will be circles.

Sec. 10.7

1. By formula (39), div $\mathbf{A} = 2$, div $\mathbf{B} = 0$. Therefore by the Ostrogradsky formula the desired flux is equal to

$$\int\limits_{(\Omega)} \operatorname{div}\mathbf{A}\,d\Omega = \int\limits_{(\Omega)} 2\,d\Omega = 2\Omega = \frac{8}{3}\pi R^3, \qquad \int\limits_{(\Omega)} \operatorname{div}\mathbf{B}\,d\Omega = 0$$

2. Since $\mathbf{A} = f(r)\,\mathbf{r}^0 = \frac{f(r)}{r}\,\mathbf{r}$, it follows that $A_x = \frac{f(r)x}{r}$,

$$\frac{\partial A_x}{\partial x} = \frac{\left[f'(r)\,\frac{x}{r}\,x + f(r)\right]r - f(r)\,x\,\frac{x}{r}}{r^2} = \frac{f'(r)}{r^2}\,x^2 + f(r)\,\frac{r^2 - x^2}{r^3}$$

In similar fashion we find $\frac{\partial A_y}{\partial y}$, $\frac{\partial A_z}{\partial z}$, whence

$$\operatorname{div}\mathbf{A} = \frac{f'(r)}{r^2}(x^2 + y^2 + z^2) + f(r)\,\frac{3r^2 - x^2 - y^2 - z^2}{r^3}$$
$$= f'(r) + \frac{2f(r)}{r}$$

3. Reasoning as in the derivation of formula (37), we obtain

$$\operatorname{div}\mathbf{A} = \frac{d[2\pi|\boldsymbol{\rho}|\,f(|\boldsymbol{\rho}|)]}{2\pi|\boldsymbol{\rho}|\;d|\boldsymbol{\rho}|} = f'(|\boldsymbol{\rho}|) + \frac{1}{|\boldsymbol{\rho}|}f(|\boldsymbol{\rho}|)$$

Here div $\mathbf{A} = 0$ if $f(|\boldsymbol{\rho}|) = \frac{C}{|\boldsymbol{\rho}|}$. This is the field (26) considered in Sec. 10.6.

Sec. 10.8

$\frac{\partial\rho}{\partial t} + \operatorname{div}(\rho\mathbf{v}) = f$, where $f = f(x, y, z, t) = \frac{dM}{d\Omega\,dt}$ and dM is the mass of neutrons entering the flux in the volume $(d\Omega)$ during time dt.

Sec. 10.9

1. $\varphi''(|\boldsymbol{\rho}|) + \frac{2}{|\boldsymbol{\rho}|}\varphi'(|\boldsymbol{\rho}|)$.
2. Partition the sphere $(\Omega)_r$ into infinitely thin "bubbles" of radius s and thickness $ds (0 \leqslant s \leqslant r)$. Then by formula (46)

$$\int\limits_{(\Omega)_r} \varphi\, d\Omega = \int\limits_0^r \Big(\int\limits_{(\sigma)_s} \varphi\, d\sigma \Big) ds = \int\limits_0^r 4\pi s^2 \varphi(O)\, ds = \frac{4}{3}\pi r^3 \varphi(O)$$

whence follows the required assertion.
3. Let O be the origin. For reasons of symmetry, we have

$$\int x\, d\sigma = \int y\, d\sigma = \int z\, d\sigma = \int xy\, d\sigma = \int xz\, d\sigma = \int yz\, d\sigma = 0$$

(all integrals are taken over $(\sigma)_r$); besides, $\int x^2\, d\sigma = \int y^2\, d\sigma = \int z^2\, d\sigma$, whence $\int x^2\, d\sigma = \frac{1}{3}\int (x^2 + y^2 + z^2)\, d\sigma = \frac{1}{3} r^2 \cdot 4\pi r^2 = \frac{4}{3}\pi r^4$. Applying an expansion of type (27) of Ch. 4 with $a = b = 0$, we get, to within fifth-order infinitesimals,

$$\int\limits_{(\sigma)_r} \varphi\, d\sigma = \varphi(O) \cdot 4\pi r^2 + \frac{4}{3}\pi r^4 \cdot \frac{1}{2}\left(\frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} + \frac{\partial^2\varphi}{\partial z^2} \right)_O$$

whence formula (47) readily follows.

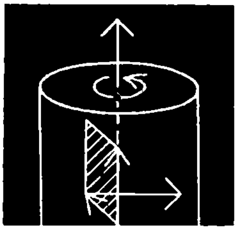
Sec. 10.10

$$\operatorname{grad} p(M) = \frac{\oint\limits_{(d\sigma)} p\, d\sigma}{d\Omega} = \lim_{(\Delta\Omega)\to M} \frac{\oint\limits_{(\Delta\sigma)} p\, d\sigma}{\Delta\Omega} \qquad (57)$$

Note that this formula holds true for any scalar field and not only for the pressure field.

Chapter 11

VECTOR PRODUCT AND ROTATION*



11.1 The vector product of two vectors

In vector algebra, besides the multiplication of a vector by a scalar (Sec. 9.1) and the scalar product of two vectors (Sec. 9.2) there is also defined a vector product of two vectors, which we now discuss.

Recall (Sec. 10.4) that a surface in space is said to be oriented if we indicate which side is the outside and which the inside. It is customary to assume that if this surface is open (has a boundary), then the orientation of the surface also generates an orientation of its contour, that is to say, a sense of traversal of the contour (boundary). Conversely, if the sense of traversal of the contour is indicated, then this leads to an orientation of the surface. The connection between the orientation of a surface and the orientation of its contour is indicated in Fig. 134. If for the basis of a system of coordinates we take the right triad of vectors **i**, **j**, **k** (such that when looking from the end of the third vector we see the shortest rotation from the first to the second counterclockwise), then we use the rule of the right-handed screw, otherwise we have the rule of the left-handed screw. For example the rule of the right-handed screw can be stated thus: if a right-handed screw (which is the one usually used in engineering and ordinary life) is rotated in the sense of traversal of the contour, then the screw will move from the inside of the surface to the outside. This can be put differently: if a tiny man walks on the outside of the surface along the contour (boundary) in the indicated direction of traversal (see Fig. 134), then the surface itself must be always on the left.**

* This chapter is a direct continuation of the two preceding chapters and relies heavily on the material of Secs. 9.1 and 9.2 and Secs. 10.1-10.4 and 10.7. Besides, we make use of the multiple integral (Sec. 4.7), determinants (Sec. 8.3) and the delta function (Sec. 6.1). In Sec. 11.12, use is made of the material of Secs. 10.5 to 10.9.

** The first three items of Fig. 134 refer to the right-handed coordinate system. From the drawing it is clear that if we turn the right-handed screw from **i** to **j**, screwing it out of the plate, the screw itself will move in the direction of **k**. The traversal of the contour in this sense corresponds to the underside of the surface (σ) being the inside, the upper side, the outside; the vector $\boldsymbol{\sigma}$ is directed along **k**.

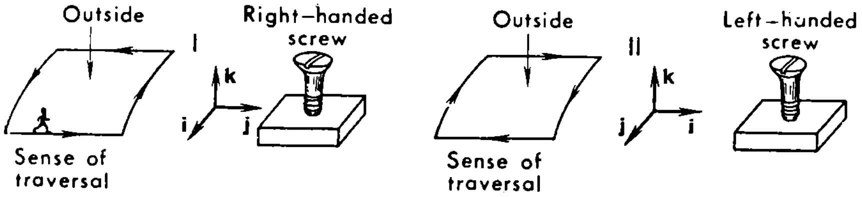


Fig. 134

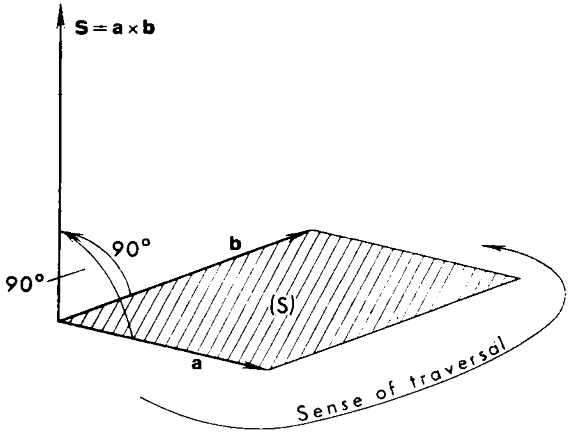


Fig. 135

Let us now take up the notion of a vector product. The *vector product* of two vectors **a** and **b** is, by definition, the vector **S** of an area (S) (see Sec. 10.4) that is obtained if **a** and **b** are referred to a single origin, a parallelogram is then constructed on these vectors, and the contour is traversed beginning with the first vector (i.e. **a**; see Fig. 135 where the rule of the right-handed screw is used; this rule will always be employed unless otherwise stated).

To summarize, the vector product of two vectors **a** and **b** is a vector directed perpendicularly to the two vectors and with modulus equal to the area of a parallelogram constructed on **a** and **b** and forming with these vectors a triad of the same sense (i.e. right-handed or left-handed) as the vectors **i**, **j**, **k**. A vector product (also sometimes called cross product) is denoted by $\mathbf{a}\times\mathbf{b}$ or $[\mathbf{ab}]$.

A few of the most important properties of a vector product are: The vector product of two nonzero vectors is equal to the zero vector (null vector) if and only if the vectors are parallel:

$$\mathbf{a}\times\mathbf{b} = \mathbf{0} \text{ is equivalent to } \mathbf{a}\parallel\mathbf{b}$$

since parallelism of the vectors amounts to the parallelogram degenerating into a line segment with area equal to zero. In particular, it is always true that $\mathbf{a}\times\mathbf{a} = \mathbf{0}$.

A vector product is anticommutative[*]

$$\mathbf{b}\times\mathbf{a} = -(\mathbf{a}\times\mathbf{b})$$

Indeed, if the order of the factors is reversed, the parallelogram remains unchanged, but the contour is traversed in the opposite sense and therefore the vector of the area is reversed.

It can be verified that a scalar factor can be taken outside the sign of a vector product:

$$(\lambda\mathbf{a})\times\mathbf{b} = \mathbf{a}\times(\lambda\mathbf{b}) = \lambda(\mathbf{a}\times\mathbf{b})$$

and that the distributive law holds:

$$(\mathbf{a}+\mathbf{b})\times\mathbf{c} = \mathbf{a}\times\mathbf{c}+\mathbf{b}\times\mathbf{c}, \quad \mathbf{c}\times(\mathbf{a}+\mathbf{b}) = \mathbf{c}\times\mathbf{a}+\mathbf{c}\times\mathbf{b}$$

When multiplying out expressions containing a vector product, watch carefully the order of the factors. For example,

$$(\mathbf{a}+2\mathbf{b})\times(2\mathbf{a}-3\mathbf{b}) = 2\mathbf{a}\times\mathbf{a}-3\mathbf{a}\times\mathbf{b}+4\mathbf{b}\times\mathbf{a}$$
$$-6\mathbf{b}\times\mathbf{b} = -7\mathbf{a}\times\mathbf{b}$$

Suppose the vectors **a** and **b** are given in expansions in terms of their Cartesian projections:

$$\mathbf{a} = a_x\mathbf{i}+a_y\mathbf{j}+a_z\mathbf{k}, \quad \mathbf{b} = b_x\mathbf{i}+b_y\mathbf{j}+b_z\mathbf{k}$$

Then, taking advantage of the following equations (verify them!),

$$\mathbf{i}\times\mathbf{j} = \mathbf{k},\ \mathbf{j}\times\mathbf{i} = -\mathbf{k},\ \mathbf{j}\times\mathbf{k} = \mathbf{i},\ \mathbf{k}\times\mathbf{j} = -\mathbf{i},$$
$$\mathbf{k}\times\mathbf{i} = \mathbf{j},\ \mathbf{i}\times\mathbf{k} = -\mathbf{j} \tag{1}$$

we get

$$\begin{aligned}\mathbf{a}\times\mathbf{b} &= (a_x\mathbf{i}+a_y\mathbf{j}+a_z\mathbf{k})\times(b_x\mathbf{i}+b_y\mathbf{j}+b_z\mathbf{k})\\ &= a_xb_y\mathbf{k}-a_xb_z\mathbf{j}-a_yb_x\mathbf{k}+a_yb_z\mathbf{i}+a_zb_x\mathbf{j}-a_zb_y\mathbf{i}\\ &= \mathbf{i}(a_yb_z-a_zb_y)+\mathbf{j}(a_zb_x-a_xb_z)+\mathbf{k}(a_xb_y-a_yb_x)\end{aligned} \tag{2}$$

(think through the structure of the last expression).

This result is easy to remember if one writes it in the form of a determinant (see formula (13) of Ch. 8):

$$\mathbf{a}\times\mathbf{b} = \begin{vmatrix}\mathbf{i} & \mathbf{j} & \mathbf{k}\\ a_x & a_y & a_z\\ b_x & b_y & b_z\end{vmatrix} \tag{3}$$

[*] Generally, an operation on any entities whatsoever is said to be commutative if it is independent of the order in which these entities are taken. The multiplication of ordinary numbers and the scalar multiplication of vectors are commutative.

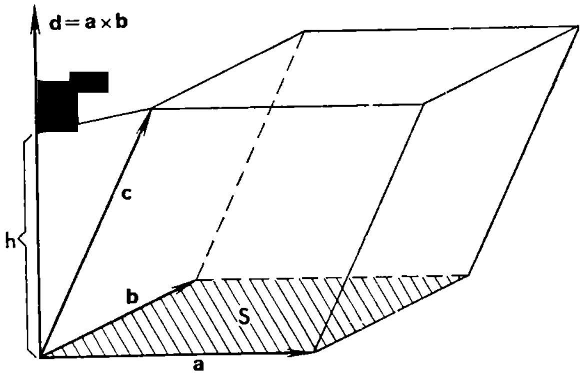


Fig. 136

Suppose we have to compute the area S of a parallelogram constructed on the vectors $\mathbf{a} = 3\mathbf{i} - 2\mathbf{j} + \mathbf{k}$ and $\mathbf{b} = -2\mathbf{i} + \mathbf{j} + 4\mathbf{k}$. Since $S = |\mathbf{a} \times \mathbf{b}|$, we calculate as follows:

$$\mathbf{S} = \mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 3 & -2 & 1 \\ -2 & 1 & 4 \end{vmatrix}$$

$$= \mathbf{i}(-8 - 1) + \mathbf{j}(-2 - 12) + \mathbf{k}(3 - 4) = -9\mathbf{i} - 14\mathbf{j} - \mathbf{k},$$

$$S = |\mathbf{a} \times \mathbf{b}| = \sqrt{9^2 + 14^2 + 1^2} = 16.7$$

Since in this example the vectors **a**, **b** are nondimensional, the area S is nondimensional as well.

Also sometimes used is the *scalar triple product* of three vectors **a**, **b**, **c**, which, by definition, is equal to the scalar quantity $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}$. The geometric meaning of this product is evident from Fig. 136:

$$(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{d} \cdot \mathbf{c} = |\mathbf{d}|\, c_d = |\mathbf{a} \times \mathbf{b}|\, c_d = Sh = V$$

Thus, the scalar triple product of three vectors is equal to the volume of a parallelepiped constructed on these vectors. In Fig. 136 the vectors **a**, **b**, **c** form a right-handed triad and we have a volume with the plus sign. For a left-handed triad the angle between **c** and **d** would be obtuse; in this case $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = -V$. (We assume here the rule of the right-handed screw.) A scalar triple product is equal to zero if and only if all three vectors are parallel to a single plane, because such parallelism means that the parallelogram has degenerated into a part of the plane and thus has zero volume.

It is easy to obtain an expression for a scalar triple product if the expansions of the factors are given in a Cartesian system of coordinates. For this purpose we have to multiply the right side of (2) by the vector $\mathbf{c} = c_x\mathbf{i} + c_y\mathbf{j} + c_z\mathbf{k}$ in the customary way to get a scalar product (via formula (5) of Ch. 9). After regrouping, this yields

$$(\mathbf{a}\times\mathbf{b})\cdot\mathbf{c} = (a_yb_z - a_zb_y)\,c_x + (a_zb_x - a_xb_z)\,c_y + (a_xb_y - a_yb_x)\,c_z$$

$$= a_x(b_yc_z - b_zc_y) - a_y(b_xc_z - b_zc_x) + a_z(b_xc_y - b_yc_x) = \begin{vmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{vmatrix}$$

We will also need a formula for the *vector triple product* $\mathbf{a}\times(\mathbf{b}\times\mathbf{c})$ of three vectors. To derive it, suppose that we have chosen the coordinate axes so that the x-axis is along the vector $\mathbf{b}$ and the y-axis lies in the plane of the vectors $\mathbf{b}$ and $\mathbf{c}$. Then the vector $\mathbf{b}$ will have a projection on the x-axis alone, or $\mathbf{b} = b_x\mathbf{i}$; similarly, $\mathbf{c} = c_x\mathbf{i} + c_y\mathbf{j}$, $\mathbf{a} = a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k}$. Using formula (3), we get

$$\mathbf{b}\times\mathbf{c} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ b_x & 0 & 0 \\ c_x & c_y & 0 \end{vmatrix} = b_xc_y\mathbf{k}, \quad \mathbf{a}\times(\mathbf{b}\times\mathbf{c}) = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_x & a_y & a_z \\ 0 & 0 & b_xc_y \end{vmatrix}$$

$$= a_yb_xc_y\mathbf{i} - a_xb_xc_y\mathbf{j}$$

This result is inconvenient in that it is "attached" to a special choice of coordinate axes. We therefore transform it (verify this):

$$\mathbf{a}\times(\mathbf{b}\times\mathbf{c}) = (a_xc_x + a_yc_y)\,b_x\mathbf{i} - a_xb_x(c_x\mathbf{i} + c_y\mathbf{j})$$

$$= \mathbf{b}(\mathbf{a}\cdot\mathbf{c}) - \mathbf{c}(\mathbf{a}\cdot\mathbf{b}) \qquad (4)$$

This formula no longer contains the coordinate projections and for this reason is independent of any choice of coordinate system. To remember this formula, use the mnemonic "abc equals bac minus cab".

Exercises

1. Find the area of a triangle with vertices

   $A(1, 0, -2)$, $B(-1, 1, 2)$, $C(1, 3, 3)$

2. What kind of triad is formed by the vectors

   $\mathbf{a} = 2\mathbf{i} - \mathbf{j}$, $\mathbf{b} = 3\mathbf{i} + 2\mathbf{k}$, $\mathbf{c} = -\mathbf{i} - \mathbf{j} + \mathbf{k}$?

3. Compute $(\mathbf{i}\times\mathbf{i})\times\mathbf{j}$ and $\mathbf{i}\times(\mathbf{i}\times\mathbf{j})$ and show that the vector produc does not possess the associative property.

11.2 Some applications to mechanics

A vector product is particularly convenient for describing rotational motion and its associated notions. Let us consider the rotation of a rigid body about a certain axis (Fig. 137) with angular velocity ω. Such rotation is customarily represented by the angular velocity

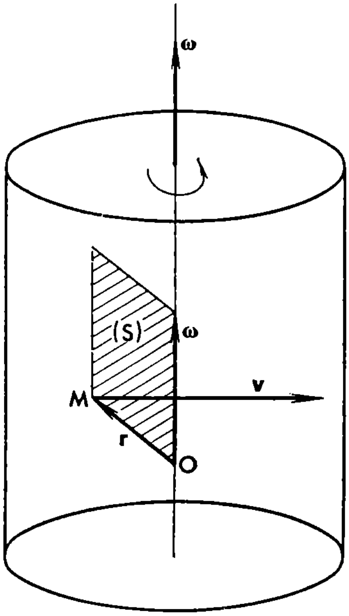


Fig. 137

vector **ω** located on the axis of rotation and directed in accordance with the sense of rotation via with the chosen rule of the screw; in Fig. 137 the sense of **ω** is chosen by the rule of the right-handed screw, as we will do in all cases. It is immaterial where precisely the vector **ω** is taken, for it is a *sliding vector*, that is, one that can be arbitrarily moved along an axis but not away from the axis .*

We assume that the origin O is chosen at any point on the axis of rotation and we seek the linear velocity **v** of an arbitrary point M with radius vector **r** (Fig. 137). It is obvious that **v** is perpendicular to both vectors **r** and **ω** and is therefore perpendicular to the entire parallelogram (S) constructed on these vectors. The numerical value of **v** is equal to the product of ω by the shortest distance between M and the axis of rotation, which is precisely the area of the indicated parallelogram. But these conditions, which are formulated for the vector **v**, are satisfied by the vector product $\boldsymbol{\omega}\times\mathbf{r}$. Thus

$$\mathbf{v} = \boldsymbol{\omega}\times\mathbf{r} \tag{5}$$

(check to see that the vector product must be taken in the order written and that the right-hand side of (5) is independent of any specific choice of point O on the axis of rotation).

* Recall that, earlier, the point of application of a vector was not fixed and it was assumed possible to have a parallel displacement of a vector to any point.

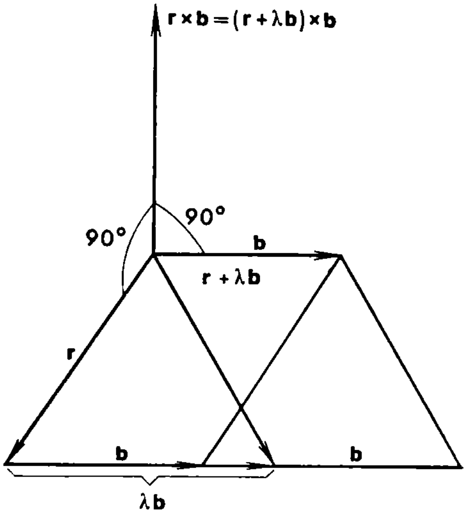


Fig. 138

The convenience of the vector of angular velocity is clearly seen in the following. Suppose that a body experiences two rotations at the same time (generally, nonparallel) with angular velocity vectors $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ the axes of rotation intersecting at the point O. Then by virtue of (5) the linear velocity of any point M is

$$\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2 = \boldsymbol{\omega}_1 \times \mathbf{r} + \boldsymbol{\omega}_2 \times \mathbf{r} = (\boldsymbol{\omega}_1 + \boldsymbol{\omega}_2) \times \mathbf{r} = \boldsymbol{\omega} \times \mathbf{r}$$

where $\boldsymbol{\omega} = \boldsymbol{\omega}_1 + \boldsymbol{\omega}_2$. Hence the body rotates with angular velocity $\boldsymbol{\omega}$ and we come to the conclusion that in a composition of rotations the vectors of angular velocity combine via the parallelogram law. It is precisely for this reason that we can call the angular velocity a vector.

Using the vector product, we can introduce such an important notion as the *moment of an arbitrary vector* $\mathbf{b}$ with origin at the point M relative to any point O: by definition, this moment is equal to $\mathbf{r} \times \mathbf{b}$, where $\mathbf{r} = \overrightarrow{OM}$. In mechanics one mostly has to do with the moment of a force $\mathbf{F}$, that is, the quantity $\mathbf{r} \times \mathbf{F}$, and the moment of the momentum $m\mathbf{v}$, that is, the quantity $m\mathbf{r} \times \mathbf{v}$.

When computing the moment of a vector, the vector may be regarded as sliding. If the vector $\mathbf{b}$ slides along itself, then this means that to $\mathbf{r}$ is added $\lambda\mathbf{b}$, where λ is a scalar. However,

$$(\mathbf{r} + \lambda\mathbf{b}) \times \mathbf{b} = \mathbf{r} \times \mathbf{b} + (\lambda\mathbf{b}) \times \mathbf{b} = \mathbf{r} \times \mathbf{b} + \mathbf{0} = \mathbf{r} \times \mathbf{b}$$

which means that the moment of the vector remains unchanged under such a sliding (Fig. 138). But if the vector is carried away from its direction, the moment changes.

Let us consider a system of particles connected in some manner, each of which has a constant mass m_i and a (generally variable) radius vector $\mathbf{r}_i = \mathbf{r}_i(t)$. Suppose each of these points is acted upon by distinct forces; we denote the resultant of all "internal" forces (the forces of interaction between the points of the system) applied to the ith particle by $\mathbf{F}_i^{\text{in}}$ and the resultant of all "external" forces by $\mathbf{F}_i^{\text{ex}}$. The peculiarity of the internal forces is that, on the basis of Newton's third law ("to every action there is an equal and opposite reaction"), for every internal force there is equal and an opposite internal force and such — this is very important — as to be located on the extension of the first force. For this reason the sum of all internal forces and also the sum of their moments about any point are equal to zero.

The basic equations of motion of a system of particles are obtained if we write Newton's second law ("force equals mass times acceleration"):

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{F}_i^{\text{ex}} + \mathbf{F}_i^{\text{in}} \tag{6}$$

Summing these equations over all particles, we get

$$\sum_i m_i \frac{d^2\mathbf{r}_i}{dt^2} = \sum_i \mathbf{F}_i^{\text{ex}} + \sum_i \mathbf{F}_i^{\text{in}} = \sum_i \mathbf{F}_i^{\text{ex}} \tag{7}$$

since the sum of all internal forces, as has already been mentioned, is equal to zero. It is convenient to introduce a point C with radius vector

$$\mathbf{r}_C = \frac{1}{M}\sum_i m_i\mathbf{r}_i, \quad \text{where } M = \sum_i m_i$$

is the total mass of the system. This point is called the *centre of mass* of system at hand. In this notation, equation (7) can be rewritten as

$$\frac{d^2}{dt^2}\sum_i m_i\mathbf{r}_i = \sum_i \mathbf{F}_i^{\text{ex}}, \text{ or } M\frac{d^2\mathbf{r}_C}{dt^2} = \sum_i \mathbf{F}_i^{\text{ex}}$$

Thus, the centre of mass moves as if it possessed the total mass of the system and were acted upon by a force equal to the sum of all external forces. In particular, if external forces are absent, then the centre of mass of a system is in rectilinear and uniform motion, $\mathbf{r}_C = \mathbf{a}t + \mathbf{b}$.

Now let us take up moments. If both sides of (6) are multiplied vectorially on the left by $\mathbf{r}_i$, we get

$$m_i\mathbf{r}_i \times \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{r}_i \times \mathbf{F}_i^{\text{ex}} + \mathbf{r}_i \times \mathbf{F}_i^{\text{in}} \tag{8}$$

Take advantage of the equation

$$\frac{d}{dt}\left(\mathbf{r}_i \times \frac{d\mathbf{r}_i}{dt}\right) = \frac{d\mathbf{r}_i}{dt} \times \frac{d\mathbf{r}_i}{dt} + \mathbf{r}_i \times \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{r}_i \times \frac{d^2\mathbf{r}_i}{dt^2}$$

which follows from the general formula of the derivative of a vector product. This formula, $\frac{d}{dt}(\mathbf{A}\times\mathbf{B}) = \frac{d\mathbf{A}}{dt}\times\mathbf{B} + \mathbf{A}\times\frac{d\mathbf{B}}{dt}$, is derived in exactly the same way as the similar formula (8) of Ch. 9 for a scalar product. From this we can rewrite (8) thus

$$\frac{d}{dt}(\mathbf{r}_i \times m_i\mathbf{v}_i) = \mathbf{r}_i \times \mathbf{F}_i^{\text{ex}} + \mathbf{r}_i \times \mathbf{F}_i^{\text{in}}$$

Summing these equations with respect to all i, we get

$$\frac{d}{dt}\sum_i (\mathbf{r}_i \times m_i\mathbf{v}_i) = \sum_i \mathbf{r}_i \times \mathbf{F}_i^{\text{ex}} + \sum_i \mathbf{r}_i \times \mathbf{F}_i^{\text{in}} = \sum_i \mathbf{r}_i \times \mathbf{F}_i^{\text{ex}} \tag{9}$$

since the sum of the moments of all internal forces is zero. The sum

$$\mathbf{G} = \sum_i \mathbf{r}_i \times m_i\mathbf{v}_i \tag{10}$$

of the angular momenta of all particles comprising a system is termed the *angular momentum* of this system with respect to the same point O, relative to which all the angular momenta are taken. The sum

$$\mathbf{L} = \sum_i \mathbf{r}_i \times \mathbf{F}_i^{\text{ex}}$$

of the moments of all external forces acting on the system is called the *total moment of the external forces.* Thus, formula (9) can be rewritten as

$$\frac{d\mathbf{G}}{dt} = \mathbf{L} \tag{11}$$

which is to say that the rate of change of the angular momentum of a system is equal to the total moment of the external forces acting on the system. In the special case of the absence of external forces or if their total moment is equal to zero, we find that the angular momentum of the system remains constant.

Exercise

When is the moment of a vector **b** about a point O equal to zero?

11.3 Motion in a central-force field

Suppose a particle is in motion acted upon by a force varying in arbitrary fashion but all the time directed towards (or away from) the coordinate origin O. This is called a *central-force field.* If at the initial time we pass a plane (P) through the origin and through the velocity vector, the point in its motion will not leave the plane since there are no forces capable of carrying the point out of (P). (The plane (P) coincides with the plane of Fig. 139.) Since the

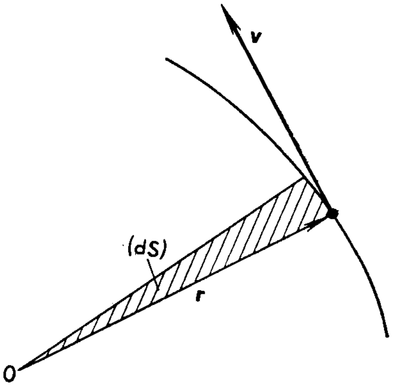


Fig. 139

external force is directed along a radius vector, its moment is equal to zero and for this reason the angular momentum of the point, $\mathbf{r}\times m\dfrac{d\mathbf{r}}{dt}$, remains constant all the time. However, the modulus of the vector $\mathbf{r}\times d\mathbf{r}$ is equal to double the area of the triangle (dS) shown in Fig. 139 (why is this?). For this reason, as the point moves, the derivative dS/dt remains constant, which means that S varies linearly with respect to time. We thus arrive at *Kepler's second law:* in the case of a central force, the area described by the radius vectors of a planet in equal times are equal. The other laws of Kepler (to be discussed later on) make essential use of the specific form of dependence of force on the length of the radius vector.*

Let $\mathbf{F} = -F(r)\,\mathbf{r}^0$. By virtue of Sec. 10.2, such a field has a potential $U(r)$, where the function $U(r)$ is the primitive of $F(r)$, i.e. $\dfrac{dU}{dr} = F(r)$. Let us introduce the polar coordinates r, φ in the (P) plane. Then the law of motion of a particle will be determined by the relations $r = r(t)$, $\varphi = \varphi(t)$.

To find these relations, we take advantage of two conservation laws: that of angular momentum (Sec. 11.2) and of total energy (Sec. 10.3). For this purpose, denote by $\mathbf{s}^0$ the vector obtained from $\mathbf{r}^0$ by a rotation through 90° in the positive sense. Then

$$\frac{d\mathbf{r}^0}{dt} = \frac{d\varphi}{dt}\,\mathbf{s}^0$$

* Kepler discovered his law empirically in studying the motions of the planets of the solar system. It was Newton who demonstrated that these laws are consequences of a definite type of gravitational force with which a central body, the sun, acts on the planets.

(This rather obvious formula is a special case of formula (14) of Ch. 9 and actually coincides with formula (10) of Ch. 5.) From this we have

$$\frac{d\mathbf{r}}{dt} = \frac{d}{dt}(r\mathbf{r}^0) = \frac{dr}{dt}\mathbf{r}^0 + r\frac{d\varphi}{dt}\mathbf{s}^0, \quad \mathbf{r}\times m\frac{d\mathbf{r}}{[dt} = mr^2\frac{d\varphi}{dt}\mathbf{r}^0\times\mathbf{s}^0,$$

$$\frac{m}{2}\left(\frac{d\mathbf{r}}{dt}\right)^2 = \frac{m}{2}\left[\left(\frac{dr}{dt}\right)^2 + \left(r\frac{d\varphi}{dt}\right)^2\right] \tag{12}$$

Since the vector $\mathbf{r}^0\times\mathbf{s}^0 = \mathbf{P}^0$ is perpendicular to the (P) plane and has a constant modulus equal to 1, the law of conservation of the angular momentum can be written as

$$mr^2\frac{d\varphi}{dt} = G \ (= \text{constant}) \tag{13}$$

The law of conservation of total energy becomes, by virtue of (12),

$$\frac{m}{2}\left[\left(\frac{dr}{dt}\right)^2 + \left(r\frac{d\varphi}{dt}\right)^2\right] + U(r) = E \ (= \text{constant}) \tag{14}$$

Here the constants G and E are determined by the initial conditions.

Express $\frac{d\varphi}{dt}$ from (13) and substitute it into (14) to get the first-order differential equation

$$\frac{m}{2}\left(\frac{dr}{dt}\right)^2 + \left[U(r) + \frac{G^2}{2mr^2}\right] = E \tag{15}$$

for $r(t)$: having found $\mathbf{r}(t)$, we can obtain $\varphi(t)$ from equation (13).

Equation (15) is quite analogous to the equation

$$\frac{m}{2}\left(\frac{dx}{dt}\right)^2 + u(x) = E \tag{16}$$

of the motion of particle m along the x-axis under the action of a force with potential $u(x)$. Let us recall the simple properties of the solutions of (16) which are considered, for example, in HM, Sec. 6.8. Suppose that the graph of the potential $u(x)$ is of the form given in Fig. 140 so that $u(\infty) = u_\infty$. Then for $u_{\min} < E < u_\infty$, say for $E = E_1$ in Fig. 140, the particle will oscillate periodically having a and b as cusps; we say that the motion is *finite*. The period of oscillations, as can readily be derived from (16), is

$$T = \sqrt{2m}\int_a^b \frac{dx}{\sqrt{E - u(x)}} \tag{17}$$

But if $E \geqslant u_\infty$, say for $E = E_2$ in Fig. 140, the particle moving leftwards will reach point c and turn back, but in its rightward motion it will go off to infinity, and the motion becomes *infinite*.

Fig. 140

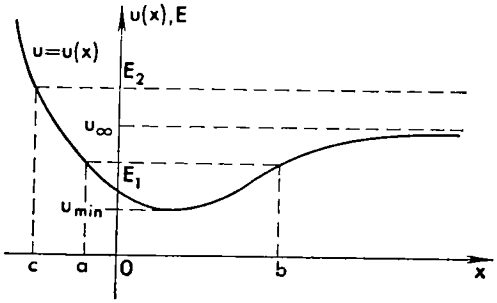
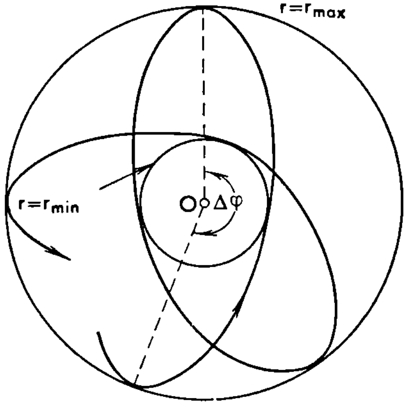


Fig. 141



Equation (15) takes the form (16) if we write x instead of $\mathbf{r}$ (which is inessential) and set

$$u(r) = U(r) + \frac{G^2}{2mr^2} \quad (0 < r < \infty) \tag{18}$$

Thus, besides the initial potential $U(r)$ there is added a *centrifugal potential* $\frac{G^2}{2mr^2}$, which depends on the initial conditions, more exactly, on the angular momentum G of the system at hand.

And so if $u_{\min} < E < u_{\infty}$ $(= U_{\infty})$, the relation $r(t)$ will be periodic. During each period of variation of r the polar angle receives one and the same increment $\Delta\varphi$. Since $\Delta\varphi$ is, generally, incommensurable with 2π, the trajectory will, as a rule, have the form of a rosette (Fig. 141) and in subsequent motion will everywhere fill in the annulus between the circles $r = r_{\min}$ and $r = r_{\max}$.

However, for the two most interesting types of central forces that we now proceed to discuss, the trajectories turn out to be closed

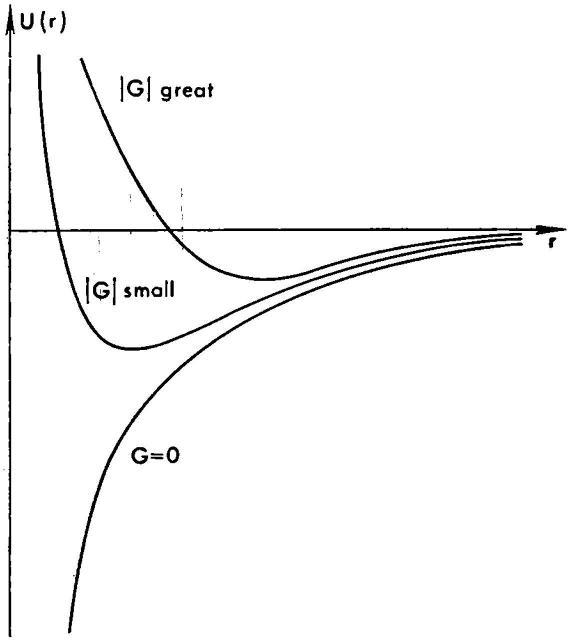


Fig. 142

paths without self-intersections — what is more, they are simply ellipses! In the first example we assume that the force is inversely proportional to the square of the distance of the particle from the origin. Such is the law of motion of celestial bodies (Newton's law) or of electrically charged particles (Coulomb's law) if the central body that gives rise to the force of attraction can be regarded as at rest (we will go into this condition later on).

In the example at hand, $F(r) = \frac{k}{r^2}$, $U(r) = -\frac{k}{r}$, and so by (18)

$$u(r) = -\frac{k}{r} + \frac{G^2}{2mr^2}$$

Fig. 142 shows the graph of this *effective potential* for $k > 0$ (as we shall assume it to be from now on) and for distinct values of G. It is clear that if $G \neq 0$, i.e. if the motion does not degenerate into rectilinear motion, then the centrifugal potential as $r \to 0$ builds up faster than the attractive potential, and so $u(+0) = \infty$. This means that a moving particle cannot fall onto the central body, at least if we assume it to be sufficiently small and thus if we disregard the possibility that the particle will "fly into" it by passing too close to the origin.

Since $\frac{du}{dr} = \frac{k}{r^2} - \frac{G^2}{mr^3}$, the minimum of the effective potential is attained when

$$\frac{k}{r^2} - \frac{G^2}{mr^3} = 0, \text{ whence } r = \tilde{r} = \frac{G^2}{mk} \text{ and } u_{\min} = -\frac{mk^2}{2G^2}.$$

Hence the solution $r \equiv \frac{G^2}{mk}$ is possible; corresponding to this solution is particle motion in a circle with centre at the origin. The corresponding angular velocity is determined from equation (13):

$$\frac{d\varphi}{dt} = \frac{G}{m\tilde{r}^2} = \pm\sqrt{\frac{k}{m\tilde{r}^3}} \tag{19}$$

It is constant, i.e. we obtain a uniform revolution of the particle about the central body. If a gravitational field is considered, then $k = \varkappa mM$, where $\varkappa$ is the gravitational constant and M is the mass of the central body. From this, by (19), we get the period of revolution:

$$T = \frac{2\pi}{d\varphi/dt} = 2\pi\sqrt{\frac{m\tilde{r}^3}{\varkappa mM}} = \frac{2\pi}{\sqrt{\varkappa M}}\tilde{r}^{3/2}$$

This is *Kepler's third law*: the squares of the periods of revolution of the planets about the sun are proportional to the cubes of their distances from the sun. Here the proof is given only for circular orbits, but with a proper refinement of the statement it proves to be valid for all orbits (see Exercise 1).

To determine the form of noncircular paths for $G \neq 0$, substitute into (15) $U(r) = -k/r$ and complete the square to get (verify this!)

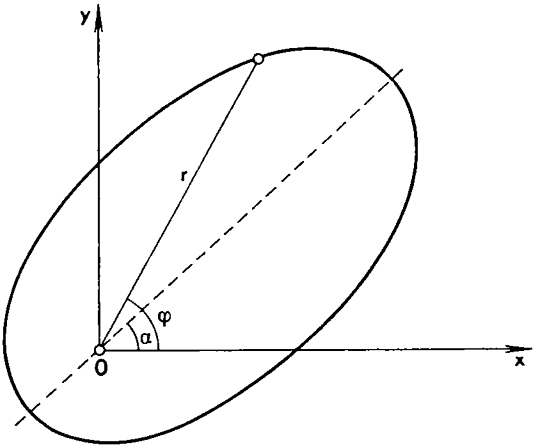
$$\left(\frac{dr}{dt}\right)^2 + \left(\frac{k}{G} - \frac{G}{mr}\right)^2 = q^2 \tag{20}$$

where for the sake of brevity we set $q = \left[\frac{2}{m}\left(E + \frac{mk^2}{2G^2}\right)\right]^{1/2}$. Here we take into account the inequality $E > -\frac{mk^2}{2G^2}$ (what does it follow from?). By (20) we can set

$$\frac{k}{G} - \frac{G}{mr} = q\cos\psi, \quad \frac{dr}{dt} = q\sin\psi, \quad \text{where } \psi = \psi(t) \tag{21}$$

Differentiating th3 first equation of (21), with respect to t, we get $(G/mr^2)\frac{dr}{dt} = -q\sin\psi\frac{d\psi}{dt}$, whence, taking into account the second equation of (21) and also (13), we find that $\frac{d\psi}{dt} = -\frac{d\varphi}{dt}$, or $\psi = \alpha - \varphi$

Fig. 143



where α = constant is determined by the initial conditions. From the first equation of (21) we get

$$r = \frac{G^2}{km}\left[1 - \frac{Gq}{k}\cos(\alpha - \varphi)\right]^{-1} = p[1 - \varepsilon\cos(\varphi - \alpha)]^{-1} \qquad (22)$$

where

$$p = \frac{G^2}{km}, \quad \varepsilon = \frac{Gq}{k}$$

From (22) it is apparent at once that for $\varepsilon < 1$ the trajectory is close to a closed path (Fig. 143), whereas for $\varepsilon \geqslant 1$ it goes off to infinity. It is easy to verify that the condition $\varepsilon < 1$ is equivalent to the inequality $E < 0$; thus (see Fig. 142) the closed lines serve as paths of bounded motions. To be more exact, the equation (22) for $\varepsilon < 1$ defines an ellipse with focus at the origin and major axis inclined at an angle α to the polar axis. (We leave it to the reader to prove this assertion by going over to Cartesian coordinates via the formulas $x' = \rho\cos(\varphi - \alpha)$, $y' = \rho\sin(\varphi - \alpha)$.) Thus, the trajectories of bounded motions are ellipses with one of the foci in the central body; therein lies *Kepler's first law*. In similar fashion it can be verified that for $\varepsilon > 1$ the motion occurs along a hyperbola, and in the boundary case of $\varepsilon = 1$, along a parabola.

It is also possible to prove that, conversely, from the Keplerian laws follows the law of gravitation $F(r) = k/r^2$. For this reason, motions that obey such a law are termed *Keplerian*.

Up to this point we have assumed the central body to be firmly "attached" to the coordinate origin. The motion of the central body

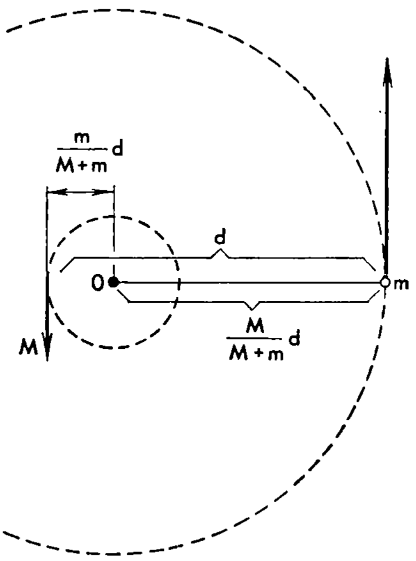


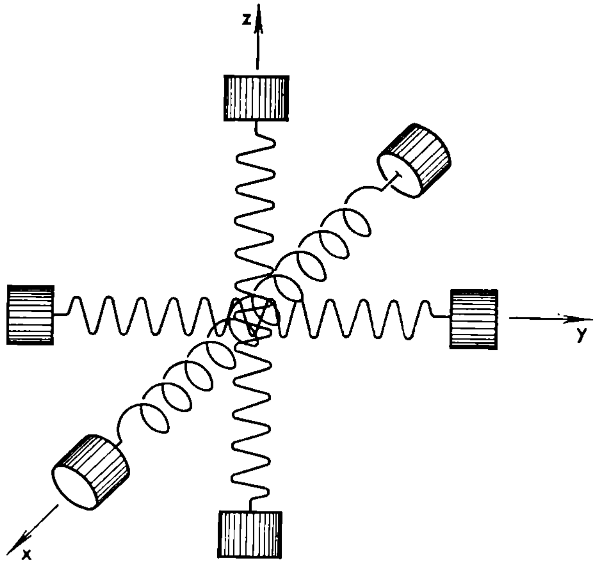
Fig. 144

can be neglected if its mass M is substantially greater than the moving mass m, $M \gg m$. It turns out, though we will not give the proof here, that the case of comparable masses M and m leads to what we have already discussed, but the second-order curves along which both masses move will then have the focus at the centre of mass of the whole system. It is interesting to note that the distribution, between the two bodies, both of the values of the kinetic energy and also the angular momenta about the centre of mass is *inversely proportional* to the masses of the bodies. Indeed, for the sake of simplicity, let us confine ourselves to the case of bodies moving in circles centred at the centre of mass O with constant angular velocity ω and denote by d the (constant) distance between the bodies (Fig. 144). Then the kinetic energy and the angular momentum of the mass M are equal, respectively, to $\frac{M}{2}\left(\frac{\omega md}{M+m}\right)^2$ and $\frac{md}{M+d}\,\frac{M\omega md}{M+m}$, whereas for the mass m these quantities are equal to $\frac{m}{2}\left(\frac{\omega Md}{M+m}\right)^2$ and $\frac{Md}{M+m}\,\frac{m\omega Md}{M+m}$, whence the inverse proportionality follows.

We leave it to the reader to consider the case of repulsion via Coulomb's law, i.e. $k < 0$. Note however that for $G \neq 0$ the trajectories are hyperbolas.

As another instance of a central force we consider the force $F =$ $= -k\mathbf{r}(k > 0)$ that is proportional to the distance of the particle

Fig. 145



from the origin. Here the potential $U(r) = \frac{k}{2} r^2$. This potential can be written in terms of Cartesian coordinates as

$$U = \frac{k}{2}(x^2 + y^2 + z^2)$$

By virtue of Sec. 4.6, such is the form of the principal (quadratic) portion of the expansion of the potential U in powers of x, y, z about the point of its minimum in the isotropic case, i.e. when the equipotential surfaces are, to within infinitesimals of higher order, not ellipsoids but spheres. Therefore the potential under consideration describes small oscillations of a particle about an arbitrary position of its stable equilibrium (without singularities of the potential) in an isotropic case. A case of this kind is, for example, realized in the scheme of Fig. 145 if all the springs have the same rigidity. From Sec. 10.6 it follows that such a potential arises also inside a homogeneous attractive sphere. We can imagine that the particle is moving inside a channel cut in accord with the computed path in the sphere, the channel being so narrow that its effect on the gravitational field can be disregarded.

If for the xy-plane we take the plane in which the particle moves, then the differential equation of motion in this instance can be written thus:

$$m\frac{d^2}{dt^2}(x\mathbf{i} + y\mathbf{j}) = -kr\mathbf{r}^0 = -k\mathbf{r} = -k(x\mathbf{i} + y\mathbf{j})$$

or, in terms of projections,

$$m\frac{d^2x}{dt^2}+kx=0,\quad m\frac{d^2y}{dt^2}+ky=0$$

Thus, in Cartesian coordinates the variables are separated, i.e. the oscillations along both axes occur independently of one another. By Sec. 7.3, the law of motion becomes

$$x=r_1\cos(\omega t+\varphi_1),\quad y=r_2\cos(\omega t+\varphi_2) \tag{23}$$

where $\omega=\sqrt{k/m}$ and the constants r_1, φ_1, r_2, φ_2 are determined by the initial conditions. We leave it up to the reader to be convinced that the equations (23) determine the particle trajectory as being an ellipse with centre at the origin.

Thus, for the force $F(r)=kr$ all paths again turn out to be closed, and the period $2\pi/\omega=2\pi\sqrt{m/k}$ of oscillations of the particle does not depend on the initial conditions. From this it follows that if, say, we dug a well through the centre of the earth and pumped out all the air, then the oscillation period of a stone thrown into the well would not depend on the amplitude and would prove to be equal to the period of revolution of a satellite at the earth's surface (air resistance disregarded).

Of course, in the second example above, the closed nature of the path was due to the fact that the variables were separated in the Cartesian axes and the oscillations along both axes were of the same period. But the surprising and remarkable thing is that in the first example, where the variables are not separated, the paths proved to be closed too! This is something in the nature of a mathematical marvel. Indeed, the form of the law $F=k/r^2$ was obtained in Sec. 10.7 directly from the condition div $\mathbf{F}(r)=0$ $(r\neq 0)$, which is natural from the standpoint of physics. In contrast, the closed nature of the paths under such a law of attraction is not at all obvious and was proved with the aid of artificial mathematical computations. This "marvel" substantially simplified the measurement and thus the analysis of comparatively small deviations of the actual trajectories from elliptic form (for instance, the motion of the perihelion of Mercury; as it will be recalled, the explanation of this motion, which follows from the general theory of relativity, served as one of the first decisive confirmations of this theory).

Exercises

1. Find the oscillation period of a particle about ellipse (22) and establish the relation of this period to the semimajor axis of the ellipse.
2. Find the period of variation of the radius of a particle for the law of attraction $\mathbf{F}=-k\mathbf{r}$ with the aid of formula (17) and explain the discrepancy with the result obtained in this section.

11.4 Rotation of a rigid body

Consider the rotation of a rigid body (Ω) about a fixed axis (the z-axis). The angular momentum $\mathbf{G}$ of this body about some point O on the axis of rotation is readily connected with the angular velocity $\boldsymbol{\omega} = \omega\mathbf{k}$ of rotation via formula (5):

$$\mathbf{G} = \int\limits_{(\Omega)} \mathbf{r}\times\mathbf{v}\,dm = \int\limits_{(\Omega)} \mathbf{r}\times(\omega\mathbf{k}\times\mathbf{r})\,\rho\,d\Omega = \omega\int\limits_{(\Omega)} \rho\mathbf{r}\times(\mathbf{k}\times\mathbf{r})\,d\Omega$$

where ρ denotes the (generally variable) density of the body. Recalling that $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$, we find, by (4),

$$\begin{aligned}\mathbf{r}\times(\mathbf{k}\times\mathbf{r}) &= (\mathbf{r}\cdot\mathbf{r})\,\mathbf{k} - (\mathbf{r}\cdot\mathbf{k})\,\mathbf{r} = (x^2+y^2+z^2)\,\mathbf{k} - z(x\mathbf{i}+y\mathbf{j}+z\mathbf{k})\\ &= -xz\mathbf{i} - yz\mathbf{j} + (x^2+y^2)\,\mathbf{k}\end{aligned}$$

whence

$$G_x = -\omega\int\limits_{(\Omega)} \rho xz\,d\Omega,\quad G_y = -\omega\int\limits_{(\Omega)} \rho yz\,d\Omega,$$

$$G_z = \omega\int\limits_{(\Omega)} \rho(x^2+y^2)\,d\Omega$$

The last integral, (Ω)

$$I_z = \int\limits_{(\Omega)} \rho(x^2+y^2)\,d\Omega \tag{24}$$

is called the *moment of inertia* of the body (Ω) about the z-axis, so that $G_z = \omega I_z$.

Suppose the body (Ω) is acted upon by certain specified forces $\mathbf{F}_i^{\text{ex}}$ and the forces $\mathbf{F}_j^{\text{sup}}$ of reaction in the supports, which are not given beforehand. If the latter forces are applied to the axis of rotation, then their moment $\mathbf{r}^{\text{sup}}\times\mathbf{F}_j^{\text{sup}} = z^{\text{sup}}\mathbf{k}\times\mathbf{F}_j^{\text{sup}}$ is perpendicular to the z-axis. To avoid considering these unknown forces, let us project (11) on the z-axis to get

$$\frac{dG_z}{dt} = \frac{d(\omega I_z)}{dt} = L_z \tag{25}$$

In the formation of the right side, i.e. the projection of the total moment of the external forces on the axis of rotation, we have to take into account only the specified forces $\mathbf{F}_i^{\text{ex}}$. We have thus passed from the vector equation (11) to the scalar equation (25).

Let us consider for instance the case where the external forces $\mathbf{F}_i^{\text{ex}}$ are absent or are directed parallel to the axis of rotation. Then $L_z = 0$ and from (25) we get $\omega I_z = \text{constant}$. If the body does not change under these circumstances, then $I_z = \text{constant}$, and for this reason $\omega =$ constant, which means the angular velocity of rotation remains unchanged. (Here we disregard the friction at the points of support, which friction leads to an additional term in the right-hand member of (25) and, as a consequence, to a damping out of the velocity of rotation.) But if I_z varies, then ω varies in inverse proportion to I_z. A dancer uses this property when pivoting on one leg. She begins the rotation with outstretched arms and then quickly brings them together (usually downwards to the initial position). In this way, by virtue of (24), she reduces the I_z of her body (very substantially, since outstretched arms make an appreciable contribution to I_z) and correspondingly increases ω. (As A. Ya. Vaganova has stated in her *Principles of Classical Dancing*, "this motion of the hands produces the extra force needed for the turn".) At the same time the dancer rises onto her toes to reduce friction. After a few turns she again extends her arms, reducing ω, and comes to a halt on both feet.

Now let us examine a *gyroscope* (actually a top), i.e. a rigid body that rotates about a fixed point, which we take for the origin of coordinates. It is assumed that the gyroscope has an axis of symmetry passing through the point of support and rapidly rotates about that axis, frictional forces being negligibly small.

Note that if a rigid body is in rotation about an axis of symmetry that passes through the origin, then the angular momentum of this body is directed along the axis of rotation and is proportional to the angular speed of rotation. Indeed, the first follows at least from reasons of symmetry, the second, directly from the definition (10). On the other hand, in mechanics proof is given that gravitational forces can be replaced by a resultant passing through the centre of gravity of the body, which clearly lies on the axis of symmetry. Therefore if a gyroscope is attached at the centre of gravity or if it is set in motion in the vertical position (in both cases the moment of the force of gravity is zero), then from (11) it follows that the angular momentum and, with it, the axis of rotation, remain constant. In the case of brief thrusts the right-hand side of (11) can for short times become nonzero. Then $\mathbf{G}$ receives a certain increment, but the greater the angular velocity, that is, the greater $|\mathbf{G}|$, the smaller is the relative value of this increment, i.e. the more stable is the gyroscope. The use of the gyroscope in engineering is based on this property.

Now let us see what will happen if the centre of gravity is located above the point of attachment (as in the case of a child's top), and the axis of rotation is nonvertical. Then the force of gravity $\mathbf{P}$ applied to the centre of gravity C creates a tilting moment $\mathbf{r}_C \times \mathbf{P}$ (Fig. 146). By equation (11), the angular momentum $\mathbf{G}$ will in time dt receive an infinitesimal increment $d\mathbf{G} = (\mathbf{r}_C \times \mathbf{P})\, dt$. Since this infinitesimal vector is horizontal and perpendicular to the vector $\mathbf{G}$, it follows that the

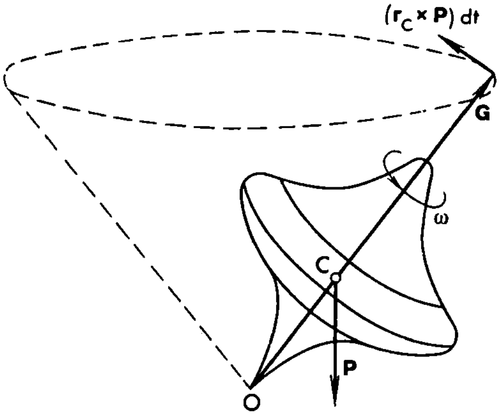


Fig. 146

vector $\mathbf{G} + d\mathbf{G}$ is obtained from $\mathbf{G}$ by a rotation about the vertical axis through an infinitesimal angle. This means that the numerical value of the angular velocity does not change but the axis of rotation of the gyroscope will turn about the vertical axis. The same reasoning can be applied to the new position and we see that because of the force of gravity the axis of symmetry of the gyroscope receives a supplementary uniform rotatory motion (called *precessional motion*) about the vertical axis. The rate of precession is the smaller, the greater ω is, since for large $\mathbf{G}$ the same $d\mathbf{G}$ implies a rotation through a smaller angle.

The foregoing conclusions are not quite exact because, due to precession, the angular momentum does not lie exactly along the axis of the gyroscope but deviates from it. That is why the motion of a gyroscope is actually more complicated than just described. But this correction is slight as long as the angular velocity is great. If the gyroscope is not kept spinning at the same rate, the unavoidable and constant friction reduces the angular velocity and the rate of precession increases, and when these rates become comparable, the nature of the motion changes appreciably and, finally, the top falls to a halt.

Exercise

Compute J_z for a homogeneous (a) cylinder of radius R and mass M with Oz as the axis of rotation, (b) rectilinear rod of length L and mass m if the z-axis serves as a perpendicular to the midpoint of the rod.

11.5 Symmetric and antisymmetric tensors

The angular momentum of a rigid body is a good domain for applying the concept of a tensor (Sec. 9.5). We will use the tensor notation given in Sec. 9.5. Let the origin O be fixed but let the coordinate axes be chosen in a variety of ways. From Sec. 11.4 it is clear that when considering the angular momentum of a body Ω an essential role is played by the integrals

$$I_{ij} = \begin{cases} -\int\limits_{(\Omega)} \rho x_i x_j \, d\Omega & (i \neq j), \\ \int\limits_{(\Omega)} \rho(x_k x_k - x_i x_j) \, d\Omega & (i = j) \end{cases} \tag{26}$$

It is easy to verify that the quantities (26) transform, under a change of coordinate basis, via the tensor rule (23) of Ch. 9 and for this reason constitute a tensor of rank two:

$$\mathbf{I} = I_{ij}\mathbf{e}_i\mathbf{e}_j \tag{27}$$

This is called the *inertia tensor* of the body (Ω) with respect to the point O. To prove this, write the quantities I_{ij} as a single formula:

$$I_{ij} = \int\limits_{(\Omega)} \rho(\delta_{ij} x_k x_k - x_i x_j) \, d\Omega$$

where the *Kronecker delta* δ_{ij} is defined for any choice of basis by

$$\delta_{ij} = \begin{cases} 0 & (i \neq j), \\ 1 & (i = j) \end{cases}$$

We leave it to the reader to convince himself that the quantities δ_{ij} form a second-rank tensor (called the *unit tensor*). Now if the basis is replaced by the formulas (19) of Ch. 9, the quantities x_i transform (as projections of the vector $\mathbf{r}$) via the formulas $x'_i = \alpha_{ij} x_j$, whence, applying the equation $x'_k x'_k = x_k x_k$ (where does it follow from?), we get

$$I'_{ij} = \int\limits_{(\Omega)} \rho(\delta'_{ij} x'_k x'_k - x'_i x'_j) \, d\Omega = \int\limits_{(\Omega)} \rho(\alpha_{ir}\alpha_{jl}\delta_{rl} x_k x_k - \alpha_{ir}\alpha_{jl} x_r x_l) \, d\Omega$$

$$= \alpha_{ir}\alpha_{jl} \int\limits_{(\Omega)} \rho(\delta_{rl} x_k x_k - x_r x_l) \, d\Omega = \alpha_{ir}\alpha_{jl} I_{rl}$$

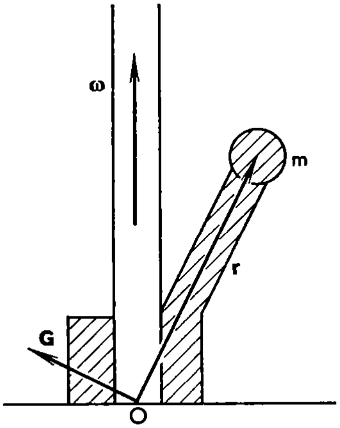
which completes the proof.

Now let the body (Ω) be in rotation with angular velocity $\boldsymbol{\omega}$ about an axis passing through the origin O. We will show how the corresponding angular momentum $\mathbf{G}$ about O is expressed. Reasoning as in Sec. 11.4 and using formula (4), we obtain

$$\mathbf{G} = \int\limits_{(\Omega)} \rho \mathbf{r} \times (\boldsymbol{\omega} \times \mathbf{r}) \, d\Omega = \int\limits_{(\Omega)} \rho[(\mathbf{r} \cdot \mathbf{r})\, \boldsymbol{\omega} - (\mathbf{r} \cdot \boldsymbol{\omega})\, \mathbf{r}] \, d\Omega$$

$$= \int\limits_{(\Omega)} \rho[x_k x_k \omega_i \mathbf{e}_i - x_j \omega_j x_i \mathbf{e}_i] \, d\Omega \tag{28}$$

Fig. 147



In order to take ω_j and $\mathbf{e}_i$ outside the brackets under the integral sign, write $\omega_i = \delta_{ij}\omega_j$ (think through the meaning of this formula!). This yields

$$\mathbf{G} = \int\limits_{(\Omega)} \rho(\delta_{ij}x_k x_k - x_i x_j)\, \omega_j \mathbf{e}_i \, d\Omega = I_{ij}\omega_j \mathbf{e}_i$$

or, to put it differently,

$$G_i = I_{ij}\omega_j \qquad (29)$$

(We can also write $\mathbf{G} = \mathbf{I} \cdot \boldsymbol{\omega}$, where the scalar product of a second-rank tensor by a vector is defined by $(\mathbf{a} \cdot \mathbf{b})\mathbf{c} = \mathbf{a}(\mathbf{b} \cdot \mathbf{c}) = (\mathbf{b} \cdot \mathbf{c})\, \mathbf{a}$, whence

$$\mathbf{I} \cdot \boldsymbol{\omega} = (I_{ij}\mathbf{e}_i\mathbf{e}_j) \cdot \omega_k \mathbf{e}_k = I_{ij}\omega_k \mathbf{e}_i(\mathbf{e}_j \cdot \mathbf{e}_k) = I_{ij}\omega_k \mathbf{e}_i \delta_{jk} = I_{ij}\omega_j \mathbf{e}_i = \mathbf{G}$$

Actually, formula (29) coincides with the formulas of Sec. 11.4, but, unlike the situation in Sec. 11.4, it is not connected with a special choice of Cartesian axes. In particular, it shows that with the exception of the spherically symmetric case (where $I_{ij} = I\delta_{ij}$), the directions of the vectors $\boldsymbol{\omega}$ and $\mathbf{G}$ are, generally, distinct. For example if for the body shown in Fig. 147 the masses of the sleeve and rod are negligibly small compared with m, the moment of rotation is, by (4), equal to

$$\mathbf{G} = \rho\mathbf{r} \times (\boldsymbol{\omega} \times \mathbf{r}) = \rho r^2 \boldsymbol{\omega} - (\omega \mathrm{pr}_{\boldsymbol{\omega}} \mathbf{r})\, \mathbf{r}$$

which means the vector $\mathbf{G}$ is in rotation with angular velocity $\boldsymbol{\omega}$. This change in the moment of rotation is compensated for, by virtue of (11), by the moment transmitted by the bush onto the axis of rotation.

We will now derive another formula for the kinetic energy T of a rotating body (Ω). To begin with, note that for any two vectors $\mathbf{a}$, $\mathbf{b}$ it will be true, on the basis of a familiar formula for the area of a parallelogram, that $|\mathbf{a}\times\mathbf{b}| = ab\sin(\widehat{\mathbf{a}, \mathbf{b}})$. On the other hand, $\mathbf{a}\cdot\mathbf{b} = ab\cos(\widehat{\mathbf{a}, \mathbf{b}})$ whence

$$|\mathbf{a}\times\mathbf{b}|^2 + (\mathbf{a}\cdot\mathbf{b})^2 = a^2b^2[\sin^2(\widehat{\mathbf{a}, \mathbf{b}}) + \cos^2(\widehat{\mathbf{a}, \mathbf{b}})] = a^2b^2 \qquad (30)$$

Form the scalar product of (28) by $\boldsymbol{\omega}$ and take advantage of formula (30) to get

$$\mathbf{G}\cdot\boldsymbol{\omega} = \int\limits_{(\Omega)} \rho[(\mathbf{r}\cdot\mathbf{r})(\boldsymbol{\omega}\cdot\boldsymbol{\omega}) - (\mathbf{r}\cdot\boldsymbol{\omega})(\mathbf{r}\cdot\boldsymbol{\omega})]\,d\Omega = \int\limits_{(\Omega)} \rho\,|\mathbf{r}\times\boldsymbol{\omega}|^2\,d\Omega = 2T$$

whence, by (29),

$$T = \frac{1}{2}\mathbf{G}\cdot\boldsymbol{\omega} = \frac{1}{2}I_{ij}\omega_i\omega_j \qquad (31)$$

The tensor (27) possesses the important property of symmetry:

$$I_{ij} = I_{ji} \qquad (32)$$

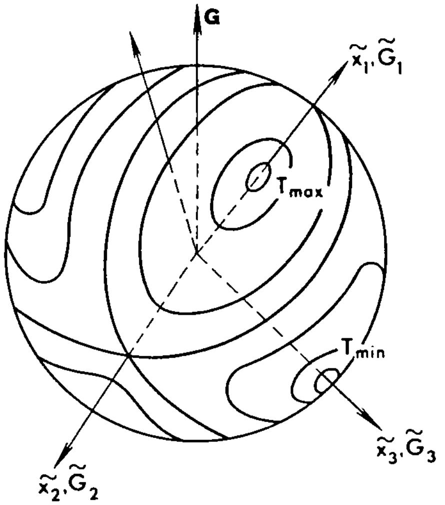
which follows immediately from the definition (26). Generally speaking, a tensor of rank 2 with this property is called a *symmetric tensor*. It is easy to prove here that if the property (32) holds for some one basis, then it holds true for any choice of basis.

The basic property of a symmetric tensor (formula (25) of Ch. 9) is the possibility of reducing it to *diagonal form*, that is, the possibility of a choice of basis $\tilde{e}_i$ for which all quantities $\tilde{p}_{ij}$ vanish when $i \neq j$. We will not give the proof of this general assertion (which holds true for symmetric tensors in a space of any number of dimensions), but two remarks are in order. First, since a diagonal tensor has the property of symmetry and this property is conserved under a change of basis, it follows that *only* symmetric tensors can be reduced to diagonal form. Second, since in the choice of Euclidean basis in three-dimensional space there are three degrees of freedom (why?), and to reduce a tensor to diagonal form, the three equalities $\tilde{p}_{12} = \tilde{p}_{13} = \tilde{p}_{23} = 0$ must be satisfied, there turn out to be exactly as many degrees of freedom as are needed.

In the axes $\tilde{x}_i$ in which the inertia tensor has diagonal form (they are called the *principal axes* of that tensor), expressions (29) for the projections of the angular momentum and (31) for the kinetic energy take the form

$$\tilde{G}_1 = \tilde{I}_{11}\tilde{\omega}_1,\quad \tilde{G}_2 = \tilde{I}_{22}\tilde{\omega}_2,\quad \tilde{G}_3 = \tilde{I}_{33}\tilde{\omega}_3,$$

$$T = \frac{1}{2}(\tilde{I}_{11}\tilde{\omega}_1^2 + \tilde{I}_{22}\tilde{\omega}_2^2 + \tilde{I}_{33}\tilde{\omega}_3^2) = \frac{1}{2}\left(\frac{1}{\tilde{I}_{11}}\tilde{G}_1^2 + \frac{1}{\tilde{I}_{22}}\tilde{G}_2^2 + \frac{1}{\tilde{I}_{33}}\tilde{G}_3^2\right) \qquad (33)$$



$\mathbf{G}$
$\tilde{x}_1, \tilde{G}_1$
T_{max}
T_{min}
$\tilde{x}_3, \tilde{G}_3$
$\tilde{x}_2, \tilde{G}_2$

Fig. 148

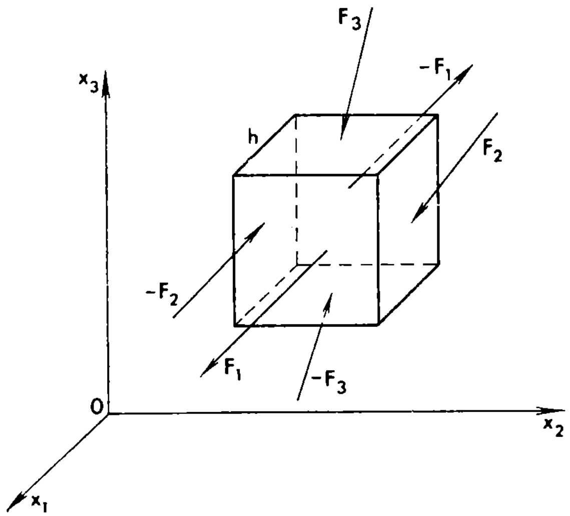
In order to get a better picture of the dependence of the kinetic energy T on the direction of the angular momentum $\mathbf{G}$, depict on a sphere $|\mathbf{G}| =$ constant rigidly attached to the body (Ω) certain lines to which correspond equal values of T; in other words, these are lines with equation (33) for distinct values of T (Fig. 148). For the sake of definiteness, suppose that $\tilde{I}_{11} < \tilde{I}_{22} < \tilde{I}_{33}$. Then for a given $|\mathbf{G}|$, the greatest value of T, equal to $\frac{1}{2\tilde{I}_{11}}|\mathbf{G}|^2$, is obtained when $\tilde{G}_2 = \tilde{G}_3 = 0$ (Fig. 148); we have the smallest value of T for $\tilde{G}_1 = \tilde{G}_2 = 0$, whereas T has a minimax for $\tilde{G}_1 = \tilde{G}_3 = 0$. In free rotation, $\mathbf{G}$ and T are constant but $\boldsymbol{\omega}$ varies both in direction and in modulus (except for rotation about one of the x_i-axes when $\boldsymbol{\omega}\|\mathbf{G}\|\,\mathbf{e}_i$). The body ($\Omega$) rotates so that the vector $\mathbf{G}$ always passes through one of the indicated lines (Fig. 148 shows the translation of $\mathbf{G}$ with respect to (Ω)).

Another important instance of a symmetric tensor is the *elastic stress tensor*. Suppose we are considering the stressed state of a rigid body caused by forces applied to it. Conceive of a cube with side h and edges parallel to the coordinate axes isolated in the medium under study. Then the elastic action of the ambient medium on each of the faces of the cube can be replaced by a force (Fig. 149) which is proportional to h^2 for small h:

$$\mathbf{F}_i = \delta_{ij}\mathbf{e}_j h^2 + \ldots$$

The quantities σ_{ij} depend on the choice of direction of the coordinate axes. It can be shown that under a change of basis they transform by the tensor rule (23) of Ch. 9 and for this reason form a second-rank

Fig. 149



tensor, which is called the elastic stress tensor and characterizes the stressed state of the medium at the point in question. The diagonal terms of this tensor define the stresses of compression or tension, and the off-diagonal terms define the shear stresses.

It is easy to verify that the elastic stress tensor is a symmetric tensor. Indeed, the total moment of external forces acting on the cube with respect to its centre is equal, up to higher-order infinitesimals, to $\sum\limits_i 2\cdot\frac{1}{2}h\mathbf{e}_i\times\mathbf{F}_i = \sum\limits_{i,j} h^3\sigma_{ij}\mathbf{e}_i\times\mathbf{e}_j$. If any $\sigma_{ij} \neq \sigma_{ji}$, then this moment would be nonzero, which is impossible since the moment of inertia of the cube is of the order of h^5 and we would have an infinite angular acceleration.

Use is also made of *antisymmetric tensors* (p_{ij}) for which

$$p_{ij} = -p_{ji} \tag{34}$$

The diagonal terms of this tensor are of necessity equal to zero. It is easy to verify that if the antisymmetry property (34) is fulfilled for some one choice of basis, then it holds true for any other choice of basis.

An antisymmetric tensor in three-dimensional space is directly connected with a certain product. Namely, suppose we have a linear mapping of space into itself with an antisymmetric matrix of the coefficients p_{ij} (see Sec. 9.5, where we stated that any second-rank tensor can, to within dimensionality, be interpreted as the matrix of a certain linear mapping of the space into itself). Then any vector $\mathbf{r} = x_i\mathbf{e}_i$ is mapped via the formula

$$Tr = p_{ij}x_j\mathbf{e}_i = p_{12}x_2\mathbf{e}_1 + p_{13}x_3\mathbf{e}_1 - p_{12}x_1\mathbf{e}_2 + p_{23}x_3\mathbf{e}_2 - p_{13}x_1\mathbf{e}_3 - p_{23}x_2\mathbf{e}_3$$

But we get the same result (verify this!) if we form the vector product of $\mathbf{p} = -(p_{12}\mathbf{e}_3 + p_{23}\mathbf{e}_1 + p_{31}\mathbf{e}_2)$ and $\mathbf{r}$; thus, $T\mathbf{r} \equiv \mathbf{p}\times\mathbf{r}$. From the last equation it is clear that the vector $\mathbf{p}$ is defined by the mapping T, and so also by the tensor p_{ij}, invariantly, which means independently of any choice of a Cartesian basis.

In conclusion consider the tensor of a linear mapping that is close to an identity mapping. It has the form $\delta_{ij} + \eta_{ij}$, where δ_{ij} is the unit tensor and η_{ij} are small coefficients. Such a tensor results, for example, when considering small deformations of an elastic body.

Represent the tensor η_{ij} as the sum of a symmetric tensor β_{ij} and an antisymmetric tensor γ_{ij} where

$$\beta_{ij} = \frac{1}{2}(\eta_{ij} + \eta_{ji}), \quad \gamma_{ij} = \frac{1}{2}(\eta_{ij} - \eta_{ji}) \tag{35}$$

Then the tensor β_{ij} assumes diagonal form in certain axes $\tilde{x}_i$ and for this reason determines a combination of small uniform tensions along these axes: $1 + \tilde{\beta}_{11}$ times along the $\tilde{x}_1$-axis (of course if $\tilde{\beta}_{11} < 0$, then we have a compression), and so on. Here the volume will increase $(1 + \tilde{\beta}_{11})(1 + \tilde{\beta}_{22})(1 + \tilde{\beta}_{33})$-fold, or, to within higher-order infinitesimals, $1 + \tilde{\beta}_{11} + \tilde{\beta}_{22} + \tilde{\beta}_{33} = 1 + \tilde{\beta}_{ii}$ times. But $\tilde{\beta}_{ii} = \beta_{ii}$ (see Sec. 9.5) and so $\tilde{\beta}_{ii} = \eta_{ii}$ by virtue of (35).

The tensor γ_{ij} defines a small rotation. Indeed, under such a rotation about the vector $\mathbf{p}^0$ through a small angle ψ, each vector $\mathbf{r}$, up to higher infinitesimals, goes into the vector $\mathbf{r} + \psi\mathbf{p}^0\times\mathbf{r}$; now we have demonstrated above how one can choose an appropriate vector $\mathbf{p} = \psi\mathbf{p}^0$ using the antisymmetric tensor γ_{ij}.

To summarize: a linear mapping that is almost an identity mapping reduces to a combination of uniform tensions along mutually perpendicular axes and a rotation. Since the volume does not change under a rotation, the sum η_{ii} (it is called the *trace* of the tensor η_{ij}) is the coefficient of the increment in volume, which means that under the given mapping the volume increases $1 + \eta_{ii}$ times.

It is to be stressed that the superposition of deformations leads to an addition of the appropriate tensors only for small deformations. Actually, this is the result of applying the formula $(1 + \alpha)(1 + \beta) \approx \approx 1 + (\alpha + \beta)$ for small α, β. Large deformations (rotations through a finite angle, for instance) are superimposed by quite a different — generally, noncommutative — law, which we will not discuss here.

Exercises

1. Prove the theorem of the possibility of reducing a symmetric tensor to diagonal form for the two-dimensional case.
2. Decompose a small shear (Fig. 111*d*) into tensions and a rotation.

11.6 True vectors and pseudovectors

There is a fundamental difference between the linear velocity vector and the angular velocity vector. There is no doubt about the direction of the linear velocity vector. By contrast, and in accord with Sec. 11.2, the angular velocity vector is laid off along the axis of rotation, but in what direction? In Sec. 11.2 we chose that direction in accordance with the rule of the right-handed screw. But we could just as easily have chosen the rule of the left-handed screw and then in Fig. 137 the vector ω would have been directed downwards. Thus, the choice of direction of the angular velocity vector along the axis of rotation is arbitrary and depends on the chosen rule of the screw, if the rule is reversed, the vector is too. Such vectors are termed *pseudovectors* to distinguish them from the *true vectors* that do not depend on the choice of the rule of the screw. To summarize: the linear velocity vector is a true vector (just like the vectors of force, acceleration, electric intensity, and so forth), while the angular velocity vector is a pseudovector. In another terminological classification, true vectors are called *polar vectors* and pseudovectors are called *axial vectors*. *

From the definition of a vector product it is clear that the vector product of two true vectors is a pseudovector because under a change in the rule of the screw the earlier outside of the parallelogram constructed on the vectors being multiplied becomes the inside (see Figs. 134, 135). Thus, the moment of a force and the angular momentum (Sec. 11.2) are pseudovectors. Similarly it can readily be verified that the vector product of a true vector by a pseudovector is a true vector (see for instance formula (5)) and the vector product of two pseudovectors is a pseudovector.

The question of the equivalence of the right-handed and left-handed coordinate systems is not at all so simple as may appear at first glance. This equivalence signifies that for any phenomenon we can have a *mirror reflection* of it under which all geometric forms are the mirror images of the originals, just like a right-hand glove is the mirror image of the left-hand one. (This is the so-called "law of conservation of parity".) Quite recently it was found that this law is not universal, and the celebrated Soviet theoretical physicist L. D. Landau (1908-1968) proposed the "principle of combined

* When considering processes that occur in time, there arises yet another classification of vectors as to their behaviour under a change of sign of *t* (compare, for instance, the vectors of velocity and displacement).

parity". According to this principle all physical phenomena admit reflections only if all particles are replaced by antiparticles.[*]

Note in conclusion that the vector multiplication of two vectors, under which the projections of the vector product are expressed in terms of the projections of the vector factors (yet at the same time the product is invariant under a choice of coordinate axes) and the ordinary rules of multiplication hold true, represents an operation that is characteristic of three-dimensional space. To put it crudely, the point is that in three-dimensional space we can agree to associate with every two unit vectors (say, $\mathbf{i}$ and $\mathbf{j}$ along the axes x and y) a third unit vector $\mathbf{k}$ along the z-axis to complete the set of three. In this way, by performing cyclic (circular) permutations, we can arrive at the formulas (1) of the vector multiplication of the unit vectors. (True because if $\mathbf{i}\times\mathbf{j}=\mathbf{k}$, then we must have $\mathbf{j}\times\mathbf{k}=\mathbf{i}$ since the vector $\mathbf{k}$ is located relative to the vectors $\mathbf{i}$, $\mathbf{j}$ in exactly the same way as $\mathbf{i}$ is relative to $\mathbf{j}$ and $\mathbf{k}$.) From the latter formulas we obtain the formulas of vector multiplication of any vectors. In n-dimensional vector space (Sec. 9.6) we have to take $n-1$ unit vectors to determine the nth missing vector. Therefore, in n-dimensional space an analogue of a vector product is a vector appropriately constructed out of $n-1$ vectors.[**] Thus, the vector product of two vectors is peculiar only to three-dimensional space. Naturally we have not enumerated all the conditions that are necessary for determining a vector product. But still we wanted to point out the difference between a vector product and a scalar product, which is defined in precisely the same way in a space of any number of dimensions.

Exercises

1. Do the formulas (1) and (3) hold true in a left-handed system of coordinates?
2. Construct a reasonable definition of a vector product in four-dimensional space.

11.7 The curl of a vector field

In calculating the work of a force field in Sec. 10.3 we already arrived at the concept of a line integral. We now consider it in the general form.

[*] At the end of 1964 new experimental findings suggested that the principle of combined parity is not exact and is sometimes violated.

[**] The analogue of a mixed product (p. 391) is made up for n vectors. This product is equal to the n-volume of a "parallelepiped" spanned by these vectors. For a plane — when $n=2$ — we have an area, which is a scalar or, to be more exact, a pseudoscalar, "pseudo" meaning "almost". What we mean here is that the quantity does not change under rotations of the coordinate system but is multiplied by -1 under an interchange of the axes x and y. An instance of a pseudoscalar in three-dimensional space was the mixed product of three true vectors equal to the volume: the sign of this quantity for the given three vectors depends on the choice of a right- or left-handed system.

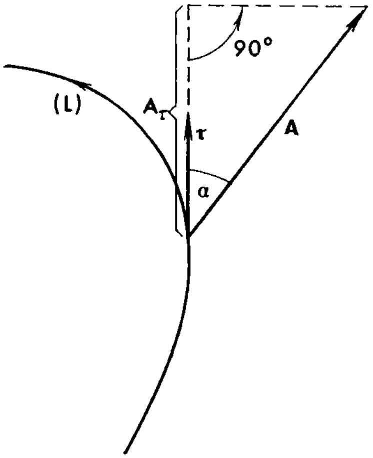


Fig. 150

Suppose an oriented line (L) is chosen in a space with a specified field of vector **A**. Then the *line integral* of **A** is called the integral

$$I = \int_{(L)} A_\tau \, dL \tag{36}$$

around the line (L), where A_τ is the projection of the vector **A** on the tangent to (L) drawn in the direction of traversal (Fig. 150). Since the vector $d\mathbf{r}$ goes along $\boldsymbol{\tau}$ and $|d\mathbf{r}| = dL$ (Sec. 9.4), it follows that the expression for the line integral can be rewritten thus:

$$I = \int_{(L)} |\mathbf{A}| \cos \alpha \, |d\mathbf{r}| = \int_{(L)} \mathbf{A} \cdot d\mathbf{r} = \int_{(L)} (A_x \, dx + A_y \, dy + A_z \, dz)$$

A line integral is a scalar quantity and has the ordinary properties of integrals. If the orientation of (L) is reversed, the integral only changes sign. If the angle α (see Fig. 150) at all points of (L) is acute, then $I > 0$, and if the angle is obtuse, then $I < 0$. The equality $I = 0$ results if the angle α is always a right angle or (this happens more often) if the integrals over parts of (L), where α is acute and α is obtuse, cancel out.

If the line (L) is closed, then the line integral

$$\Gamma = \oint_{(L)} \mathbf{A} \cdot d\mathbf{r} = \oint_{(L)} (A_x \, dx + A_y \, dy + A_z \, dz) \tag{37}$$

is called the *circulation* of the vector **A** around the line (L). The circulation possesses the following important property of additivity. Suppose an oriented open surface (S) is split into a number of parts,

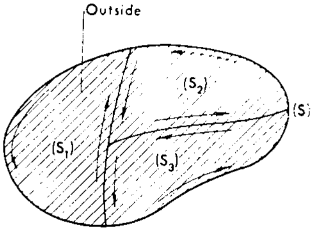


Fig. 151

say three — (S_1), (S_2), and (S_3) — as in Fig. 151. Denote the contours of (S) and of these parts by (L), (L_1), (L_2), and (L_3), in accordance with the orientation of (S), and the corresponding circulations by Γ, Γ_1, Γ_2, Γ_3. Then,

$$\Gamma = \Gamma_1 + \Gamma_2 + \Gamma_3$$

Indeed, if all circulations on the right are represented in the form of a sum of line integrals around the separate arcs shown in Fig. 151, then the integrals around the arcs interior to (S) will all cancel out (since each such arc is traversed twice in opposite directions), and the integrals around the contour arcs of (S) are additive and yield the circulation in the left member.

This additivity property enables us to say that the circulation (37) is "generated" on the surface (S) and, hence, to speak of a "density of generation of the circulation", that is to say, a circulation generated by an infinitesimal piece of surface and referred to unit area of this piece. The advisability of considering such a density is also suggested by the following circumstance. It is easy to verify that the circulation of a constant vector is always zero:

$$\oint_{(L)} (C_1\,dx + C_2\,dy + C_3\,dz) = (C_1x + C_2y + C_3z)\Big|_{(L)}$$

where the last symbol indicates that we have to take the increment of the result of integration when the point traverses the contour (L). But after such a traversal the last expression in brackets returns to its original value and so the increment is zero. But now we can reason as at the end of Sec. 10.7, namely that by virtue of Taylor's formula the vector $\mathbf{A}$ inside an infinitesimal contour may be represented as the sum of a constant vector and of first-order terms. Then an almost complete compensation occurs: the integral of a constant vector is zero, while the integral of first-order terms yields a quantity of second order of smallness. Thus, the circulation around an infinitesimal contour is proportional not to the length of the contour but to the area embraced by the contour.

To compute the indicated density of generation of the circulation, compute the circulation of the vector $\mathbf{A}$ around an infinitely

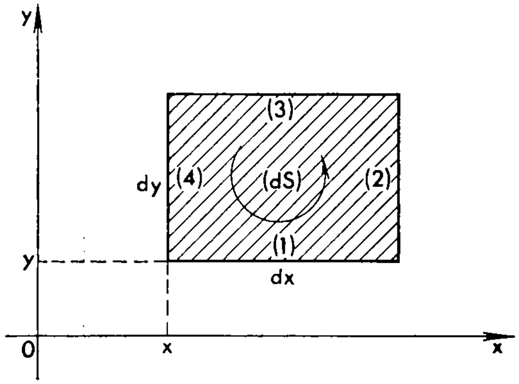


Fig. 152

small contour. First assume that this contour lies in the plane z-constant. Besides, since the shape of the contour is inessential when computing density, for this contour take a rectangle with sides parallel to the coordinate axes (see Fig. 152, where the size of the rectangle is somewhat enlarged). By formula (37) the appropriate circulation is

$$d\Gamma = \int\limits_{(1)} A_x\,dx + \int\limits_{(2)} A_y\,dy + \int\limits_{(3)} A_x\,dx + \int\limits_{(4)} A_y\,dy \qquad (38)$$

(the numerals indicate the successive sides of the rectangle in Fig. 152), since only one variable on each side varies, while the other differentials are equal to zero. Taking into account the sense of traversal of the indicated sides, we get from (38)

$$\begin{aligned} d\Gamma &= (A_x)_1\,dx + (A_y)_2\,dy - (A_x)_3\,dx - (A_y)_4\,dy \\ &= [(A_y)_2 - (A_y)_4]\,dy - [(A_x)_3 - (A_x)_1]dx \qquad (39) \end{aligned}$$

where the numerical subscript indicates the side on which the appropriate projection is taken. However, to within higher-order infinitesimals,

$$(A_y)_2 - (A_y)_4 = \frac{\partial A_y}{\partial x}\,dx, \quad (A_x)_3 - (A_x)_1 = \frac{\partial A_x}{\partial y}\,dy$$

and so the formula (39) yields

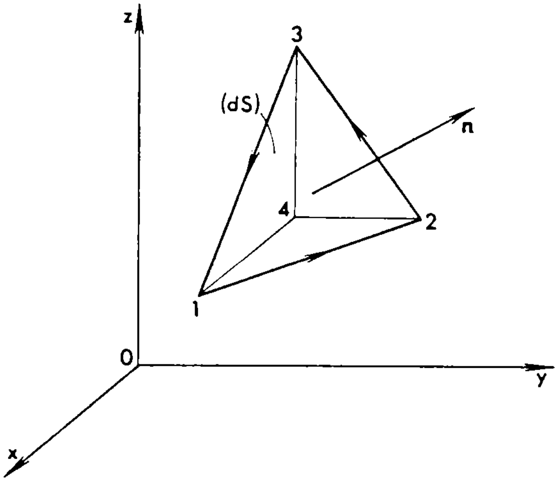
$$d\Gamma = \frac{\partial A_y}{\partial x}\,dx\,dy - \frac{\partial A_x}{dy}\,dy\,dx = \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)dS$$

Thus, for an infinitely small closed contour,

$$\frac{d\Gamma_{xy}}{dS_{xy}} = \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \qquad (40)$$

(the subscripts in the left-hand member indicate that the contour is parallel to the xy-plane). Also, the contour is traversed in the posi-

Fig. 153



tive sense, otherwise the sign has to be reversed or, what is the same thing, we have to regard $dS_{xy} < 0$.

Cartesian coordinates in space are completely equivalent and so any formula containing these coordinates can produce any other correct formula if x, y, z are replaced respectively by y, z, x or z, x, y. (This — see Sec. 11.6 — is called a *cyclic*, or *circular*, permutation, under which a right-handed system of coordinates remains right-handed.) For this reason, it follows from (40) that

$$\frac{d\Gamma_{yz}}{dS_{yz}} = \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}, \quad \frac{d\Gamma_{zx}}{dS_{zx}} = \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x} \qquad (41)$$

Now let us consider an infinitesimal oriented area (dS) with an arbitrary inclination relative to the coordinate axes. In order to compute the circulation it is most convenient to take this area in the form of a triangle, as in Fig. 153. On this triangle construct a tetrahedron with faces parallel to the coordinate planes and label the vertices of the tetrahedron with numbers as indicated in Fig. 153. Then it is easy to see that

$$d\Gamma = d\Gamma_{123} = d\Gamma_{124} + d\Gamma_{234} + d\Gamma_{431}$$

since on the right side the integrals around the segments 41, 42, and 43 cancel out. But the right side can be computed by formulas (40) and (41):

$$d\Gamma = \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right) dS_{124} + \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right) dS_{234} + \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right) dS_{431} \qquad (42)$$

where the numerical subscripts indicate the areas in question.

The result becomes more surveyable if we introduce via the following formula a vector called the *curl* (or *rotation*) of the field **A** and denoted by curl **A**:

$$\operatorname{curl}\mathbf{A} = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right)\mathbf{i} + \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right)\mathbf{j} + \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_z}{\partial y}\right)\mathbf{k} \qquad (43)$$

If we also note that by virtue of the formula (18) of Ch. 10,

$$dS_{234}\mathbf{i} + dS_{431}\mathbf{j} + dS_{124}\mathbf{k} = d\mathbf{S}_{234} + d\mathbf{S}_{431} + d\mathbf{S}_{124} = d\mathbf{S} = \mathbf{n}\, dS$$

and if we divide by dS, then (42) can be rewritten more simply thus:

$$\frac{d\Gamma}{dS} = (\operatorname{curl}\mathbf{A})\cdot\mathbf{n} = \operatorname{curl}_n \mathbf{A} \qquad (44)$$

Here the subscript n indicates that we take the projection of the curl on the normal **n**. This formula gives the circulation around an infinitesimal contour referred to the unit area embraced by this contour.*

Thus, the projection of the curl of the field on any direction **n** is equal to the ratio of the circulation of the field around an infinitesimal contour perpendicular to **n** to the area embraced by this contour. From this it is evident for one thing that the curl, whose definition (43) is attached to a chosen coordinate system, actually is invariantly connected with the field (it forms a new vector field since at every point of space, generally, it has its own value): it does not depend on the choice of coordinate system since the left-hand side of (44) does not depend on this choice, and a knowledge of the projection of the vector on every direction determines this vector in unique fashion. At the same time the curl of a true vector is a pseudovector (see Sec. 11.6), since under a change of the rule of the screw for the same orientation of the area (dS), i.e. with the same vector **n**, the traversal of its contour is reversed and so the circulation changes sign. Note, incidentally, that obtaining a new vector field by determining the curl of another vector field is a specific feature of three-dimensional space. Yet obtaining a vector field as a gradient of a scalar field in space of any number of dimensions occurs in the same way. The relationship here is the same as between a vector product and a scalar product (see the end of Sec. 11.6).

Now suppose instead of an infinitesimal surface we have in space a finite oriented surface (S) with contour (L). Wc have already seen

* From this it follows that the direction of the vector curl **A** in space is determined by the direction of the normal to an area for which $\frac{d\Gamma}{dS}$ is a maximum. This definition is similar to the definition of the direction of the gradient of a scalar φ as the direction of a line (l), on which direction $\frac{d\varphi}{dl}$ attains a maximum.

that the circulations corresponding to separate parts of the surface are additive, i.e. the total circulation is

$$\Gamma = \int\limits_{(S)} d\Gamma$$

And so from (37) and (44) it follows that

$$\oint\limits_{(L)} \mathbf{A} \cdot d\mathbf{r} = \int\limits_{(S)} (\text{curl}_n\, \mathbf{A})\, dS = \int\limits_{(S)} \text{curl}\, \mathbf{A} \cdot d\mathbf{S} \tag{45}$$

that is, the circulation of a field around a closed contour is equal to the flux (see Sec. 10.7) of the curl of this field through the surface bounded by the indicated contour. This important formula is called *Stokes' theorem.*

There is another useful integral formula involving the curl that transforms the integral

$$\mathbf{I} = \oint\limits_{(S)} \mathbf{A} \times d\mathbf{S}$$

over a closed surface (S) (oriented in natural fashion, i.e. with outside pointing to infinity) into an integral over the volume (Ω) bounded by (S). To derive it, note that

$$I_x = \mathbf{I} \cdot \mathbf{i} = \oint\limits_{(S)} (\mathbf{A} \times d\mathbf{S}) \cdot \mathbf{i} = \int\limits_{(S)} (\mathbf{i} \times \mathbf{A}) \cdot d\mathbf{S}$$

Here we make use of the fact that by virtue of the geometric meaning of a scalar triple product this product does not change under a cyclic permutation of the factors. However,

$$\mathbf{i} \times \mathbf{A} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 0 & 0 \\ A_x & A_y & A_z \end{vmatrix} = -A_z \mathbf{j} + A_y \mathbf{k}$$

or

$$I_x = \oint\limits_{(S)} (-A_z \mathbf{j} + A_y \mathbf{k}) \cdot d\mathbf{S}$$

Transforming this integral by the Ostrogradsky formula (Sec. 10.7), we get

$$I_x = \int\limits_{(\Omega)} \text{div}\,(-A_z \mathbf{j} + A_y \mathbf{k})\, d\Omega = \int\limits_{(\Omega)} \left(-\frac{\partial A_z}{\partial y} + \frac{\partial A_y}{\partial z} \right) d\Omega$$

In similar fashion we find

$$I_y = \int\limits_{(\Omega)} \left(-\frac{\partial A_x}{\partial z} + \frac{\partial A_z}{\partial x} \right) d\Omega, \quad I_z = \int\limits_{(\Omega)} \left(-\frac{\partial A_y}{\partial x} + \frac{\partial A_x}{\partial y} \right) d\Omega$$

whence

$$\oint\limits_{(S)} \mathbf{A} \times d\mathbf{S} = \mathbf{I} = I_x\mathbf{i} + I_y\mathbf{j} + I_z\mathbf{k} = \int\limits_{(\Omega)} \left[\left(-\frac{\partial A_z}{\partial y} + \frac{\partial A_y}{\partial z} \right) \mathbf{i} \right.$$

$$\left. + \left(-\frac{\partial A_x}{\partial z} + \frac{\partial A_z}{\partial x} \right) \mathbf{j} + \left(-\frac{\partial A_y}{\partial x} + \frac{\partial A_x}{\partial y} \right) \mathbf{k} \right] d\Omega = -\int\limits_{(S)} \operatorname{curl} \mathbf{A} \, d\Omega$$

Exercises

1. Use the last formula to obtain an invariant definition (not related to any choice of coordinate system) of the curl similar to the definitions of divergence (formula (33) of Ch. 10) and gradient (formula (57) of Ch. 10).
2. Use the Stokes theorem to prove the Cauchy theorem on the integral of an analytic function (Sec. 5.8).
   Hint. Pass to real integrals and take advantage of the Cauchy-Riemann conditions (17) of Ch. 5.

11.8. The Hamiltonian operator del

Let us write down the basic differential operations that can be performed on a scalar field u and a vector field $\mathbf{A} = A_x\mathbf{i} + A_y\mathbf{j} + A_z\mathbf{k}$:

$$\operatorname{grad} u = \frac{\partial u}{\partial x}\mathbf{i} + \frac{\partial u}{\partial y}\mathbf{j} + \frac{\partial u}{\partial z}\mathbf{k},$$

$$\operatorname{div} \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z},$$

$$\operatorname{curl} \mathbf{A} = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \right) \mathbf{i} + \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x} \right) \mathbf{j} + \left(\frac{\partial A_y}{\partial x} + \frac{\partial A_x}{\partial y} \right) \mathbf{k}$$

The Irish mathematician William R. Hamilton noticed that these three operations can be written more simply if one introduces the symbol

$$\nabla = \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}$$

called *del.* Taken by itself, this symbol is an operation sign, that is, an operator. It is a vector differential operator that preserves the features of a vector and of a differentiation operator. (The general notion of an operator is discussed in Sec. 6.2.)

"Multiplication" (which is an operation) of the del operator by a scalar (by a scalar field, to be more exact) u and by a vector $\mathbf{A}$ takes place by the following natural rules:

$$\nabla u = \left(\mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}\right)u = \mathbf{i}\frac{\partial u}{\partial x} + \mathbf{j}\frac{\partial u}{\partial y} + \mathbf{k}\frac{\partial u}{\partial z} = \operatorname{grad} u,$$

$$\nabla \cdot \mathbf{A} = \left(\mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}\right)\cdot(\mathbf{i}A_x + \mathbf{j}A_y + \mathbf{k}A_z)$$

$$= \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} = \operatorname{div} \mathbf{A},$$

$$\nabla \times \mathbf{A} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ A_x & A_y & A_z \end{vmatrix} = \mathbf{i}\left(\frac{\partial A_z}{\partial y} + \frac{\partial A_y}{\partial z}\right) + \text{etc.} = \operatorname{curl} \mathbf{A}$$

Del, being a differential operator, operates only on the factor that stands immediately to the right of it: for example,

$$(\nabla u)\, v = (\operatorname{grad} u)\, v = v \operatorname{grad} u, \quad \nabla(uv) = \operatorname{grad}\,(uv)$$

Therefore if there is no factor following del, then it is an operator: for example,

$$\mathbf{A}\cdot\nabla = (\mathbf{i}A_x + \mathbf{j}A_y + \mathbf{k}A_z)\cdot\left(\mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}\right)$$

$$= A_x\frac{\partial}{\partial x} + A_y\frac{\partial}{\partial y} + A_z\frac{\partial}{\partial z}$$

is a scalar differential operator which can operate on a scalar field or a vector field. It finds use, in particular, in the formula for the rate of change of a field along a trajectory (Sec. 4.1):

$$\frac{du}{dt} = v_x\frac{\partial u}{dx} + v_y\frac{\partial u}{\partial y} + v_z\frac{\partial u}{\partial z} + \frac{\partial u}{\partial t} = (\mathbf{v}\cdot\nabla)\, u + \frac{\partial u}{\partial t} \qquad (46)$$

Repeating the derivation of this formula, we can readily verify that there is an analogous formula for the rate of change of a vector field along a trajectory:

$$\frac{d\mathbf{A}}{dt} = (\mathbf{v}\cdot\nabla)\,\mathbf{A} + \frac{\partial\mathbf{A}}{\partial t}$$

The separate terms here have the same meaning as in (46). This enables us, for one thing, to rewrite the fluid-flow equation (54) of Ch. 10 in the form

$$\rho(\mathbf{v}\cdot\nabla)\,\mathbf{v} + \rho\frac{\partial\mathbf{v}}{\partial t} = -\operatorname{grad} p + \mathbf{f}$$

from which it is now quite easy to pass to the coordinate form.

When operating with the del operator, one uses the rules of vector algebra and the rules of differentiation. For instance,

$$\begin{aligned}\operatorname{curl}(\mathbf{A}+\mathbf{B}) &= \nabla\times(\mathbf{A}+\mathbf{B}) = \nabla\times\mathbf{A}+\nabla\times\mathbf{B}\\ &= \operatorname{curl}\mathbf{A}+\operatorname{curl}\mathbf{B},\\ \operatorname{div}(\lambda\mathbf{A}) &= \nabla\cdot(\lambda\mathbf{A}) = \lambda(\nabla\cdot\mathbf{A}) = \lambda\operatorname{div}\mathbf{A}\quad(\lambda=\text{constant})\end{aligned} \tag{47}$$

since multiplication by a vector and also differentiation have these properties of linearity. At the same time, we cannot regard λ in (47) as depending on a point in space (that is, as being a scalar field), for then it would mean that we had taken a variable quantity outside the sign of differentiation. In order to embrace this case, note that in the ordinary formula for the derivative of a product,

$$(uv)' = u'v + uv' \tag{48}$$

we get the first term if in the process of differentiation we take v to be constant, and the second term if in this process we take u to be constant, and so the differentiation (48) can be carried out as follows:

$$(uv)' = (u_c v)' + (uv_c)' = u_c v' + u'v_c = uv' + u'v$$

where the subscript c indicates that in differentiating we regard the given quantity as a constant (if of course the quantity stands outside the differentiation sign, then the subscript c can be dropped). Thus,

$$\begin{aligned}\operatorname{div}(u\mathbf{A}) &= \nabla\cdot(u\mathbf{A}) = \nabla\cdot(u_c\mathbf{A}) + \nabla\cdot(u\mathbf{A}_c)\\ &= u(\nabla\cdot\mathbf{A}) + (\nabla u)\cdot\mathbf{A} = u\operatorname{div}\mathbf{A} + \operatorname{grad}u\cdot\mathbf{A}\end{aligned}$$

(this formula was derived in a different way in Sec. 10.8).

After applying the differential operation to a field we get a new field, to which these operations can again be applied. To illustrate, consider the "compound" operation curl grad u. We can write it in the form $\nabla\times(\nabla u)$. But for an "ordinary" vector $\mathbf{a}$ and an "ordinary" scalar u it is always true that

$$\mathbf{a}\times(\mathbf{a}u) = \mathbf{0} \tag{49}$$

(why?). This means that if for $\mathbf{a}$ on the left we substitute its expansion along Cartesian axes and carry out the computations by the formal rules of vector algebra, we get zero. But the computation of the combination $\nabla\times(\nabla u)$ is carried out by the same formal rules as in (49), only instead of a_x, a_y, a_z we have to take $\frac{\partial}{\partial x}$, $\frac{\partial}{\partial y}$, $\frac{\partial}{\partial z}$. This means that we again get zero; that is, it is always true that

$$\operatorname{curl}\operatorname{grad}u = \mathbf{0} \tag{50}$$

Similarly we find that it is always true that

$$\operatorname{div}\operatorname{curl}\mathbf{A} = 0 \tag{51}$$

(verify this !). This simple property has an important consequence. Namely, for any field **A** we can, besides vector lines (Sec. 10.7), consider *vortex lines*, that is to say, vector lines of the field curl **A**. Formula (51) states that the vortex lines cannot have either sources or sinks, which means they cannot originate or terminate.

Finally, the combination

$$\text{div grad} = \nabla \cdot \nabla = \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

(the Laplacian operator Δ) was considered in Sec. 10.9.

Exercises

1. Derive formulas for curl (u**A**), div (**A**×**B**).
2. On the basis of formulas (33) and (57) of Ch. 10 and the solution of Exercise 1, Sec. 11.7, write the symbolic expression ∇ in the form of an integral.

11.9 Potential fields

In Sec. 10.3 we considered the problem of the existence of a potential (potential energy) in the case of field of force. We now consider this question in the general form. A vector field **A** is said to be a *potential field* if it is the gradient of some scalar field; if we denote this field by $-\varphi$, then

$$\mathbf{A} = -\text{grad}\ \varphi \tag{52}$$

(cf. formula (9) of Ch. 10). Here the field φ is called the *potential* of the field **A**.

Since the gradient of a constant scalar field is equal to zero, the potential of any field **A** (if the potential exists) is determined to within an arbitrary additive constant. By apt choice of this constant, it is possible to make the value of thc potential equal to zero at any specified point. Customarily, the value of the potential at infinity is taken to be zero. The difference in the values of the potential at two points no longer depends on this slight arbitrariness in the choice of the potential, since the constant additives in such a difference cancel out.

Not every field by far is a potential field, namely, from (52) it immediately follows that

$$\text{curl}\ \mathbf{A} = -\text{curl grad}\ \varphi = \mathbf{0} \tag{53}$$

(see formula (50)), that is, a potential field is necessarily an irrotational field. Taking into account expression (43) for the curl, the condition (53) can be rewritten as

$$\frac{\partial A_z}{\partial y} = \frac{\partial A_y}{\partial z}, \quad \frac{\partial A_x}{\partial z} = \frac{\partial A_z}{\partial x}, \quad \frac{\partial A_y}{\partial x} = \frac{\partial A_x}{\partial y} \tag{54}$$

Conditions (54) can also be derived in a different way. In coordinate form, (52) is

$$A_x\mathbf{i} + A_y\mathbf{j} + A_z\mathbf{k} = -\left(\frac{\partial\varphi}{\partial x}\,\mathbf{i} + \frac{\partial\varphi}{\partial y}\,\mathbf{j} + \frac{\partial\varphi}{\partial z}\,\mathbf{k}\right)$$

whence it follows that

$$A_x = -\frac{\partial\varphi}{\partial x}, \quad A_y = -\frac{\partial\varphi}{\partial y}, \quad A_z = -\frac{\partial\varphi}{\partial z}$$

But then

$$\frac{\partial A_z}{\partial y} = -\frac{\partial^2\varphi}{\partial z\,\partial y}, \quad \frac{\partial A_y}{\partial z} = -\frac{\partial^2\varphi}{\partial y\,\partial z}$$

and from the equation of mixed partial derivatives (Sec. 4.1), we get the first relation of (54). The others are derived in similar fashion.

Conversely, suppose a field specified throughout the space be irrotational.* Then this field is of necessity noncirculatory, i.e.

$$\oint\limits_{(L)} \mathbf{A}\cdot d\mathbf{r} = 0$$

for any closed contour (L). Indeed, span the contour (L) with a surface (S). Applying the Stokes theorem (18) to (S), we find that if curl $\mathbf{A} = \mathbf{0}$, then

$$\oint\limits_{(L)} \mathbf{A}\cdot d\mathbf{r} = \int\limits_{(S)} \operatorname{curl}\mathbf{A}\cdot d\mathbf{S} = 0$$

In other words, for an irrotational field the circulation around any infinitely small contour is zero, and so the circulation around any finite contour is also equal to zero.

In Sec. 10.3 we show that a noncirculatory field of necessity has a potential, which is constructed via the formula (see (8) of Ch. 10)

$$\varphi(M) = \int\limits_{\overset{\frown}{MM_0}} \mathbf{A}\cdot d\mathbf{r} \tag{55}$$

where M_0 is an arbitrarily fixed point and the choice of the path of integration is immaterial, since the line integral in a noncirculatory field depends solely on the position of the beginning and end of the path of integration. True, in Sec. 10.3 we spoke of a force field and interpreted the potential as work, but from the mathematical point of view such a concrete interpretation is inessential since we can speak about any vector field and its potential.

* If the field is considered only in a portion of the whole space, there may be certain complications that will be discussed in Sec. 11.13.

To summarize, when considering a vector field throughout a space, the requirements that this field be potential, irrotational or noncirculatory are completely equivalent so that fulfilment of one of these requirements implies the fulfilment of all the others. It again follows from formula (55), by virtue of the arbitrary character of the point M_0, that the potential is determined to within an additive constant.

If the vector field $\mathbf{A}$ is not only irrotational but is devoid of sources of vector lines (see Sec. 10.7); that is, if

$$\operatorname{curl} \mathbf{A} = \mathbf{0} \text{ and } \operatorname{div} \mathbf{A} = 0$$

then from the former it follows that $\mathbf{A} = -\operatorname{grad} \varphi$ and from the latter, that

$$\operatorname{div} \operatorname{grad} \varphi = -\operatorname{div} \mathbf{A} = 0, \text{ or } \nabla^2 \varphi = 0$$

Thus, in this case the potential of the field must satisfy the Laplace equation (which was examined in Sec. 10.9) or, to put it differently, it must be a harmonic function.

If an irrotational field considered throughout a space is without sources not only at a finite distance but also at infinity, then the vector of such a field is simply identically zero, that is, there is actually no field. This is associated with the following property of harmonic functions: a function that is harmonic throughout a space and is equal to zero at infinity is identically zero. A still more general fact can be proved, though we will not do so: a function that is harmonic throughout a space and is bounded at infinity is identically equal to a constant, which means that such a potential is also associated with a zero field. Thus, so that the theory can have content, we must either allow for field sources at infinity, that is, the unlimitedness of a potential at infinity, or allow for sources at a finite distance, which means not assuming the potential to be harmonic throughout the whole space. For example, to a field $\mathbf{A}$ = constant there corresponds a potential $\varphi = -\mathbf{A} \cdot \mathbf{r}$ (check this!) that is unbounded at infinity, while to the Newtonian field (Sec. 10.3) there corresponds a potential that is everywhere harmonic except at one point.*

Let us return to the force field $\mathbf{F}$ and use the letter A to denote the work, as in Sec. 10.3. This force can have a potential but it can also fail to correspond to any potential. This latter fact can be determined in two ways: either by verifying that $\operatorname{curl} \mathbf{F} \neq \mathbf{0}$, that is, that at least one of the conditions

$$\frac{\partial F_z}{\partial y} = \frac{\partial F_y}{\partial z}, \quad \frac{\partial F_x}{\partial z} = \frac{\partial F_z}{\partial x}, \quad \frac{\partial F_y}{\partial x} = \frac{\partial F_x}{\partial y}$$

* At this point $\Delta\varphi$ is infinite, proportional to the delta function, so that although the point is "small" compared with an infinite volume, this point must not be forgotten.

Fig. 154

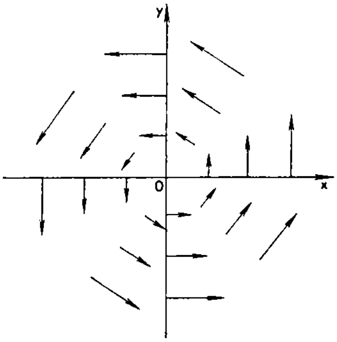
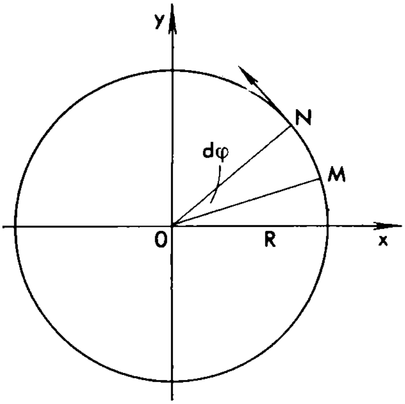


Fig. 155



(see the conditions (54)) is not fulfilled, or by verifying that the field is not noncirculatory, which is to say that at least for one closed contour the work of the force **F** is different from zero.

Take the following example. Suppose a force **F** lies in the *xy*-plane, i.e., it forms a plane field and at each point is perpendicular to the straight line joining this point with the coordinate origin (it is perpendicular to the radius vector) and is directed counterclockwise. Now suppose the magnitude of the force is proportional to the distance r of the point to the origin, that is, $|\mathbf{F}| = ar$, where a is a constant of proportionality. (Fig. 154 shows such a force for several points of the plane.) Suppose a body is in motion about a circle of radius R under the action of such a force (Fig. 155). A very small path from M to N (this corresponds to a rotation through the angle $d\varphi$) can be taken approximately to be rectilinear. It is equal to $R\,d\varphi$, as the arc length of a circle corresponding to the central angle $d\varphi$. The force is directed along the path and so the work performed on the section MN is

$$dA = R\,d\varphi \cdot |\mathbf{F}| = R\,d\varphi \cdot aR = aR^2\,d\varphi$$

Hence if the point traverses the entire circle, the work done is

$$A = \int_0^{2\pi} aR^2\,d\varphi = aR^2 \cdot \int_0^{2\pi} d\varphi = \pi R^2 \cdot 2a$$

The work is proportional to the area of the circle and is by no means zero. (It can be proved that for the force at hand the work done in traversing a closed curve is proportional to the area bounded by the curve of motion for any shape whatsoever of that curve.) It is clear that such a force cannot correspond to any potential.

The last conclusion can also be done by formally computing the curl. Since the field of force that results in this case is the same as the field of linear velocities in the case of a rotation about the z-axis with angular velocity a, we can take advantage of formula (5) to get

$$\mathbf{F} = (a\mathbf{k}) \times \mathbf{r} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 0 & 0 & a \\ x & y & 0 \end{vmatrix} = -ay\mathbf{i} + ax\mathbf{j} \tag{56}$$

(this can easily be obtained directly from Fig. 154). From this we have

$$\operatorname{curl} \mathbf{F} = \nabla \times \mathbf{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ -ay & ax & 0 \end{vmatrix} = a\mathbf{k} + a\mathbf{k} = 2a\mathbf{k} \neq \mathbf{0}$$

Thus, the curl is not equal to zero and so the field is not a potential field.

Exercise

Prove that the field

$\mathbf{A} = 2xz\mathbf{i} + y^2\mathbf{j} + x^2\mathbf{k}$

is a potential field and construct its potential.

11.10 The curl of a velocity field

The concept of a curl is especially pictorial when considering a field of linear velocities $\mathbf{v}$ of particles of a continuous medium. Let us examine some examples.

Let the medium be in translational motion like a rigid body. Then $\mathbf{v}$ = constant and since the curl is expressed with the aid of operations of differentiation, it follows that curl $\mathbf{v} = \mathbf{0}$. Here, there is no circulation of velocity.

Suppose the medium is rotating about an axis with angular velocity ω like a rigid body. We take advantage of the fact that the curl is invariantly connected with the field and, hence, when calculating the curl we can arrange the coordinate axes in any way we wish for the sake of convenience. And so we send the z-axis along the axis of rotation. Then computations similar to those carried out at the end of Sec. 11.9 yield

$$\mathbf{v} = -\omega y\mathbf{i} + \omega x\mathbf{j},$$
$$\operatorname{curl} \mathbf{v} = 2\omega\mathbf{k} = 2\boldsymbol{\omega}$$

Hence, the curl here is constant throughout the space and is equal to twice the vector of angular velocity. For this reason, the circulation of linear velocity around a small contour is a maximum if this contour is perpendicular to the axis of rotation (then the curl is projected on the normal to the corresponding small area in full length) and if the plane of the contour is parallel to the axis of rotation, then the circulation is zero.

It can be demonstrated that any motion of a rigid body at every instant of time is the result of a superposition of translational and rotational motions. By virtue of the preceding two paragraphs, at every instant the curl of the linear velocity of a rigid body is the same at all points of the body and is equal to twice the instantaneous vector of the angular velocity.

Now let us take up the motion of a medium in which the distances between its points vary. Suppose we have a "pure" compression of gas by a piston along the x-axis to the plane $x = 0$; then the velocity field is of the form

$$\mathbf{v} = -\lambda x\mathbf{i}$$

where λ is the proportionality constant. Calculating the curl, we get

$$\operatorname{curl} \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ -\lambda x & 0 & 0 \end{vmatrix} = \mathbf{0}$$

Cauchy proved that any motion of a small volume of a continuous medium under strain (of a fluid or a solid under strain) is at any time the result of a superposition of translational and rotational motions and also of "pure" compressions and tensions. (Note that small portions of an incompressible liquid may experience simultaneously "pure" compressions and tensions in different directions.) Since a nonzero curl is obtained only for rotational motion, we see that for an arbitrary motion of a medium the curl of the field of linear velocity $\mathbf{v}$ of particles is equal at each point to twice the angular velocity vector of the corresponding particle. True, in the general case, the curl at different points is different. Thus, in the case of fluid flow the fact

Fig. 156

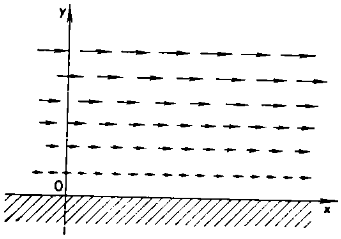
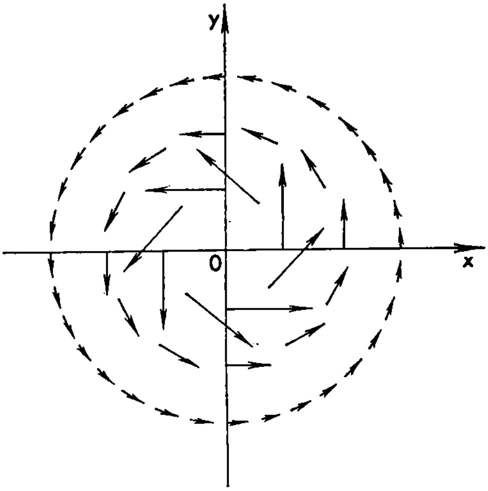


Fig. 157



that the curl of the field of linear velocity is different from zero indicates the presence of a vorticity, a rotation, whence the term rotation (curl).

Now let us examine the "shear" flow represented by arrows in Fig. 156. This is what happens in the flow of a viscous fluid along a solid wall. Here, $\mathbf{v} = ay\mathbf{i}$, where a is the proportionality factor, whence curl $\mathbf{v} = -a\mathbf{k}$ (check this!). This is a vortex-type flow and every particle of the fluid rotates clockwise with angular velocity $a/2$.

Finally, let us examine a simple "vortex line", i.e. a plane-parallel field defined in the xy-plane by the equation

$$\mathbf{v} = (p\mathbf{k}) \times \frac{\mathbf{r}^0}{r} = (p\mathbf{k}) \times \frac{\mathbf{r}}{r^2}$$

This field, which is shown in Fig. 157, is somewhat reminiscent of the field in Fig. 154, but here the modulus of the field vector is inversely (not directly) proportional to the distance from the origin. Here the z-axis is the axis of the vortex line.

As in (56), let us find the expression of velocity in coordinate form:

$$\mathbf{v} = -\frac{py}{x^2+y^2}\mathbf{i} + \frac{px}{x^2+y^2}\mathbf{j} \tag{57}$$

whence it is easy to calculate directly that curl $\mathbf{v} = \mathbf{0}$ off the z-axis (check this!). Thus, each particle of the fluid traverses the vortex line, undergoing deformation without performing rotational motion about its own axis. At the same time, since the circulation of vector $\mathbf{v}$ around any circle with centre at the origin is

$$\frac{p}{r}\cdot 2\pi r = 2\pi p$$

it follows that at the origin there is a "source of circulation" whose density in the xy-plane is equal to

$$\frac{d\Gamma}{dS} = 2\pi p\delta(\mathbf{r})$$

(where δ is the delta function,* see Sec. 6.3). Since the curl of a plane-parallel field is perpendicular to the plane of the field, in this example we get

$$\text{curl}\,\mathbf{v} = 2\pi p\delta(x\mathbf{i} + y\mathbf{j})\,\mathbf{k} = 2\pi p\delta(x)\,\delta(y)\,\mathbf{k} \tag{58}$$

(regarding the last representation, see Sec. 6.3).

Exercise

Find the directions in a particle of a vortex line that remain invariant under an infinitesimal displacement.
Hint. Proceed from the fact that the vector $d\mathbf{r}$ must not turn.

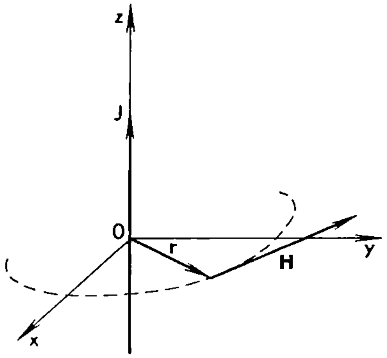
11.11 Magnetic field and electric current

The concepts of vector product and curl are extensively employed in discussions of magnetic fields. For the sake of simplicity we consider the fields in a vacuum.

A magnetic field is completely characterized at every point of space by the "magnetic-field intensity" vector $\mathbf{H}$. This vector is largely analogous to the vector $\mathbf{E}$ of electric-field intensity (Secs. 10.5 and 10.9) but does not act on fixed charges, like $\mathbf{E}$, but on permanent magnets or (what will soon be seen to be the same thing) on moving charges. A magnetic field can also be established by permanent magnets or moving charges.

* Here we have a two-dimensional delta function: $\delta(\mathbf{r}) = \delta(x)\,\delta(y)$ since the integration is over a plane, $\Gamma = \int \text{curl}\,\mathbf{A}\cdot d\mathbf{S}$.

Fig. 158



Let us first consider the simplest scheme. Suppose a magnetic field is established by a current J flowing along an infinite (practically speaking, a very long) rectilinear conductor which we assume to be coincident with the z-axis. Clearly, we then have a plane-parallel field, which it suffices to consider solely in the xy-plane. Experiment tells us that the vector **H** is then obtained as indicated in Fig. 158; it lies in the xy-plane and is perpendicular to the radius vector **r**, the direction of **H** being determined by the rule of the right-handed screw. The intensity is directly proportional to the current and inversely proportional to the distance of the point from the conductor, i. e. $H = a\frac{J}{r}$, where a is a proportionality factor; for a certain choice of units of H and J, it turns out equal to $2/c$, where c is the velocity of light, so that $H = \frac{2J}{cr}$.

The field **H** under consideration has precisely the form shown in Fig. 157, which is to say that it has the formula (57), where for p we have to substitute $\frac{2J}{c}$. From this, by formula (58), we get

$$\operatorname{curl}\mathbf{H} = 2\pi\frac{2J}{c}\,\delta(x)\,\delta(y)\,\mathbf{k}$$

On the other hand, the product $J\delta(x)\,\delta(y)\,\mathbf{k}$ is the current density shown in Fig. 158. We will denote the current density by **j** (do not confuse this with the unit vector of the y-axis!). In this example we then have

$$\operatorname{curl}\mathbf{H} = \frac{4\pi}{c}\mathbf{j} \tag{59}$$

By this formula, **j** is expressed in terms of **H**. We can obtain the inverse, in which **H** is expressed in terms of **j**. To do this we introduce the "vector potential" **A** of the field **H** by analogy with the scalar potential of an electric field (see Sec. 10.5),

$$\mathbf{A}(\mathbf{r}) = \frac{1}{c}\int \frac{\mathbf{j}(\mathbf{r}_0)}{|\mathbf{r} - \mathbf{r}_0|}\, d\Omega_0 \tag{60}$$

In the case at hand this potential is equal to

$$\mathbf{A}(\mathbf{r}) = \frac{1}{c}\int\limits_{-\infty}^{\infty} \frac{J\mathbf{k}\, d\zeta}{\sqrt{x^2 + y^2 + (\zeta - z)^2}} = \frac{J}{c}\int\limits_{-\infty}^{\infty} \frac{d\zeta}{\sqrt{x^2 + y^2 + (\zeta - z)^2}}\,\mathbf{k}$$

Although the last integral diverges, we can differentiate it with respect to parameters (cf. Sec. 10.6), whence, in particular,

$$\operatorname{curl}\mathbf{A} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ 0 & 0 & \dfrac{J}{c}\displaystyle\int \end{vmatrix}$$

$$= \frac{J}{c}\left(\frac{\partial}{\partial y}\int\limits_{-\infty}^{\infty} \frac{d\zeta}{\sqrt{x^2 + y^2 + (\zeta - z)^2}}\,\mathbf{i} - \frac{\partial}{\partial x}\int\limits_{-\infty}^{\infty} \frac{d\zeta}{\sqrt{x^2 + y^2 + (\zeta - z)^2}}\,\mathbf{j}\right)$$

$$= -\frac{J}{c}\int\limits_{-\infty}^{\infty} [x^2 + y^2 + (\zeta - z)^2]^{-3/2}\, d\zeta \cdot (y\mathbf{i} - x\mathbf{j})$$

This integral can easily be evaluated by the substitution $\zeta - z =$ $= \sqrt{x^2 + y^2}\tan s$ and we get (check this!)

$$\operatorname{curl}\mathbf{A} = -\frac{J}{c}\cdot\frac{2}{x^2 + y^2}(y\mathbf{i} - x\mathbf{j})$$

We have again arrived at the formula (57) with $p = \frac{2J}{c}$, i.e. at **H**:

$$\mathbf{H} = \operatorname{curl}\mathbf{A} = \frac{1}{c}\operatorname{curl}\int \frac{\mathbf{j}(\mathbf{r}_0)}{|\mathbf{r} - \mathbf{r}_0|}\, d\Omega_0 \tag{61}$$

When considering systems of infinite rectilinear conductors, their magnetic fields **H** and also the current densities are additive, which is to say that (59) and (61) hold true all the same. We will also assume that they are valid for any stationary, not necessarily rectilinear, distribution of currents in space. (This is rather clear for formula (59) since it connects the values of **H** and **j** in a single point, and formula (61) can be derived from (59), but we will not go into that here.) This assumption is justified by the fact that the consequences derived from it are in agreement with one another and find experimental corroboration.

An important consequence follows from (59). To obtain it, we apply the Stokes theorem to a computation of the circulation of the vector **H** around any closed contour (L) bounding the surface (σ)

$$\oint\limits_{(L)} \mathbf{H}\cdot d\mathbf{r} = \int\limits_{(\sigma)} (\operatorname{curl}\mathbf{H})_n\,d\sigma = \frac{4\pi}{c}\int\limits_{(\sigma)} j_n\,d\sigma \qquad (62)$$

The integral on the right has a simple physical meaning (cf. Sec. 10.4): this is the quantity of electricity passing outwards through (σ) in unit time, i. e. the total current J through (σ). Thus, the circulation of the vector **H** around any closed contour is proportional to the total current through the surface bounded by that contour.

Another important consequence is obtained if we take the divergence of both sides of (59). We obtain

$$\operatorname{div}\mathbf{j} = \frac{c}{4\pi}\operatorname{div}\operatorname{curl}\mathbf{H} = 0$$

(see (51)). This "law of conservation of electricity" in the stationary case could have been derived independently of formula (59): it means that in unit time just as much electricity flows into any volume as flows out of it, i. e. the total flux of the vector **j** through any closed surface is equal to zero (cf. similar reasoning in Sec. 10.8).

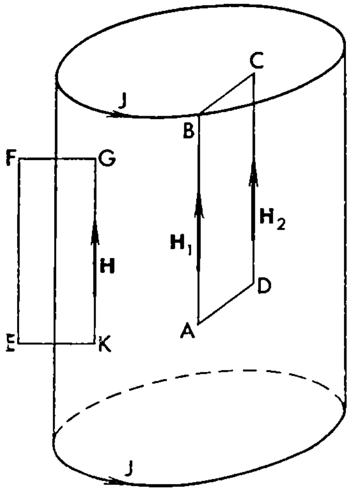
If we take the divergence of both sides of (61), we see that

$$\operatorname{div}\mathbf{H} = \operatorname{div}\operatorname{curl}\mathbf{A} = 0 \qquad (63)$$

Thus (see Sec. 10.7), unlike an electric field, a magnetic field does not have sources of vector lines, which means that magnetic lines of force cannot originate or terminate anywhere. (Since we relied solely on (51), it follows that in the general case, too, a vector field can be the curl of some "vector potential" only if the divergence of this field is zero, which means it does not have sources of vector lines. Such fields are customarily called *solenoidal fields.*)

To illustrate, let us compute the magnetic field of an infinite cylindrical coil having n turns in the winding per unit length and carrying a current J. We will disregard the thickness of the winding (later on we will see that it is inessential) and for the sake of simplicity will regard the winding not as a helical winding but as a circular one. From formula (60) we see that in the given case the vector potential **A** is parallel to the plane of the turns and, by the meaning of the problem, forms a plane-parallel field. But then from formula (61) and from the definition (43) of curl it immediately follows that the field **H** is everywhere parallel to the axis of the coil and does not vary along the coil, which is to say it is also plane-parallel. Now apply formula (62) to the rectangular contour $ABCD$ shown in Fig. 159.

Fig. 159



Since the integrals along the line segments BC and DA are zero and since no current flows through the surface bounded by this contour, we get

$$H_1 h - H_2 h = 0$$

(h is the altitude of the rectangle), whence $H_1 = H_2$. Thus, a magnetic field inside a solenoid (coil) is constant not only in height but also in cross section. Similarly, we find that the field outside the coil is also constant, and since at infinity it is zero, it is zero everywhere outside the coil. Finally, applying (62) to the rectangular contour $EKGF$, we get $Hh = \frac{4\pi}{c} Jnh$, whence we find the magnetic-field intensity inside the coil to be

$$H = \frac{4\pi}{c} Jn$$

We see that it does not depend on the radius of the coil but only on the number of ampere turns per unit length. For one thing, from this there follows the earlier-mentioned independence of the field of the winding thickness. We leave it to the reader to obtain an expression for the potential $\mathbf{A} = \frac{H}{2} r\mathbf{e}_\varphi (r < r_0)$, $\mathbf{A} = \frac{Hr_0^2}{2r} \mathbf{e}_\varphi (r > r_0)$, where $\mathbf{r} = x\mathbf{i} + y\mathbf{j}$, $r = |\mathbf{r}|$, $\mathbf{e}_\varphi = \mathbf{k} \times \mathbf{r}^0$, and r_0 is the radius of the coil (do not confuse r_0 with $\mathbf{r}^0 = \mathbf{r}/r$).

The problem of constructing a magnetic field from a given system of currents can be approached differently. We assume that the magnetic field at any point M is created by each current element $d\mathbf{J} = \mathbf{j}\, d\Omega$ in accord with the *Bio-Savart law*

$$d\mathbf{H} = \frac{1}{c(M_0M)^3}(d\mathbf{J}\times\overrightarrow{M_0M})$$

where M_0 is the point at which the current element is located. The total vector $\mathbf{H}$ at M is found by means of integrating over all elements of current. The fact that the formula for $d\mathbf{H}$ is written in precisely the way it is, is justified by the fact that all consequences of this formula are in agreement with experiment. This approach is equivalent to our earlier approach: from the Bio-Savart law we can derive formulas (59) and (61), and conversely.

Exercise

Obtain from the Bio-Savart law the intensity of the magnetic field: (a) of a straight infinite conductor; (b) on the axis of a solenoid from one turn.

11.12 Electromagnetic field and Maxwell's equations

Actually, in the general case, there exist in one and the same region of space both an electric and a magnetic field; the result of this is an *electromagnetic field*, which at every point in space (at any rate, in a vacuum, and that is what we assume here) is characterized by two vectors: the electric vector $\mathbf{E}$ and the magnetic vector $\mathbf{H}$. The differential equations connecting these vectors are called *Maxwell's equations* and play a very important role in physics.

Actually, the appropriate equations have already been written out for a stationary electromagnetic field. These are first of all the equations

$$\operatorname{curl}\mathbf{E} = \mathbf{0}, \quad \operatorname{div}\mathbf{E} = 4\pi\rho$$

of which the former follows from the potentiality of a steady-state electric field (Secs. 10.5 and 11.9) and the latter is the equation (42) of Ch. 10. Besides, they are also the equations

$$\operatorname{curl}\mathbf{H} = \frac{4\pi}{c}\mathbf{j}, \quad \operatorname{div}\mathbf{H} = 0,$$

(see (59) and (63)).

For the stationary case the interrelationship between the electric and the magnetic field is not apparent. The situation is quite different in the nonstationary case that we now take up. It turns out that every change in the electric field affects the magnetic field and every change in the magnetic field acts on the electric field, so that there is no way of considering these fields separately.

When considering the former action it is exceedingly useful to apply the concept of displacement currents. To derive the appropriate

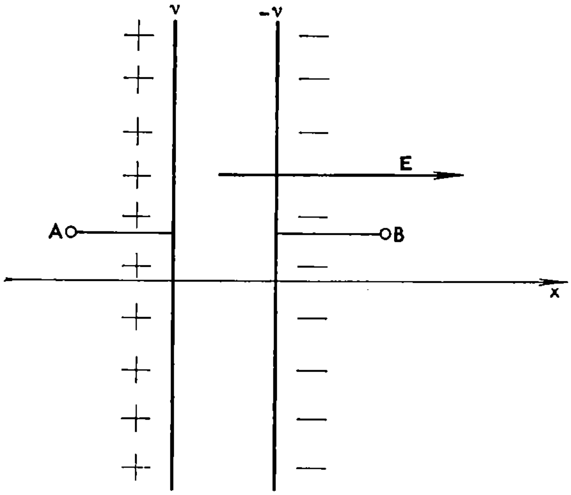


Fig. 160

formulas let us first consider a capacitor (Fig. 160), the plates of which are charged with surface density $+\nu$ for the left one and $-\nu$ for the right one. Each of these plates generates in the space between them an electric field, which can be calculated by the formulas (29) of Ch. 10 so that the total field is

$$\mathbf{E} = 4\pi\nu\mathbf{i} \tag{64}$$

If ν increases, then the left-hand plate is approached from without by positive charges that settle on it at the rate of $J = \frac{d(S\nu)}{dt}$, where S is the area of one plate; negative charges settle at the same rate on the right-hand plate. If for a moment we imagine that the points A and B are connected by a conductor on which charges do not settle, then by the law of conservation of electricity there should pass through any cross section between A and B one and the same current J. Actually, the charges do not pass between the plates of the capacitor, but the law of constancy of currents can be preserved if we introduce the idea of a *displacement current* J that flows, as it were, from A to B. By formula (64) the density of this displacement current is

$$\mathbf{j}_{\text{dis}} = \frac{J}{S}\mathbf{i} = \frac{d\nu}{dt}\mathbf{i} = \frac{1}{4\pi}\frac{\partial\mathbf{E}}{\partial t}$$

It turns out that in the formulas relating the magnetic vector $\mathbf{H}$ and the current $\mathbf{j}$, the change in electric field has to be replaced by the displacement current whose density is computed from the same formula, $\mathbf{j}_{\text{dis}} = \frac{1}{4\pi} \frac{\partial \mathbf{E}}{\partial t}$. And so for a nonstationary electric field, instead of (59) we have to write the equation

$$\operatorname{curl} \mathbf{H} = \frac{4\pi}{c} (\mathbf{j} + \mathbf{j}_{\text{dis}}) = \frac{4\pi}{c} \mathbf{j} + \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t} \tag{65}$$

Similarly, also in the formula (61), to construct **H** we have to substitute $\mathbf{j} + \mathbf{j}_{\text{dis}}$ for **j**. But then, arguing as in Sec. 11.11, we can derive the relation

$$\operatorname{div} \mathbf{H} = 0 \tag{66}$$

(see (63)).

Now let us turn to a variable magnetic field. Experiments (Faraday) show that any change in the magnetic field induces an electric field, which, unlike the field generated by charges, is a rotational field. We now show that the corresponding law of induction is of the form

$$\operatorname{curl} \mathbf{E} = -\frac{1}{c} \frac{\partial \mathbf{H}}{\partial t} \tag{67}$$

By virtue of Sec. 11.9 the electric field here does not have a potential, i.e. we cannot speak of a difference of potentials in the field, and the work done by the field depends not only on the beginning and end of the path, but on the entire trajectory.

In order to derive formula (67), imagine placed in a field a closed conductor (L) bounding a surface (σ). According to the results of Faraday's experiments, the change in magnetic flux $\int\limits_{(\sigma)} \mathbf{H} \cdot d\boldsymbol{\sigma}$ gives rise to an emf in the circuit (L) that is proportional to the rate of change of the flux. However, the indicated emf is equal to the sum of the elementary emf in small portions of the circuit, and these small emf are equal, as it is easy to see, to $\mathbf{E} \cdot d\mathbf{r}$. Thus

$$\oint\limits_{(L)} \mathbf{E} \cdot d\mathbf{r} = -k \frac{d}{dt} \int\limits_{(\sigma)} \mathbf{H} \cdot d\boldsymbol{\sigma} = -k \int\limits_{(\sigma)} \frac{\partial \mathbf{H}}{\partial t} \cdot d\boldsymbol{\sigma}$$

Here the minus sign is taken by the Lenz law, according to which the induced emf is counter to the change in flux; k is a proportionality factor, which can be shown to be equal to $1/c$; the differentiation with respect to t has been brought under the integral sign as differentiation with respect to a parameter (Sec. 3.6). Transforming the left-hand side by the Stokes theorem, we get

$$\int\limits_{(\sigma)} \operatorname{curl} \mathbf{E} \cdot d\boldsymbol{\sigma} = -\frac{1}{c} \int\limits_{(\sigma)} \frac{\partial \mathbf{H}}{\partial t} \cdot d\boldsymbol{\sigma}$$

whence follows (67) because (σ) is arbitrary.

The case of a force that does not correspond to a potential is realized in a transformer. The current flowing in the primary winding of a transformer sets up a magnetic field. If the current in the primary winding varies with time, a variable magnetic field is set up which gives rise to an electric field of precisely that type (a so-called rotational field), as shown in Fig. 154. This means that the force acting on a charged body, say an electron, located inside the secondary winding does not correspond to any potential at all.

An electron in motion round a circle can build up more and more energy. This of course does not mean there is a violation of the law of conservation of energy, since the accelerated motion of the electron will in turn affect the primary winding, giving rise to an additional consumption of energy in the current sources supplying it. But for a single electron the law that "the sum of the kinetic and potential energies of an electron is a constant" does not hold true because the force acting on the electron does not correspond to any potential energy. This principle of accelerating electrons is used in an electron accelerator called a betatron. Here, the electric field accelerates the electrons and the magnetic field appropriately bends their paths, making them move in circles.

Since a rotational electric field appears only during a change in the magnetic field, such an electric field will not be constant for a long time, for when the magnetic field reaches a maximum, the rate of its change becomes zero, after which the electric field vanishes as well.

Because of the short duration of the electric field, heavy particles (protons, say) do not have time to build up sufficient energy in a betatron. Electrons are light and so they build up energy very well. There are betatrons in which electrons acquire an energy of tens of millions of electron volts, which is the energy they would acquire by passing between a potential difference of tens of millions of volts. Fast electrons emitted in radioactive transformations were called beta rays in the early days of radioactive research when the physical nature of these rays had not yet been elucidated. Whence also the name "betatron" for the instrument in which fast electrons are obtained by accelerating slow electrons without the phenomenon of radioactivity occurring.

The equation

$$\operatorname{div} \mathbf{E} = 4\pi\rho \tag{68}$$

which connects the electric field and charges (see formula (42) of Ch. 10), holds true in the nonstationary case as well. All four equations, (65) to (68), form the system of Maxwell's equations. To them we can add the continuity equation

$$\frac{\partial\rho}{\partial t} + \operatorname{div} \mathbf{j} = 0 \tag{69}$$

which is derived in exactly the same way as the similar equation (41), Ch. 10, in hydrodynamics, and also one or another relation connecting **j** and **E** (the generalized Ohm law, which in the simplest cases is of the form $\mathbf{j} = \lambda\mathbf{E}$, where λ is the coefficient of electric conductivity) and including in the general case the action of external forces.

Exercise

Derive equation (69) from equations (65) and (68).

11.13 Potential in a multiply connected region

Consider the force field **F**. From both conditions for the existence of a potential — the condition $\operatorname{curl} \mathbf{F} = \mathbf{0}$ (Sec. 11.9) and the condition that the work done in traversing a closed curve is zero (Sec. 10.3) — it is evident that the possibility of disrupting these conditions, that is, the possibility of the existence of forces of a nonpotential type, involves considering a function of two or three variables instead of one. Indeed, in the case of one variable (motion in a straight line) it is possible to return to the original point only by covering each section of the path twice: once in one direction and then in the return direction. And so (if the force does not depend on the time or on the velocity but only on the position of the body) in the case of motion in a straight line, the work of the force is zero if the path ends at the point from which it started.

In arbitrary motion in a plane or in space, it is possible to start out from the initial point along a certain curve, reach a terminal point, and then return to the beginning via a different curve. It may then happen that the work is not equal to zero. Therein lies the difference between motion along a single straight line when the potential energy $U(x)$ corresponds to any force $\mathbf{F}(x)$, and motion in a plane or in space where the potential energy may not exist at all.

Now let us examine the motion of a body in a plane or in space. If we force the body to move along a single definite open curve, then we again have the same motion in a straight line. Indeed, in that case we will describe the position of the body on the curve by the path traversed along that curve (the path is reckoned from some chosen point on the curve). Then we actually have to do with one variable, the path length of the curve. Therefore, when considering the open second winding of a transformer, we can speak of a difference of potentials (or, better, of an electromotive force) at the ends. But if we place a closed circle in the first winding, then there will not be any definite potential difference in the circle between the two points A and B (Fig. 161). The work done in carrying a charge from A to B depends on which curve is traversed, ADB or ACB.

Now suppose the body is not compelled to move along a fixed curve but can move to the side as well. Then, for a potential to exist, it is required that the work of the force in moving the body

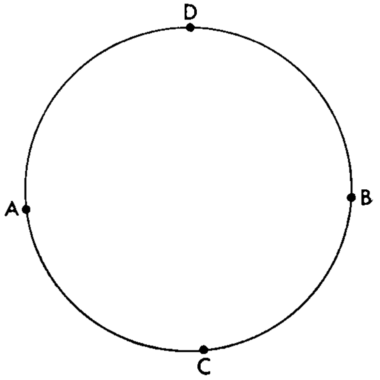


Fig. 161

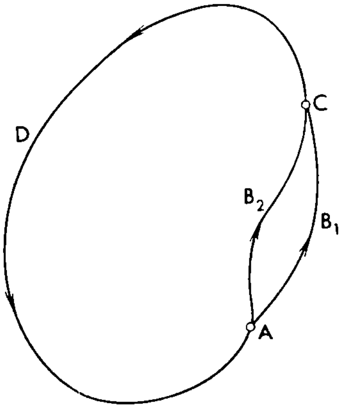
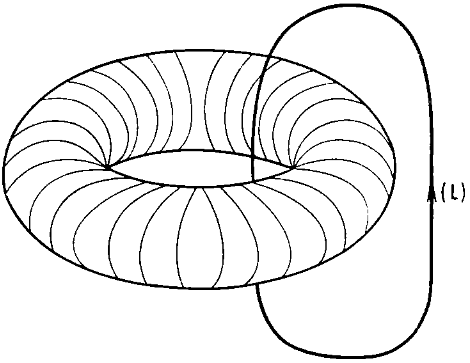


Fig. 162

around any infinitesimal closed curve be equal to zero. By virtue of Sec. 11.7, this is equivalent to the condition curl $\mathbf{F} = \mathbf{0}$ (since such work is equal to the circulation of the vector $\mathbf{F}$ around the given contour), which means that the field $\mathbf{F}$ must be irrotational.

But let curl $\mathbf{F} = \mathbf{0}$ not throughout the space but only in a certain portion (region) (G). Then the work done around a finite closed contour (L) lying in (G) need not necessarily equal zero! We can only assert that, given an infinitesimal deformation of the closed contour, the work does not change since the work done around the contour AB_1CDA (Fig. 162) differs from the work done around the contour AB_2CDA by the amount of work done round the contour AB_1CB_2A, which is equal to zero. By carrying out such a deformation, we can obtain a contour lying in (G) and substantially differing from the original contour, although the work of the force $\mathbf{F}$ done around these contours is the same.

Fig. 163



From this it is clear that if the contour (L) can, within the limits of (G), be contracted to a point via a continuous deformation, then the work of the force around (L) is zero. This is true because after such a contraction the work will clearly be zero (for there will be no motion). But since the work did not change during the deformation of the contour, it was zero for the original contour as well.

The region (G) can have the property of being *simply connected,* which means that any closed curve in such a region can, by means of a continuous deformation, be contracted into a point without touching the boundary of the region. For example, the interior of a circular cylinder, the interior or exterior of a sphere are all simply connected regions. By contrast, the exterior of an infinite circular cylinder, the interior or exterior of a torus (the surface of a doughnut) are *nonsimply connected* regions.

If the region (G) is simply connected, then it follows from the foregoing that in this case the work of a force over a finite closed curve lying in (G) is of necessity equal to zero. If we disregard the specific physical interpretation of a field, then we can draw the following general mathematical conclusion (cf. Sec. 11.9): in a simply connected region, an irrotational field is necessarily potential, that is, noncirculatory.

In contrast, if the region is nonsimply connected, then the circulation of an irrotational field may be different from zero if the contour at hand cannot be contracted into a point by a continuous deformation while remaining within the region, which is to say that an irrotational field may be circulatory. Such for instance are: a magnetic field inside a toroidal coil (with the planes of the turns passing through the axis of rotation) carrying direct current; a velocity field with irrotational fluid flowing round a closed channel, and so forth. If we construct the potential using formula (55), it will be ambiguous: if we traverse a closed contour, then the circulation of the field around this contour will be subtracted from the potential. When we speak of a potential field, we have in mind only a single-valued potential; thus, an irrotational field in a nonsimply connected region need not necessarily be potential.

Of considerable interest is a toroidal transformer with a circular iron core. When an alternating current is passed through the winding, an irrotational electric field is formed in the outer space, this field has a nonzero circulation around the contour (L) coupled to the torus (Fig. 163).

Exercise

Prove that the plane field defined by formula (57) outside the circle (L) with centre at the origin is an irrotational field. What will happen if the potential is constructed by formula (55)?

ANSWERS AND SOLUTIONS

Sec. 11.1

1. $\overrightarrow{AB} = -2\mathbf{i} + \mathbf{j} + 4\mathbf{k}$, $\overrightarrow{AC} = 3\mathbf{j} + 5\mathbf{k}$, $\overrightarrow{AB}\times\overrightarrow{AC} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ -2 & 1 & 4 \\ 0 & 3 & 5 \end{vmatrix} =$ $= -7\mathbf{i} + 10\mathbf{j} - 6\mathbf{k}$, $S_{\Delta ABC} = \frac{1}{2}\,|\overrightarrow{AB}\times\overrightarrow{AC}| = \frac{1}{2}\sqrt{7^2 + 10^2 + 6^2} =$ $= \frac{1}{2}\sqrt{185} = 6.8$.

2. Since $(\mathbf{a}\times\mathbf{b})\cdot\mathbf{c} = \begin{vmatrix} 2 & -1 & 0 \\ 3 & 0 & 2 \\ -1 & -1 & 1 \end{vmatrix} = 9 > 0$, the triad is right-handed.

3. $(\mathbf{i}\times\mathbf{i})\times\mathbf{j} = \mathbf{0}$, $\mathbf{i}\times(\mathbf{i}\times\mathbf{j}) = -\mathbf{j}$, that is, $(\mathbf{i}\times\mathbf{i})\times\mathbf{j} \neq \mathbf{i}\times(\mathbf{i}\times\mathbf{j})$.

Sec. 11.2

If (1) $\mathbf{b} = \mathbf{0}$, or (2) $\mathbf{b}$ is applied to point O, or (3) $\mathbf{b}$ is directed exactly to O or from O so that the straight line on which the vector $\mathbf{b}$ lies passes through O. The third case embraces the first two cases.

Sec. 11.3

1. $T = \int dt = \int \frac{dr}{dr/dt}$; the substitution of $\frac{dr}{dt}$ from equation (15) leads to the integral $T = 2\int\limits_{r_{\min}}^{r_{\max}} \sqrt{\frac{m}{2}}\left(E + \frac{k}{r} - \frac{G^2}{2mr^2}\right)^{-1/2} dr$. Calculating for $-\frac{mk^2}{2G^2} < E < 0$ yields $T = \pi k\sqrt{-\frac{m}{2E^3}}\cdot$ The semi-

major axis a of the ellipse is equal to $\frac{1}{2}\left(\frac{p}{1-\varepsilon}+\frac{p}{1+\varepsilon}\right)$ or, after transformations, $-\frac{k}{2E}$. For $k = \varkappa mM$ we get $T = \frac{2\pi}{\sqrt{\varkappa M}}\, a^{3/2}$. This is the statement of Kepler's third law in the general case.

2. Using formula (17) gives the value

$$\sqrt{2m}\int\limits_{r_{\min}}^{r_{\max}}\left(E-\frac{k}{2}r^2-\frac{G^2}{2mr^2}\right)^{-1/2}dr=\pi\sqrt{\frac{m}{k}}$$

for the period of variation of the radius. The difference is due to the fact that in the type of motion under consideration here the period of variation of the radius $r(t)$ is half the period of revolution of the particle.

Sec. 11.4

(a) Passing to the polar coordinates (α, φ) (see end of Sec. 4.7), we get

$$(J_z)_{\text{cyl}}=\int\limits_0^h dz\int\limits_0^{2\pi}d\varphi\int\limits_0^R \rho\alpha^2\alpha\,d\alpha=2\pi\rho\frac{R^4}{4}h=\frac{MR^2}{2}$$

(b) Denote the linear density of the rod by λ; then

$$(J_z)_{\text{rod}}=2\int\limits_0^{L/2}x^2\lambda\,dx=2\lambda\frac{1}{3}\left(\frac{L}{2}\right)^3=\frac{mL^2}{12}$$

Incidentally, these two results make it possible to analyze more precisely the rotation of a dancer.

Sec. 11.5

1. In the two-dimensional case, if the basis $\mathbf{e}'_i$ is obtained from the basis $\mathbf{e}_i$ by a rotation through the angle φ, then $\alpha_{11}=\alpha_{22}=\cos\varphi$, $\alpha_{12}=-\alpha_{21}=\sin\varphi$. If the tensor p_{ij} is symmetric, then

$$\begin{aligned}p'_{12}=\alpha_{1k}\alpha_{2l}p_{kl}&=(p_{22}-p_{11})\sin\varphi\cos\varphi+p_{12}(\cos^2\varphi-\sin^2\varphi)\\&=\frac{1}{2}(p_{22}-p_{11})\sin2\varphi+p_{12}\cos2\varphi\end{aligned}$$

Choosing $\tan 2\varphi = 2p_{12}/(p_{11}-p_{22})$, we get $p'_{12}=0$, which is what is required.

2. Here, $\eta_{12}=k$ (small) and the others are $\eta_{ij}=0$; therefore $\beta_{12}=\beta_{21}=\gamma_{12}=-\gamma_{21}=k/2$, the others are $\beta_{ij}=\gamma_{ij}=0$. Applying the solution of Problem 1, we find that to the tensor β_{ij} is associated a tension of $1+\frac{k}{2}$ times along the straight line

$x_1 = x_2$ and a compression of $1 - \frac{k}{2}$ times along the straight line $x_1 = -x_2$; by virtue of the solution of the third exercise of Sec. 9.5, to the tensor γ_{ij} there corresponds a rotation through the angle $\alpha = -\frac{k}{2}$. The entire picture can be regarded as a plane picture (in the plane x_1, x_2) or as a plane-parallel picture.

Sec. 11.6

1. Yes.

2. The basic formula is: $[\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3] = \mathbf{e}_4$. From it there follow three other formulas via circular permutations. For an interchange of two factors, the result is multiplied by -1. The formula for the product of any three vectors is similar to (3), but the determinant in it is of the fourth order. (It can be verified that the vector product is perpendicular to all the vectors being multiplied, is equal in absolute value to the volume of a parallelepiped constructed on them, and forms with them a quadruple of the "same sense" as the quadruple $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$, $\mathbf{e}_4$.)

Sec. 11.7

1. $$\operatorname{curl} \mathbf{A}(M) = -\frac{\oint\limits_{(dS)} \mathbf{A}\times d\mathbf{S}}{d\Omega} = -\lim_{(\Delta\Omega)\to M} \frac{\oint\limits_{(\Delta S)} \mathbf{A}\times d\mathbf{S}}{\Delta\Omega}.$$

2. $\oint\limits_{(L)} f(z)\,dz = \oint\limits_{(L)} (u + iv)\,(dx + i\,dy) = \oint\limits_{(L)} (u\,dx - v\,dy) + i\oint\limits_{(L)} (v\,dx + u\,dy)$. If we consider the field $\mathbf{A} = u\mathbf{i} - v\mathbf{j}$, then the integral $\oint\limits_{(L)} (u\,dx - v\,dy)$ is equal to the circulation of this field around (L) and, for this reason, by the Stokes theorem $\oint\limits_{(L)} (u\,dx - v\,dy) = \int\limits_{(S)} \operatorname{curl} \mathbf{A}\cdot d\mathbf{S}$. But $\mathbf{A} = A_x(x, y)\,\mathbf{i} + A_y(x, y,)\,\mathbf{j}$, that is, by formula (43),

$$\operatorname{curl} \mathbf{A} = \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)\mathbf{k} = \left(-\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)\mathbf{k} = 0$$

by virtue of the second formula of (17), Ch. 5. Which means this last integral is equal to zero. The integral $\oint\limits_{(L)} (v\,dx + u\,dy)$ is investigated in similar fashion.

Sec. 11.8

1. $u \operatorname{curl} \mathbf{A} + \operatorname{grad} u \times \mathbf{A}$, $\mathbf{B} \cdot \operatorname{curl} \mathbf{A} - \mathbf{A} \cdot \operatorname{curl} \mathbf{B}$.
2. $\nabla \cdot = (d\Omega)^{-1} \oint\limits_{(d\sigma)} \cdot\, d\boldsymbol{\sigma}$.

Sec. 11.9

$$\operatorname{curl} \mathbf{A} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ 2xz & y^2 & x^2 \end{vmatrix} = \mathbf{0}.$$ Applying formula (55), we choose $M_0 = (0, 0, 0)$ and the path MM_0 consisting of line segments connecting the points (x, y, z) and $(0, y, z)$, $(0, y, z)$ and $(0, 0, z)$, and $(0, 0, z)$ and $(0, 0, 0)$. Whence

$$\varphi(x, y, z) = \int\limits_x^0 2xz\, dx + \int\limits_y^0 y^2\, dy + \int\limits_z^0 x^2 \Big|_{x=0} dz = -x^2 z - \frac{y^3}{3}$$

Sec. 11.10

Since during time dt the vector $d\mathbf{r}$ will pass into the vector $d\mathbf{r} + d\mathbf{v}\, dt$, the condition of invariance of the direction $d\mathbf{r}$ is of the form $d\mathbf{v} \parallel d\mathbf{r}$. Since $\mathbf{v} = (p\mathbf{k}) \times \frac{\mathbf{r}}{r^2}$, it follows that $d\mathbf{v} = = (p\mathbf{k}) \times \left(\frac{d\mathbf{r}}{r^2} - \frac{2\mathbf{r} \cdot d\mathbf{r}}{r^4} \mathbf{r}\right)$, and from $d\mathbf{v} \parallel d\mathbf{r}$ we get $\frac{\mathbf{r}}{r^2} - \frac{2\mathbf{r} \cdot d\mathbf{r}}{r^4} \mathbf{r} \perp d\mathbf{r}$, that is, $(r^2\, d\mathbf{r} - 2(\mathbf{r} \cdot d\mathbf{r})\, \mathbf{r}) \cdot d\mathbf{r} = 0$. And so $r^2 (d\mathbf{r})^2 - 2(\mathbf{r} \cdot d\mathbf{r})^2 = 0$, $\cos \widehat{(\mathbf{r}, d\mathbf{r})} = \frac{\mathbf{r} \cdot d\mathbf{r}}{r |d\mathbf{r}|} = \pm \frac{\sqrt{2}}{2}$, that is, $d\mathbf{r}$ must form an angle of ± 45° with $\mathbf{r}$.

Sec. 11.11

(a) Let the current J flow along the z-axis and compute the vector $\mathbf{H}$ at the point in the xy-plane with radius vector $\mathbf{r}$. Then

$$d\mathbf{H} = \frac{1}{c(r^2 + \zeta^2)^{3/2}} (J\, d\zeta \mathbf{k} \times (\mathbf{r} - \zeta \mathbf{k})) = \frac{Jr\, d\zeta}{e(r^2 + \zeta^2)^{3/2}} (\mathbf{k} \times \mathbf{r}^0),$$

$$\mathbf{H} = \int\limits_{-\infty}^{\infty} \frac{Jr\, d\zeta}{c(r^2 + \zeta^2)^{3/2}} (\mathbf{k} \times \mathbf{r}^0) = \int\limits_{-\pi/2}^{\pi/2} \frac{J \cos s\, ds}{rc} (\mathbf{k} \times \mathbf{r}^0) = \frac{2J}{cr} (\mathbf{k} \times \mathbf{r}^0)$$

(here we put $\zeta = r \tan s$). We thus arrive at a law that has already been considered.

(b) Let the current J flow around a circle (in the xy-plane) of radius R with centre at the origin and compute the intensity at the point $(0, 0, z)$. Then, representing the circle in parametric form, $x = R \cos \psi$, $y = R \sin \varphi$, we get

$$\mathbf{H} = \int \frac{1}{c(z^2 + R^2)^{3/2}} (J\, d\mathbf{r} \times (z\mathbf{k} - \mathbf{r}))$$

$$= \int_0^{2\pi} \frac{J}{c(z^2 + R^2)^{3/2}} ((-R \sin \varphi\, \mathbf{i} + R \cos \varphi\, \mathbf{j})\, d\varphi\, (z\mathbf{k} - R \cos \varphi\, \mathbf{i}$$

$$- R \sin \varphi\, \mathbf{j})) = \frac{JR}{c(z^2 + R^2)^{3/2}} \int_0^{2\pi} (z \cos \varphi\, \mathbf{i} + z \sin \varphi\, \mathbf{j} + R\mathbf{k})\, d\varphi$$

$$= \frac{2\pi J R^2}{c(z^2 + R^2)^{3/2}} \mathbf{k}$$

Incidentally, it is easy from this to derive a formula for a magnetic field of an infinite solenoid.

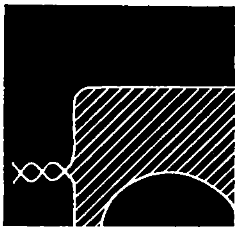
Sec. 11.12

Take the divergence of both sides of (65).

Sec. 11.13

The equality curl $\mathbf{v} = \mathbf{0}$, is verified via formula (43). Since the left-hand side of (57) is equal to grad (p arctan (y/x)), then by formula (55) we get $\varphi(x, y) = -p$ arctan (y/x) + constant. This "potential" is multi-valued; in a traversal of the circle (L) it receives an increment $-2\pi p$, which means the original irrotational field is not a potential field.

Chapter 12
CALCULUS OF VARIATIONS



Basic courses in mathematical analysis (see, for example, HM, Sec. 2.6) take up the question of seeking the extrema of a function of one variable, and in Sec. 4.6 we considered the problem of finding the extrema of a function of several variables, in other words, we considered problems with a finite number of degrees of freedom. The basic aim of the calculus of variations is to obtain general methods for finding extrema in problems involving an infinite number of degrees of freedom. In this chapter we will need certain facts from the theory of functions of several variables (mainly from Secs. 4.1 and 4.6).

12.1 An instance of passing from a finite number of degrees of freedom to an infinite number

Let us consider a chain of material particles stretched between two fixed points, the particles being successively connected by the same kind of springs. Suppose the particles are acted upon by small transverse forces that deviate the chain from the unloaded state of equilibrium (Fig. 164); for the sake of simplicity we assume that the displacements of the particles are perpendicular to the line of the unloaded state, i.e., the x-axis. Let us find the loaded state of equilibrium that is characterized by deviations $y_1, y_2, \ldots, y_{n-1}$ of the particles from the x-axis.

To do this, we determine the potential energy U of the chain in any (not necessarily equilibrium) deviated state, reckoning U from the unloaded (but already taut) state of the chain. This energy is made up of two parts:

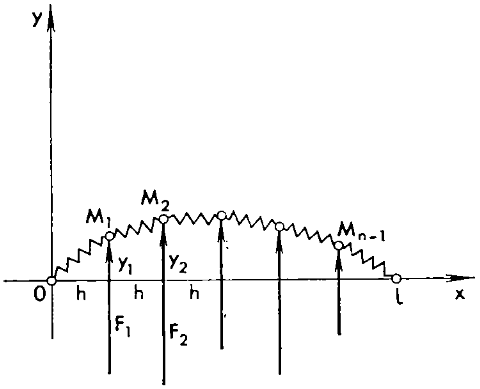
$$U = U_{\text{el}} + U_{\text{ext}} \tag{1}$$

The first represents the work spent on overcoming the elasticity of the springs, the second, the work to overcome external forces. The first is proportional to the elongation Δl of the chain:

$$U_{\text{el}} = P\Delta l \tag{2}$$

where P is the stretching force of the chain; we will assume that this force remains constant under deviations. The elonga-

Fig. 164



tion of the spring connecting particles M_i and M_{i+1} is equal to (see Fig. 164)

$$\Delta l_i = \sqrt{h^2 + (\Delta y_i)^2} - h = h\left(\sqrt{1 + \frac{(\Delta y_i)^2}{h^2}} - 1\right)$$
$$= h \cdot \frac{1}{2} \, \frac{(\Delta y_i)^2}{h^2} = \frac{(y_{i+1} - y_i)^2}{2h}$$

Here we took advantage of the approximate formula $\sqrt{1 + \alpha} = = 1 + \frac{1}{2}\,\alpha$ which holds true for small $|\alpha|$, i.e. in our case for small $\frac{(\Delta y_i)^2}{h^2}$. By (2),

$$U_{\text{el}} = P \sum_i \frac{(y_{i+1} - y_i)^2}{2h} = \frac{P}{2h} \sum_i (y_{i+1} - y_i)^2$$

It is still easier to find

$$U_{\text{ext}} = \sum_i y_i(-F_i) = -\sum_i F_i y_i$$

Thus, by (1),

$$U = \frac{P}{2h} \sum_i (y_{i+1} - y_i)^2 - \sum_i F_i y_i \tag{3}$$

In the equilibrium position, the potential energy must have a "stationary" value, that is, infinitesimal changes in the coordinates must lead to higher order changes in the potential energy; in other words, derivatives of the potential energy with respect to the coordinates in the equilibrium position must equal zero. And it is easy to verify that the derivative $\frac{\partial U}{\partial y_i}$ is equal (with sign reversed) to the total force acting on the ith particle. Thus, if the derivative were different from zero, this would indicate an uncompensated force, which is impossible under equilibrium. Now the potential energy must have a minimum in the stable position of equilibrium. This is necessary so that a displacement from the equilibrium position can generate a compensating force that will return the system to the equilibrium position; in other words, so that such a displacement should definitely require expending a positive amount of work.

Since on the right-hand side of (3) the coordinate y_i appears in three terms:

$$\frac{P}{2h}[(y_i - y_{i-1})^2 + (y_{i+1} - y_i)^2] - F_i y_i$$

it follows that

$$\frac{\partial U}{\partial y_i} = \frac{P}{2h}\Big[2(y_i - y_{i-1}) - 2(y_{i+1} - y_i)\Big] - F_i$$
$$= -\frac{P}{h}(y_{i+1} - 2y_i + y_{i-1}) - F_i$$

and the steady-state condition yields

$$\frac{P}{h}(y_{i+1} - 2y_i + y_{i-1}) + F_i = 0 \quad (i = 1,\ 2,\ \ldots,\ n-1) \qquad (4)$$

What we have is a system of $n - 1$ algebraic equations of the first degree in $n - 1$ unknown coordinates $y_1, y_2, \ldots, y_{n-1}$. Note that for $i = 1$ we have to put y_0 equal to zero in (4) (this is a fixed point), and for $i = n - 1$, y_n enters into (4) and is also put equal to zero. Solving system (4), we get the desired position of equilibrium.

Consider the following example. Let all forces $F_i = F$ be the same. Rewrite system (4) as

$$\begin{aligned} y_2 - y_1 \qquad\ \ &= -\frac{hF}{P} + y_1, \\ y_3 - 2y_2 + y_1 &= -\frac{hF}{P}, \\ y_4 - 2y_3 + y_2 &= -\frac{hF}{P}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

Check to see that by adding the first and second equations, the first, second, and third, and so forth, we get

$$\begin{aligned} y_2 - y_1 &= -\ \ \frac{hF}{P} + y_1, \\ y_3 - y_2 &= -\,2\frac{hF}{P} + y_1, \\ y_4 - y_3 &= -\,3\frac{hF}{P} + y_1, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

In the first equation, by transposing y_1 to the right side and then carrying out the same procedure of addition, we get, using the easily verifiable equation $1 + 2 + ... + k = \frac{1}{2} k(k + 1)$,

$$y_2 = -\frac{hF}{P} + 2y_1,$$

$$y_3 = -\frac{2 \cdot 3}{2} \frac{hF}{P} + 3y_1,$$

$$y_4 = -\frac{3 \cdot 4}{2} \frac{hF}{P} + 4y_1,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

In the general form,

$$y_i = -\frac{(i-1)\,i}{2} \frac{hF}{P} + iy_1 \tag{5}$$

Since y_n must turn out equal to zero, it follows that

$$-\frac{(n-1)\,n}{2} \frac{hF}{P} + ny_1 = 0, \quad \text{whence} \quad y_1 = \frac{(n-1)\,hF}{2P}$$

Substituting into (5), we finally get

$$y_i = -\frac{(i-1)\,i}{2} \frac{hF}{P} + i\frac{(n-1)\,hF}{2P} = \frac{i(n-i)\,hF}{2P} \tag{6}$$

$$(i = 1, 2, ..., n - 1)$$

the formula holds true for $i = 0$ and for $i = n$ as well.

The system at hand is determined by the deviations of $n - 1$ of its particles, which means that it has $n - 1$ degrees of freedom. By increasing n for a given length l of the chain, we at the same time increase infinitely the number of degrees of freedom, and in the limit we obtain from the chain a continuous string, which constitutes a system with an infinite number of degrees of freedom (because when considering a string, we can arbitrarily specify the deviation of any number of its points). To summarize, then: from a discrete pointwise system of particles we obtain, in the limit, a continuous medium with a continuously distributed mass.

Now let us see how the expression for potential energy and the condition of static equilibrium transform in the limit. We will assume that the outer transverse force is distributed along the string with a certain density $f(x)$ so that there is a force $f(x)h$ for the small length h. Rewrite (3) for the potential energy as

$$U = \sum_i \left[\frac{P}{2}\left(\frac{y_{i+1} - y_i}{h}\right)^2 - f_i y_i\right] h$$

For large n, that is, for small h, we can replace $\frac{y_{i+1} - y_i}{h}$ by y'_i and the potential assumes the form

$$U = \sum_i \left[\frac{P}{2} y'^2 - fy\right]_i h$$
$$= \sum_i \left[\frac{P}{2} y'^2 - fy\right]_i \Delta x$$

since $h = \Delta x$. But this is an integral sum, and in the limit, as $h \to 0$, this sum becomes the integral

$$U = \int_0^l \left[\frac{P}{2} (y')^2 - f(x)\, y\right] dx \qquad (7)$$

This is the expression for the potential energy in the case of a continuous string. And so the problem of finding the shape of equilibrium of a taut loaded string can be formulated mathematically thus: to find the function $y(x)$ (which describes the shape of the string) that satisfies the conditions of attachment,

$$y(0) = 0, \quad y(l) = 0 \qquad (8)$$

and that imparts a minimal value to the integral (7). Since there are an infinitude of degrees of freedom in the choice of such a function, this problem no longer belongs to the theory of functions of several variables, but to the calculus of variations.

Now let us see what the condition (4) of static equilibrium goes into in the limit; (4) can now be rewritten as

$$P\,\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + f_i = 0 \qquad (9)$$

Recall that the ratio (Sec. 2.2)

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}$$

— the so-called *second divided difference* — is close to the value of the second derivative y'' for small h. Thus in the limit, as $h \to 0$, we get from (9)

$$Py'' + f(x) = 0, \text{ that is, } y'' = -\frac{1}{P} f(x) \qquad (10)$$

and we have to find the solution $y(x)$ of this equation that satisfies the conditions (8). We solved this problem in Sec. 8.8.

And so the variational problem of finding the minimum of an integral under the conditions (8) reduced to solving the differential equation (10) under the same conditions. It is to be stressed that in deriving equation (10) we did not make use of the fact that it was precisely the minimum of integral (7) that was being sought, because when deriving the conditions (4) we only made use of the condition that expression (3) be stationary. This means that (10) is the condition of static equilibrium in the problem at hand. However, it is easy to verify that in the given example the equilibrium is stable and thus the potential energy (7) does not merely assume a stationary value for the solution but a minimal value. Indeed, from physical reasoning it is clear that there is some position of the string that corresponds to a minimum of potential energy. But in Sec. 8.8 we saw that under conditions (8) the equation (10) has only a unique solution, which means that this solution is the one that makes the integral (7) a minimum.

Let us consider an example. Let $f(x) \equiv f_0 =$ constant. Then equation (10) is readily integrated:

$$y' = -\frac{f_0}{P}x + C_1, \quad y = -\frac{f_0}{P}\frac{x^2}{2} + C_1x + C_2$$

From the conditions (8) we get

$$C_2 = 0, \quad C_1 = \frac{f_0}{P}\frac{l}{2}$$

whence, finally,

$$y = -\frac{f_0}{P}\frac{x^2}{2} + \frac{f_0}{P}\frac{l}{2}x = \frac{f_0x(l-x)}{2P}$$

(Obtain this same expression from the solution (6) for a discrete case by putting $F = f_0h,\ n = \frac{1}{h},\ i = \frac{x}{h}.$)

We conclude with a more general case where the particles M_i (and, in the limit, the points of the string) are acted upon also by an elastic force tending to return them to the nonloaded state of equilibrium, with coefficient of elasticity K (cushion). Here, the work

$$\sum_i \int_0^{y_i} (Ky)\,dy = \sum_i \frac{Ky_i^2}{2}$$

is done to overcome the elasticity of the cushion, and so the expression for the potential energy becomes

$$U = \frac{P}{2h}\sum_i (y_{i+1} - y_i)^2 + \frac{K}{2}\sum_i y_i^2 - \sum_i F_iy_i$$

(instead of (3)). The condition of stationarity, i.e. static equilibrium, yields

$$\frac{P}{h}(y_{i+1} - 2y_i + y_{i-1}) - Ky_i + F_i = 0$$

$$(i = 1, \ldots, n-1)$$

When passing to a continuous string it is natural to introduce the concept of "linear density of the coefficient of elasticity" k so that the force of elasticity per length h of the string is equal to khy. Then instead of (7) and (10) we get the relations

$$U = \int_0^l \left[\frac{P}{2}(y')^2 + \frac{k}{2}y^2 - f(x)\,y\right]dx, \tag{11}$$

$$Py'' - ky + f(x) = 0 \tag{12}$$

Thus, here too the solution of the variational problem reduces to the solution of the boundary-value problem for the differential equation.

Exercise

Solve the system (4) under the conditions (8) in the general case. Obtain the solution for a continuous string as $n \to \infty$.

12.2 Functional

The next variational problem, which arose at the end of the 17th century and was solved by Leibnitz, L'Hospital and Newton independently, demonstrated the force of the burgeoning mathematical analysis.

Suppose a particle M acted upon by gravity rolls without friction with a zero initial velocity from point A to point B along a curve (L) (Fig. 165). How can this curve be chosen so that the descent occurs in minimum time? For an analytic statement of the problem, denote the unknown equation of the curve by $y = y(x)$; then the function $y(x)$ must first of all satisfy the conditions:

$$y(a) = y_a, \quad y(b) = y_b \tag{13}$$

where y_a and y_b are the ordinates of the specified points A and B. The velocity of M at any time is readily determined by proceeding from the law of conservation of energy:

$$\frac{mv^2}{2} = mg(y_a - y), \quad \text{whence} \quad v = \sqrt{2g(y_a - y)}$$

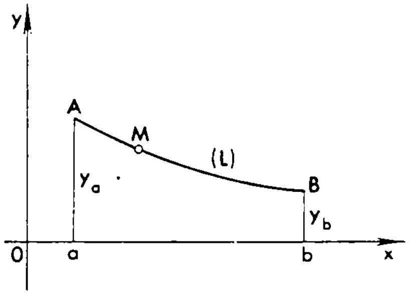
The horizontal component of the velocity is

$$\frac{dx}{dt} = \frac{dx}{ds}\,\frac{ds}{dt} = v\,\frac{dx}{ds} = v\,\frac{dx}{\sqrt{dx^2+dy^2}} = \sqrt{2g(y_a - y)}\,\frac{1}{\sqrt{1 + y'^2}}$$

Expressing dt in these terms and integrating, we get the total time of descent:

$$T = \int_a^b \frac{\sqrt{1 + y'^2}}{\sqrt{2g(y_a - y)}}\,dx \tag{14}$$

Fig. 165



Thus, it is required to choose, from among all functions that satisfy the conditions (13), that function for which the integral (14) assumes the smallest possible value.

The foregoing problem, like the problems examined in Sec. 12.1 for a continuous string, are typical problems of the calculus of variations. Two features characterize such problems. First of all, they are problems that involve an extremum (maximum or minimum), which is to say, they are problems in which it is required to make a certain numerical parameter an extremum (U in the problems in Sec. 12.1, T in the problem of the curve of quickest descent). We have already solved extremum problems; for functions of one variable the desired value is a certain number, namely the value of the independent variable, and for functions of several variables the desired element is a set of numbers. In contrast to this, in the foregoing problems of this chapter we did not seek a number or a set of numbers but a function $y(x)$, such that the indicated numerical parameter is expressed in terms of all values of this function (via formula (7) or (11) of Sec. 12.1 and (14) in the problem just discussed).

There are many other problems involving an extremum in which we seek a function (geometrically speaking, a curve or a surface) or a set of functions. All such problems constitute the subject matter of the calculus of variations.

Calculus of variations problems can also be examined from the following point of view. When selecting a value of the independent variable there is one degree of freedom. If we seek a set of n values of such variables (as in Chapter 4), then there are n degrees of freedom. Now if we vary the functional relationship in an arbitrary manner, then there are an infinitude of degrees of freedom; indeed, it is possible, in an arbitrary fashion, to specify the values of a function for any number of values of the independent variable. This means that the problems of the calculus of variations are problems involving extrema for the case of an infinite number of degrees of freedom in the choice of the desired object.

The following scheme is characteristic of variational problems. There is a certain scalar parameter I (this is U in Sec. 12.1 and T

in the last problem) that is expressed by a definite formula (7), (11) or (14) in terms of an unknown function $y(x)$ (so far we confine ourselves to functions of one variable) that has to be chosen. This function is more or less arbitrary, although it satisfies certain conditions: for instance, the conditions (8) in Sec. 12.1 and (13) in this section, and also the requirement of continuity.

A law of this nature, according to which with every function of a definite class of functions there is associated the value of a certain scalar parameter, is called a *functional*. Thus, the basic calculus of variations problem is a problem involving finding the extremum of a given functional.

To clarify the concept of a functional, recall once again that of a function. Consider, say, the formula $y = x^2$. By this formula, every value x is associated with a specific value y: for $x = 2$ we have $y = 4$, for $x = -\frac{1}{3}$ we have $y = \frac{1}{9}$, etc. We have a law according to which certain numbers x are associated with certain numbers y. The formula $y = x^3$ defines a different law, i.e. a different function, the formula $y = \sin\, x$ a third law, etc.

The simplest example of a functional is the definite integral. Consider, for example, the formula

$$I = \int_0^1 y^2\, dx \quad (y = y(x)) \tag{15}$$

Substituting various concrete functions in place of $y(x)$, we get concrete numerical values for I. For instance, choosing $y = x^2$, we get

$$I = \int_0^1 (x^2)^2\, dx = \int_0^1 x^4\, dx = \frac{x^5}{5}\Big|_0^1 = \frac{1}{5} = 0.2$$

Taking $y = x^3$, we get $I = \frac{1}{7} = 0.143$; choosing $y = \sin x$, we get $I = \frac{2 - \sin^2 1}{4} = 0.273$, and so forth. Thus, formula (15) specifies a law by which every function $y(x)$ is associated with a value I, which means formula (15) defines a functional. The formula

$$I = \int_0^1 xy'\, dx \quad \left(y = y(x),\quad y' = \frac{dy}{dx}\right) \tag{16}$$

defines a different functional, and the formula

$$I = \int_{-1}^3 y^2\, dx \quad (y = y(x))$$

a third functional (note the limits of integration), and so forth.

To summarize; if numbers are associated with numbers, we speak of a function. If functions are associated with numbers, we speak of a functional. Recall (Sec. 6.2) that if functions are associated with functions, we have an *operator*.

The functional (16) is called a *linear functional*; this means that when the functions $y(x)$ are added, the values of the functional I are also added:

$$\int_0^1 x(y_1 + y_2)'\,dx = \int_0^1 xy_1'\,dx + \int_0^1 xy_2'\,dx$$

The functional (15) is a *nonlinear (quadratic) functional.*

In investigating a functional it is sometimes important to determine how its value varies under a small change in the function on which it is dependent. Let us consider this problem using (15). Let the right-hand member be represented by a function $y(x)$, and then by a new function $y(x) + \delta y(x)$, where $\delta y(x)$ — the *variation* of y — is an arbitrary function that assumes small values. (For example, we could first have $y = x^2$, and then $y = x^2 + \alpha x^3$, where the constant α is small.) Then the value of the functional also changes slightly and becomes

$$\int_0^1 [y + \delta y]^2\,dx = \int_0^1 y^2\,dx + 2\int_0^1 y\delta y\,dx + \int_0^1 (\delta y)^2\,dx$$

Thus, the increment in this value is

$$\Delta I = 2\int_0^1 y\delta y\,dx + \int_0^1 (\delta y)^2\,dx \tag{17}$$

If for the time being we fix the function $y(x)$ and change its variation, then, depending on δy, the first term in the right-hand member of (17) is a linear term, whereas the second one is a quadratic term (in the general case we also have terms of higher powers as well). Since the values of δy are small, the main role in the right member is played by the linear term, while the quadratic term is of higher order. This linear term in the increment of the functional is called the *variation of the functional* and is denoted by δI; in the case of (15) we thus have

$$\delta I = 2\int_0^1 y\delta y\,dx \tag{18}$$

Thus, to within higher order terms, we have

$$\Delta I \approx \delta I \tag{19}$$

In those cases when second and higher order terms can be neglected, we simply say that the variation of the functional is an infinitesimal increment obtained through an infinitesimal change (variation) in the function upon which the value of the functional depends. (However, in the case of substantial δy, it is of course true that $\Delta I \neq \delta I$!) Recall the fundamentals of differential calculus and you will see a complete analogy between the notions of the differential of a function and the variation of a functional.

Exercises

1. Find the variation of the functionals $I = \int\limits_0^1 \frac{x}{y^2}\, dx$;

$$y^2(0) + \int\limits_0^1 (xy + y'^2)\, dx.$$

2. For the functional $I = \int\limits_0^1 y^2\, dx$ put $y = 2x$, $\delta y = \alpha x^2$ and compare δI with $\Delta \delta$ for $\alpha = 1,\ -0.1,\ 0.01$.

12.3 Necessary condition of an extremum

Let a function $y(x)$ realize an extremum of the functional I. In other words, the value of the functional for the function $y(x)$ is greater (in the case of a maximum) and less (in the case of a minimum) than the values of this functional for all functions sufficiently close to $y(x)$. (The last stipulation has to do with the fact that, generally, a functional can have several extrema, just as a function can have several extreme points.) Then ΔI will be negative for a maximum and positive for a minimum for all the indicated functions, that is, in both cases there is no change of sign. But then it follows from this that

$$\delta I = 0 \tag{20}$$

Indeed, from (19) it follows that if $\delta I \neq 0$, then ΔI and δI have the same sign, but due to the linear dependence of δI on δy when δy changes sign, δI will also change sign (see, for example, formula (18)), which runs counter to the foregoing.

Thus, condition (20) is a necessary condition for an extremum. Incidentally, it is quite analogous to the necessary condition for the extremum of a function that we find in the differential calculus. If the function $y(x)$ attains an extremum in comparison with all close-lying functions, then it follows that δy can be arbitrary in the left member of (20) (it is not even required that this function be small because, due to the linearity of δI, if $\delta I = 0$ for small δy, then this is also valid for any δy). If an extremum is attained in comparison with the functions of a certain class (see, for instance, the conditions (8) and (13)), then δy must be such that $y + \delta y$ belongs to this class.

The condition (20) is the condition of stationarity of the functional I for a change in the type of function $y(x)$, upon which this functional depends. In many problems we may be interested in all stationary values and not only in the extremal values of the functional. For instance, in the problems of Sec. 12.1 we say that if for the functional we consider the potential energy of a continuous medium, then to every stationary value there corresponds a static position of equilibrium, whereas to the minimal values of the functional there correspond stable states of equilibrium.

Let us illustrate the application of condition (20) in an elementary problem involving finding the extremum of a functional of the form

$$I = \int_a^b F(x, y)\, dx \tag{21}$$

where for y we substitute any function of x. To find δI we substitute $y + \delta y$ instead of y and expand the result in a Taylor series:

$$I + \Delta I = \int_a^b F(x, y + \delta y)\, dx$$

$$= \int_a^b F(x, y)\, dx + \int_a^b F'_y(x, y)\, \delta y\, dx + \int_a^b F''_{yy} \frac{(\delta y)^2}{2!}\, dx + \dots$$

The variation of the functional, that is, the linear portion of the increment, is equal to

$$\delta I = \int_a^b F'_y(x, y)\, \delta y\, dx$$

From this, by virtue of the arbitrary nature of the choice of $\delta y(x)$ it follows that

$$F'_y(x, y) = 0 \tag{22}$$

Indeed, if we put $\delta y = \alpha F'_y(x, y(x))$ (α small), then the last integral is equal to

$$\alpha \int_a^b (F'_y)^2\, dx$$

But it must be equal to zero, whence follows our assertion. (If a continuous function does not assume any negative values, and the integral of it is zero, then the function is identically zero.)

We can arrive at condition (22) in a different way. If the integral (21) must have an extremal (say, minimal) value, then the integrand must also be minimal. In other words, if we add independent terms, then every term must be minimal if we want the sum to be minimal. Thus, for every x the value of y has to be chosen from the condition of the function F being a minimum, which immediately leads to (22). Hence, there is nothing fundamentally new here as compared with problems arising from the essentials of differential calculus.

For one thing, we arrive at functionals of the type (21) if in the reasoning of Sec. 12.1 we assume that the points of a string are not elastically related in any way and are subjected solely to the action of an external force and the elasticity of the cushion. It is clear that in that case every point of the string will take up a location irrespective of the positions of the remaining points, so that there is nothing fundamentally new here in comparison with the statics of a point. In this instance, the functional has the form

$$U = \int_0^l \left[\frac{k}{2} y^2 - f(x)\, y\right] dx$$

(which is (11) for $P = 0$) and the condition of equilibrium (22) yields

$$ky - f(x) = 0, \text{ or } y = \frac{1}{k} f(x)$$

The same result is obtained from (12) for $P = 0$. It is interesting that this solution is, generally, discontinuous both due to the possibility of discontinuities in the function $f(x)$ (which, for example, is the case if an external load is applied only to a portion of the string) and due to the boundary conditions (8). Of course, that is as it should be if the points of the string are in no way related to one another. Actually, it is hard to speak of a "string" in such a situation.

Exercise

Find the functions $y(x)$ that make the functional minimal:

$$\text{(a)} \int_0^1 (1 - x)\,(y - 2x)^2\, dx, \qquad \text{(b)} \int_0^2 (1 - x)\,(y - 2x)^2\, dx.$$

12.4 Euler's equation

Take another look at the problem in Sec. 12.1 on the equilibrium of an elastic string under the action of an external force. We found that problem reduces to finding an extremum (or even simply a stationary value) of a functional in the class of functions satisfying the boundary conditions (8). Let us try to find a solution with the aid of condition (20); to do this, we first try to compute δU. Let y have an increment δy; then y' receives the increment $\delta(y')$. But it is easy to verify that $\delta(y') = (\delta y)'$: indeed, if $\delta y = Y(x) - y(x)$, then $\delta(y') = Y'(x) - y'(x) = [Y(x) - y(x)]' = (\delta y)'$, whence

$$\Delta U = \int_0^l \left[\frac{P}{2}(y' + \delta y')^2 - f(x)(y + \delta y)\right] dx - \int_0^l \left[\frac{P}{2}(y')^2 - f(x)\, y\right] dx$$

$$= \int_0^l \left[Py'\delta y' + \frac{P}{2}(\delta y')^2 - f(x)\,\delta y\right] dx$$

Dropping the second-order term, we get

$$\delta U = \int_0^l [Py'\delta y' - f(x)\,\delta y]\, dx$$

The necessary condition (20) yields

$$P\int_0^l y'\delta y'\, dx - \int_0^l f(x)\,\delta y\, dx = 0 \tag{23}$$

This equation should hold for any variation δy that satisfies the conditions

$$\delta y(0) = 0, \quad \delta y(l) = 0 \tag{24}$$

which are necessary for $y + \delta y$ to satisfy the same conditions (8) as y.

The new element in the condition (23) is that we have $\delta y'$ along with δy, and these two quantities cannot be considered to be independent. So let us integrate the first term by parts, applying (24):

$$0 = Py'\delta y\Big|_{x=0}^{l} - P\int_0^l y''\delta y\, dx - \int_0^l f(x)\,\delta y\, dx$$

$$= -\int_0^l [Py'' + f(x)]\,\delta y\, dx$$

Taking advantage of the arbitrariness of δy and reasoning as in Sec. 12.3, we get

$$Py'' + f(x) = 0$$

which is equation (10).

Thus the presence of y' in the expression of the functional relates the adjacent values of y (cf. the example involving springs, which we discussed in Sec. 12.1). In this case it is impossible to select a value of y

independently of adjacent values, as was done for the functional (21), and instead of a finite equation, as in Sec. 12.3, we get a differential equation.

Now let us derive an analogous differential equation for the functional of a more general type:

$$I = \int_a^b F(x, y, y')\, dx \left(y = y(x), \quad y' = \frac{dy}{dx}\right) \tag{25}$$

where F is a given function, and a and b are given limits of integration. Suppose we are seeking the extremum of this functional for certain given boundary conditions:

$$y(a) = y_a, \; y(b) = y_b \tag{26}$$

Reasoning as we did in the foregoing problem and utilizing the general formula (2) of Ch. 4, in which the role of x, dx, y, dy are played, respectively, by y, δy, y', $\delta y'$, we get the variation of the functional (25)

$$\delta I = \int_a^b [F'_y(x, y, y')\, \delta y + F'_{y'}(x, y, y')\, \delta y']\, dx$$

(This same pattern is used for setting up the variation and functionals of a different form.) Now, if a certain function $y(x)$ that satisfies the conditions (26) realizes the extremum of the functional (25) compared with all close-lying functions that satisfy those same conditions, then by the criterion (20) it must be true that

$$\int_a^b [F'_y(x, y, y')\, \delta y + F'_{y'}(x, y, y')\, \delta y']\, dx = 0 \tag{27}$$

This equation should hold true for any variation δy that satisfies the relations

$$\delta y(a) = 0, \quad \delta y(b) = 0 \tag{28}$$

which are needed so that $y + \delta y$ can satisfy the same conditions (26) as y.

Can we satisfy equation (27) by putting $F'_y = 0, F'_{y'} = 0$? We thus obtain two equations for a single desired function $y(x)$ (recall that the expression $F(x, y, y')$ is given). In the general case, two equations with one desired function do not have a general solution. The solution to our problem does exist, however, for the simple reason that δy and $\delta y'$ are independent quantities. It is possible to find $y(x)$ or, to put it differently and better, an equation for $y(x)$ such that for any δy and $\delta y'$ corresponding to it, one term, i.e. $\int F'_y \delta y\, dx$, is exactly compensated for by another term, $\int F'_{y'}\delta y'\,dx$, although they are not equal to zero when taken separately. This is how it is done.

Integrating the second term in (27) by parts and applying the relations (28), we get

$$0=\int_a^b F'_y\delta y\,dx+F'_{y'}\delta y\Big|_a^b-\int_a^b \frac{d}{dx}(F'_{y'})\,\delta y\,dx$$

$$=\int_a^b\left[F'_y-\frac{d}{dx}(F'_{y'})\right]\delta y\,dx$$

From this, by virtue of the arbitrary nature in the choice of δy, it follows that the last square bracket is identically zero. True enough, if we take δy equal to this bracket and only near the points a and b make it quickly diminish to zero (this is necessary so as to satisfy the relations (28)), then the last integral will be roughly equal to $\int_a^b\left[F'_y-\frac{d}{dx}F'_{y'}\right]^2dx$. But it must be equal to 0, whence follows the assertion.

We have thus arrived at the so-called *Euler equation*:

$$F'_y(x,\ y,\ y')-\frac{d}{dx}F'_{y'}(x,\ y,\ y')=0 \tag{29}$$

Note that when differentiating $\frac{d}{dx}$ here the quantity y is regarded as a function of x; in detail, by the rule for differentiating a composite function, the Euler equation can be rewritten thus:

$$F'_y(x,\ y,\ y')-F''_{xy'}(x,\ y,\ y')-F''_{yy'}(x,\ y,\ y')\,y'\\ -F''_{y'y'}(x,y,y')\,y''=0 \tag{30}$$

Here, the partial derivatives of $F(x,\ y,\ y')$ are taken disregarding the dependence of y and y' on x. It is clear that this is a differential equation of the second order and, for this reason its general solution contains two arbitrary constants that are defined with the help of the boundary conditions (26). The Euler equation generalizes the equations of static equilibrium (10) and (12) examined in Sec. 12.1.

Now let us examine another derivation of the Euler equation (29) that does not make use of the formal procedure of integration by parts. To do this, we partition the interval $a\leqslant x\leqslant b$ by means of points $x_0=a<x_1<x_2<\ldots<x_n=b$ into subintervals of equal length h and write down approximately the integrals of each of the terms in

the left-hand member of (27) in the form of sums, the first sum taken over integral points of division and the second sum over half-integral points of division (Secs. 2.1, 2.2). If we replace the derivative by the difference quotient, then instead of (27) we get the condition

$$\sum_k (F'_y)_k \delta y_k + \sum_k (F'_{y'})_{k+{}^1/_2} \frac{\delta y_{k+1} - \delta y_k}{h} = 0 \tag{31}$$

At first glance, it might appear that the coefficient of δy_k in the second sum is appreciably greater than the corresponding coefficient in the first sum, since for small h, $|(F'_{y'})_{k+{}^1/_2}|/h \gg |(F'_y)_k|$. But we have to take into account that in the second sum δy_k participates in two adjacent terms: with $(F'_{y'})_{k+{}^1/_2}$ and with $(F'_{y'})_{(k-1)+{}^1/_2}$. For this reason, the total coefficient of δy_k in the left-hand member of (31) is

$$(F'_y)_k - \frac{1}{h}[(F'_{y'})_{k+{}^1/_2} - (F'_{y'})_{k-{}^1/_2}] \approx (F'_y)_k - \left[\frac{d}{dx}(F'_{y'})\right]_k \tag{32}$$

Thus, h cancels out and both terms in the right member of (32) are of the same order.

Because δy_k is arbitrary, from condition (31) we find that this coefficient must be zero, i.e., we arrive at (29).

To illustrate, we take the problem of Sec. 12.2 on the curve of fastest descent. In this case (see (14)) it is necessary to put

$$F = \frac{\sqrt{1 + y'^2}}{\sqrt{2g(y_a - y)}} \tag{33}$$

Since x does not appear directly in the right member, it follows that the second term is absent on the left side of (30), and after multiplying both sides of (30) by y' it can be rewritten as $(F - F'_{y'}\, y')' = 0$. Then, integrating, we get

$$F - F'_{y'}y' = C_1 \tag{34}$$

which for the concrete example of (33) yields

$$\frac{\sqrt{1 + y'^2}}{\sqrt{2g(y_a - y)}} - \frac{y'}{\sqrt{1 + y'^2}\,\sqrt{2g(y_a - y)}}\, y' = C_1$$

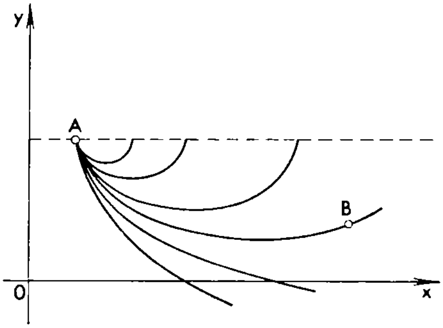
or

$$\frac{1}{\sqrt{1 + y'^2}\,\sqrt{2g(y_a - y)}} = C_1$$

In order to integrate this equation we introduce an artificial parameter t by the formula

$$y_a - y = r(1 - \cos t) \quad \text{where} \quad r = 1/4g\, C_1^2$$

Fig. 166



After the appropriate manipulations we get

$$\frac{1}{\sqrt{1+y'^2}} = C_1\sqrt{2g(y_a - y)} = \sqrt{\frac{1-\cos t}{2}};$$

$$1 + y'^2 = \frac{2}{1-\cos t}; \quad y'^2 = \frac{1+\cos t}{1-\cos t} = \frac{\sin^2 t}{(1-\cos t)^2};$$

$$\frac{dy}{dx} = \pm\frac{\sin t}{1-\cos t}; \quad dx = \pm\frac{1-\cos t}{\sin t}dy = \pm r(1-\cos t)\,dt;$$

$$x = \pm r(t - \sin t) + C_2$$

This expression together with the preceding expression for y give the parametric equations for what is called a *cycloid* (see, for example, HM, Sec. 1.8), which is a curve described by a point of a circle of radius r rolling down without sliding along a horizontal straight line drawn through A. Thus, the curve of quickest descent is a cycloid with cusp at the point A; here, the radius r must be chosen so that the first arch of the cycloid passes through the terminal point. Fig. 166 shows several such arches, of which the fourth passes through the terminal point. It is interesting to note that this arch partially passes below the terminal point, which is something that might not have been foreseen. Incidentally, a glance at the answer makes this clear at once: in order to cover a large horizontal path it is best to fall lower in order to gain velocity and then, at the end of the path, rise to the point of destination. The point B must of course be lower than A, otherwise there will be no solution at all.

Let us return to the general functional (25) for the boundary conditions (26). Suppose, having fixed these conditions, we obtain a certain solution $y(x)$ to the extremum problem. Now suppose that these conditions can change. Then the extremal value of the functional I will depend on these conditions:

$$I = I(a, b, y_a, y_b) \tag{35}$$

Here we substitute into I only those functions that make I extremal for the fixed boundary conditions. If this were not done, then I would depend not only on the boundary conditions but also on an arbitrary function $y(x)$.

It may become necessary to study the functional relationship (35) of the value of I under a change of the boundary conditions $y(a) = y_a$, $y(b) = y_b$. Since the function (35) already depends on a finite number of independent variables, this problem is solved with the tools of ordinary analysis, by calculating derivatives. As an example, here is how the derivative $\frac{\partial I}{\partial y_a}$ is calculated. If a, b, y_b are fixed and y_a varies, then the solution $y(x)$ of the extremum problem will change and will receive the increment $\delta y(x)$. Then, to within higher order infinitesimals,

$$\Delta I = \int_a^b (F'_y \delta y + F'_{y'} \delta y')\, dx = F'_{y'} \delta y \Big|_a^b + \int_a^b \left[F'_y - \frac{d}{dx}(F'_{y'})\right] \delta y\, dx$$

But since $y(x)$ was a solution of the extremum problem, it satisfies the Euler equation and so the last integral vanishes. Besides, $\delta y(b) = 0$ (since y_b remains unchanged) and we get

$$\Delta I = -(F'_{y'})_a\, \delta y(a), \quad \frac{\partial I}{\partial y_a} = -(F'_{y'})_a$$

The other derivatives of the function (35) are calculated in similar fashion.

Exercises

1. Solve the problems: (a) $I = \min \int_0^1 (y^2 + y'^2)\, dx$, $y(0) = 0$, $y(1) = 1$;

   (b) $I = \min \int_0^1 yy'^2\, dx$, $y(0) = p > 0$, $y(1) = q > 0$.

2. In (b) find $\frac{\partial I}{\partial p}$ directly and via the formula derived in the text.

3. Using the calculus of variations, derive the equation of a straight line as the shortest distance between two specified points.

12.5 Does a solution always exist?

Condition (20) is only a necessary condition for an extremum. The sufficient conditions are rather involved, but in most practical problems they are not required. For example, in the problem examined in Sec. 12.4 we did not verify that for the constructed arch of the cycloid

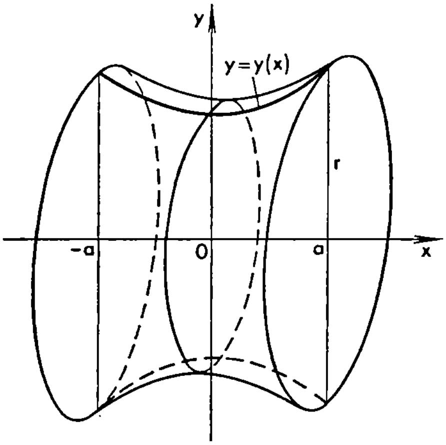


Fig. 167

it is precisely a minimal time of descent that is accomplished. But from physical reasoning it is clear that some such solution to the problem of the minimum for $y_a > y_b$ exists and only that one has to be found. And since using the necessary condition (Euler's equation) yielded only one solution, that is the desired solution. Of course, if we did not give thought to the problem, we could pose the question of seeking the curve of maximum time of descent and we could arrive at the same Euler equation and the same solution. But that answer would be erroneous, for we now see that the constructed solution (which is the only solution) produces precisely the minimum time of descent. As to the corresponding maximum problem, it has no solution at all and it is easy to construct curves with arbitrarily large times of descent.

Thus, an extremum problem may not have a solution; in certain cases this is more or less clear at once from the very statement of the problem (as in the foregoing example when considering the maximum problem), whereas in other cases it follows from the results of computations. An example will serve to illustrate this point.

Let it be required to find the shape of a film stretched over two equal rings located perpendicular to the lines joining their centres. This is the shape of a soap film stretched over the rings (Fig. 167, where the heavy lines represent the axial cross-section of the film). Since it is clear from considerations of symmetry that the desired surface is a surface of revolution and the area of the surface of revolution, as we know (see HN, Sec. 4.7), is

$$S = 2\pi \int_{-a}^{a} y\,dl = 2\pi \int_{-a}^{a} y\sqrt{1 + y'^2}\,dx \tag{36}$$

the matter reduces to finding the function $y(x)$ that makes the integral (36) a minimum under the boundary conditions

$$y(-a)=r,\quad y(a)=r \tag{37}$$

Here too the x does not enter directly under the integral sign so that we can take advantage of the intermediate integral (34), which gives us

$$2\pi y\sqrt{1+y'^2}-\frac{2\pi yy'}{\sqrt{1+y'^2}}\,y'=C_1,\ \text{i.e.}\ \frac{y}{\sqrt{1+y'^2}}=\frac{C_1}{2\pi}$$

Denoting the right-hand member by $\frac{1}{k}$ and carrying out certain manipulations, we get

$$\frac{y^2}{1+y'^2}=\frac{1}{k^2},\quad k^2y^2=1+\left(\frac{dy}{dx}\right)^2,\quad \frac{dy}{dx}=\pm\sqrt{k^2y^2-1},$$

$$\frac{dy}{\sqrt{k^2y^2-1}}=\pm\,dx,\ \int\frac{dy}{\sqrt{k^2y^2-1}}=\pm\,(x+C) \tag{38}$$

The last integral for $k=1$ is given in HM, p. 476 (No. 31) and is equal to $\ln\,(y+\sqrt{y^2-1})$. Let us try to compute the derivative of the function $\ln\,(ky+\sqrt{(ky)^2-1})$; it is equal to $k/\sqrt{k^2y^2-1}$ (check this!). And so from (38) we get $\frac{1}{k}\ln\,(ky+\sqrt{k^2y^2-1})=\pm\,(x+C)$, whence, after some simple manipulations which we leave to the reader, $y=\frac{e^{k(x+C)}+e^{-k(x+C)}}{2k}$ (the $\pm$ sign is not essential here).

For the curve to be symmetric about $x=0$, it must be true that $C=0$, and so, finally,

$$y=\frac{e^{kx}+e^{-kx}}{2k}$$

This curve is called the *catenary*, which is the shape of a chain suspended at the ends (see Exercise 2 in Sec. 12.8). To summarize, then: the desired surface of a soap film results from rotating a catenary.

For the boundary conditions (37) to be satisfied, it is required that

$$\frac{e^{ka}+e^{-ka}}{2k}=r,\ \text{or}\ \frac{e^{ak}+e^{-ak}}{2}=rk \tag{39}$$

From this we can find the as yet unknown value of the parameter k. Let us try to solve this problem graphically. For given a and r, depict the graphs of variation of the left and right sides of equation (39) (see Fig. 168) and find the point of intersection of the two curves. We are surprised to find that instead of the expected one solution there are two (for relatively small a, i.e. when the rings are close together) or none (when the rings are far apart). What shape does the film take after all?

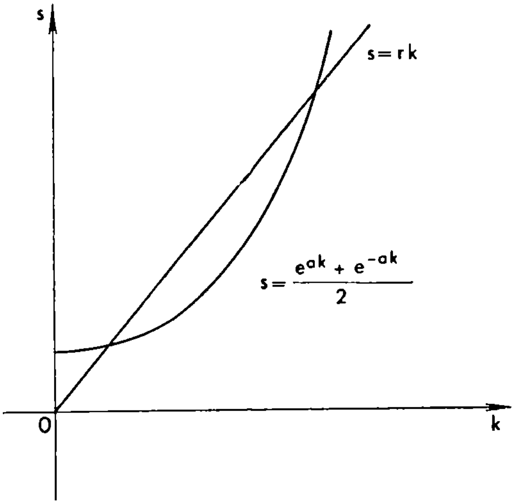


Fig. 168

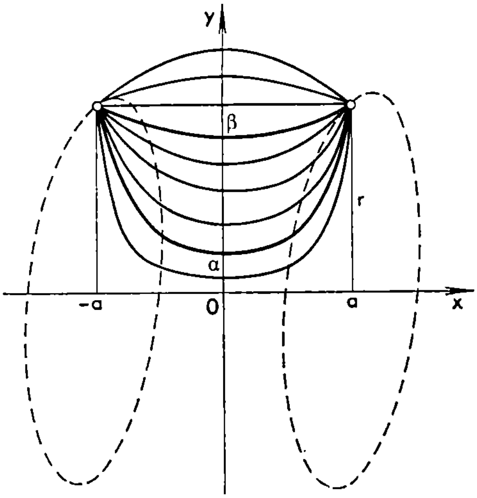


Fig. 169

First let a be comparatively small, which means there are two solutions as shown by the heavy lines in Fig. 169 (the rings are indicated here by dashed lines). If we imagine the other shapes of the film shown in Fig. 169 by the light lines, then the areas of the corresponding surfaces of revolution will have the values represented graphically in Fig. 170. We see that of the two solutions, the minimum area is realized for the upper one and the maximum area for the lower one. This is why the upper solution yields a stable form of equilibrium of the film, and the lower solution gives an unstable form.

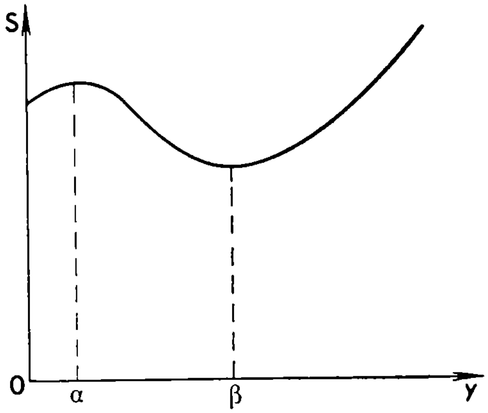


Fig. 170

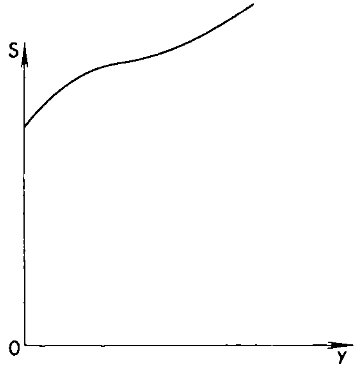


Fig. 171

Now if a is increased (for a given r), that is, if we move the rings away from each other, the extremum points α and β come together and for sufficiently large a the graph of the areas assumes the form shown in Fig. 171. Thus, in this case the film tends to reduce its area and contracts to the line of centres, separates, and assumes the form of two separate circles, each of which spans its ring. (Incidentally, the film behaves in the same way for small a if it begins to deform and has too thin a waist.) Thus, in this case there will be no single surface. The shape of the graphs in Figs. 170 and 171 is confirmed by the fact that for $y = 0$ we will have $S = 2\pi r^2$ and for $y = r$ it will be true that $S = 4\,\pi r a$, so that for small a the last value is smaller than $S|_{y=0}$ and for large a it is greater. For large a, the shape consisting of two separate circles may be regarded as a generalized solution of our problem. This solution is in the nature of a "terminal" or "cuspidal" minimum, for which the Euler equation does not hold (cf. Sec. 12.9).

The calculations have shown that we arrive at conclusions which could not have been foreseen.

It is easy to calculate what value of a/r is critical in the sense that equilibrium is possible for smaller values and is not possible for greater values. This occurs when both graphs in Fig. 168 touch each other, that is to say, the derivatives of both functions with respect to k are equal to

$$\frac{ae^{ak} - ae^{-ak}}{2} = r$$

Combining this equation with (39) and setting $\frac{a}{r} = \lambda$ (this is the critical value), we get

$$\frac{e^{ak} - e^{-ak}}{2} = \frac{1}{\lambda}; \quad e^{ak} = rk + \frac{1}{\lambda}; \quad e^{-ak} = rk - \frac{1}{\lambda};$$

$$1 = \left(rk + \frac{1}{\lambda}\right)\left(rk - \frac{1}{\lambda}\right) = r^2k^2 - \frac{1}{\lambda^2};$$

$$k = \frac{\sqrt{1+\lambda^2}}{r\lambda} = \frac{\sqrt{1+\lambda^2}}{a}; \quad \frac{e^{\sqrt{1+\lambda^2}} - e^{-\sqrt{1+\lambda^2}}}{2} = \frac{1}{\lambda}$$

And so λ is found from the last equation. A rough numerical calculation gives the approximate value of $\lambda = 0.7$.

It is interesting to note that if the upper solution in Fig. 169 makes for an absolute minimum of the area of the surface of revolution, then the lower solution maximizes only in the represented family of surfaces. If we go outside the limits of this family, the lower solution will be stationary since it satisfies the Euler equation, but by no means will it be maximal; it will be in the nature of a minimax (cf. Sec. 4.6). If we choose any two sufficiently close points in the lower solution and regard them as boundary conditions, then the arc between them is stable, which means it realizes a solution of the minimum problem. This means that for any change of the lower solution over a sufficiently small interval the area will increase. The maximum problem for a surface of revolution does not have a solution. (This is also clear from the possibility of crimping or corrugating any surface.)

The idea that a small section of a stationary solution is not only stationary but also extremal turns out to be useful in many problems.

Exercise

Consider the minimum problem of the functional $I = \int_0^a (y'^2 - y^2)\,dx$ ($a > 0$) for the boundary conditions $y(0) = y(a) = 0$. These conditions are satisfied, for instance, by the functions $y = Cx\,(a - x)$ and $y = C \sin \frac{\pi x}{a}$. What conclusion can be drawn from this concerning the existence of a solution to the problem?

12.6 Variants of the basic problem

There are large numbers of other types of variational problems that are discussed in courses of the calculus of variations. Some of these problems are investigated in the manner that we considered (in Sec. 12.4) the extreme-value problem of the functional (25) under the condition (26). For example, derivatives of the desired function of order higher than second can appear under the integral sign; in that case the Euler equation is of a higher order than second (namely, twice as high as the highest order of the derivatives that enter into the functional). Several unknown functions can appear under the integral sign; then they are sought via a system of differential Euler equations, the number of equations in the system being equal to the number of desired functions, for it is necessary to equate to zero the variations of the functional obtained by varying each of the functions. Such, for example, is the case of seeking a curve in space, for such a curve is determined by two functions, say, $y(x)$ and $z(x)$.

Now let us consider a variational problem for a function of several variables (for the sake of definiteness, we take two independent variables). Suppose we are seeking the function $z(x, y)$ that gives a maximum to the integral

$$\iint\limits_{(\sigma)} F(x, y, z, z'_x, z'_y)\, dx\, dy \tag{40}$$

where (σ) is a certain specified region with boundary (L) with the boundary condition

$$z|_{(L)} = \varphi \text{ (given)} \tag{41}$$

Arguing in the manner of Sec. 12.4 leads to the Euler equation, which in this case is of the form

$$F'_z - \frac{\partial}{\partial x} F'_{z'_x} - \frac{\partial}{\partial y} F'_{z'_y} = 0 \tag{42}$$

Bear in mind that when computing $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$, the z is regarded as a function of the variables x and y. Thus, to obtain a solution we get a second-order partial differential equation with the boundary condition (41). The solution of such equations is beyond the scope of this text, but one physically important example will be examined (another one is given in Sec. 12.12).

We seek the equation for the equilibrium form of a membrane stretched over a rigid frame (contour). We assume the membrane to be homogeneous (the same at all points), isotropic (the same in all directions) and stretched with a force T per unit length. The potential energy of the membrane that results from stretching it over the contour is due to an increase in its area as compared with the horizontal position. In integral calculus, proof is given that the area Q of the surface described by the equation $z = z(x, y)$ is

$$\iint\limits_{(\sigma)} \sqrt{1 + (z'_x)^2 + (z'_y)^2}\, dx\, dy$$

so that the increase ΔQ in the area of the membrane is equal to

$$\iint\limits_{(\sigma)} \sqrt{1 + (z'_x)^2 + (z'_y)^2}\, dx\, dy - \sigma = \iint\limits_{(\sigma)} [\sqrt{1 + (z'_x)^2 + (z'_y)^2} - 1]\, dx\, dy$$

Regarding the deflection as small (this requires that the given contour of the membrane deflect only slightly from the plane $z = 0$) and the quantities z'_x and z'_y as small too, we expand the integrand in a series and discard the higher order terms:

$$\Delta Q = \iint\limits_{(\sigma)} \left\{ \left[1 + \frac{1}{2} [(z'_x)^2 + (z'_y)^2] + \text{higher order terms} \right] - 1 \right\} dx\, dy$$

$$= \frac{1}{2} \iint\limits_{(\sigma)} [(z'_x)^2 + (z'_y)^2]\, dx\, dy$$

Assuming that the tension of the membrane remains unchanged in the process of stretching, we find the work done in this process to be

$$\Delta A = \frac{T}{2} \iint\limits_{(\sigma)} [(z'_x)^2 + (z'_y)^2]\, dx\, dy \tag{43}$$

Hence that will also be the accumulated potential energy.

However, from physics it is known that, from among all possible forms, a membrane chooses that form of equilibrium for which the potential energy is a minimum. Hence we have the problem of minimizing the integral (43). By virtue of the general Euler equation (42) we get $0 - \frac{\partial}{\partial x}\left(\frac{T}{2}\, 2z'_x\right) - \frac{\partial}{\partial y}\left(\frac{T}{2}\, 2z'_y\right) = 0$ which, after cancelling, yields

$$z''_{xx} + z''_{yy} = 0 \tag{44}$$

Thus, the form of equilibrium of the membrane is described by the function $z = z(x, y)$ that satisfies the Laplace equation (see Sec. 5.7). To find this form in a specific instance, it is necessary to find the solution to equation (44) for the boundary condition (41).

Exercises

1. Derive the Euler equation for the functional $I = \int_a^b F(x, y, y', y'')\,dx$ under the boundary conditions $y(a) = y_a$, $y'(a) = y'_a$, $y(b) = = y_b$, $y'(b) = y'_b$.
2. Write out equation (42) in full, like (30).

12.7 Conditional extremum for a finite number of degrees of freedom

Let us go back to the extremum problems for a system with a finite number of degrees of freedom. In the problems considered in Sec. 4.6 the independent variables were not connected by any relations; such extrema are called *absolute*. There are also problems involving *conditional extrema* in which the independent variables are related by definite equations. Let us begin with functions of two variables.

Suppose we are seeking the maximum or minimum of the function $z = f(x, y)$ under the condition that the variables x and y are not independent but are connected by the relation

$$F(x, y) = h \tag{45}$$

This means that the values of the function f are considered and compared only for points (in the plane of the arguments) lying on the line given by the equation (45). For example, Fig. 172 depicts level lines of a certain function $f(x, y)$ having an absolute maximum at the point K; the heavy line (L) given by the equation (45) is also shown. Going along (L), we come to the level line with the highest label at point A and then pass immediately into a region with lower labels, a region of lower altitudes. This means that there is a conditional maximum of the function f at the point A, but it is not an absolute maximum because if we move from (L) towards K, we could find higher values of f near A. In similar fashion, we can check that at the point B we have a conditional minimum, and at the point C another conditional maximum; in other words, there are three conditional extrema here. Thus, an absolute maximum is like a mountain peak, whereas a conditional maximum is the highest point of a given mountain path (the projection of this path on the xy-plane has the equation (45)).

If it is possible to express y in terms of x from the constraint equation (45), then this result can be substituted into the expression for z,

$$z = f[x, y(x)] \tag{46}$$

to obtain z as a function of one independent variable. Since there is no longer any condition (it has been taken into account by the substitution $y = y(x)$), it follows that the problem of seeking the

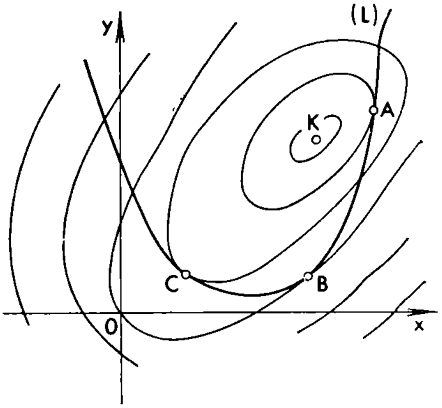


Fig. 172

extremum of z becomes an absolute-extremum problem. A similar result is obtained if equation (45) can be solved for x or if the equation of the line (45) can be represented in parametric form.

But such a solution of (45) is not always possible and advisable. Then we can reason as follows. The constraint equation (45) defines fundamentally a certain relationship $y = y(x)$, although it is implicit and not known. Thus, z is a composite function (46) of the independent variable x, and the necessary condition for an extremum yields, by the formula of the derivative of a composite function,

$$\frac{dz}{dx} = f'_x + f'_y \frac{dy}{dx} = 0 \tag{47}$$

Here, dy/dx signifies the derivative of the implicit function $y(x)$ defined from the condition (45). By the rules of Sec. 4.3, we find that $F'_x + F'_y \frac{dy}{dx} = 0$, or $\frac{dy}{dx} = -F'_x/F'_y$. Substituting this expression into (47), we get, at the point of the conditional extremum,

$$f'_x - \frac{F'_x}{F'_y} f'_y = 0 \quad \text{that is,} \quad -\frac{F'_x}{F'_y} = -\frac{f'_x}{f'_y} \quad \text{or} \quad \frac{f'_x}{F'_x} = \frac{f'_y}{F'_y}$$

(The middle equation signifies, by virtue of Sec. 4.3, that at the point of the conditional extremum the curve (45) touches the level line of the function f, cf. Fig. 172.) Denote the last relation at the point under consideration by λ. Then at the point of the conditional extremum we have

$$\frac{f'_x}{F'_x} = \frac{f'_y}{F'_y} = \lambda \tag{48}$$

or

$$f'_x - \lambda F'_x = 0, \quad f'_y - \lambda F'_y = 0 \tag{49}$$

Set

$$f^*(x, y; \lambda) = f(x, y) - \lambda F(x, y) \tag{50}$$

where λ is an unknown parameter called a *Lagrange multiplier*. Then (49) can be written thus:

$$f_x^{*\prime} = 0, \quad f_y^{*\prime} = 0 \tag{51}$$

Thus, we have the same equations as in the case of the absolute extremum (see (28) of Ch. 4); however, they are set up not for the function f itself, but rather for the changed function f^* defined by formula (50). Equations (51) together with the constraint equation (45) form a system of three equations in three unknowns x, y, λ. These equations yield the conditional extremum points.

The Lagrange multiplier λ has a simple meaning. To determine it, denote the coordinates of the point of conditional extremum and the extremal value itself by $\bar{x}$, $\bar{y}$, and $\bar{z}$, respectively. Up to now we considered h to be fixed, but if we vary h, then these three quantities will depend on h. Let us determine at what rate the extremal value of z will change as h is varied. Since $\bar{z}(h) = f(\bar{x}(h), \bar{y}(h))$, it follows that

$$\frac{d\bar{z}}{dh} = f'_x \frac{d\bar{x}}{dh} + f'_y \frac{d\bar{y}}{dh} \tag{52}$$

On the other hand, by (45),

$$F'_x \frac{d\bar{x}}{dh} + F'_y \frac{d\bar{y}}{dh} = 1 \tag{53}$$

From (52), (48) and (53) we get

$$\frac{d\bar{z}}{dh} = \lambda F'_x \frac{d\bar{x}}{dh} + \lambda F'_y \frac{d\bar{y}}{dh} = \lambda \left(F'_x \frac{d\bar{x}}{dh} + F'_y \frac{d\bar{y}}{dh} \right) = \lambda$$

Thus, the multiplier λ is equal to the rate of change of the extremal value as the parameter h varies in the condition. The Lagrange method is remarkable in that the derivative $d\bar{z}/dh$ can be found without writing out the function $\bar{z}(h)$, which may be very intricate, in explicit form.

The statement of a conditional-extremum problem is typical of problems in economics: we seek the maximum quantity of goods z for specified expenditures h; we know the dependence of z and h on the type of action described by the variables x, y. For optimal action, to every h there corresponds one definite $\bar{z}$. The derivative $dh/d\bar{z}$ is the cost price of the (surplus) product in an ideally adjusted economic system where the outlay h has already been made and $\bar{z}$ has been produced, and it is now required to increase the production.

Investigation of a conditional extremum is carried out in similar fashion for functions of any number of variables and for any number of constraints. For example, if we seek the extremum of the function $f(x, y, z, u, v)$ under the conditions

$$F_1(x, y, z, u, v) = 0, \quad F_2(x, y, z, u, v) = 0, \qquad (54)$$
$$F_3(x, y, z, u, v) = 0$$

then we have to proceed as if we were seeking the absolute extremum of the function $f^* = f - \lambda_1 F_1 - \lambda_2 F_2 - \lambda_3 F_3$, where λ_1, λ_2, λ_3 are unknown Lagrange multipliers. The necessary condition of an extremum for f^* yields $f^{*\prime}_x = 0, f^{*\prime}_y = 0, f^{*\prime}_z = 0, f^{*\prime}_u = 0, f^{*\prime}_v = 0$, which, together with (54), produces $5 + 3$ equations in $5 + 3$ unknowns $x, y, z, u, v, \lambda_1, \lambda_2, \lambda_3$.

Exercise

Find the conditional extremum of the function $u(x, y, z) = x^2 - y^2 + z^2 - 2x$ (a) under the condition $x + 2y - z = 3$; (b) under the conditions $x + y - z = 0$, $x + 2y = 1$.

12.8 Conditional extremum in the calculus of variations

Conditional-extremum problems in the calculus of variations are posed and solved in a manner similar to that done in Sec. 12.7 for problems with a finite number of degrees of freedom. To illustrate, let us consider the problem of finding the extremum of the functional (25) under the boundary conditions (26) and the accessory integral condition

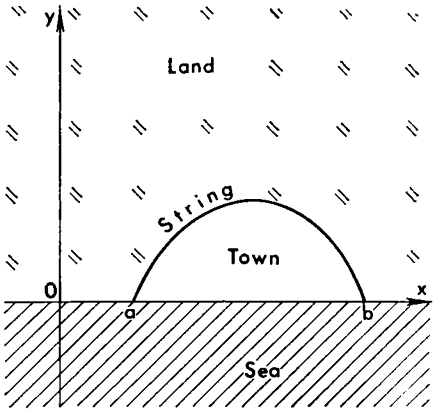
$$\int_a^b G(x, y, y')\, dx = K \qquad (55)$$

where G is a given function and K is a given number. If we partition the interval of integration into a large number of small subintervals and replace the integrals (25) and (55) by sums depending on the values of y_i of the desired function at the division points (that is, if we perform a transition just the reverse of that carried out in Sec. 12.1), then we arrive at a conditional-extremum problem of a function of a finite number of variables y_i. By virtue of the results of Sec. 12.7, this problem is equivalent to the absolute-extremum problem of the expression

$$\int_a^b F(x, y, y')\, dx - \lambda \int_a^b G(x, y, y')\, dx = \int_a^b (F - \lambda G)\, dx$$

(here, we passed from sums to integrals), where λ is a constant Lagrange multiplier not known beforehand. Thus, instead of (29) we have to write a similar Euler equation for the function $F^* = F - \lambda G$. After solving this equation we find the two constants of inte-

Fig. 173



gration and the constant λ from the three conditions (26) and (55), and, by virtue of Sec. 12.7, the quantity λ is of great importance since it is equal to

$$\lambda = \frac{dI_{\text{extrem}}}{dK} \tag{56}$$

This equation makes it possible to determine the character of the dependence of the extremal (generally, stationary) value I_{extrem} of the functional if the parameter K in (55) can vary (this value I_{extrem} then of course depends on K).

By way of an example, let us consider the famous *Dido problem.* As the story goes, there lived a Princess Dido of Phoenicia, who, pursued by the ruler of a neighbouring country, fled to North Africa where she bargained with a local chieftain for some land, agreeing to pay a fixed sum for as much land as could be encompassed by a bull's hide. When her request was granted, she proceeded to cut the hide into very thin strips, tied them end to end, and then to the amazement of the onlookers she enclosed an enormous portion of land. Dido then founded the celebrated ancient city of Carthage.

Ancient scholars already posed the problem of how Dido should have arranged her string in order to encompass a maximum area, in other words, how to choose from all curves of given length that one which embraces the largest area. It turned out that the solution of this *isoperimetric problem* is a circle (if the curve is closed) or an arc of a circle (if the curve is not closed, say, if Dido chose land along the sea so that the ends of the string lay on the shore; see Fig. 173). However, it is not at all easy to prove this in a rigorous manner.

Let us see how Dido's problem can be formulated analytically. For the sake of simplicity, assume the shoreline to be straight and put the coordinate axes as in Fig. 173. We consider a variant in which the position of the endpoints of the string is fixed. Then the shape of the string is given by the equation $y = y(x)$; the function $y(x)$ is not known beforehand but it must be such as to satisfy the conditions

$$y(a) = 0, \quad y(b) = 0 \tag{57}$$

Besides, the given length of the curve is

$$\int_a^b \sqrt{1 + y'^2}\, dx = L \tag{58}$$

(the formula for the length of a curve is derived in integral calculus; see HM, Sec. 4.5). And it is required that the area, i.e.

$$S = \int_a^b y\, dx \tag{59}$$

be a maximum. We thus have the following problem: from among all functions $y = y(x)$ that satisfy the conditions (57) and (58) choose the one for which the integral (59) has the largest possible value.

The problem of Dido was first solved by geometrical methods involving ingenious but not altogether obvious reasoning. Using the apparatus of the calculus of variations, we can solve it by a standard procedure. By virtue of what was said at the beginning of this section, it is required merely to solve the Euler equation for the function

$$F^* = y - \lambda\sqrt{1 + y'^2} \tag{60}$$

Using the intermediate integral (34), we get

$$y - \lambda\sqrt{1 + y'^2} + \lambda \frac{y'}{\sqrt{1 + y'^2}}\, y' = C_1$$

Transforming, we obtain

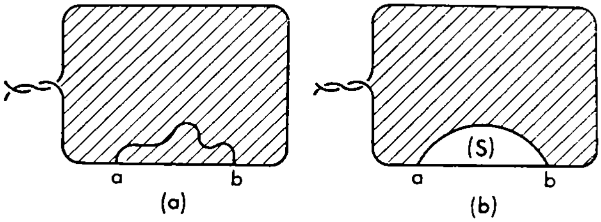
$$y - \frac{\lambda}{\sqrt{1 + y'^2}} = C_1; \quad \lambda^2 = (1 + y'^2)(C_1 - y)^2; \quad y'^2 = \frac{\lambda^2}{(C_1 - y)^2} - 1;$$

$$\frac{dy}{dx} = \pm \frac{\sqrt{\lambda^2 - (y - C_1)^2}}{y - C_1}; \qquad \frac{(y - C_1)\, dy}{\pm \sqrt{\lambda^2 - (y - C_1)^2}} = dx; \tag{61}$$

$$\pm \sqrt{\lambda^2 - (y - C_1)^2} = x + C_2; \quad (x + C_2)^2 + (y - C_1)^2 = \lambda^2$$

We get the equation of a circle, hence the solution of Dido's problem is an arc of a circle, as we have already pointed out. Since the length L of the arc of the circle and its endpoints are given, it is easy to find the centre of the circle and to represent the arc itself.

Fig. 174



Dido's problem has variants. If the length L is given, but the endpoints a, b of the arc are not given, then of all arcs of the given length it is required to find the one that bounds maximum area. It can be verified that the desired arc of the circle has to approach the shoreline perpendicularly; in particular, if the shoreline is straight, then we have to take a semicircle. If there is no sea and the line has to be closed, then, as pointed out above, we get a circle. For the last case, let us verify the relation (56). Here, $I_{\text{extrem}} = \pi R^2$, $K = L = 2\pi R$ (R is the radius of the circle), whence

$$\frac{dI}{dK} = \frac{d(\pi R^2)}{d(2\pi R)} = R$$

This corresponds precisely to formula (61), from which it is evident that the radius of the circle is equal to $|\lambda|$ (below we will see that $\lambda = R$, so that the signs in (56) are also in agreement).

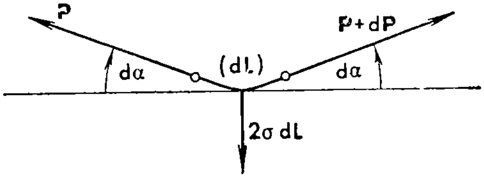
The problem of Dido has a simple physical explanation. Imagine a rigid horizontal frame over which is stretched a soap film and let a thin string rest on the film with ends attached to the frame (Fig. 174*a*) at the points a and b. If the film is punctured with a needle inside the encompassed area, then, due to surface tension, the film will stretch (see Fig. 174*b*) so that the area becomes the smallest possible, which means the film-free area becomes the largest possible. This is precisely the problem of Dido.

This physical picture enables one to interpret the Euler equation as an equation of static equilibrium (as in Sec. 12.1). If the length L of the string is given, then the potential energy U of the system is proportional to the surface of the film, or

$$U = 2\sigma(S_0 - S)$$

where σ is the "coefficient of surface tension" (the factor 2 is due to the fact that the film has two sides) and S_0 is the area embraced by the frame, so that $S_0 - S$ is the area of the film. As we have seen, the condition of static equilibrium reduces to the stationarity of the potential energy U, which amounts to the stationarity of S (the maximal nature of S signifies the minimal nature of U, which means equilibrium stability). We thus arrive at the Euler equation for the function F^*.

Fig. 175



The solution that we have found can be obtained from physical considerations. Let us consider (Fig. 175) an element (dL) of string acted on by the force $2\sigma\, dL$ of surface tension and the forces of tension of the string P and $P + dP$ applied to its endpoints. Projecting these forces on the tangent line to the element and discarding all terms higher than first order, we get

$$-P + P + dP = 0$$

(bear in mind that the cosine of a small angle differs from unity by a second-order quantity). From this $dP = 0$, that is, $P =$ constant along the string. Projecting the forces on the normal, we get

$$2P \sin d\alpha = 2P\, d\alpha = 2\sigma\, dL \qquad (62)$$

whence

$$\frac{2\, d\alpha}{dL} = \frac{2\sigma}{P} = \text{constant}$$

The ratio $\frac{2\, d\alpha}{dL}$ constitutes the curvature k of the string (Sec. 9.4). This means that the string at equilibrium has a constant curvature, i.e. it constitutes an arc of a circle of radius $R = \frac{1}{k} = \frac{P}{2\sigma}$. (Incidentally, the constancy of P may also be obtained from formula (62), which can be rewritten as $P = 2\sigma R$ by proceeding from the solution (61) of the conditional-extremum problem; here $R = \lambda$, or $P = 2\sigma\lambda$.)

We can approach the soap-film problem in a different way. We assume the tension P of the string to be given, and its length L unknown; in other words, we consider the scheme shown in Fig. 176. In this case the potential energy of the system is equal to

$$U = \text{constant} - 2\sigma S + PL = \text{constant} - 2\sigma\left(S - \frac{P}{2\sigma} L\right)$$

since the load rises by ΔL when L is increased by ΔL, that is, U increases by $P\Delta L$. We have thus arrived at a problem involving the absolute extremum of the functional

$$S - \frac{P}{2\sigma} L = \int_a^b \left(y - \frac{P}{2\sigma}\sqrt{1 + y'^2}\right) dx$$

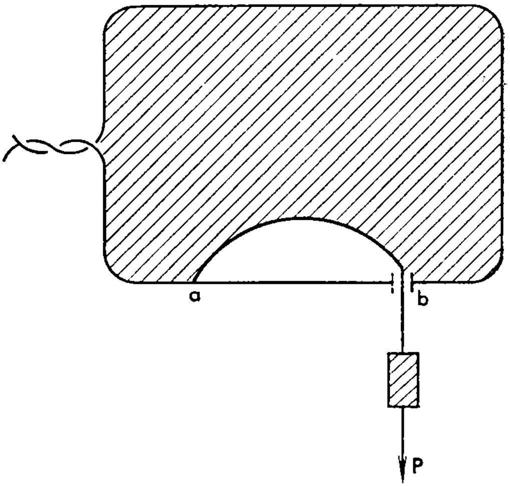


Fig. 176

We now have to set up the Euler equation for the function (60) with $\lambda = \frac{P}{2\sigma}$. As we have seen, the solution of this equation is a curve given by the equation (61), that is, an arc of a circle of radius

$$R = |\lambda| = \frac{P}{2\sigma} \tag{63}$$

Thus, the solution of a conditional-extremum problem in the new interpretation may be regarded as a solution of the absolute-extremum problem; this transition has a physical as well as a formal meaning.

It is interesting to note that in the second interpretation the problem has two solutions: from (63) we get the value of the radius of the curved string, but there are two possible positions of the string with a given radius (Fig. 177)! If we consider the family of circles passing through the points a and b, the curve of U as a function of h is of the form shown in Fig. 178. Thus, of the two possible positions of the weight, the lower one is stable and the upper one is unstable. Here, as at the end of Sec. 12.5, the minimum is of an absolute nature, whereas the maximum is extremal only among the arcs of the circles. Actually, for the higher position the energy has a minimax.

The problem of Dido can be generalized by assuming that it is required to embrace the most valuable portion of land, the cost of unit area of land not being a constant but, say, dependent on the distance from the sea. This generalized problem can no longer be solved in an elementary fashion. But it can be solved with the aid of Euler's equation by noting that here instead of (59) we have

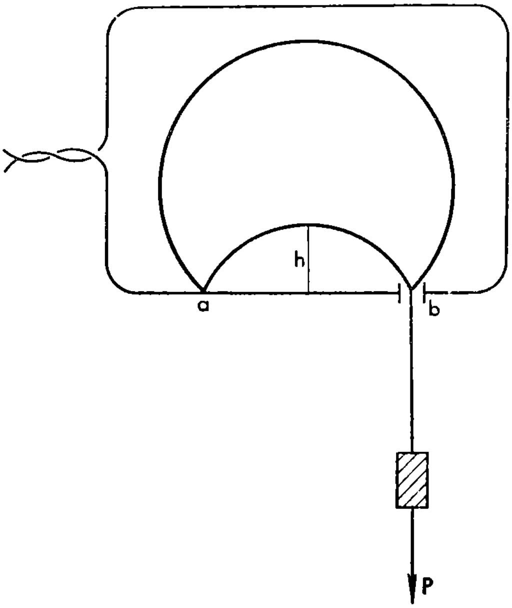


Fig. 177

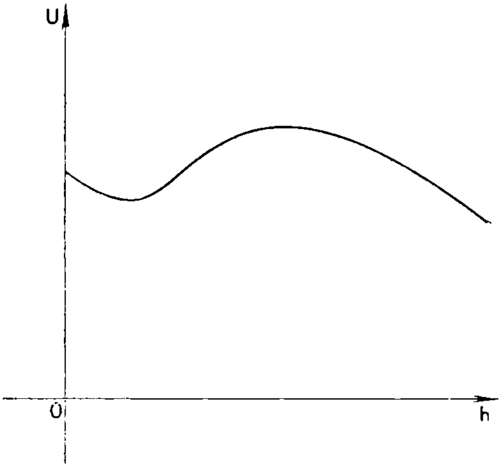


Fig. 178

to maximize the integral $\int\limits_a^b F(y)\,dx$, where $F(y) = \int\limits_0^y \varphi(\eta)\,d\eta$ and $\varphi(\eta)$ is the cost of unit area at a distance η from the sea. (If $\varphi(\eta) \equiv 1$, then we come back to (59).)

Let us consider another conditional-extremum problem: the problem of the distribution of charges in a conductor. Imagine an isolated conductor (Ω) of arbitrary shape charged with q electricity. This quantity is arranged on the conductor with density ρ, which it is required to find.

To solve this problem, recall formula (23) of Ch. 10 for a potential generated by a distributed charge, which we rewrite as

$$\varphi(\mathbf{r}) = \int\limits_{(\Omega)} \frac{\rho(\mathbf{r}_0)}{|\mathbf{r} - \mathbf{r}_0|}\,d\Omega_0$$

Therefore the total potential energy of the charge is

$$U = \frac{1}{2}\int\limits_{(\Omega)} \varphi(\mathbf{r})\,dq = \frac{1}{2}\int\limits_{(\Omega)} \varphi(\mathbf{r})\,\rho(\mathbf{r})\,d\Omega$$

$$= \frac{1}{2}\int\limits_{(\Omega)}\int\limits_{(\Omega)} \frac{\rho(\mathbf{r}_0)\,\rho(\mathbf{r})\,d\Omega\,d\Omega_0}{|\mathbf{r} - \mathbf{r}_0|}$$

By the general principle of static equilibrium, this integral must assume a minimal (or, at any rate, a stationary) value provided

$$\int\limits_{(\Omega)} \rho(\mathbf{r})\,d\Omega = q\ (= \text{constant})$$

since charges can only be redistributed but cannot be created or made to vanish. What we have, therefore, is a conditional-extremum problem. To solve it, equate to zero the variation $\delta(U - \lambda q)$, where λ is a Lagrange multiplier.

But by the formula for the differential of a product,

$$\delta U = \frac{1}{2}\int\limits_{(\Omega)}\int\limits_{(\Omega)} \frac{\delta\rho(\mathbf{r}_0)\,\rho(\mathbf{r})\,d\Omega\,d\Omega_0}{|\mathbf{r} - \mathbf{r}_0|} + \frac{1}{2}\int\limits_{(\Omega)}\int\limits_{(\Omega)} \frac{\rho(\mathbf{r}_0)\,\delta\rho(\mathbf{r})\,d\Omega\,d\Omega_0}{|\mathbf{r} - \mathbf{r}_0|}$$

If in the first integral we denote $\mathbf{r}$ by $\mathbf{r}_0$ and $\mathbf{r}_0$ by $\mathbf{r}$, then it is equal to the second integral, whence

$$0 = \delta(U - \lambda q) = \int\limits_{(\Omega)}\int\limits_{(\Omega)} \frac{\rho(\mathbf{r}_0)\,\delta\rho(\mathbf{r})\,d\Omega\,d\Omega_0}{|\mathbf{r} - \mathbf{r}_0|} - \lambda\int\limits_{(\Omega)} \delta\rho(\mathbf{r})\,d\Omega$$

$$= \int\limits_{(\Omega)} [\varphi(\mathbf{r}) - \lambda]\,\delta\rho(\mathbf{r})\,d\Omega$$

Since $\delta\rho(\mathbf{r})$ is arbitrary, we get $\varphi(\mathbf{r}) - \lambda = 0$, or $\varphi(\mathbf{r}) = \lambda = \text{constant}$.

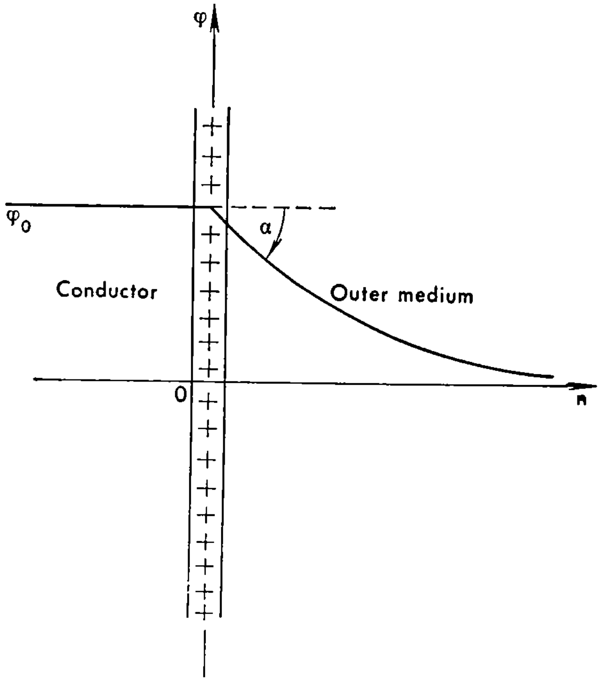


Fig. 179

Summarizing, we see that a free charge in a conductor distributes itself so that the potential φ_0 is the same at all points of the conductor. It is also evident that the Lagrange multiplier in this problem is equal to φ_0 (derive this last result from the formula (56)). The constancy of the potential could also have been derived by physical reasoning: if the potential were different at distinct points, then the freely moving charges in the conductor would flow from high-potential regions to regions of lower potential, so that under these conditions static equilibrium is impossible.

Since the potential is connected with charges by the familiar Poisson equation (43) of Ch. 10, we can draw the conclusion from the constancy of the potential that $\rho = 0$ inside the conductor, which means that the entire charge is located on the surface of the conductor. Incidentally, this conclusion could have been made on the basis of physical reasoning.

When we pass through the surface of the conductor, the potential has a "corner" on the graph (Fig. 179), which means a discontinuity in the first derivative. If we choose the coordinate axes so that the origin lies at a certain point of the conductor and the z-axis is in the direction of the outer normal to the surface of the conductor, then the derivative $\frac{\partial\varphi}{\partial z}$ will have a jump $-\tan\alpha$. Therefore $\frac{\partial^2\varphi}{\partial z^2}$, and with it $\Delta\varphi$, will have a delta-like singularity of the form $-\tan\alpha\delta(z)$. On the other hand, if the charge is located on the surface with density σ, then near the point at hand, $\rho = \sigma\delta(z)$. Therefore the equation (43) of Ch. 10 yields $-\tan\alpha\delta(z) = -4\pi\sigma\delta(z)$, or $\tan\alpha = 4\pi\sigma$.

It is to be noted in conclusion that solving a conditional-extremum problem always signifies the solution of a certain "companion" conditional extremum problem. For example, in the problem of Dido we get the same answer if we seek the minimum length of a string enclosing a given area. Indeed, if a noncircular contour (L_1) enclosed the same area as the circle (L), $L_1 < L$, then by proportionally increasing the contour (L_1), we could obtain the contour (L_2), $L_2 = L$, enclosing a greater area than (L), but this runs counter to the solution of Dido's problem. Similarly, in the potential problem we could seek the maximum charge that fits on a conductor for a given value of the potential energy. In the general case, it turns out that the problem of the stationary value of A, provided B is constant, and the problem of the stationary value of B, provided A is constant, lead to the same solutions. This is quite true because they reduce to problems of the stationarity of $A + \lambda B$ or $B + \mu A = \mu(A + \lambda_1 B)$ $(\lambda_1 = 1/\mu)$. That is to say, since μ is constant, of $A + \lambda_1 B$. Hence the two problems are equivalent.

Exercises

1. Find a solid of revolution of the largest volume having a given surface area.
2. Find the shape of equilibrium for a homogeneous heavy string suspended at the endpoints.
   Hint. The string must lie so that its centre of gravity is as low as possible.
3. In the foregoing problems, indicate the meaning of the Lagrange multiplier.

12.9 Extremum problems with restrictions

We deal here with another class of extremum problems that have come to the fore in recent years. These are problems in which the desired quantities or the desired function are restricted by accessory conditions in the form of inequalities.

We start with the simplest case. Suppose we are seeking the extremum of the function $y = f(x)$, the independent variable x being restricted to the range given by the inequalities $a \leqslant x \leqslant b$ (Fig. 180). We then have the possibility of inner extrema (a maximum at $x = c$ in Fig. 180 and a minimum at $x = d$) and also of *endpoint extrema* (the endpoint minimum at $x = a$ and the endpoint maximum at $x = b$). For finding inner extrema, use can be made of the statio-

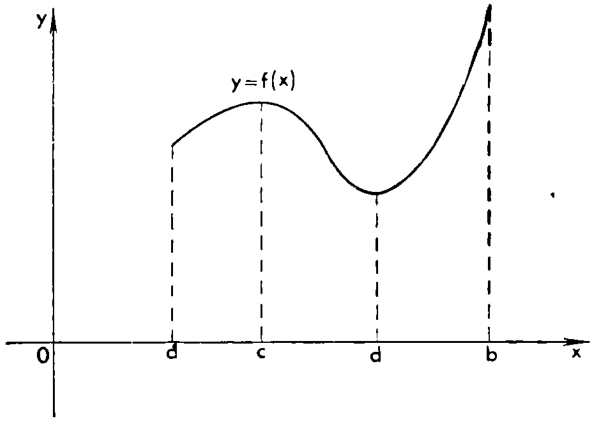


Fig. 180

narity condition $y' = 0$, whereas for the endpoint extrema this condition does not hold, which is to say that the latter are in the nature of "cuspidal" extrema (cf. HM, Sec. 4.2.).

Suppose we are now looking for the extremum of a function of two variables $z = f(x,y)$ and the independent variables are connected by the restriction $F(x,y) \geqslant 0$. We assume that this inequality defines, in the xy-plane of the arguments, a certain finite region (S) with boundary (L) (Fig. 181) on which $F \equiv 0$. The function f can have both inner extrema within (S) and boundary extrema on (L). To find the former, we can use the conditions of stationarity discussed in Sec. 4.6. However, as in the one-dimensional case, these conditions fail for the boundary extrema. To find the boundary extrema, note that if the function f has an extremum (say, a maximum) at some point M (Fig. 181), then the value of $f(M)$ is less than all values of f on (L) near M. Therefore, at the same time, f is a minimum at M provided $F = 0$, and such points can be sought out by the method of Sec. 12.7. And so the boundary extrema are found by the rule for finding conditional extrema.

Extremum problems with restrictions are also encountered in the calculus of variations. By way of an illustration we consider the problem of the curve of quickest descent that was solved in Sec. 12.4, but let us impose the restriction that the particle in the descent is not allowed to fall below the terminal point, that is, the desired solution is restricted by the inequality $y \geqslant y_b$. We draw a cycloid with cusp at the initial point A (Fig. 182), and the sought-for curve must lie entirely in the hatched area. It is clear that if the terminal point lies to the left of $\overline{B}$, say, at B_1, then the desired curve is the arch AB_1 of the cycloid, since this solution is the best of all curves and also satisfies the stated inequality. But what will happen if the terminal point lies to the right of $\overline{B}$, say at B_2?

Fig. 181

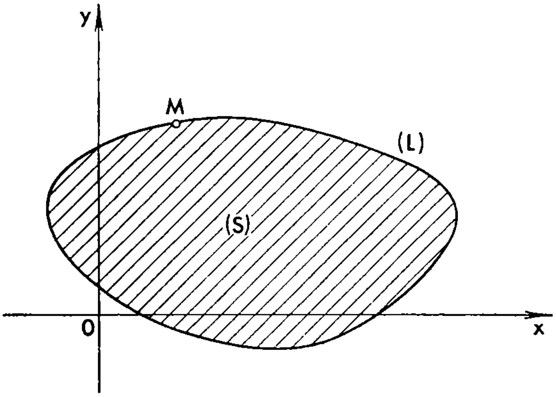
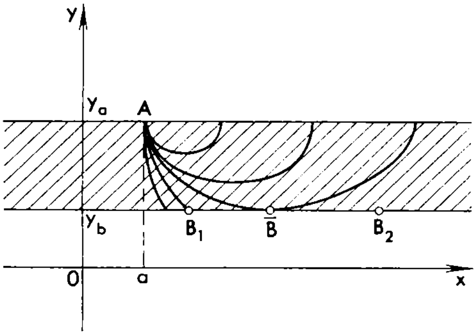


Fig. 182



First of all, observe that the portion of the desired curve which lies strictly inside (not on the boundary) of the hatched area satisfies the Euler equation and therefore is a cycloid. This is true, for if the entire curve serves as the solution of the problem, then any arc of the curve is the curve of quickest descent between the endpoints.* Thus, the desired curve may consist of the integral curves of the Euler equation (which are called extremals) and the curves lying on the boundary of the hatched area. But now we see that the desired curve consists of the arc $A\bar{B}$ of the cycloid and the line segment $\bar{B}B_2$. Indeed, take any other possible curve, say $AB_1\bar{B}B_2$. Then on its segment $AB_1\bar{B}$ as well it must be a solution of the minimum problem, which is not so, for the arc of the cycloid $A\bar{B}$ yields a shorter time of

* An arc AC between A and any point C situated between A and B is the curve of quickest descent with zero initial velocity, which is to say it solves the same problem as was posed in determining AB. The arc between the two intermediate points C' and C'' solves the problem of quickest descent for a given nonzero initial velocity at C' corresponding to the difference in the altitudes of A and C'.

descent. It is readily seen that we have to take the arc of precisely that cycloid for which B is the most rightwards one, i.e. the cycloid tangent to the lower straight line.

Thus, the solution of a variational problem with restrictions may only partially satisfy the condition of stationarity (Euler's equation) or not at all. As in the case of a function of one variable, this solution is in the nature of a cuspidal extremum.

Extremum problems with restrictions also arise in control processes. For example, the problem may be one of controlling the motion of a vehicle in space by turning rudders so as to set the vehicle on a given flight path in the shortest possible time. Since what is sought is the law of rotation of the rudders in time, we have a variational problem. And if the rudders have turning limiters (designed, say, for restricting centrifugal forces), then we have a problem with restrictions. In an optimal control process obtained by solving a variational problem, there are time intervals during which the rudders lie on the limiters, that is, the solution passes along the boundary of the region in which it can reside. On approaching this boundary, the solution must satisfy definite conditions that we will not consider here.

Exercises

1. Find the maximal and minimal values of the function $z = x^2 - y^2 - x + y$ in a triangle bounded by the straight lines $x = 0$, $y = 0$, $x + y = 2$.
2. Find the position of a heavy homogeneous string of length $L > 2$ suspended at the points $(-1, 1)$ and $(1, 1)$ if the string is located in the upper half-plane $y \geqslant 0$.

12.10 **Variational principles. Fermat's principle in optics**

Besides separate problems of a variational nature, examples of which have been discussed in the foregoing sections, there are variational principles, each of which is applicable to analyzing a broad class of phenomena. Each of these principles ordinarily asserts that of all states, processes, and the like, that are admissible for the given constraints or restrictions in the physical system under consideration, only that state or process, etc., is materialized for which a certain functional (quite definite for the given principle) assumes a stationary value. Each of these principles makes possible a uniform consideration of many problems embraced by a single theory, and for this reason it often happens that such a principle is a substantial scientific attainment.

One of the most important and general principles of physics has already been demonstrated in a large number of cases. It consists in the following: of all possible positions of a physical system admissible under given constraints, the equilibrium position is that for which the potential energy of the system at hand has a stationary value (the value is minimal for a stable equilibrium of the system). If the system has a finite number of degrees of freedom, the potential energy can be expressed as a function of a finite number of generalized coordinates (see Sec. 4.8), so that the application of the indicated principle reduces to finding the minimum of a function of several variables (see Sec. 4.6). But if we consider a continuous medium whose state is described by a certain function of the coordinates (or by a system of such functions), then the potential energy is a functional, and application of this principle is carried out on the basis of the methods of the calculus of variations, just as we have already done.

Historically, one of the first variational principles was that of Fermat in optics. It states that of all conceivable paths joining two specified points, a light ray chooses the path it can traverse in minimum time. To write down the Fermat principle mathematically, recall that the velocity of light c in a vacuum has a very definite value, whereas in a transparent medium the velocity is $\frac{c}{n}$, where $n > 1$ is the index of refraction of the medium, and generally depends not only on the medium but also on the wavelength of the light. For the sake of simplicity, we consider a case where n does not depend on the wavelength and so will have a definite value in every medium. Since light covers a path dL in time $dt = $ $= dL : \frac{c}{n} = \frac{n}{c}\, dL$, it follows that the time t during which the given path (L) is traversed is found from

$$t = \frac{1}{c} \int\limits_{(L)} n\, dL \tag{64}$$

Thus, we have a functional that depends on the choice of path connecting two points, so that the problem of finding the shape of the light ray is a variational problem.

The basic laws of propagation of light can be derived from the Fermat principle.

For example, in the case of the propagation of light in a medium with a constant *optical density*, i.e. with a constant velocity of light, the Fermat principle leads us to conclude that light is propagated in a straight line (Fig. 183, where S denotes the light source and A the point of observation), since the straight line SA is the shortest line connecting S and A.

From this, in turn, it follows that if the light path consists of several sections (for example, between successive reflections, refractions, and so forth), each of which is in a medium of constant optical density, then each of these sections is a straight-line seg-

Fig. 183

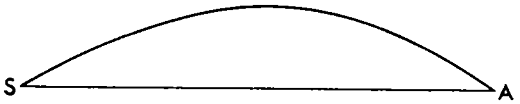
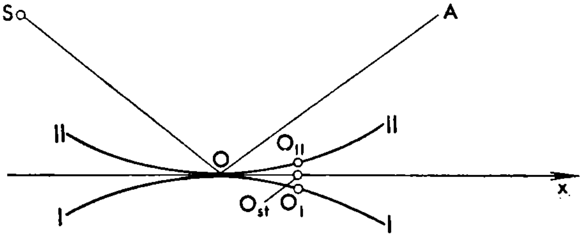


Fig. 184



ment and the entire path is a broken line. To determine the path length, it is required to find the coordinates of the vertices of the broken line, which is to say that there is only a finite number of degrees of freedom. For this reason such problems are solved by the tools of differential calculus. In particular, by solving the minimum problem, it is easy in this way to derive the laws of reflection of light from a plane surface and the refraction of light passing through a plane interface of two media (see HM, Sec. 4.1).

We see that Fermat's principle does indeed lead to conclusions that are in full agreement with experiment as concerns the reflection and refraction of light.

Now let us consider the reflection of light from a curved surface tangent to the plane at the point O (Fig. 184). The figure shows two examples of surfaces bent in opposite directions: IOI, concave downwards, IIOII, concave upwards from the x-axis. (We consider cylindrical surfaces with generatrices perpendicular to the plane of the drawing).

It can be shown that it suffices here to consider the rays that lie in the plane of the drawing, and sections of reflecting surfaces by the plane of the drawing. For this reason, in the future we will not speak about a reflecting surface but about a reflecting straight line, and not about reflecting curved surfaces, but about the reflecting curves IOI and IIOII in the plane of Fig. 184.

We can consider the problem without any concrete computations. It is known (see HM, Sec. 3.17) that the distance between a curve and a tangent line is proportional to $(x - x_0)^2$, where x_0 denotes the abscissa of the point of tangency O.

Let us consider the lengths of the broken lines $L_{st}(SO_{st}A)$, $L_I(SO_IA)$, and $L_{II}(SO_{II}A)$. These broken lines are not shown in Fig. 184 so as not to clutter up the drawing; the points O_{st}, O_I, O_{II}, as can be seen in the drawing, lie to the right of the point of tangency O for one and the same value of x; here, O_{st} lies on the straight line, O_{I} on the lower line I, O_{II} on the upper line II. If we see the points S, A, O_{st}, O_{I}, O_{II} in the drawing, it is easy to imagine the broken lines.

The abscissas of O_{st}, O_{I}, O_{II} are the same. The ordinates of O_{st}, O_{I} and O_{II} differ by a quantity proportional to $(x - x_0)^2$. Hence, the lengths of L_{st}, L_{I}, L_{II} also differ by a quantity proportional to $(x - x_0)^2$. Let us write the expansions of L_{st}, L_{I} and L_{II} in a Taylor series in powers of $(x - x_0)$:

$$L_{\mathrm{st}}(x) = L_{\mathrm{st}}(x_0) + (x - x_0)\,L'_{\mathrm{st}}(x_0) + \frac{(x - x_0)^2}{2}\,L''_{\mathrm{st}}(x_0) + \ldots,$$

$$L_{\mathrm{I}}(x) = L_{\mathrm{I}}(x_0) + (x - x_0)\,L'_{\mathrm{I}}(x_0) + \frac{(x - x_0)^2}{2}\,L''_{\mathrm{I}}(x_0) + \ldots,$$

$$L_{\mathrm{II}}(x) = L_{\mathrm{II}}(x_0) + (x - x_0)\,L'_{\mathrm{II}}(x_0) + \frac{(x - x_0)^2}{2}\,L''_{\mathrm{II}}(x_0) + \ldots$$

Since L_{st}, L_{I} and L_{II} differ only by a quantity of the order of $(x - x_0)^2$, this means that

$$L_{\mathrm{st}}(x_0) = L_{\mathrm{I}}(x_0) = L_{\mathrm{II}}(x_0), \tag{65}$$

$$L'_{\mathrm{st}}(x_0) = L'_{\mathrm{I}}(x_0) = L'_{\mathrm{II}}(x_0)\ (= 0), \tag{66}$$

$$L''_{\mathrm{st}}(x_0) \neq L''_{\mathrm{I}}(x_0) \neq L''_{\mathrm{II}}(x_0) \tag{67}$$

The first of these equations, (65), are obvious consequences of the fact that all three curves — the straight line st, the curve I, and the curve II — pass through the single point O, $x = x_0$.

The second set of equations, (66), result from the fact that at the point O the three indicated curves are tangent to one another, and the angle of incidence of a ray is equal to the angle of reflection.

We know that if the derivative is zero, this means that L_{st}, L_{I} and L_{II}, as functions of x, can have a minimum or a maximum at $x = x_0$. Whether this is a minimum or a maximum depends on the sign of the second derivative.

For the straight line, we have a minimum, $L''_{\mathrm{st}} > 0$. However, from (67) we see that we cannot draw this conclusion for the curves I and II. In particular, for sufficient curvature of curve II, its length L_{II} has precisely a maximum at O for $x = x_0$ and not a minimum. Because the curve II rises to the left and right of O, the path $SO_{\mathrm{II}}A$ is shorter than SOA and the path SOA is the longest of all adjacent paths from point S to any point of the curve II and thence to point A.

Experiment shows that in the case of curve II as well the reflection occurs at the point O, that is, at the point where the length of the path has a maximum; it is obvious that at this point the angle of incidence is still equal to the angle of reflection.

The fact that the path length may not be a minimum but a maximum is very important for understanding the meaning and origin of the Fermat principle. It is obvious that this principle is not the consequence of any "striving" on the part of light to choose shortest routes: if there were such a "desire" on the part of light, it would not be indifferent to whether we were dealing with a minimum or a maximum.

Actually, Fermat's principle is a consequence of the fact that the propagation of light is a propagation of waves in which the electric and magnetic fields rapidly change sign (10^{15} times per second for visible light). Accordingly, the wavelength of light is very small. At a given instant, the signs of the field are opposite at points a half-wavelength apart. Suppose at a certain time the field has a negative sign at the light source S. The sign of the field at A depends on how many times the sign can change over the route from S to A. If we consider two neighbouring routes from S to A, then for the same path length those fields that arrive at A from S along these routes will have the same sign; they will combine and reinforce each other.

If the path lengths are distinct, then the fields can either have the same sign or different signs and, on the average, they cancel out. Herein lies the reason why, in the propagation of light, an important role is played by beams of rays of the same wavelength and with the same time of propagation.

That the derivative of the path length with respect to the coordinate of the point of reflection is equal to zero means that at least to within terms of the order of $(x - x_0)^2$ the path lengths are the same, the waves that go through point O and the adjacent waves are reinforced. From this point of view it is obviously immaterial whether we deal with a minimum or a maximum, $L'' > 0$ or $L'' < 0$. What is more, it is clear that the best conditions for reflection are obtained when the path length is the same over as large a portion of the reflecting surface as possible. Which means that it is desirable that $L'' = 0$; then the dependence of L on x for small x is still weaker — only in the terms $(x - x_0)^3$ and higher. Physically speaking, $L'' = 0$ corresponds to a curvature of the reflecting line for which a mirror collects (focusses) at the point A the rays emanating from S.

In geometrical optics, rays are regarded as geometric lines. We have just seen, in the example of reflection, that the laws of geometrical optics are a consequence of the wave nature of light. Instead of one single ray we considered a beam of neighbouring rays.

We can give an estimate of the thickness of this beam. The amplitudes of rays whose path length from S to A differs by less than one half of the wavelength combine. We set up the conditions: $|L(x) - L(x_0)| = \frac{1}{2}\lambda$. Near the point of reflection, $L'(x_0) = 0$, $|L(x) - L(x_0)| = \frac{1}{2}(x - x_0)^2 |L''(x_0)|$, whence $|x - x_0| = \sqrt{\frac{\lambda}{|L''(x_0)|}}$. Thus, in reality the light is not reflected in a single point of the mirror, but on a spot, the dimensions of which, as may be seen from the formula, only tend to zero as the wavelength approaches zero.

Using Fermat's principle, one can consider the more complicated problem of the shape of a light ray in a medium with a gradually varying optical density, for example, the light ray coming from a star to an observer when account is taken of the gradual variation in the density of the atmosphere. In this case, the ray will proceed by a curvilinear path and will be determined from the condition of the stationary value (ordinarily a minimum) of the integral (64). In certain cases, integration of the appropriate Euler equation can be carried to quadrature. Consider for example the propagation of light in the xy-plane if the velocity of light depends only on the altitude y, that is, $n = n(y)$. Rewriting (64) as

$$ct = \int_{(L)} n(y)\sqrt{1 + y'^2}\, dx$$

we can take advantage of the intermediate integral (34) of Euler's equation, which yields

$$n(y)\sqrt{1 + y'^2} - n(y)\frac{y'}{\sqrt{1 + y'^2}}\, y' = C_1, \quad \text{or} \quad \frac{n(y)}{\sqrt{1 + y'^2}} = C_1$$

From this it is easy to find

$$y' = \frac{dy}{dx} = \frac{\sqrt{[n(y)]^2 - C_1^2}}{C_1}, \qquad \frac{C_1\, dy}{\sqrt{[n(y)]^2 - C_1^2}} = dx,$$

$$C_1 \int \frac{dy}{\sqrt{[n(y)]^2 - C_1^2}} = x + C_2$$

If it is possible to take the last integral for some special form of the function $n(y)$, we get the equation of the light ray in closed form, otherwise the integral can be computed in approximate fashion (Ch. 1). It is interesting to note that for $n = \frac{n_0}{\sqrt{y_0 - y}}$ we have a problem that was solved in Sec. 12.4 (light is propagated along cycloids) and for $n = n_0 y$ we have a problem solved in Sec. 12.5 (light is propagated along catenaries).

Note that in the general case where the refractive index depends on the light frequency, it is necessary to distinguish between the group velocity and the phase velocity of light in the medium.

The *phase velocity of light*, c_{ph}, gives the following relationship between the wavelength and the oscillation period:

$$\lambda = c_{\text{ph}}T$$

Here, all fields are harmonically dependent on x and t:

$$H,\ E \propto \cos\left(\frac{2\pi t}{T} - \frac{2\pi x}{\lambda}\right) = \cos\left[\frac{2\pi}{\lambda}(c_{\text{ph}}t - x)\right]$$

($\propto$ is the variation symbol). The quantity under the cosine symbol may be called the phase, whence the term "phase velocity". The refractive index gives a direct description of the phase velocity, $c_{\text{ph}} = \frac{c}{n}$, where c is the velocity of light in vacuum.

We must distinguish between c_{ph} and the so-called *group velocity of light*, c_{gr}. The group velocity is involved in the answer to the question: What is the time lapse for a light ray travelling from point 1, after a shutter is opened, to point 2 in a homogeneous medium? The answer is $t = r_{12}/c_{\text{gr}}$, where c_{gr} is expressed in terms of the refractive index thus:

$$c_{\text{gr}} = \frac{c}{n + \omega\dfrac{dn}{d\omega}}$$

Hence the expression c_{gr} is equal to c_{ph} only in the particular case where n does not depend on ω, $c_{\text{gr}} = c_{\text{ph}}$; in the general case they differ.

For example, the refractive index is less than 1 for X-rays in the medium, $n = 1 - \frac{a}{\omega^2}$. Hence $c_{\text{ph}} > c$. But this does not mean that a signal can be propagated with a velocity greater than that of light! For it is easy to see that

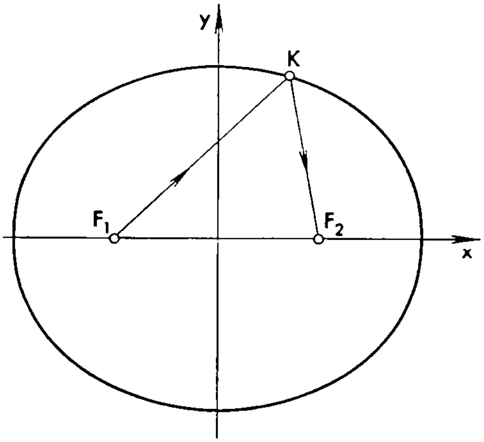
$$n + \omega\frac{dn}{d\omega} = 1 - \frac{a}{\omega^2} + \omega\frac{2a}{\omega^3} = 1 + \frac{a}{\omega^2}$$

so that $c_{\text{gr}} < c$.

Experiment shows that Fermat's principle involves just n and c_{ph}; for this reason it is once again clear that the Fermat principle is not the result of a tendency on the part of light to get from one point to another in the shortest possible time. The matter lies in the condition under which the waves combine, and this depends on the phase of the wave.

The difference between minimality and stationarity of the value of a functional in a real process is illustrated very well in the following simple example taken from a paper entitled "The Principle

Fig. 185



of Least Action" [13] by the celebrated German theoretical physicist Max Planck (1858-1947). The path of a free particle moving on a surface without application of external forces is the shortest route connecting the initial and terminal points of the path. On a sphere, this is an arc of a great circle. But if the path is longer than half the circumference of a great circle, then, as can readily be seen, the length, though stationary, will not be minimal compared with the lengths of neighbouring routes. Neither will the value be a maximum, for by introducing zigzags it is possible to leave the initial and terminal points and yet increase the path length. This value is in the nature of a minimax (see Sec. 4.6). Recalling the once current idea that nature is governed by God's handiwork and that underlying every phenomenon of nature is a conscious intention directed at a specific purpose and that this purpose is attained by the shortest route and via the best means, Planck ironically observed: "Hence Providence is no longer operative beyond the limits of a semicircle."

Exercises

1. The equation of an ellipse is known to be derived from the condition that the sum of the distances of any point K from two points $F_1(x_1 = -c,\ y = 0)$ and $F_2\ (x_2 = c,\ y = 0)$, that is, the sum $F_1K + F_2K = L$, is the same for all points (Fig. 185). Find the tangent and the normal to an ellipse at an arbitrary point K. Find the angles between the lines F_1K, F_2K and the normal to the point K. Show that all rays emanating from F_1 pass through F_2 after reflection and, hence, come to a focus at F_2.
2. Consider the reflection from the parabola $y = x^2$ of a parallel beam of light going downward along the axis of ordinates

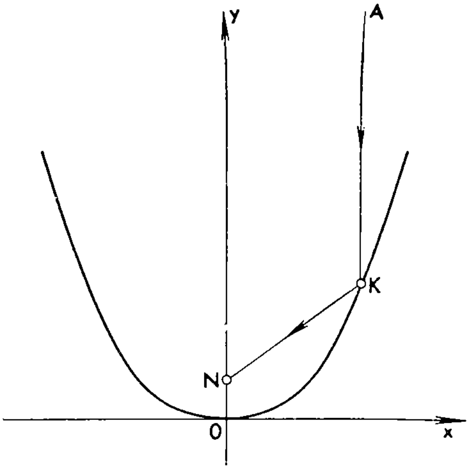


Fig. 186

(Fig. 186). Assuming that the light moves from A, $x_A = 0$, $y_A = Y$ and regarding Y as large, replace by $Y - y_K$ the distance AK equal to $\sqrt{x_K^2 + (Y - y_K)^2}$. Find the length $L = AKN$ and find the point N for which $\frac{dL}{dx_K} = 0$. Find the tangent and the normal to the parabola at the point K and convince yourself that for N the condition of the angle of incidence being equal to the angle of reflection holds true.

3. The points S and A are located 1 cm above a mirror and are separated by 2 cm. Find $L''(x_0)$ and, for $\lambda = 5.10^{-5}$ cm, find the dimensions of the reflecting region.

12.11 Principle of least action

The success of the universal principle of the minimum of potential energy used to determine the position of equilibrium of a system stimulated searches for an analogous universal principle with the aid of which it might be possible to determine possible motions of a system. This led to the discovery of the "principle of least action".

Let us first consider a special case. Suppose a particle of mass m is in motion along the x-axis under the action of a force with potential $U(x)$. As we know (see HM, Sec. 6.8), the equation of motion is then of the form

$$m\frac{d^2x}{dt^2} = -U'(x), \quad \text{or} \quad m\frac{d^2x}{dt^2} + U'(x) = 0 \tag{68}$$

It is easy to choose a functional for which the last equation is just the Euler equation (Sec. 12.4). Denoting $\frac{dx}{dt} = \dot{x}$, we rewrite (68) in the form

$$\frac{dU}{dx} + \frac{d}{dt}(m\dot{x}) = 0, \text{ or } \frac{d(-U)}{dx} - \frac{d}{dt}\left[\frac{d}{dx}\left(\frac{m\dot{x}^2}{2}\right)\right] = 0 \tag{69}$$

In form, the last equation resembles the Euler equation, which in the case of the desired function $x(t)$ should have the form (cf. (29))

$$\frac{\partial}{\partial x} F(t, x, \dot{x}) - \frac{d}{dt}\left[\frac{\partial}{\partial \dot{x}} F(t, x, \dot{x})\right] = 0 \tag{70}$$

However, in (70) the same function is differentiated in both members, whereas this is not so in equation (69). So let us have (69) take on the aspect of (70) by adding, under the derivative signs, terms whose derivatives are equal to zero:

$$\frac{\partial}{\partial x}\left[\frac{m\dot{x}^2}{2} - U(x)\right] - \frac{d}{dt}\left\{\frac{\partial}{\partial \dot{x}}\left[\frac{m\dot{x}^2}{2} - U(x)\right]\right\} = 0$$

We see (cf. formulas (25) and (29)) that the desired functional is

$$S = \int_{t_1}^{t_2} \left[\frac{m\dot{x}^2}{2} - U(x)\right] dt$$

Observe that the term $\frac{1}{2} m\dot{x}^2$ is just equal to the kinetic energy T of the moving particle; denote the integrand as

$$L = T - U \tag{71}$$

This is the so-called *Lagrangian function*. Then the variational problem consists in seeking the stationary value of the integral

$$S = \int_{t_1}^{t_2} L\, dt \tag{72}$$

which is called the *action*. Here, t_1 and t_2 are the initial and terminal time of the motion and we compare all conceivable forms of motions with the same initial conditions and the same terminal conditions. It can be verified that in a large number of cases, for instance if the time interval $t_2 - t_1$ is sufficiently small, the integral (72) has a minimal value and not merely a stationary value for actual motion. For this reason, the possibility of finding that motion by proceeding from the variational problem for the integral (72) is called the *principle of least action.*

It turns out that the variational principle of least action is of a universal nature and holds true for any closed systems not involving dissipation of energy, for example, via friction; incidentally, a system with dissipation may in a certain sense be considered open. According to this principle, of all conceivable (under the given constraints) modes of passing from one state at time t_1 to another state at time t_2, the system chooses that mode for which the action, that is, the integral (72), assumes a stationary (minimal, as a rule) value. Here the Lagrangian function L is the difference (71) between the kinetic energy and the potential energy of the system, each of these energies being expressed in terms of the generalized coordinates of the system (see Sec. 4.8) and the time derivatives. The principle of least action is applicable both to systems with a finite number of degrees of freedom and to continuous media, and not only to mechanical but also electromagnetic and other phenomena.

Let us apply the principle of least action to deriving the equation of small transverse oscillations of a taut membrane that was considered at the end of Sec. 12.6. Since the kinetic energy of an element $d\sigma$) of the membrane is equal to

$$\frac{1}{2}\rho\, d\sigma\left(\frac{\partial z}{\partial t}\right)^2$$

where ρ is the surface density of the membrane, the total kinetic energy of the membrane is

$$E=\frac{1}{2}\rho\iint\limits_{(\sigma)}\left(\frac{\partial z}{\partial t}\right)^2 dx\, dy$$

Using the expression (43) for the potential energy, we get the Lagrangian function and the action:

$$L=\frac{1}{2}\iint\limits_{(\sigma)}\left\{\rho\left(\frac{\partial z}{\partial t}\right)^2-T\left[\left(\frac{\partial z}{\partial x}\right)^2+\left(\frac{\partial z}{\partial y}\right)^2\right]\right\}dx\, dy,$$

$$S=\frac{1}{2}\int\limits_{t_1}^{t_2}\iint\limits_{(\sigma)}\left\{\rho\left(\frac{\partial z}{\partial t}\right)^2-T\left[\left(\frac{\partial z}{\partial x}\right)^2+\left(\frac{\partial z}{\partial y}\right)^2\right]\right\}dx\, dy\, dt$$

Applying the Euler equation with respect to the three independent variables (cf. Sec. 12.6), we get the equation of oscillations of a membrane:

$$-\frac{\partial}{\partial t}\left(\rho\frac{\partial z}{\partial t}\right)+\frac{\partial}{\partial x}\left(T\frac{\partial z}{\partial x}\right)+\frac{\partial}{\partial y}\left(T\frac{\partial z}{\partial y}\right)=0$$

or

$$\frac{\partial^2 z}{\partial t^2}=\frac{T}{\rho}\left(\frac{\partial^2 z}{\partial x^2}+\frac{\partial^2 z}{\partial y^2}\right)=a^2\left(\frac{\partial^2 z}{\partial x^2}+\frac{\partial^2 z}{\partial y^2}\right)\quad\left(a=\sqrt{\frac{T}{\rho}}\right)$$

At first glance, expression (71) appears to be formal and unnatural for the Lagrangian function because it is not clear what meaning the difference of energies can be endowed with. The situation is clarified by a simple example. The trajectory of a particle with given energy E moving in a given stationary potential field U is determined by the minimum condition $\int \mathbf{p} \cdot d\mathbf{r}$ along the path, where $\mathbf{p} = m\mathbf{v}$ is the momentum and $\mathbf{r}$ is the radius vector. And so we have

$$T - U = 2T - (T + U) = 2T - E$$

Hence

$$\int_{t_1}^{t_2} L\, dt = \int_{t_1}^{t_2} 2T\, dt - \int_{t_1}^{t_2} E\, dt = \int_{t_1}^{t_2} 2T\, dt - E(t_2 - t_1)$$

where the term $E(t_2 - t_1)$ is constant. Writing $2T = 2\frac{mv^2}{2} = m\mathbf{v} \cdot \mathbf{v}$, $\mathbf{v}\, dt = d\mathbf{r}$, we then find that the extremum condition $\int_{t_1}^{t_2} L\, dt$ coincides with the extremum condition $\int \mathbf{p} \cdot d\mathbf{r}$ for given initial and terminal points of the path and a given energy. Here, $p^2 = 2m\ (E - U)$.

At the same time the stationarity condition $\int \mathbf{p} \cdot d\mathbf{r}$ is fully analogous to the Fermat principle. The point is that in quantum mechanics the wavelength associated with the motion of a body is equal to $\lambda = 2\pi\hbar/p$, where $\hbar$ is Planck's constant, $\hbar \cong 10^{-27}$g-cm^2/s. Here the probability of finding a particle at a given point is proportional to the square of the amplitude of the wave, and waves that traverse different paths combine. It is precisely the condition of stationarity $\int \mathbf{p} \cdot d\mathbf{r}$ along the trajectory computed in classical mechanics that is the condition that waves with the same phase traversing paths close to the trajectory are additive.

The analogy between the mechanical principles of least action and Fermat's principle played a big role in the development of quantum mechanics.

There are many other variational principles in diverse fields, including those far removed from physics, such as economics. In economic problems we usually have certain resources (money, materials, etc.) at our disposal that are to be utilized in order to obtain maximum usefulness. What we have is a maximization problem in which the sought-for element is an optimal plan for utilizing resources and maximizing usefulness. Depending on the nature of the problem, the choice of a plan may involve either a finite number of degrees of freedom, which means we seek a set of parameters defining the plan, or an infinite number of degrees of freedom, so that what we seek is a function. In the former case, the problem is sometimes solved with the tools of differential calculus (Secs. 4.6, 4.7) and sometimes, if the number of desired quantities and the restrictions imposed on them are great, with the tools of a new mathematical discipline called the "theory of mathematical programming". In the latter instance, the problem belongs to the calculus of variations. Similar extremal principles are constantly being used in our activities, although we are not always aware of this fact nor do we always apply them in the most apt fashion.

Exercise

Proceeding from the expression (11) for the potential energy of a homogeneous string on a cushion, derive the equation of motion of the string disregarding the kinetic energy of the cushion.

12.12 Direct methods

For a long time almost the only way to solve variational problems was to pass to the differential equations of Euler. However, the solution of the resulting boundary-value problems for the differential equations very often turned out to be a complicated matter. To obtain an explicit solution in quadratures is a rare event and one has to make use of approximate methods. Rarer still is an explicit solution of the Euler equations for variational problems involving several independent variables.

Of late, frequent use is made of a number of effective approximate methods for the direct solution of variational problems; these methods do not involve passing to differential equations and go by the name of *direct methods of the calculus of variations*. Most of them are based on passing to extremum problems of a function of several variables, which is to say, to problems with a finite number of degrees of freedom. We will now consider two such methods.

The first is based on a process which is the reverse of that described in Sec. 12.1, where we passed from a problem with a finite number of degrees of freedom to the variational problem by means of refining the partition of the interval, as a result of which the sum (3) passed into the integral (7). Using the reverse process, we can pass from the integral to a sum and thus reduce the problem of extremizing a functional to the problem of extremizing a function of several variables.

For example, suppose we are considering the extremum of the functional (25) under the boundary conditions (26). Partitioning the interval of integration into n equal subintervals by means of the division points $x_0 = a,\ x_1,\ x_2, \dots, x_n = b$ and setting $y(x_i) = y_i$, we approximately replace $y'_i = \frac{y_{i+1} - y_i}{h}$ $\left(h = \frac{b-a}{n}\right)$:

$$\int_a^b F(x, y, y')\, dx \approx \sum_{i=0}^{n-1} F(x_i, y_i, y'_i)\, h$$

$$\approx h \sum_{i=0}^{n-1} F\left(x_i, y_i, \frac{y_{i+1} - y_i}{h}\right) \qquad (73)$$

(As in Secs. 1.1 and 2.2, we could make use of the more exact formulas of numerical differentiation and integration, but we will not dwell on that here.) Since the values $y_0 = y_a$ and $y_n = y_b$ are given, the problem reduces to finding the extremum of the function of $n - 1$ variables $y_1, y_2, \dots, y_{n-1}$ in the right member of (73). As was demonstrated in Sec. 12.1, this problem can sometimes be carried to completion with relative ease.

One of the most common direct methods of the calculus of variations is the so-called *Ritz method*[*] which consists in the fact that the desired function is sought in a form that involves several arbitrary constants (parameters). For example, in the case of one independent variable in the form

$$y = \varphi(x;\ C_1, C_2, \dots, C_n) \qquad (74)$$

Here the right member is chosen so that the given boundary conditions are satisfied for any values of the parameters. Substituting (74) into the expression for the given functional, we find that the value of the functional turns out to be dependent on $C_1, C_2, \dots, C_n$. Thus, the problem of extremizing the functional reduces to the problem of extremizing a function of n independent parameters $C_1, C_2, \dots, C_n$, and this problem can be solved by the methods of Ch. 4. True, the solution thus obtained is only an approximate solution of the original problem, since by far not every function can be represented in the form (74). The larger the number of parameters C_i that have been introduced, the more "flexible" is the formula (74), that is, the more exactly we can represent the desired solution by this formula, but the more complicated are the computations. For this reason, in practice, the number of such parameters is restricted to a few and at times it is sufficient to take $n = 1$.

[*] This method was proposed in 1908 by the German physicist and mathematician F. Ritz. In 1915 the Russian mechanician B. G. Galerkin (1871—1945) used a more general method that is applicable to boundary-value problems that are not necessarily of a variational origin. For this reason, the method is sometimes called the *method of Ritz-Galerkin.*

Some examples. Suppose we seek to minimize the functional

$$I = \int_0^1 (y'^2 + y^2)\, dx \tag{75}$$

given the boundary conditions

$$y(0) = 0, \quad y(1) = 1 \tag{76}$$

We seek an approximate solution in the form

$$y = x + Cx(1 - x) \tag{77}$$

(The first of these terms satisfies the conditions (76), the second, the corresponding homogeneous conditions that continue to hold when multiplied by an arbitrary constant, so that the entire sum also satisfies the conditions (76); the same pattern is used for setting up the function (74) in other examples as well.) Substituting (77) into (75) yields, after calculations,

$$\begin{aligned} I &= \int_0^1 [(1 + C - 2Cx)^2 + (x + Cx - Cx^2)^2]\, dx \\ &= \left(1 + C^2 + 4C^2\frac{1}{3} + 2C - 4C\frac{1}{2} - 4C^2\frac{1}{2}\right) \\ &\qquad + \left(\frac{1}{3} + C^2\frac{1}{3} + C^2\frac{1}{5} + 2C\frac{1}{3} - 2C\frac{1}{4} - 2C^2\frac{1}{4}\right) \\ &= \frac{4}{3} + \frac{1}{6}C + \frac{11}{30}C^2 \end{aligned}$$

To find the minimum of this function of C, equate the derivative to zero:

$$\frac{1}{6} + \frac{11}{15}C = 0$$

whence $C = \frac{5}{22} = -0.227$, which means the approximate solution (77) is of the form

$$y_{\text{I}} = x - 0.227x(1 - x) = 0.773x + 0.227x^2$$

A more exact result is obtained if we seek the approximate solution in the form $y = x + C_1x\ (1 - x) + C_2x^2(1 - x)$, which includes two parameters. Substituting into (75) and equating to zero the derivatives with respect to C_1 and C_2 leads (we leave it up to the reader to carry out the computations) to $C_1 = -0.146$, $C_2 = -0.163$, that is, to the approximate solution

$$\begin{aligned} y_{\text{II}} &= x - 0.146x(1 - x) - 0.163x^2(1 - x) \\ &= 0.854x - 0.017x^2 + 0.163x^3 \end{aligned}$$

It is easy to obtain an exact solution here. Indeed, the Euler equation for the functional (75) is of the form

$$2y - \frac{d}{dx}(2y') = 0, \quad y'' - y = 0$$

and has the general solution (Sec. 7.3)

$$y = Ae^x + Be^{-x}$$

where A and B are arbitrary constants. Under the conditions (76) we get

$$y = \frac{e^x - e^{-x}}{e - e^{-1}}$$

This exact solution can be compared with both approximate solutions (see table)

x	y_{I}	y_{II}	y_{exact}	x	y_{I}	y_{II}	y_{exact}
0.0	0.000	0.000	0.000	0.6	0.546	0.541	0.542
0.1	0.080	0.085	0.085	0.7	0.652	0.645	0.645
0.2	0.164	0.171	0.171	0.8	0.764	0.756	0.754
0.3	0.252	0.259	0.259	0.9	0.880	0.874	0.873
0.4	0.346	0.349	0.350	1.0	1.000	1.000	1.000
0.5	0.443	0.443	0.444				

Thus, even in the simplest variant the Ritz method yields a very high accuracy in this case.

The Ritz method produces still higher accuracy if it is required to find not the extremizing function but the extremal value itself of the functional. True enough, for a small variation of the function near the stationary value of the functional leads to a still smaller variation of the value of the functional. (Compare the change in the independent and dependent variables near the stationary value of the function, for instance, the change in x and y near the value $x = c$ in Fig. 180.) For this reason, the error in the value of the functional will be of higher order compared with the error in the extremizing function. Thus, in the foregoing example, when we substitute the function $y_1(x)$, the functional (75) produces an approximate minimal value of 1.314, whereas the exact value is 1.313. The error comes to 0.1%. And to detect the error when substituting $y_{\text{II}}(x)$ would require computations of much higher accuracy.

If we perform similar computations for the functional

$$I = \int_0^1 (y'^2 + xy^2)\, dx \tag{78}$$

(this the reader can do for himself), then for just about the same amount of computation we obtain an approximate solution with high accuracy, although in this case the exact solution cannot be expressed in terms of elementary functions. This is no obstacle to direct methods.

Here is another example of an extremal problem for a function of two variables. Let it be required to find the extremum of the functional

$$I = \int_{-1}^{1}\int_{-1}^{1}\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + 2u\right]dx\,dy \tag{79}$$

from among functions that vanish on the boundary of a square bounded by the straight lines $x = \pm 1$, $y = \pm 1$. Here, the Euler equation cannot be solved exactly by means of elementary functions. We seek an approximate solution by the Ritz method in the form

$$u = C(1 - x^2)(1 - y^2)$$

Substitution into (79) gives

$$I = \int_{-1}^{1}\int_{-1}^{1} \{[-2Cx(1 - y^2)]^2 + [-2Cy(1 - x^2)]^2$$
$$+ 2C(1 - x^2)(1 - y^2)\}\,dx\,dy$$
$$= 4C^2\frac{2}{3}\cdot 2\left(1 - \frac{2}{3} + \frac{1}{5}\right) + 4C^2\frac{2}{3}\cdot 2\left(1 - \frac{2}{3} + \frac{1}{5}\right)$$
$$+ 2C\cdot 2\left(1 - \frac{1}{3}\right)\cdot 2\left(1 - \frac{1}{3}\right) = \frac{256}{45}C^2 + \frac{32}{9}C$$

Equating the derivative to zero produces a minimum at $C = -\frac{5}{16}$. Hence the solutions of the minimum problem of the functional (79) for given boundary conditions is approximately of the form

$$u = -\frac{5}{16}(1 - x^2)(1 - y^2)$$

A comparison with the exact formula, which has the form of an infinite series, shows that the error of this approximate solution is equal, on the average, to 1.5%; the error in the value of the functional I is about 0.2%.

Exercises

1. Find an approximate solution to the problem of the minimum of the functional (75) under the conditions (76), choosing the approximate solution in the form $y = x + C \sin \pi x$. Compare the resulting values of I and y (0.5) with the exact values.

2. Find an approximate solution to the problem of the minimum of the functional (78) under the conditions (76), choosing the approximate solution in the form (77).
3. Find an approximate solution of the problem discussed in the text on the minimum of the functional (79), taking the approximate solution in the form $u = C \cos \frac{1}{2}\pi x \cos \frac{1}{2}\pi y$. Compare the resulting values of I and $u(0, 0)$ with those found in the text.

ANSWERS AND SOLUTIONS

Sec. 12.1

Reasoning as in the text, we find

$$y_i = -\frac{h}{P}[(i-1)F_1 + (i-2)F_2 + \dots + F_{i-1}] + iy_1,$$

$$0 = -\frac{h}{P}[(n-1)F_1 + (n-2)F_2 + \dots + F_{n-1}] + ny_1$$

whence

$$y_i = \frac{h(n-i)}{Pn}[1F_1 + 2F_2 + \dots + (i-1)F_{i-1}] + \frac{hi}{Pn}[(n-i)F_i + (n-i-1)F_{i+1} + \dots + 1F_{n-1}]$$

Putting $n = \frac{1}{h}$, $i = \frac{x_i}{h}$, $F_i = f(x_i)\,h$, $h \to 0$, we get, in the limit,

$$y = \frac{l-x}{Pl}\int_0^x \xi f(\xi)\,d\xi + \frac{x}{Pl}\int_x^l (l-\xi)\,f(\xi)\,d\xi \qquad \text{(cf. Sec. 6.2)}$$

Sec. 12.2

1. $\delta I = -\int_0^1 \frac{2x\delta y}{y^3}\,dx,\ 2y(0)\,\delta y(0) + \int_0^1 (x\delta y + 2y'\delta y')\,dx.$

2. In this example, $\Delta I = \int_0^1 (2x + \alpha x^2)^2\,dx - \int_0^1 (2x)^2\,dx = \alpha + \frac{\alpha^2}{5}$, $\delta I = \alpha$.

Computations:

α	ΔI	δI	relative error
1	1.2	1	−17%
−0.1	−0.098	−0.1	− 2%
0.01	0.01002	0.01	− 0.2%

Sec. 12.3

In both cases, equation (22) is of the form $(1 - x)\,2(y - 2x) = 0$, whence $y = 2x$. This solution gives the functional a stationary value of $I = 0$, which is minimal only in the case (a). In the case (b) the solution has the nature of a minimax (cf. Sec. 4.6); the minimum problem has no solution.

Sec. 12.4

1. (a) The Euler equation has the form $2y - \frac{d}{dx}(2y') = 0$, or $y'' - y = 0$, whence $y = C_1e^x + C_2e^{-x}$ (see Sec. 7.3). From the boundary conditions we get $0 = C_1 + C_2$, $1 = C_1e + C_2e^{-1}$, whence $C_1 = \frac{1}{e - e^{-1}}$, $C_2 = -\frac{1}{e - e^{-1}}$ and, finally, $y = \frac{e^x - e^{-x}}{e - e^{-1}}$; (b) by formula (34), $yy'^2 - 2yy' \cdot y' = C$, $yy'^2 = -C$, $\sqrt{y}\,dy = \pm\sqrt{-C}\,dx$, $y^{3/2} = \pm\frac{3}{2}\sqrt{-C}x + C_2 = C_1x + C_2$. From the boundary conditions $p^{3/2} = C_2$, $q^{3/2} = C_1 + C_2$, whence $y^{3/2} = (q^{3/2} - p^{3/2})\,x + p^{3/2}$ and, finally, $y = [(q^{3/2} - p^{3/2})\,x + p^{3/2}]^{2/3}$.
2. Substitution of the last solution into I yields (after calculations) the value $\frac{4}{9}(q^{3/2} - p^{3/2})^2$, whence $\frac{\partial I}{\partial p} = -\frac{4}{3}p^{1/2}(q^{3/2} - p^{3/2})$. By the formula derived in the text, $\frac{\partial I}{\partial p} = -(2yy')\Big|_{x=0} = -2p\frac{2}{3}[(q^{3/2} - p^{3/2})\,x + p^{3/2}]^{-1/3}(q^{3/2} - p^{3/2})\,\big|_{x=0} = -\frac{4}{3}p^{1/2}$ $(q^{3/2} - p^{3/2})$, which is the same thing.
3. Since the length of the graph $y = f(x)$ $(a \leqslant x \leqslant b)$ is equal to $\int_a^b \sqrt{1 + y'^2}\,dx$, the problem reduces to determining the curve that minimizes the functional under the conditions (26). From (34) we get $\sqrt{1 + y'^2} - \frac{2y'y'}{2\sqrt{1 + y'^2}} = C$, whence $y' = C_1$, $y = C_1x + C_2$. This is the equation of the straight lines.

Sec. 12.5

From the Euler equation it turns out that for $a \neq \pi, 2\pi, 3\pi, \ldots$ the only function that satisfies the boundary conditions and gives the functional a stationary value is $y = 0$. Substituting $y = Cx(a - x)$, we get $I = C^2\frac{a^3}{30}(10 - a^2)$, whence we see that for $a > \sqrt{10}$ the minimum problem has no solution.

Substituting $y = C \sin \frac{\pi x}{a}$, we get $I = \frac{C^2}{a}(\pi^2 - a^2)$, whence it is apparent that for $a > \pi$ the minimum problem does not have a solution. It can be demonstrated that such will be the case for $a = 2\pi, 3\pi, \dots$ as well, and for $a < \pi$ the function $y \equiv 0$ serves as a solution to the minimum problem.

Sec. 12.6

1. As for (27), we arrive at the equation

$\int_a^b (F'_y \delta y + F'_{y'} \delta y' + F'_{y''} \delta y'')\, dx = 0$. Integrating the second term by parts once and the third twice, we get

$$\int_a^b \left(F'_y - \frac{d}{dx} F'_{y'} + \frac{d^2}{dx^2} F'_{y''}\right) \delta y\, dx = 0$$

whence we arrive at the desired equation:

$$F'_y - \frac{d}{dx} F'_{y'} + \frac{d^2}{dx^2} F'_{y''} = 0$$

2. For brevity, set $z'_x = p$, $z'_y = q$; we get the equation

$$F'_z - F''_{px} - F''_{pz} \frac{\partial z}{\partial x} - F''_{pp} \frac{\partial^2 z}{\partial x^2} - 2F''_{pq} \frac{\partial^2 z}{\partial x \partial y} - F''_{qy} - F''_{qz} \frac{\partial z}{\partial y} - F''_{ql} \frac{\partial^2 z}{\partial y^2} = 0$$

Sec. 12.7

(a) $u^* = x^2 - y^2 + z^2 - 2x - \lambda(x + 2y - z)$, whence the stationarity conditions yield

$$2x - 2 - \lambda = 0, \quad -2y - 2\lambda = 0, \quad 2z + \lambda = 0$$

Then, expressing x, y, z in terms of λ and substituting into the constraint equation, we get $\lambda = -2$, whence the coordinates of the point of conditional extremum are $x = 0$, $y = 2$, $z = 1$; (b) $u^* = x^2 - y^2 + z^2 - 2x - \lambda_1(x + y - z) - \lambda_2(x + 2y)$. From this $2x - 2 - \lambda_1 - \lambda_2 = 0$, $-2y - \lambda_1 - 2\lambda_2 = 0$, $2z + \lambda_1 = 0$. Invoking the constraint equations, we get $x = \frac{1}{2}$, $y = \frac{1}{4}$, $z = \frac{3}{4}$.

Sec. 12.8

1. Let OX be the axis of rotation, and $y = y(x) \geqslant 0$ $(a \leqslant x \leqslant b)$ the equation of the contour of the axial cross section; then the volume and the surface of the solid of revolution are expressed by the formulas (see HM, Sec. 4.7) $V = \pi \int_a^b y^2\, dx$,

$S = 2\pi \int_a^b y \sqrt{1+y'^2}\,dx$. The intermediate integral (34) for the function $\pi(y^2 - 2\lambda y\sqrt{1+y'^2})$ is of the form $\pi(y^2 - 2\lambda y\sqrt{1+y'^2}) + \frac{2\pi\lambda y y'}{\sqrt{1+y'^2}} y' = C_1$, or $\pi\left(y^2 - \frac{2\lambda y}{\sqrt{1+y'^2}}\right) = C_1$. Since y can be zero, then $C_1 = 0$ and we get $y - \frac{2\lambda}{\sqrt{1+y'^2}} = 0$, whence, integrating, we find $(x+C)^2 + y^2 = 4\lambda^2$. The desired solid is a sphere.

2. Since the length L of a suspended string (L) with equation $y = y(x)$ is given and the height of the centre of gravity, as follows from Sec. 11.2, is determined by the formula $y_{\text{c.g.}} = \frac{1}{L}\int_{(L)} y\,dL$, the problem reduces to minimizing the integral $B = \int_{(L)} y\,dL = \int_{(L)} y\sqrt{1+y'^2}\,dx$ for a given $L = \int_{(L)} \sqrt{1+y'^2}\,dx$. The Euler equation has to be written for the function $y\sqrt{1+y'^2} - \lambda\sqrt{1+y'^2} = (y-\lambda)\sqrt{1+y'^2}$, which, after the substitution $y - \lambda = \bar{y}$, goes into the function investigated in the example of Sec. 12.5. The solution of the problem is a catenary with equation $y = \frac{1}{2k}\left[e^{k(x+C)} + e^{-k(x+C)}\right] + \lambda$, where k, C, λ are arbitrary constants.

3. In the first example, $\lambda = \frac{1}{2R}$ where R is the radius of a sphere. In the second example, λ is equal to the difference between the ordinate of the extreme lower point of the catenary and the radius of curvature at this point.

Sec. 12.9

1. The function z has a stationary value $z = 0$ at the point $\left(\frac{1}{2}, \frac{1}{2}\right)$. On the legs of the triangle there is one conditional stationary value $z = \frac{1}{4}$ at the point $\left(0, \frac{1}{2}\right)$ and another one $z = -\frac{1}{4}$ at the point $\left(\frac{1}{2}, 0\right)$. At the vertices the values are 0, -2, and 2. Hence, $z_{\max} = 2$ is attained at the point $(2, 0)$, $z_{\min} = -2$ at the point $(0, 2)$.

2. In Exercise 2 of Sec. 12.8 we obtained the equation of the hanging portions of the string. Two cases are possible here. If the string is sufficiently short, $L < L_0$, so that it does not reach the "floor", i.e. the straight line $y = 0$, we get $y = \frac{e^{kx} + e^{-kx}}{2k} - \frac{e^k + e^{-k}}{2k} + 1$.

Here, $L = \frac{e^k - e^{-k}}{k}$ so that for a given L it is possible to find k numerically.

Since $y'_{\min} = \frac{2 + 2k - e^k - e^{-k}}{2k}$, it follows that L_0 is determined from the $k > 0$ for which $2 + 2k = e^k + e^{-k}$. For $L > L_0$, the curve consists of two hanging portions and one lying on the floor. The right hanging portion has the equation $y = \frac{e^{k(x-C)}}{2k} + \frac{e^{-k(x-C)}}{2k} - \frac{1}{k}$, and $\frac{e^{k(1-C)} + e^{-k(1-C)}}{2k} - \frac{1}{k} = 1$, whence we can express C in terms of k. The entire curve is of length $L = 2C + \frac{e^{k(1-C)} - e^{-k(1-C)}}{k}$.

Sec. 12.10

1. The equation of the ellipse is $\sqrt{(x + c)^2 + y^2} + \sqrt{(x - c)^2 + y^2} = L$; after manipulations, we have $4(L^2 - 4c^2)x^2 + 4L^2y^2 = L^2(L^2 - 4c^2)$. The equations of the tangent and the normal at an arbitrary point (x_0, y_0) are, respectively,

$$y - y_0 = -\frac{(L^2 - 4c^2)\,x_0}{L^2 y_0}(x - x_0)$$

$$\text{and } y - y_0 = \frac{L^2 y_0}{(L^2 - 4c^2)\,x_0}(x - x_0)$$

The desired angles are found from the formula $\tan(\beta - \alpha) = \frac{\tan\beta - \tan\alpha}{1 + \tan\alpha\tan\beta}$. For the first angle we have to put $\tan\alpha = \frac{y_0}{x_0 + c}$, $\tan\beta = \frac{L^2 y_0}{(L^2 - 4c^2)\,x_0}$, whence, after some manipulating, we get $\tan(\beta - \alpha) = \frac{4cy_0}{L^2 - 4c^2}$. Such also is the tangent of the second angle, whence follows the last assertion in Exercise 1.

2. If the point N has coordinates $(0, y_N)$, then $L = Y - y_K + \sqrt{x_K^2 + (y_K - y_N)^2} = Y - x_K^2 + \sqrt{x_K^2 + (x_K^2 - y_N)^2}$. From the condition $dL/dx = 0$ we get $y_N = \frac{1}{4}$. The equations of the tangent and the normal at the point K are, respectively, of the form

$$y - x_K^2 = 2x_K(x - x_K) \quad\text{and}\quad y - x_K^2 = -\frac{1}{2x_K}(x - x_K)$$

and the equation of the straight line NK has the form $y - x_K^2 = \frac{x_K^2 - 1/4}{x_K}(x - x_K)$. The equality of the required angles is proved as in Exercise 1.

3. Let the points S and A have the coordinates $(-1, 1)$ and $(1, 1)$. Then $L = \sqrt{(x+1)^2+1} + \sqrt{(x-1)^2+1}$, $x_0 = 0$, $L''(x_0) = \frac{1}{\sqrt{2}}$ cm, whence the reflecting region has diameter equal to $2\sqrt{\frac{5\cdot 10^{-5}}{1/\sqrt{2}}} = 1.7\cdot 10^{-2}$ cm $= 0.17$ mm.

Sec. 12.11

In the given example,

$$L = T - U = \frac{1}{2}\int_0^l \rho\left(\frac{\partial y}{\partial t}\right)^2 dx - \int_0^l \left[\frac{P}{2}\left(\frac{dy}{dx}\right)^2 + \frac{k}{2}y^2 - f(x)\,y\right] dx$$

Writing down the action integral and then applying the Euler equation of Sec. 12.6, we get the desired equation in the form $-ky + f(x) - \frac{\partial}{\partial t}\left(\rho\,\frac{\partial y}{\partial t}\right) + \frac{\partial}{\partial x}\left(P\,\frac{\partial y}{\partial x}\right) = 0$, that is, $\frac{\partial^2 y}{\partial t^2} = \frac{P}{\rho}\frac{\partial^2 y}{\partial x^2} - \frac{k}{\rho}y + \frac{1}{\rho}f(x)$.

Sec. 12.12

1. Substituting $y = x + C\sin\pi x$ into (75), we get $I = \frac{4}{3} + \frac{2}{\pi}C + \frac{\pi^2+1}{2}C^2$. From the condition $\frac{dI}{dC} = 0$ we find $C = \frac{-2}{\pi(\pi^2+1)} = -0.0588$. The values of I and $y(0.5)$ turn out equal to 1.315 (error: 0.2%) and 0.441 (error: 0.7%).

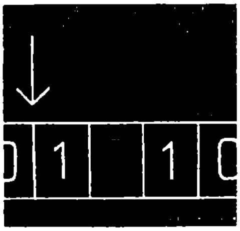
2. Substituting, we get $I = \frac{5}{4} + \frac{C}{10} + \frac{7C^2}{20}$.

   From the minimality condition we have $C = -\frac{1}{7}$, $I = \frac{87}{70} = 1.247$.

3. Substituting, we get $I = \frac{\pi^2 C^2}{2} + \frac{32C}{\pi^2}$. From the minimality condition we have $C = -\frac{32}{\pi^4} = -0.331$. The values of I and $u(0, 0)$ come out equal to -0.537 and -0.331 instead of -0.556 and -0.312.

Chapter 13

THEORY OF PROBABILITY



13.1 Statement of the problem

In nature, engineering and elsewhere we frequently encounter phenomena whose study requires knowledge of a special branch of mathematics called the *theory of probability*.

The most elementary example and the one almost invariably used in this connection is coin tossing. Flip a coin and it will fall heads or tails. If a coin is flipped so that it turns several times in the air, it will, on the average, come up heads the same number of times as it does tails. A coin tossed a very large number of times N will fall approximately $\left(\frac{1}{2}\right)N$ times heads and $\left(\frac{1}{2}\right)N$ times tails. The factors of N are the same, equal to $\frac{1}{2}$, both for the case of heads and for the case of tails. They are called *probabilities* in this case. We say that in a single tossing of a coin the probability of falling heads, w_h, is equal to $\frac{1}{2}$, the probability of falling tails, w_t, is also equal to $\frac{1}{2}$.

Let us take another instance. Consider a die, one face of which is white and the others black. If the die is thrown a large number of times N, we approximately get $(1/6)N$ times white and $(5/6)N$ times black (it is assumed that the die is made of homogeneous material). Here we say that the probability of the white face turning up is $w_{wh} = 1/6$ and the probability of a black face turning up is $w_b = 5/6$. It is clear that if there are two mutually exclusive outcomes in a separate trial (heads or tails in the first example, and a white or black face in the second one), then the sum of their probabilities is 1.

Let us examine the question of radioactive decay (this was considered in some detail in HM, Ch. 5). We intend to observe a separate atom of the radioactive material. The probability that the atom will decay in a small time τ that has elapsed from the start of observations is equal to $w\tau$, where w is a constant, which means the probability is proportional to the time interval τ. And hence the probability that radioactive decay during this time will not occur is equal to $1 - w\tau$. The constant w characterizes the given radioactive substance. It has the dimensions of 1/s and is connected with the mean lifetime T of the given element by the relation $w = \frac{1}{T}$.

Clearly, the probability of decay may be equal to $w\tau$ only for very small intervals of time τ. Indeed, for large τ we can obtain, say, $w\tau > 1$, which is obviously meaningless. And so we have to consider an extremely small time interval dt. Then we will be dealing with a very small probability of decay $w\,dt$ and with a probability extremely close to $(1 - w\,dt)$ that the atom will not experience decay at all. From this, we will get, in Sec. 13.5, the probability a of not decaying in a finite time t, $a = (1 - w\,dt)^{t/dt} = e^{-wt}$ and the probability b of decaying in this time, $b = 1 - a = 1 - e^{-wt}$.

There are cases where several (more than two) outcomes of a single trial are possible. For instance, in throwing an ordinary die (a cube with numbers ranging from 1 to 6 on the faces) we have six possible outcomes: 1, 2, 3, 4, 5, or 6 turning up. The probability of each outcome is equal to 1/6.

Finally, the result of a separate trial may be described by a quantity that assumes a continuous sequence of values. Take, for example, the case of fishing with a rod; we let p describe the weight of each fish caught. We can divide the fish roughly into three classes: small (up to 100 grams), medium (from 100 grams to 1 kilogram), and large (exceeding 1 kg). Then the probable outcome of catching a single fish will be described by three numbers: the probability of catching a small fish w_{sm}, a medium fish w_{med}, and a big fish, w_{big}. Then

$$w_{\text{sm}} + w_{\text{med}} + w_{\text{big}} = 1$$

Of course this is a very crude description of fishing. For example, "catching a big fish" may mean one weighing 1.1 kg or 20 kg.

The probability of catching a fish weighing between p and $p + dp$, where dp is a very small increment in weight, will be denoted by dw. This probability is naturally proportional to dp. Besides, dw depends on p as well.

Indeed, there are no grounds to assume that the probability of catching a fish weighing between 100 and 110 grams is the same as that of catching a fish weighing between 1000 and 1010 grams. We therefore put

$$dw = F(p)\,dp$$

The function $F(p)$ here is called the *probability distribution function*.

We know that the sum of all the probabilities of catching a fish of one kind or another is equal to 1, which yields

$$\int_0^\infty F(p)\,dp = 1$$

We can modify the problem and instead of regarding a catch as a separate trial we can take casting the line. The result is either a catch or (more often unfortunately) a miss. We can introduce the probability of pulling in an empty line k and the probability of catching a fish (of any weight) equal to $1 - k$, and then subdivide the cases of catching one or another type of fish in accordance with the function $F(p)$. Alternatively, we might regard an empty line as a "catch of weight zero", thus relieving us of saying no fish were caught. Then we have a new probability distribution function $f(p)$. This will be a function for which zero is a singular point. The integral $\int_0^\infty f(p)\,dp = 1$, but the point $p = 0$ will make a finite contribution to this integral. This means that $f(p)$ contains a term of the form $k\delta(p)$ (see the delta function discussed in Ch. 6). The new function $f(p)$, which refers to a single throw of the line, is connected with the old function $F(p)$ (which referred to a single real — with $p > 0$ — caught fish) by the relation

$$f(p) = k\delta(p) + (1 - k)\,F(p)$$

The foregoing examples illustrate the material dealt with by the theory of probability. We now consider certain questions that arise in this connection.

For example, let a coin be tossed five times. What is the probability that it will fall heads in all five cases? What is the probability that it will fall heads in four cases and tails in one, the order in which they appear being immaterial? Here we assume that the probability of obtaining heads in a single toss is equal to $\frac{1}{2}$. The solution of problems of this type requires a certain amount of manipulative skill, and for very large numbers of tosses one has to resort to methods of higher mathematics, for otherwise the computations are practically impossible to carry out. The next range of questions has to do with radioactive decay. Starting with the probability of decay in time dt, it is required to find the probability of decay in time t, which may be of any duration not necessarily small. For this we will need the apparatus of higher mathematics. (In particular, we will obtain a number of the results given in HM,. Ch. 5, although the reasoning will be somewhat different.)

In Ch. 5 of HM we considered a large number of atoms N and obtained a law of the variation of N with time. Actually, we considered only mean values there. The r.ew approach will enable us to solve much more subtle problems such as: What is the probability of observing, in a certain instrument, a number of disintegrations that differs from the mean?

In the fishing problem, we can pose the question of the probability of obtaining a certain overall catch as the result of catching 2, 3, ..., n fish or as the result of casting the line 2, 3, ..., n times. The latter problem calls for rather involved mathematical techniques.

In deriving the laws of probability theory we will need the Stirling formula (Sec. 3.3), and also the formula $\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$ that was derived in Sec. 4.7.

Exercise

In the card game of "dunce"' we use a pack of 36 cards of four suits. What is the probability of the first card dealt being a spade? a queen? a queen of spades? the trump card (the trump card is the suit of the first card obtained after dealing)?

13.2 Multiplication of probabilities

The basis for solving the problems posed in Sec. 13.1 is the *law of compound probabilities of independent events.*

This can be illustrated in the following simple example. Suppose a die has one white face and five black faces so that the probability of a white face appearing is $w_{\text{wh}} = 1/6$, the probability of a black face turning up is $w_{\text{b}} = 5/6$. Suppose we have another die with two faces green and four red. For this die, the probabilities of a green face and a red face turning up are, respectively, equal to

$$w_{\text{g}} = \frac{2}{6} = \frac{1}{3}, \quad w_{\text{r}} = \frac{4}{6} = \frac{2}{3}$$

Put both dice in a glass, juggle them and then throw them on the table. Four distinct outcomes are possible: one white, the other green, one white, the other red, one black, the other green, and, finally, one black and the other red. These outcomes will be denoted as: wg, wr, bg, br. The number of cases in which these outcomes are realized are denoted by N_{wg}, N_{wr}, N_{bg}, N_{br}, and the corresponding probabilities, by w_{wg}, w_{wr}, w_{bg}, w_{br}. We now pose the problem of determining these probabilities.

Suppose we have carried out a very large number of trials N. We divide them into two groups: those in which the first die exhibits a white face (w) (irrespective of the colour of the second die) and those in which the first die exhibits a black face (b) (irrespective of the colour of the face that turns up on the second die). Thus, $N = N_w + N_b$. Since $w_w = \frac{N_w}{N}$, it follows that $N_w = w_w \cdot N = \frac{1}{6} N$. Similarly, $N_b = w_b \cdot N = \frac{5}{6} N$.

On the other hand, it is clear that $N_w = N_{wg} + N_{wr}$. We will now need the notion of the independence of events. We assume that the fact that the first die turned up white has no effect on the probability of green or red turning up on the second die. In other words we consider the two events — a definite colour turning up on the first die and a definite colour turning up on the second die — to be *independent.* The outcome of one of these events cannot in any way affect the outcome of the other. For this reason, the probability of green turning up on the second die is $w_g = \frac{N_{wg}}{N_w}$, while the probability of red turning up on the second die is $w_r = \frac{N_{wr}}{N_w}$. From this we find $N_{wg} = w_g \cdot N_w = w_g \cdot w_w \cdot N$. On the other hand, $w_{wg} = \frac{N_{wg}}{N}$ and so

$$w_{wg} = w_w \cdot w_g$$

Thus, the probability of a compound event (white turning up on one die and green on the other one) is equal to the product of the probabilities of simple independent events.

It is now clear that

$$w_{wr} = w_w \cdot w_r, \quad w_{bg} = w_b \cdot w_g, \quad w_{br} = w_b \cdot w_r$$

We have already said that only the following four outcomes are possible in a simultaneous throw of the two dice we have been discussing: wg, wr, bg, br.

And so we must have

$$w_{wg} + w_{wr} + w_{bg} + w_{br} = 1$$

It is easy to see that this is so, for

$$\begin{aligned} w_{wg} + w_{wr} + w_{bg} + w_{br} &= w_w \cdot w_g + w_w \cdot w_r + w_b \cdot w_g + w_b \cdot w_r \\ &= w_w \cdot (w_g + w_r) + w_b \cdot (w_g + w_r) \\ &= (w_g + w_r)\,(w_w + w_b) \end{aligned}$$

Each of the last two parentheses contains the sum of the probabilities of two simple events, each of which precludes the other and some one of which must definitely occur. (If white turns up, then black cannot turn up, and vice versa. And either black or white will turn up definitely.) It is clear that such a sum of probabilities is equal to unity:

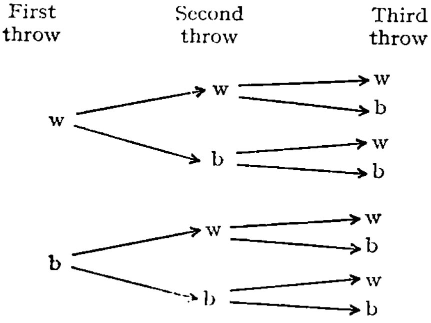
$$w_w + w_b = 1, \quad w_g + w_r = 1$$

whence it follows that

$$w_{wg} + w_{wr} + w_{bg} + w_{br} = 1$$

Now let us take up the next example. Suppose we have a die with a certain number (precisely what number is immaterial) of white faces and a certain number of black faces. We denote the probability of white and black occuring in one throw of the die by α and β respectively. Suppose an experiment consists in three throws of the die. Since there are two possible outcomes in each throw, there will be $2^3 = 8$ distinct possible outcomes in the experiment.*

We enumerate them in the array given below.



Outcome of three throws	Probability of outcome
1w, 2w, 3w	α^3
1w, 2w, 3b	$\alpha^2\beta$
1w, 2b, 3w	$\alpha^2\beta$
1w, 2b, 3b	$\alpha\beta^2$
1b, 2w, 3w	$\alpha^2\beta$
1b, 2w, 3b	$\alpha\beta^2$
1b, 2b, 3w	$\alpha\beta^2$
1b, 2b, 3b	β^3

The arrows indicate all eight possible outcomes. In the second to the last column, the numeral indicates the number of the throw and the letter the outcome of that throw. For example, 1b, 2w, 3b means black in the first throw, white in the second, and black again in the third. The probabilities of each of these eight outcomes are given in the last column. These probabilities are readily computed by the law of compound probabilities.

In the foregoing table we distinguish, say, between the cases 1w, 2w, 3b and 1b, 2w, 3w; in both cases there was one black face and two whites, only in the former case the black face turned up in the third throw and in the latter case it turned up in the first throw. Ordinarily, in specific problems of this type we are only interested in the total number of occurrences of a white face and the total number of occurrences of black, the order of the occurrences being immaterial.

* If the experiment consists in throwing a die n times, then there are 2^n possible distinct outcomes.

From this standpoint, the eight cases considered above fall into four groups:

$$\mathrm{w} = 3,\ \mathrm{b} = 0,\ \mathrm{w} = 2,\ \mathrm{b} = 1,\ \mathrm{w} = 1,\ \mathrm{b} = 2,\ \mathrm{w} = 0,\ \mathrm{b} = 3^{*}$$

This notation is clear enough: w = 2, b = 1 stands for a group consisting of all cases in which white turned up twice and black once. (The order of occurrence is immaterial.) Now let us set up a table indicating all the groups, all cases into which each group falls, the probabilities of each case, and the probabilities of each group.

To compute the probability of the group, note the following.

If some group of events combines several disjoint cases, then the probability of the group, that is, the probability of some one case occurring, is equal to the sum of the probabilities of the independent cases that constitute the group. An example will illustrate this. Suppose we toss a die with white, black, and red faces, the probabilities of occurrence of the faces of these colours being equal, respectively, to

$$w_{\mathrm{w}},\ w_{\mathrm{b}},\ w_{\mathrm{r}}\ (w_{\mathrm{w}} + w_{\mathrm{b}} + w_{\mathrm{r}} = 1)$$

Consider a group of events consisting in the fact that either a white or a black face turns up. We denote the probability of this group by $w_{\mathrm{w+b}}$. Let us find $w_{\mathrm{w+b}}$.

Since $w_{\mathrm{w}} = \frac{N_{\mathrm{w}}}{N}$, $w_{\mathrm{b}} = \frac{N_{\mathrm{b}}}{N}$, then it follows that $N_{\mathrm{w}} = w_{\mathrm{w}} \cdot N$, $N_{\mathrm{b}} = w_{\mathrm{b}} \cdot N$, where N is the total number of throws of the die. It is clear that

$$w_{\mathrm{w+b}} = \frac{N_{\mathrm{w}} + N_{\mathrm{b}}}{N} = \frac{w_{\mathrm{w}} \cdot N + w_{\mathrm{b}} \cdot N}{N} = w_{\mathrm{w}} + w_{\mathrm{b}}$$

And so $w_{\mathrm{w+b}} = w_{\mathrm{w}} + w_{\mathrm{b}}$.

We now have the following table:

Group	Case	Probability of case	Probability of group
w = 3, b = 0	1w, 2w, 3w	α^3	α^3
	1w, 2w, 3b	$\alpha^2\beta$	
w = 2, b = 1	1w, 2b, 3w	$\alpha^2\beta$	$3\alpha^2\beta$
	1b, 2w, 3w	$\alpha^2\beta$	
	1w, 2b, 3b	$\alpha\beta^2$	
w = 1, b = 2	1b, 2w, 3b	$\alpha\beta^2$	$3\alpha\beta^2$
	1b, 2b, 3w	$\alpha\beta^2$	
w = 0, b = 3	1b, 2b, 3b	β^3	β^3

[*] If the trial consists in tossing a die n times, then the 2^n cases fall into $n + 1$ groups:

$$\mathrm{w} = n,\ \mathrm{b} = 0;\ \mathrm{w} = n - 1,\ \mathrm{b} = 1;\ \mathrm{w} = n - 2,\ \mathrm{b} = 2;\ \ldots;\ \mathrm{w} = 0,\ \mathrm{b} = n$$

It is easy to see that in n throws of a die, the probability of white occurring m times and black k times ($m + k = n$) for the given law of alternation of white and black faces is equal to $\alpha^m\beta^k$. The group w $= m$, b $= k$ includes all cases in which white occurred m times and black occurred k times with different order of alternation of white and black.

The probability of the group w $= m$, b $= k$ is equal to the term in the binomial expansion

$$(\alpha + \beta)^n = \alpha^n + n\alpha^{n-1}\beta + \frac{n!}{(n-2)!\,2!}\,\alpha^{n-2}\beta^2 + \dots + \beta^n$$

that contains the factor $\alpha^m\beta^k$, which means it is equal to $\frac{n!}{m!\,k!}\,\alpha^m\beta^k$.

Indeed, what is the coefficient of $\alpha^m\beta^k$ in the binomial expansion of $(\alpha + \beta)^n$? It is the number of ways of taking m factors α and k factors β by multiplying out the product

$$\underbrace{(\alpha + \beta)\ (\alpha + \beta)\ (\alpha + \beta) \dots (\alpha + \beta)}_{n \text{ times}}$$

In exactly the same way, the number of cases in a group is the number of ways of obtaining m white faces and k black faces by alternating white and black faces in different ways. Consequently, the coefficient of $\alpha^m\beta^k$ in the expansion of $(\alpha + \beta)^n$ by the binomial theorem is equal to the number of cases in the group. The probability of each definite case w $= m$, b $= k$ is equal to $\alpha^m\beta^k$ by the law of compound probabilities. Therefore, the probability of a group is equal to the corresponding term in the binomial expansion.

It is clear that the sum of the probabilities of all groups must be equal to 1, since all groups embrace all possible cases. Let us verify this.

Since $\alpha + \beta = 1$, it follows that $(\alpha + \beta)^n = 1$. On the other hand,

$$\begin{aligned}(\alpha + \beta)^n &= \alpha^n + n\alpha^{n-1}\beta + \frac{n!}{(n-2)!\,2!}\,\alpha^{n-1}\beta^2 + \dots + \beta^n \\ &= w(\mathrm{w} = n,\ \mathrm{b} = 0) + w(\mathrm{w} = n - 1,\ \mathrm{b} = 1) \\ &\qquad + w(\mathrm{w} = n - 2,\ \mathrm{b} = 2) + \dots + w(\mathrm{w} = 0,\ \mathrm{b} = n)\end{aligned}$$ [*]

Thus

$$w(\mathrm{w} = n,\ \mathrm{b} = 0) + w(\mathrm{w} = n - 1,\ \mathrm{b} = 1) + \dots + w(\mathrm{w} = 0,\ \mathrm{b} = n) = 1$$

[*] The notation $w(\mathrm{w} = m,\ \mathrm{b} = k)$ stands for the probability of the group in which white turns up m times, and black k times.

Exercises

1. What is the probability that in two tosses of a coin heads will come up twice?
2. A coin is tossed three times. What is the probability that it will come up heads in all three cases? twice heads and once tails?
3. Four faces of a die are black and two white. What is the probability that in two throws white will occur twice and black twice? white once and black once?
4. The die in Exercise 3 is thrown three times. What is the probability that white will turn up twice and black once? black twice and white once?
5. A gun is fired at a target. In each shot, the probability of a hit is 0.1, of a miss, 0.9. Two shots are fired. What is the probability that one is a hit and the other a miss?
6. Under the conditions of Exercise 5 three shots are fired. What is the probability that one is a hit and two are misses? two are hits and one is a miss?
7. The situation is as in Exercise 5. What is the probability of hitting the target once if four shots are fired? if five shots are fired?

13.3 Analysing the results of many trials

In the preceding we obtained general formulas for the case of n identical trials in each of which two distinct outcomes are possible with probabilities α and β, respectively.

For small values of n these formulas are sufficiently simple and pictorial.

But if n is great, the formulas cease to be pictorial. Extra work has to be done to cast the formulas in a convenient and clear-cut form. Only after such a strenuous job is it possible to extract the precious grains from the rock* and polish them into brilliant results that will call forth further generalizations.

Let us first consider the simplest case of $\alpha = \beta = \frac{1}{2}$, corresponding to, say, flipping a coin or throwing a die with three white and three black faces.

Suppose we take $n = 100$. It is clear that the probability of obtaining 100 heads (or 100 tails) in 100 tosses of a coin is extremely small. By the results of Sec. 13.2, it is equal to $\left(\frac{1}{2}\right)^{100} = \frac{1}{2^{100}} \approx \frac{1}{10^{30}}$. If a machine performs 100 tosses of the coin per second, it will require on the average of 10^{28} seconds $\approx 3 \cdot 10^{20}$ years to obtain a single case of

* As the poet Mayakovsky might have put it: "Science is like extracting radium, with one gram of the precious mineral to show for a year of arduous labour — a thousand tons of verbal ore to get a single unified formula".

a run of 100 heads. It is quite clear that in the majority of cases 100 tosses will yield roughly 50 heads and 50 tails. But it is not at all obvious what the probability will be of *exactly* 50 heads and 50 tails in a series of 100 tosses. Neither is it clear what the probability will be of an outcome that differs from the average. For example, what is the probability of 55 heads turning up and 45 tails, or 60 heads and 40 tails?

Let us try to answer these questions.

To start with, note that the probability of every specific occurrence of 50 tails and 50 heads, say, in alternation is equal to $\frac{1}{2^{100}} \approx \frac{1}{10^{30}}$, as was established in Sec. 13.2. Which means that this probability is equal to the probability of heads turning up one hundred times in a row. The considerably greater probability of the occurrence of 50 heads and 50 tails (in any order) is due to the fact that this event (this group), as was pointed out in Sec. 13.2, consists of a very large number of different cases of different alternations of heads and tails.

In Sec. 13.2 we found out that this probability is equal to

$$\frac{n!}{m!\,k!}\,\alpha^m\beta^k$$

where α is the probability of heads, β is the probability of tails, n is the number of tosses, m is the number of occurrences of heads, and k is the number of occurrences of tails ($m + k = n$). We will consider an even number n (in our case, $n = 100$). We are interested in the occurrence of $\frac{n}{2}$ tails and $\frac{n}{2}$ heads. Therefore

$$\frac{n!}{m!\,k!} = \frac{n!}{\left[\left(\frac{n}{2}\right)!\right]^2}$$

and the probability of the event that interests us is

$$w\left(\text{heads} = \frac{n}{2},\ \text{tails} = \frac{n}{2}\right) = \frac{n!}{\left[\left(\frac{n}{2}\right)!\right]^2}\left(\frac{1}{2}\right)^n$$

For large n this expression is terribly unwieldy. Let us simplify it by using the approximate formula of Stirling for the expression $n!$ (see Sec. 3.3).

By Stirling's formula we have

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n, \quad \left(\frac{n}{2}\right)! = \sqrt{\pi n}\left(\frac{n}{2e}\right)^{\frac{n}{2}}$$

and so we get

$$w\left(\text{heads} = \frac{n}{2},\ \text{tails} = \frac{n}{2}\right) = \sqrt{\frac{2}{\pi n}}$$

For $n = 100$ we then have $w = \sqrt{\frac{2}{100\pi}} \approx 0.08$. The probability of 50 heads and 50 tails occurring in a definite order is roughly equal to 10^{-30}. And so the number of distinct cases in which the same result is realized in a different order is equal to

$$\frac{0.08}{10^{-30}} = 8 \cdot 10^{28}$$

Now let us try to determine the probability of the outcome of a trial that differs but slightly from the most probable outcome (in the example at hand, the most probable outcome is 50 times heads and 50 times tails). Denote by δ the deviation of an outcome from the most probable outcome, $\delta = m - \frac{n}{2}$. For example, $\delta = 5$ corresponds to 55 heads and 45 tails, $\delta = -5$ corresponds to 45 heads and 55 tails, $\delta = 3$ to 53 heads and 47 tails, etc. Denote by $w(\delta)$ the probability of the appropriate outcome of a trial. The most probable result is $\delta = 0$ so that

$$w(0) = \sqrt{\frac{2}{\pi n}} \tag{1}$$

We now compute the probability $w(\delta)$. This is the probability of occurrence of $m = \frac{n}{2} + \delta$ heads and $k = \frac{n}{2} - \delta$ tails, and so it is equal to

$$w(\delta) = \frac{n!}{\left(\frac{n}{2} + \delta\right)!\left(\frac{n}{2} - \delta\right)!}\,\alpha^{\frac{n}{2}+\delta}\beta^{\frac{n}{2}-\delta}\text{, where } \alpha = \beta = \frac{1}{2}$$

Consequently

$$w(\delta) = \frac{n!}{\left(\frac{n}{2} + \delta\right)!\left(\frac{n}{2} - \delta\right)!}\left(\frac{1}{2}\right)^n$$

Now increase δ by unity and compute $w(\delta + 1)$ to get

$$w(\delta + 1) = \frac{n!}{\left(\frac{n}{2} + \delta + 1\right)!\left(\frac{n}{2} - \delta - 1\right)!}\left(\frac{1}{2}\right)^n$$

Therefore

$$\frac{w(\delta + 1)}{w(\delta)} = \frac{\left(\frac{n}{2} + \delta\right)!\left(\frac{n}{2} - \delta\right)!}{\left(\frac{n}{2} + \delta + 1\right)!\left(\frac{n}{2} - \delta - 1\right)!}$$

Note that

$$\left(\frac{n}{2}+\delta+1\right)! = \left(\frac{n}{2}+\delta+1\right)\left(\frac{n}{2}+\delta\right)!$$

$$\left(\frac{n}{2}-\delta\right)! = \left(\frac{n}{2}-\delta\right)\left(\frac{n}{2}-\delta-1\right)!$$

so that

$$\frac{w(\delta+1)}{w(\delta)} = \frac{\frac{n}{2}-\delta}{\frac{n}{2}+\delta+1} = \frac{1-\frac{2\delta}{n}}{1+\frac{2(\delta+1)}{n}}$$

Taking logs of the right and left sides, we get

$$\ln w(\delta+1) - \ln w(\delta) = \ln\left(1-\frac{2\delta}{n}\right) - \ln\left(1+\frac{2(\delta+1)}{n}\right) \qquad (2)$$

We consider a large number of trials n and assume, besides, that the quantity $\delta \ll \frac{n}{2}$, which means we are studying small deviations from the most probable result. Therefore the quantities $\frac{2\delta}{n}$ and $\frac{2(\delta+1)}{n}$ are small and, consequently, the logarithms on the right of (2) can be expanded in series. In this expansion, we confine ourselves to the first terms of the series:

$$\ln\left(1-\frac{2\delta}{n}\right) = -\frac{2\delta}{n}, \quad \ln\left(1+\frac{2(\delta+1)}{n}\right) = \frac{2(\delta+1)}{n}$$

Formula (2) assumes the form

$$\ln w(\delta+1) - \ln w(\delta) = -\frac{4\left(\delta+\frac{1}{2}\right)}{n}$$

Note that $\ln w(\delta+1) - \ln w(\delta)$ may be replaced approximately by the value of the derivative of the function $\ln w(z)$ computed at the midpoint of the interval, i.e. for a value of the argument equal to $\delta+\frac{1}{2}$ so that

$$\ln w(\delta+1) - \ln w(\delta) = \frac{d}{d\delta}\left[\ln w\left(\delta+\frac{1}{2}\right)\right]$$

Of course, $w\left(\delta+\frac{1}{2}\right)$ is not the probability that the deviation of the number of heads from the most probable number will be equal to $\delta+\frac{1}{2}$, for this deviation is of necessity integral. This is the result of interpolating (Sec. 2.1) the function $w(\delta)$ with integral values of the argument to half-integral values. And so we have

$$\frac{d}{d\delta}\left[\ln w\left(\delta+\frac{1}{2}\right)\right]=\frac{4\left(\delta+\frac{1}{2}\right)}{n}$$

Here, put $\delta+\frac{1}{2}=z$; then $d\delta=dz$ and we get

$$\frac{d}{dz}\ln w(z)=-\frac{4z}{n}$$

Integrate this relation from $z=0$ to $z=\delta$ to get $\ln w(\delta)-\ln w(0)=$ $=-\frac{2\delta^2}{n}$ or, taking antilogarithms,

$$w(\delta)=w(0)\,e^{-\frac{2\delta^2}{n}}$$

Using (1), we finally get

$$w(\delta)=\sqrt{\frac{2}{\pi n}}\,e^{-\frac{2\delta^2}{n}}\text{ *} \tag{3}$$

It is easy to see that this result agrees with the requirement

$$\int_{-\infty}^{+\infty} w(\delta)\,d\delta=1 \tag{4}$$

Knowing the type of dependence $w\propto e^{-2\delta^2/n}$, we could have determined $w(0)$ from condition (4) without resorting to Stirling's formula. Actually, δ is an integer and varies only over the range $-\frac{n}{2}\leqslant\delta\leqslant\frac{n}{2}$ but for large n $w(\delta)$ varies so slowly and at the end-points is so small that no noticeable mistake is introduced by replacing the sum by an integral and integrating from $-\infty$ to $+\infty$.

The graph of w versus δ is precisely of the same type that was considered in Sec. 6.1 in connection with the definition of a delta function. When $n=2$ we get the graph $y_{(2)}(x)$ shown in Fig. 68, and for other n the graph $y_{(2)}(x)$ has to be compressed $\sqrt{\frac{n}{2}}$ times towards the x-axis and stretched $\sqrt{\frac{n}{2}}$ times away from the y-axis. The curve corresponding to formula (3) is called the *probability curve*. It is a bell-shaped curve. From (3) it is evident that $w(-\delta)=$ $=w(\delta)$, which is what was to be expected: the probability of 55

* By similar arguments, denoting $f(n)=\ln(n!)$ and proceeding from the equation $f(n+1)-f(n)=\ln(n+1)$, we can derive the Stirling formula up to a constant factor. (Do this!)

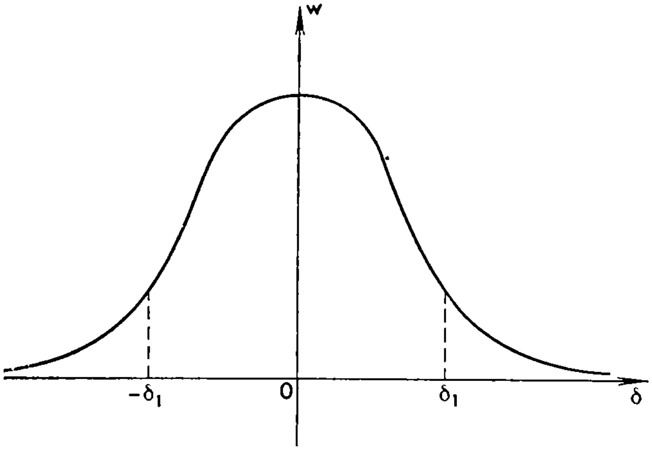


Fig. 187

heads and 45 tails ($\delta = 5$) is equal to the probability of 45 heads and 55 tails occurring.

Using formula (3), we find for $n = 100$ (when, as we know, $w(0) = 0.08$):

$$\begin{aligned}
w(1) &= w(0) \cdot e^{-0.02} = 0.98\, w(0),\\
w(2) &= w(0) \cdot e^{-0.08} = 0.92\, w(0),\\
w(5) &= w(0) \cdot e^{-0.5} = 0.61\, w(0),\\
w(10) &= w(0) \cdot e^{-2} = 0.14\, w(0),\\
w(20) &= w(0) \cdot e^{-8} = 0.00034\, w(0)
\end{aligned}$$

Let us agree to use the term "expected" for those outcomes of trials for which $\frac{1}{e} w(0) \leqslant w(\delta) \leqslant w(0)$. Cases with smaller probability will occur rarely. Thus, expected events are those for which $-\delta_1 \leqslant \delta \leqslant \delta_1$, where δ_1 is found from the condition $w(\delta_1) = \frac{1}{e} w(0)$ (Fig. 187). This yields $w(0)\, e^{-\frac{2\delta^2}{n}} = \frac{1}{e} w(0)$ or $\frac{2\delta_1^2}{n} = 1$, whence we finally get $\delta_1 = \frac{1}{2} \sqrt{2n}$. Hence, expected outcomes are those in which

$$-\frac{1}{2}\sqrt{2n} \leqslant \delta \leqslant \frac{1}{2}\sqrt{2n}$$

In our example, the most probable outcome is that corresponding to $\delta = 0$ (50 heads and 50 tails). But its probability is slight, being equal to 0.08 and is only just a bit greater than the probability of close-lying outcomes. For instance, the probability of obtaining

51 heads and 49 tails (or 49 heads and 51 tails) is almost the same: $0.98 \cdot 0.08 = 0.078$.

The probability of obtaining 57 heads and 43 tails (or 43 heads and 57 tails) is much less. This probability is equal to $\frac{0.08}{e} = 0.029$. We can therefore consider as expected the probability of obtaining a number of heads ranging from 43 to 57, that is, $50 \pm \delta$ where $0 \leqslant \delta \leqslant \delta_1 = 7$.

The quantity δ_1 is proportional to $\sqrt{n}$, and so the larger n is, the broader the range of the expected outcome. For instance, when $n = 10\,000$ we get $\delta_1 = \frac{\sqrt{20\,000}}{2} \approx 70$, so we should expect a result with heads turning up from 4930 times to 5070 times. However, the portion of δ_1 relative to the number of trials n diminishes with increasing n, since the quantity $\frac{\delta_1}{n}$ is proportional to $\frac{1}{\sqrt{n}}$, i.e. it is the smaller, the greater n is.

Suppose we toss a coin to find experimentally the probability of heads turning up. And suppose we do not know whether the coin is unbiased or not (bent so that one side turns up more often than the other). Suppose that the coin is even (unbiased) and the probability of heads turning up is $w_h = 0.5$.

In 100 throws we most likely will get from 43 to 57 heads, or $0.43 \leqslant w_h \leqslant 0.57$. The error in determining the probability will not exceed 0.07 either way. This means that if we get 44 tails and 56 heads in 100 tosses, one should not conclude that there is a greater probability of heads. The experiment is not sufficiently exact; the deviation from 50 lies within the limits of statistical error, as we say. All we can say is that the probability of heads lies within the limits $w = 0.56 \pm 0.07$, which is to say between 0.49 and 0.63: $0.49 \leqslant w \leqslant 0.63$; $w = 0.50$ is not in the least excluded. To make this more precise, we have to increase the number of tosses. In the case of 10 000 tosses, we will most likely have 4930 to 5070 heads, which is $0.4930 \leqslant w_h \leqslant 0.5070$. Here the error in determining the probability will not exceed 0.007 either way. It is clear that the error $\frac{\delta_1}{n}$ in determining the probability is proportional to $\frac{1}{\sqrt{n}}$. Which means that to reduce the error 10 times in determining the probability we have to increase the number of trials 100 times.

Formula (3) is obtained on the assumption that δ is small. It is to be expected therefore that for large δ this formula will yield considerable errors. To illustrate, for the case $n = 100$ let us compute the probability $w(50)$, which is the probability that heads turn up all 100 times and tails not once. Using (3) we find

$$w(50) = w(0)\, e^{-\frac{2 \cdot 50^2}{100}} = 0.08 \cdot e^{-50} \approx 10^{-23}$$

On the other hand, we have already calculated this probability by the exact formula and found that it is equal to 10^{-30}. Thus, the approximate formula (3) does indeed produce substantial errors for large δ. But the errors are not important since for large δ these values of the probabilities are practically negligible.

What is the probability w_{exp} of obtaining an expected result, that is, a result for which δ lies in the range from $-\delta_1$ to $+\delta_1$? This probability is obviously equal to

$$w_{\text{exp}} = \int_{-\delta_1}^{+\delta_1} w(\delta)\, d\delta$$

or, using (3),

$$w_{\text{exp}} = \sqrt{\frac{2}{\pi n}} \int_{-\frac{1}{2}\sqrt{2n}}^{+\frac{1}{2}\sqrt{2n}} e^{-\frac{2\delta^2}{n}}\, d\delta$$

In this last integral we make the change of variable

$$\delta = \frac{t\sqrt{n}}{2} \left(t = \frac{2\delta}{\sqrt{n}}\right)$$

then $d\delta = \frac{\sqrt{n}}{2} dt$. The integral will lie within the range from $-t_1$ to $+t_1$, where

$$t_1 = \frac{2\delta_1}{\sqrt{n}} = 2\frac{\sqrt{2n}}{2\sqrt{n}} = \sqrt{2}$$

Therefore

$$w_{\text{exp}} = \sqrt{\frac{2}{\pi n}}\, \frac{\sqrt{n}}{2} \int_{-\sqrt{2}}^{+\sqrt{2}} e^{-\frac{t^2}{2}} dt = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{2}} e^{-\frac{t^2}{2}} dt$$

Tables have been compiled for the function $\Phi(x) = \frac{2}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$, which is called the *probability integral*. (See the table at the end of this chapter.) Using the table, we find

$$w_{\text{exp}} = \Phi(\sqrt{2}) = \Phi(1.414) = 0.842$$

Hence those cases we called expected cases* constitute 84% of all possible outcomes of trials, the rest making up only 16%.

* It is clear that the expression "expected" is used here in a conventional sense since the probability of obtaining a result that is not expected (we could hardly say "unexpected") is not so small, 16%, under the accepted definition.

In some problems one wants the probability of a result for which the quantity δ does not exceed a given δ_0. This probability is $w = \sqrt{\frac{2}{n\pi}} \int\limits_{-\infty}^{\delta_0} e^{-\frac{2\delta^2}{n}} d\delta$.

Making the same change of variable as in the preceding case, we get

$$w = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{t_0} e^{-\frac{t^2}{2}} dt = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{t_0} e^{-\frac{t^2}{2}} dt, \text{ where } t_0 = \frac{2\delta_0}{\sqrt{n}}$$

so that

$$w(\delta < \delta_0) = \frac{1}{2} + \frac{1}{2}\Phi(t_0) = \frac{1}{2} + \frac{1}{2}\Phi\left(\frac{2\delta_0}{\sqrt{n}}\right)$$

where Φ is taken from the table.

Now let us consider the case $\alpha \neq \beta$. If n throws of a die are made (with the probability of a white face turning up in a single throw being equal to α and of a black face, β), the probability of m white and k black faces occurring $(m + k = n)$ is, by Sec. 13.2, $w = \frac{n!}{m!\,k!} \alpha^m \beta^k$.

Let us find the maximum of w as a function of m. It is more convenient to seek the maximum of $\ln w$ instead of w. (It is clear that w and $\ln w$ attain a maximum for the same value of m.) We have $\ln w = \ln n! - \ln m! - \ln k! + m \ln \alpha + k \ln \beta$ or, recalling that $k = n - m$, we get

$$\ln w = \ln n! - \ln m! - \ln(n - m)! + m \ln \alpha + (n - m) \ln \beta$$

We will assume that the number of throws n is fixed and so $\ln w$ depends only on m. We find the derivative $\frac{d \ln w}{dm}$ without being upset by the fact that, by the meaning of the problem, m only assumes integral values. We get

$$\frac{d \ln w}{dm} = -\frac{d \ln m!}{dm} - \frac{d \ln(n - m)!}{dm} + \ln \alpha - \ln \beta$$

The first term on the right is computed thus:

$$\frac{d \ln m!}{dm} = \frac{\ln(m + 1)! - \ln m!}{(m + 1) - m} = \ln(m + 1) = \ln m + \ln\left(1 + \frac{1}{m}\right)$$

Since in a large number of trials n, the number m is most often also large, the quantity $\ln\left(1 + \frac{1}{m}\right) \approx \frac{1}{m}$ can be neglected in comparison with $\ln m$. Hence

$$\frac{d \ln m!}{dm} = \ln m$$

And so

$$\frac{d\ln w}{dm} = -\ln m + \ln(n - m) + \ln\alpha - \ln\beta = \ln\frac{\alpha(n - m)}{m\beta} \tag{5}$$

The maximum condition — that the derivative be zero — yields

$$\ln\frac{\alpha(n - m)}{m\beta} = 0, \quad \text{or} \quad \frac{\alpha}{\beta}\,\frac{n - m}{m} = 1$$

Finally,

$$\frac{m}{n - m} = \frac{\alpha}{\beta}$$

This last result is easy to grasp: the most probable outcome of a trial is that in which the number of whites (m) is to the number of blacks ($n - m$) as the probability of obtaining white (α) in a single trial is to the probability of obtaining black (β) in a single trial.

From the last relation we get $m\beta = \alpha(n - m)$, whence $m(\alpha + \beta) = \alpha n$. Since $\alpha + \beta = 1$, it follows that $m = \alpha n$. Hence, $k = n - m = n(1 - \alpha) = n\beta$. To summarize, the most probable outcome is that in which white turns up αn times and black βn times.*

Denote the probability of this outcome by $w(n\alpha)$. Then

$$w(n\alpha) = \frac{n!}{(\alpha n)!\,(\beta n)!}\,\alpha^{n\alpha}\cdot\beta^{n\beta} \tag{6}$$

By Stirling's formula we get

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^{n}, \quad (\alpha n)! = \sqrt{2\pi n\alpha}\left(\frac{\alpha n}{e}\right)^{\alpha n}$$

$$(\beta n)! = \sqrt{2\pi n\beta}\left(\frac{\beta n}{e}\right)^{\beta n}$$

and so

$$\frac{n!}{(\alpha n)!\,(\beta n)!} = \frac{1}{\sqrt{2\pi n\alpha\beta}}\;\frac{1}{\alpha^{n\alpha}}\;\frac{1}{\beta^{n\beta}}$$

Using formula (6), we find

$$w(n\alpha) = \frac{1}{\sqrt{2\pi n\alpha\beta}} \tag{7}$$

Let us pose the problem of determining the probability of an outcome that deviates but slightly from the most probable outcome. To be more precise, we seek to determine the probability of the number of white faces being equal to $m = n\alpha + \delta$ where δ is small in comparison with n.

* Note that the numbers αn and βn may prove to be nonintegral. Then the most probable number of occurrences of white is taken to be equal to the integer closest to αn.

We proceed from formula (5) in which we put $m = n\alpha + \delta$, $dm = d\delta$. Then (5) takes the form

$$\frac{d \ln w(n\alpha + \delta)}{d\delta} = \ln \frac{\alpha(\beta n - \delta)}{(\alpha n + \delta)\beta}$$

Transform the right-hand side as follows:

$$\ln \frac{\alpha(\beta n - \delta)}{(\alpha n + \delta)\beta} = \ln \frac{\alpha\beta n\left(1 - \frac{\delta}{\beta n}\right)}{\alpha n\left(1 + \frac{\delta}{\alpha n}\right)\beta} = \ln\left(1 - \frac{\delta}{n\beta}\right) - \ln\left(1 + \frac{\delta}{n\alpha}\right)$$

Since $\frac{\delta}{n\beta}$ and $\frac{\delta}{n\alpha}$ are small in comparison with unity, we can put

$$\ln\left(1 - \frac{\delta}{n\beta}\right) = -\frac{\delta}{n\beta}, \quad \ln\left(1 + \frac{\delta}{n\alpha}\right) = \frac{\delta}{n\alpha}$$

whence we get

$$\frac{d \ln w(n\alpha + \delta)}{d\delta} = -\frac{\delta}{n\beta} - \frac{\delta}{n\alpha} = -\frac{\delta}{n\alpha\beta}$$

Thus

$$\frac{d \ln w(n\alpha + \delta)}{d\delta} = -\frac{\delta}{n\alpha\beta}$$

Integrating this equation from 0 to δ, we get

$$\ln w(n\alpha + \delta) - \ln w(n\alpha) = -\frac{\delta^2}{2n\alpha\beta}$$

whence

$$w(n\alpha + \delta) = w(n\alpha)\, e^{-\frac{\delta^2}{2n\alpha\beta}}$$

Recalling that $w(n\alpha) = \frac{1}{\sqrt{2\pi\alpha\beta n}}$, we finally obtain

$$w(n\alpha + \delta) = \frac{1}{\sqrt{2\pi\alpha\beta n}}\, e^{-\frac{\delta^2}{2\alpha\beta n}} \qquad (8)$$

This is just the probability of the outcome in which the number of white faces of a die differs from the most probable outcome by δ.

We again call expected those outcomes for which $\frac{1}{e} w(n\alpha) \leqslant w(n\alpha + \delta) \leqslant w(n\alpha)$, that is, for which $-\delta_1 \leqslant \delta \leqslant \delta_1$, where δ_1 is determined from the condition $w(n\alpha + \delta_1) = \frac{1}{e} w(n\alpha)$. Using (7) and (8), we get

$$\frac{1}{\sqrt{2\pi\alpha\beta n}}\, e^{-\frac{\delta_1^2}{2\alpha\beta n}} = \frac{1}{e\sqrt{2\pi\alpha\beta n}}, \quad \text{or} \quad \frac{\delta_1^2}{2\alpha\beta n} = 1$$

and, finally, $\delta_1 = \sqrt{2\alpha\beta n}$.

Using (8), we can readily prove that $\int\limits_{-\infty}^{+\infty} w(n\alpha+\delta)\,d\delta = 1$. It is also easy to verify that when $\alpha = \beta = 1/2$ we get the formulas derived at the beginning of this section.

Exercises

1. A coin is tossed 1000 times. What is the probability of obtaining exactly 500 heads? exactly 510 heads?
2. A coin is tossed 1000 times. What is the probability of obtaining at least 500 heads? at least 510 heads?
3. One hundred shots are fired with a probability of hitting the target equal to 0.1, the probability of a miss being equal to 0.9. What is the probability that the target will be hit exactly 10 times? exactly eight times?
4. Under the conditions of Exercise 3, what is the probability that the target will be hit at least 8 times? at least 10 times? at least 12 times?
5. A total of 1000 shells are fired with a probability of a hit equal to 0.01. What is the probability that at least 8 shells will hit the target? at least 11 shells?

13.4 Entropy

Thus we have seen that the range of expected outcomes is proportional to $\sqrt{n}$, whereas the range of all conceivable outcomes is equal to n. With our definition of expected outcomes — when it was required that their probability exceed $\frac{1}{e}$ of the maximum — they constitute 84% of all outcomes for a large number of trials. We could give an alternative definition of expected outcomes, for example, by replacing the coefficient $\frac{1}{e}$ by 0.001. We would then find that δ_1 is still proportional to $\sqrt{n}$ (with a different constant of proportionality), and the expected outcomes would constitute about 99.98% of the total, etc. At the same time, the interval of expected outcomes for very large n constitutes only a minute fraction of the interval of all conceivable outcomes, since $\frac{\sqrt{n}}{n} \to 0$ as $n \to \infty$.

This law has many important applications in physics. It states that the values of a *random variable* (in this case, the number of occurrences of a white face) for a sufficiently large number of trials concentrate about the most probable mean value with an arbitrarily high relative accuracy. Here it is common to use the principle of replacing a large number of trials involving a single entity by a single trial involving a *statistical array*, which is a system consisting of a large number of the same entities. For example, instead of throwing a die many times we can consider the simultaneous throwing of many dice, and the number of white faces that turn up will obey the very same law, which is to say that they will concentrate about the most probable value.

Let us first consider a simple imaginary example. Suppose there is an enormous pack of cards (say, $N = 10^{12}$ cards) including αN red cards and βN black cards ($\alpha + \beta = 1$). Cards of a single colour are taken to be indistinguishable. If we assume the state of the pack to be the order in it of the red and black cards, then it is easy to see that the total number $\Omega = \Omega(N, \alpha)$ of possible states is equal to the number of combinations of N elements taken αN at a time, or

$$\Omega = \frac{N!}{(\alpha N)!\,(\beta N)!}$$

Taking advantage of the Stirling formula, we get (see the computations on page 531)

$$\Omega = \frac{1}{\sqrt{2\pi N\alpha\beta}}\,\frac{1}{\alpha^{\alpha N}}\,\frac{1}{\beta^{\beta N}}$$

We have seen that by far not all these states are equally probable. For example, it is easy to calculate that the probability that for $N = 10^{12}$, $\alpha = 0.5$, the mean density of red cards in the first 10^{10} cards will exceed (in the case of decent shuffling) by more than 1% their mean density throughout the pack (equal to 0.5). And since 10^{10} is only a small part of 10^{12}, we can take it that the somewhat elevated frequency of occurrence of red cards in this first portion of the pack does not alter the probability of 0.5 that the next card in each case is red. Thus, the desired probability is equal to the probability considered in Sec. 13.3 that, in 10^{10} tosses of a coin, heads turns up $\geqslant 1.01\cdot 0.5\cdot 10^{10} = (0.5\cdot 10^{10} + 0.5\cdot 10^{8})$ times. By virtue of (3), the sought-for probability is equal to

$$\int_{0.5\cdot 10^{8}}^{\infty} \sqrt{\frac{2}{\pi\cdot 10^{10}}}\, e^{-2\delta^{2}\cdot 10^{-10}}\, d\delta = \int_{10^{3}}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-t^{2}/2}\, dt = \frac{1}{2}\left[1-\Phi(10^{3})\right]$$

(Here we put $2\cdot 10^{-5}\,\delta = t$.)

Our table of values of the probability integral does not contain entries for such large values of the argument, but using the method described on page 69 makes it easy to obtain, for large x,

$$1 - \Phi(x) = \frac{2}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-t^{2}/2}\, dt \approx \frac{2}{\sqrt{2\pi}x}\, e^{-x^{2}/2}$$

And so the desired probability is equal to $\frac{1}{\sqrt{2\pi}} 10^{-3} e^{-10^6/2} \approx 10^{-2\cdot 10^5}$. This is an unbelievably small number. We can be quite sure that if the pack is shuffled and checked thousands of millions of times every second, then the foregoing increase in the mean density of red cards in the first portion of the pack will never be recorded during the whole lifetime of the solar system. The state of the pack must be such that any portion of it consisting of a very large number n of cards contains roughly αn red cards and βn black cards with the range of error of the order of $\sqrt{n}$. Of course, if n is small, $n = 10$ or 100, then the mean density of red cards in some portion may prove to be substantially different from α, that is, the density may have *local fluctuations.*

We use the term *entropy* of the pack of cards for the number $S = \ln \Omega$, that is, the logarithm of the number of its possible states. From the foregoing we get

$$S \approx \alpha N \ln \frac{1}{\alpha} + \beta N \ln \frac{1}{\beta} - \frac{1}{2} \ln(2\pi\alpha\beta N) \approx N \left(\alpha \ln \frac{1}{\alpha} + \beta \ln \frac{1}{\beta}\right)$$

On the right-hand side we neglected the term of the order of $\ln N \ll N$. Referred to a single card, the entropy is equal to $\frac{S}{N} = \alpha \ln \frac{1}{\alpha} + \beta \ln \frac{1}{\beta}$.

Now imagine two packs of cards with characteristics N_1, α_1 and N_2, α_2, respectively, and no interchange of cards between them (the packs) being possible. Such a system has $\Omega = \Omega_1 \Omega_2$ possible states (any state of the first pack may be combined with any state of the second pack), whence the appropriate entropy is

$$S = \ln \Omega = \ln \Omega_1 + \ln \Omega_2 = S_1 + S_2$$

or the entropy of a system of several noninteracting components is equal to the sum of the entropies of the components. Now if these two packs are shuffled together, then they constitute a single common pack, the number of cards being $N = N_1 + N_2$, of which there are $\alpha_1 N_1 + \alpha_2 N_2 = \alpha N$ red cards, where $\alpha = \frac{\alpha_1 N_1 + \alpha_2 N_2}{N}$. The appropriate entropy is equal to

$$\widetilde{S} = N \left(\alpha \ln \frac{1}{\alpha} + \beta \ln \frac{1}{\beta}\right)$$
$$= (\alpha_1 N_1 + \alpha_2 N_2) \ln \frac{N}{\alpha_1 N_1 + \alpha_2 N_2} + (\beta_1 N_1 + \beta_2 N_2) \ln \frac{N}{\beta_1 N_1 + \beta_2 N_2}$$

It is easy to verify that for $\alpha_2 = \alpha_1$ (and hence $\beta_2 = \beta_1$) it will be true that $\widetilde{S} = S$. This is also evident from the fact that for $\alpha_2 = \alpha_1$ the interaction between packs (i.e. their joint shuffling) does not yield anything new. But if $\alpha_2 \neq \alpha_1$, then, as we will now verify, it must be true that $\widetilde{S} > S$. Indeed,

$$\widetilde{S} - S = (\alpha_1 N_1 + \alpha_2 N_2) \ln \frac{N}{\alpha_1 N_1 + \alpha_2 N_2} + (\beta_1 N_1 + \beta_2 N_2) \ln \frac{N}{\beta_1 N_1 + \beta_2 N_2}$$
$$- N_1 \left(\alpha_1 \ln \frac{1}{\alpha_1} + \beta_1 \ln \frac{1}{\beta_1}\right) - N_2 \left(\alpha_2 \ln \frac{1}{\alpha_2} + \beta_2 \ln \frac{1}{\beta_2}\right)$$

If we consider N_1, N_2, and α_1 as given, and α_2 as variable, and if we note that $\beta_2 = 1 - \alpha_2$, then it is easy to compute directly that

$$\frac{d(\widetilde{S} - S)}{d\alpha_2} = N_2 \ln \frac{(\beta_1 N_1 + \beta_2 N_2)\, \alpha_2}{(\alpha_1 N_1 + \alpha_2 N_2)\, \beta_2}$$

$$\frac{d^2(\widetilde{S} - S)}{(d\alpha_2)^2} = \frac{\alpha_1 N_1 N_2}{\alpha_2(\alpha_1 N_1 + \alpha_2 N_2)} + \frac{\beta_1 N_1 N_2}{\beta_2(\beta_1 N_1 + \beta_2 N_2)}$$

For $\alpha_2 = \alpha_1$ we have $\widetilde{S} - S = 0$, $\dfrac{d(\widetilde{S} - S)}{d\alpha_2} = 0$, and since from the expression of the second derivative it follows that it is positive for all α_2 and for this reason the graph of $\widetilde{S} - S$ as a function of α_2 is convex downwards, then for $\alpha_2 \neq \alpha_1$ we will have $\widetilde{S} - S > 0$, i.e. $\widetilde{S} > S$, which is what was asserted.

If we imagine that the two packs with characteristics N_1, α_1 and N_2, α_2 are in juxtaposition but, due to the large volume, exhibit only "local shuffling" of comparatively small portions of the combined pack, then for $\alpha_2 > \alpha_1$ there will occur (as the shuffling takes place) a gradual diffusion of red cards from the second pack into the first, and the mean density of red cards will gradually even out. If at any time in this process we put up a barrier between the first N_1 cards and the last N_2 cards, we can stop the diffusion and compute the entropy in the intermediate state after such a separation. This entropy increases as the evening out of the mean density of red cards takes place.

After the evening out process, the entropy remains practically constant. Of course, we can imagine that, due to the constant shuffling, the mean density of red cards in one part of the pack will become substantially greater than in the other. We have already demonstrated, however, that in the case of a large number of cards the probability of such an event is inconceivably small. Thus, the process of entropy increase is *irreversible.*

The pack of cards we have just been discussing is the simplest model of a statistical physical system. The model is one-dimensional (the cards are arranged in a row) and each card can exist only in two states: red or black. (For an entity that can be in several states with probabilities p_i, the entropy referred to a single item turns out equal to $\sum_i p_i \ln \frac{1}{p_i}$). Similar properties are exhibited by more complicated actual statistical systems like, say, the system of molecules of a gas contained in a specific volume.

According to quantum mechanics, every value of energy E is associated with a definite and very large number Ω of the quantum states of the system at hand, the internal energy of which does not exceed E. To get an idea of the order of this number, let us take advantage of the approximate estimate, in quantum mechanics, of the number Ω_1 of states of a single molecule in a volume V, whose momentum does not exceed a certain value p: $\Omega_1 = p^3V/6\pi^2\hbar^3$, where $\hbar$ is Planck's constant, $\hbar \approx 10^{-27}$ erg-s. The mean momentum of a single molecule is connected with the temperature τ of the gas by the formula $p = \sqrt{3km\tau}$, where k is the Boltzmann constant, $k = 1.38 \cdot 10^{-16}$ erg/degree, and m is the mass of a molecule equal to $\frac{\mu}{N_0}$, where μ is the molecular weight and $N_0 = 6.02 \cdot 10^{23}$ (Avogadro's number) is the number of molecules in a gram-molecule of gas. For oxygen ($\mu = 32$), in 1 litre $= 10^3$ cm^3 at $\tau = 0°\text{C} = 273$ K, we get

$$\Omega_1 = \left(3 \cdot 1.38 \cdot 10^{-6} \frac{32}{6.02 \cdot 10^{23}} \cdot 273\right)^{3/2} \cdot 10^3/6\pi^2(10^{-27})^3 \approx 2.5 \cdot 10^{29}$$

But in 1 litre of gas under normal conditions there are about $N = 3 \cdot 10^{22}$ molecules. And so we can approximately take

$$\Omega = \Omega_1^N = (2.5 \cdot 10^{29})^{3 \cdot 10^{22}} \approx 10^{10^{24}}$$

for the total number of quantum states of such a portion of oxygen.

This is a fantastically large number. Incidentally, as will be seen from what follows, this number does not actually occur in calculations, but rather its logarithm (which too is very great) and, moreover, even the difference of the values of this logarithm for a system under different conditions.

The quantity $S = k \ln \Omega$ is called the entropy of the system: here, k is a fundamentally inessential proportionality factor equal to the Boltzmann constant. The appearance of this factor is due to the fact that the concept of entropy was originally introduced for other reasons, in connection with the study of thermal processes. For this reason, the unit for measuring entropy is chosen so that it retains the same value as in thermodynamics. In other words, the coefficient k enables us to obtain the temperature in degrees instead of ergs.

As E varies, so does Ω and, hence, S as well; and in the case of a constant volume of gas the ratio $\frac{dE}{dS} = \tau$ is called its *temperature*. In a joint consideration of two systems a and b, as long as they are not interacting, we have $\Omega = \Omega_a\Omega_b$ for the combined system, and so $S = S_a + S_b$. If the systems can exchange energy but are otherwise independent (the portions of gas are separated by a thin partition), then the same relations hold true but the participating quantities will vary in the course of time. If the system as a whole is isolated from the ambient medium, then the energy is conserved: $E = E_a + E_b =$ constant, whence $\dot{E}_a + \dot{E}_b = 0$, where the dot denotes the time derivative. From this we have

$$\dot{S} = \dot{S}_a + \dot{S}_b = \frac{1}{\tau_a}\dot{E}_a + \frac{1}{\tau_b}\dot{E}_b = \left(\frac{1}{\tau_a} - \frac{1}{\tau_b}\right)\dot{E}_a$$

Hence, for $\tau_a = \tau_b$ we have $\dot{S} = 0$; the entropy does not vary at equal temperatures of the interacting portions of gas. But if, say, the temperature of the system a is less than that of the system b, $\tau_a < \tau_b$, then $\dot{E}_a > 0$ and $\dot{S} > 0$, which means the system a acquires energy from b and the entropy of the overall system increases. A similar increase in entropy occurs in the case $\tau_a > \tau_b$ as well. This process of pumping energy and increasing entropy continues until the temperatures even out, after which the entropy remains practically constant. It is logically possible to conceive of a case where in the exchanges of energy the temperature at some time in one portion of the system again becomes greater than in another portion, but this is immeasurably less probable than the earlier discussed variation in the density of red cards. Incidentally, extremely small portions of gas can exhibit fluctuations of temperature and of the associated particle velocities, as witness the familiar phenomenon of the Brownian motion.

Since the entropy of a statistical system can only increase or remain constant, the processes in which the entropy increases are termed irreversible processes. These processes become impossible if, when filmed, they are played back in reverse order. Mathematically, inversion of a process reduces to replacing the time t by $-t$. Thus, if the law of development of a process is recorded in the form of a linear differential equation not containing the time t explicitly, then the condition of reversibility consists in the presence, in the equation, of derivatives with respect to t solely of even order. We can illustrate this in the case of the linear oscillator that was considered in Sec. 7.3: the term $h\frac{dy}{dt}$ in equation (21) of Ch. 7 corresponded to friction in the system and friction leads to a dissipation of energy and, thus, to irreversibility of the process.

Note in conclusion that the principle of entropy increase (the law of nondecreasing entropy) or the so-called second law of thermodynamics is of a probabilistic nature and is ultimately connected with the fact that the more probable a state of nature, the more often it occurs. To put it differently, processes associated with decrease of entropy are not actually observed not because of their logical or formal contradictory nature but because of their extremely small probability.

13.5 Radioactive decay. Poisson's formula

Let us consider the phenomenon of radioactive decay. As we know (see HM, Sec. 5.3), the probability of a single atom decaying in an extremely small time t_1 is equal to wt_1; the probability that decay will not occur in this time is equal to $1 - wt_1$. (Here, w is a constant describing the given radioactive material.)

Consider a long time interval t. Let us find the probability $w(t)$ that no decay will occur in that time interval. To do this, partition the interval t into small subintervals of duration $t_1, t_2, \ldots, t_n$. The probability that decay will not occur during time t_i is equal to

$$1 - wt_i$$

The probability $w(t)$ is equal to the product of the probabilities that decay does not occur in any one of the time intervals $t_1, t_2, \ldots, t_n$. Therefore

$$w(t) = (1 - wt_1) \cdot (1 - wt_2) \ldots (1 - wt_n)$$

Consider $\ln w(t)$. It is clear that

$$\ln w(t) = \ln(1 - wt_1) + \ln(1 - wt_2) + \ldots + \ln(1 - wt_n)$$

Since the quantities $wt_1, wt_2, \ldots, wt_n$ are small in comparison with 1, the logarithms on the right can be expanded in a series. Confining ourselves to the first term of the expansion, we get

$$\ln w(t) = -wt_1 - wt_2 - \ldots - wt_n = -wt$$

Taking antilogarithms, we get

$$w(t) = e^{-wt}$$

We thus obtain a familiar result: the ratio of the number of atoms that have not decayed in time t to the original number of atoms is e^{-wt}.

If the probability that an atom will not decay during time t is denoted by β, then $\beta = e^{-wt}$. The probability α that during time t an atom will decay is $\alpha = 1 - e^{-wt}$.

If there are n atoms, then the probability $w(m, n)$ that m of them will decay and $k = n - m$ will not decay is given by

$$w(m, n) = \frac{n!}{m!\,k!}\alpha^m\beta^k \tag{9}$$

where $\alpha = 1 - e^{-wt}$, $\beta = e^{-wt}$ (see Sec. 13.2).

Let us consider an important special case: let the total number of radioactive atoms n be extremely great and the probability of decay during time t extremely small so that the most probable number of disintegrations during t is a finite number, which we denote by μ. Then $\mu = \alpha n$, as was established in Sec. 13.3.

To summarise, we want to know the form of (9) if n increases without bound, α approaches zero without bound, and their product $\alpha n = \mu$ (and m as well) remains a finite number.

We consider the factor $\frac{n!\,\alpha^m}{k!} = \frac{n!\,\alpha^m}{(n-m)!}$ which we write as follows:

$$\frac{n!\,\alpha^m}{(n-m)!} = n(n-1)(n-2)\dots(n-m+1)\cdot\alpha^m$$
$$= (n\alpha)(n\alpha-\alpha)(n\alpha-2\alpha)\dots[n\alpha-(m-1)\alpha]$$
$$= \mu(\mu-\alpha)(\mu-2\alpha)\dots[\mu-(m-1)\alpha]$$

Therefore, for fixed m and for very large n, we have

$$\frac{n!\,\alpha^m}{(n-m)!} \approx \mu^m$$

Now let us consider the quantity $\beta^k = \beta^{n-m} = (1-\alpha)^{n-m}$. Since α is an extremely small quantity, it follows that β is close to unity. But the exponent $n - m$ is great and so, replacing β^{n-m} by 1, we risk committing a substantial error.

Let us do as follows:

$$\beta^{n-m} = (1-\alpha)^{n-m} = \frac{(1-\alpha)^n}{(1-\alpha)^m} \approx (1-\alpha)^n$$

since $(1-\alpha)^m$ is close to 1. Recalling that $n\alpha = \mu$, we get

$$(1-\alpha)^n = \left(1-\frac{\mu}{n}\right)^n = \left[\left(1-\frac{\mu}{n}\right)^{-\frac{n}{\mu}}\right]^{-\mu} \approx e^{-\mu}$$

since $\left(1-\frac{\mu}{n}\right)^{-\frac{n}{\mu}}$ is the closer to e, the greater n is.

Finally, from (9) we find, as $n \to \infty$,

$$w(m, n) = w_\mu(m, n) \to w_\mu(m) = \frac{\mu^m}{m!}e^{-\mu} \tag{10}$$

The new notation $w_\mu\ (m)$ signifies the probability of observing m disintegrations if the most probable number of disintegrations is μ, and the number of atoms n is very great so that number of disintegrations is a small portion of the number of atoms.

The law expressed by (10) is called the *Poisson distribution.*

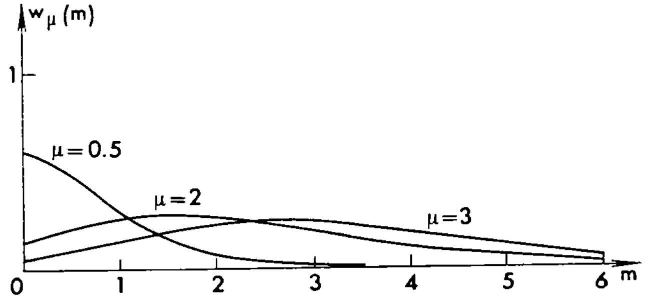
Let us convince ourselves that the sum of the probabilities $w_\mu(m)$ for all values of m is equal to 1, that is, that

$$\sum_{n=0}^{\infty} w_\mu(m) = 1$$

Indeed,

$$\sum_{m=0}^{\infty} w_\mu(m) = \sum_{m=0}^{\infty}\frac{\mu^m}{m!}e^{-\mu} = e^{-\mu}\sum_{m=0}^{\infty}\frac{\mu^m}{m!} = e^{-\mu}\cdot e^{\mu} = 1$$

Fig. 188



The Poisson distribution shows what the probability is of observing m disintegrations if the most probable number of disintegrations is μ and the separate disintegrations are independent, which is to say that the fact that a certain number of disintegrations have occurred does not alter the probability of obtaining another disintegration (for this purpose we stipulated that the total number of radioactive atoms n is great so that $\mu \ll n$, $m \ll n$).

In contrast to the problems of preceding sections, here the total number of disintegrations is not restricted in any way. In preceding sections we considered cases where a definite — albeit very great — number of trials, n, are carried out. We specified this number n and it entered into the final results. In this section, n (the number of atoms or, what is the same thing, the number of trials, a trial consisting in seeing whether an atom decays or not) is assumed to be unlimited. Hence, so is the number of acts of decay (disintegration). Fundamentally, it is possible (though highly unlikely) to observe any (arbitrarily large) number of disintegrations for one and the same most probable number of disintegrations μ.

If μ is small, $\mu \ll 1$, then, using (10), we find that the probability of not observing a single instance of disintegration is equal to $e^{-\mu} \approx 1 - \mu$, which is extremely close to unity. The probability of observing a single disintegration is markedly less, namely, $\mu \cdot e^{-\mu} \approx$ $\approx \mu(1 - \mu) \approx \mu$. The probabilities of observing two, three, etc. disintegrations diminish very rapidly (they are equal to $\mu^2/2$, $\mu^3/6$, etc.).

If μ is great, then it is most probable to observe μ disintegrations. Indeed, let us find for what m (μ is constant!) the quantity $w_\mu(m)$ has a maximum. It is more convenient to seek the maximum of $\ln w_\mu(m)$. Since $\ln w_\mu(m) = -\mu + m \ln \mu - \ln m!$, it follows that

$$\frac{d \ln w_\mu(m)}{dm} = \ln \mu - \frac{d}{dm}(\ln m!) = \ln \mu - \ln m$$

(see Sec. 13.3). Therefore the equation $\frac{d \ln w_\mu(m)}{dm} = 0$ yields $m = \mu$.

In Fig. 188 we have the Poisson distribution for $\mu = 0.5$, $\mu = 2$, $\mu = 3$. Note that Fig. 188 is not exact: the quantity w is actually given only for integer values of m, so that the curve of $w(m)$ for all m, i.e. for fractional m, say, is meaningless. The way to do it would be to give for every μ a series of quantities $w_\mu(0)$, $w_\mu(1)$, $w_\mu(2)$, etc. depicted by vertical lines erected from integral points on the axis of abscissas. The curve $w_\mu(m)$ passes through the tops of these lines.

The aspect of the function $w_\mu(m)$ for large μ will be described in Sec. 13.8.

Exercises

1. Consider the ratio of $w_\mu(m)$ to $w_\mu(m - 1)$ and draw a conclusion about the maximum of $w_\mu(m)$ for a given μ.
2. Suppose a certain quantity x assumes in a series of trials the values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$. Prove that the mean value $\bar{x}$ of this quantity in a single trial is equal to $x_1p_1 + + \ldots + x_np_n$. On this basis, prove that the mean value $\overline{m}$ of the number of disintegrations during time t is just equal to μ.

13.6 An alternative derivation of the Poisson distribution

Here we present an alternative derivation of the Poisson distribution that is based on arguments that differ from those used in Sec. 13.5. Imagine a large number of instruments (counters) each of which records disintegrations in similar samples containing a long-half-life material.

For the sake of calculational convenience, we assume that on the average there is one disintegration in each sample in unit time. Then, on the average, there will be t disintegrations in time t. Denote by x_0 the number of counters that do not record a single disintegration ($m = 0$), by x_1 the number of counters that record a single disintegration ($m = 1$), by x_2 the number of counters that record two disintegrations ($m = 2$), etc. It is clear that x_0, x_1, x_2, ... depend on the time that has elapsed from the start of the experiment. Suppose at a definite time t we know the quantities $x_0(t)$, $x_1(t)$, $x_2(t)$, ..., indicating the number of counters that have recorded 0, 1, 2, ... disintegrations.

How will these numbers of counters change over a small time interval dt?

In any group of n counters there will be n disintegrations in unit time and there will be $n\,dt$ disintegrations in time dt. And so in this group, $n\,dt$ counters will record one disintegration each.

Let us consider the group of counters that have not recorded a single disintegration. There will be one disintegration in $x_0(t)\,dt$ of these counters, and these counters will pass into another group, namely the group x_1. Hence the number of counters that have not recorded a single disintegration during the time $t + dt$ is

$$x_0(t + dt) = x_0(t) - x_0(t)\,dt$$

whence

$$x_0(t + dt) - x_0(t) = -x_0(t)\,dt, \text{ or } \frac{dx_0}{dt} = -x_0$$

Consider the group $x_1(t)$. As before, $x_1(t)\,dt$ of these counters will go into the group x_2, but $x_0(t)\,dt$ counters from group x_0 will go into the group under consideration. Therefore

$$x_1(t + dt) = x_1(t) - x_1(t)\,dt + x_0(t)\,dt$$

whence

$$\frac{dx_1}{dt} = x_0 - x_1$$

Continuing this reasoning, we arrive at a chain of differential equations:

$$\left.\begin{array}{l} \dfrac{dx_0}{dt} = -\,x_0, \\ \dfrac{dx_1}{dt} = x_0 - x_1, \\ \dots\dots\dots\dots\dots \end{array}\right\} \qquad (11)$$

Clearly, at the initial time $t = 0$ there will be $x_0 = N$, where N is the total number of counters, $x_1 = x_2 = x_3 = \dots = 0$. The equations (11) are easily solved one after the other (see Sec. 7.2), and we get

$$\left.\begin{array}{l} x_0 = Ne^{-t}, \\ x_1 = Nte^{-t}, \\ x_2 = N\dfrac{t^2}{2}e^{-t}, \\ x_3 = N\dfrac{t^3}{3!}e^{-t}, \\ \dots\dots\dots\dots\dots \end{array}\right\}$$

Suppose μ units of time have elapsed since the inception of the process. During this time there will have occurred an average of μ disintegrations in each sample, and in each counter of group m, whose number is equal to $x_m(\mu) = N\frac{\mu^m}{m!}e^{-\mu}$, there will have occurred m disintegrations. For this reason, the probability that there will be m disintegrations in a randomly chosen counter is equal to the ratio of the number of counters of the group, $x_m(\mu)$, to the total number of counters, or

$$w_\mu(m) = \frac{x_m(\mu)}{N} = \frac{\mu^m}{m!}e^{-\mu}$$

We have obtained the same result as in Sec. 13.5.

Exercise

Consider the following generalization of the problem that has been analyzed. Suppose a certain entity can be in the states ..., C_{-2}, C_{-1}, C_0, C_1, C_2, ..., and at time $t = 0$ it was in the state C_0 and during time dt it passed with probability $\omega\, dt$ into the next state or, with probability $\alpha\, dt$, into the preceding state. Indicate a system of equations that satisfy the probability $p_i(t)$ of being in the state C_i at time t, find approximate expressions for $p_i(t)$ by the method of successive approximations, and find the mean value of the number of the state at time t. For what values of α and ω do we get the problem discussed in the text?

13.7 Continuously distributed quantities

Let us return to the fishing problem discussed in Sec. 13.1. Let the probability of catching a fish of weight between p and $p + dp$ be equal to $f(p)\,dp$ (no fish caught will be assumed to be equal to a fish of weight zero — that's so that everyone will be satisfied). The function $f(p)$ is called the *distribution function*. It satisfies the condition $\int\limits_0^{+\infty} f(p)\,dp = 1$, since this integral is the sum of the probabilities of all the events that can take place.

Note that actually in any body of water there can be no fish whose weight exceeds some number P. However, we put $+\infty$ as the upper limit of integration instead of P. This can be done by taking the function $f(p)$ as being equal to 0 for $p > P$ or, at any rate, as being a rapidly decreasing function as p increases and as having such small values for $p > P$ that the value of the integral is for all practical purposes unaffected by these values*.

It is important to note the following. We consider that only a very small portion of all the fish in the pond are caught. For this reason, the probability of catching a fish whose weight lies between the limits of p and $p + dp$ does not depend on how many fish have already been caught and what they weigh. In other words, the function $f(p)$ does not depend on what fish have already been caught. This function describes a given rather large body of water.

We pose the following problem. Suppose a large number n of fish have been caught (to be more precise, the fish hook has been pulled out of the water n times). What is the mean weight of a single fish? The probability of catching a fish weighing between p and

* In other problems the quantity at hand (in the given case the weight of a fish) can assume negative values as well. In this case $f(p)$ satisfies the condition $\int\limits_{-\infty}^{+\infty} f(p)\,dp = 1.$

$p + dp$, where dp is small, is equal to $f(p)\,dp$. For this reason, of the total number n, there will be $n \cdot f(p)\,dp$ cases in which a fish of weight p will have been caught.* The weight of all such fish is equal to $p \cdot nf(p)\,dp$. Integrating the last expression over all p, that is, from $p = 0$ to $p = +\infty$, we get the total weight of the catch after casting the line n times:

$$P_n = n \int_0^\infty pf(p)\,dp$$

Dividing the total weight of the catch P_n by the number of times n the line is cast, we get the mean weight of a fish referred to a single cast of the line:

$$\overline{p}_1 = \frac{P_n}{n} \int_0^\infty pf(p)\,dp \qquad (12)$$

We now pose a more complicated problem. Suppose n fish have been caught. The probability that the total weight of the catch will lie within the range from p to $p + dp$ is $F_n(p)\,dp$. The function $F_n(p)$ is the distribution function over the weight of a catch consisting of n fish. Let us attempt to find this function. To do this we first set up an equation connecting $F_{n+1}(p)$ and $F_n(p)$.

Suppose we have the distribution $F_n(p)$ after pulling out the line n times. How is it possible, after $n + 1$ extractions of the line, to obtain a total weight of the catch lying within the range from p to $p + dp$?

If the weight of the last, $(n + 1)$st, fish lies in the range from μ to $\mu + d\mu$, where $d\mu$ is much less than dp, then for the sum of the weights of n fish and the $(n + 1)$st fish to fall within the given range from p to $p + dp$ it is necessary that the weight of n fish lie within the range from $p - \mu$ to $p + dp - \mu$ (here we disregard the small quantity $d\mu$).

The probability of catching the $(n + 1)$st fish weighing from μ to $\mu + d\mu$ is, as we know, equal to $f(\mu)\,d\mu$. The probability of the weight of the first n fish lying in the range from $p - \mu$ to $p - \mu + dp$ is $F_n(p - \mu)\,dp$.

The probability of an event consisting in the weight of the first n fish falling in the indicated range and the weight of the $(n + 1)$st

* More precisely, a fish whose weight is very close to p, namely, within the interval dp about p. Note that for a quantity that varies continuously and is not obliged to assume definite integral (generally discrete) values, it is meaningless to ask what the probability is that it will assume an *exactly* definite value: the probability is clearly zero. Only the probability of falling in a definite interval is meaningful, and this probability is proportional to the length of the interval when the interval is small.

fish also being in this range is equal to the product of the probabilities of these separate events and, hence, is equal to

$$f(\mu)\,d\mu \cdot F_n(p-\mu)\,dp \tag{13}$$

Note that the total weight of all fish caught that lies in the indicated range can be obtained in a multitude of way since the weight μ of the last fish can assume any value from 0 to $+\infty$. And so the total probability that the weight of all fish caught lies in the indicated range is equal to the sum of the expressions (13) written for distinct values of μ. And since μ assumes all possible values, in place of the sum we have the integral. Thus, this probability is

$$\int_0^\infty f(\mu)\,d\mu F_n(p-\mu)\,dp = dp\int_0^\infty f(\mu)F_n(p-\mu)\,d\mu \tag{14}$$

However, by definition, the probability that the total weight of $(n+1)$ fish lies in the range from p to $p+dp$, is equal to $F_{n+1}(p)\,dp$. Equating this last expression to the right member of (14) and cancelling out dp, we get

$$F_{n+1}(p) = \int_0^\infty f(\mu)\,F_n(p-\mu)\,d\mu \tag{15}$$

Knowing the function $F_1(\mu) = f(\mu)$ (since $F_1(p)$ refers to the case of a catch consisting of a single fish) and F_n, formula (15) enables us to find $F_{n+1}(p)$, which is to say, to pass successively from subscript n to $n+1$.*

Let us consider a simple example.

Suppose $f(p) = \frac{1}{q}$ if $0 < p < q$ and $f(p) = 0$ for all other values of p. This means that the body of water does not have any fish weighing more than q, and the probability of catching any fish weighing less than q is the same. For this reason there are no "zero" fish (in $f(p)$ there is no delta term). The mean weight of a caught fish is

$$\bar{p}_1 = \int_0^q p\,\frac{1}{q}\,dp = \frac{q}{2}$$

* Actually, in (15) the integration is performed from 0 to p since $F_n(p-\mu)$ is equal to zero for $\mu > p$, and therefore μ can assume values only between 0 and p. If instead of the weight of a fish we consider a quantity that can assume values of both signs, then in (15) the integration is carried out from $-\infty$ to $+\infty$.

It is clear that the condition $\int\limits_0^\infty f(p)dp = 1$ is fulfilled. Indeed,

$$\int\limits_0^\infty f(p)\,dp = \int\limits_0^q f(p)\,dp = \int\limits_0^q \frac{1}{q}\,dp = \frac{1}{q}\cdot q = 1$$

As has been pointed out above, $F_1(p) = f(p)$. Let us find $F_2(p)$. Using (15), we get

$$F_2(p) = \int\limits_0^\infty f(\mu)\cdot f(p-\mu)\,d\mu$$

Since $f(\mu)$ is different from zero and is equal to $\frac{1}{q}$ only when $0<\mu<q$, then $F_2(p) = \int\limits_0^q \frac{1}{q} f(p-\mu)\,d\mu$. Put $p - \mu = t$ in this last integral; then $dt = -\,d\mu$ and we get

$$F_2(p) = -\frac{1}{q}\int\limits_p^{p-q} f(t)\,dt = \frac{1}{q}\int\limits_{p-q}^{p} f(t)\,dt$$

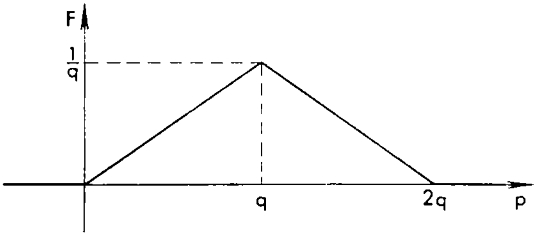
Now let us consider separately the case $0 < p < q$ and the case $q < p < 2q$. If $0 < p < q$, then $p - q < 0$. Taking into account that the function $f(t)$ is different from zero (and is equal to $1/q$) only when $0 < t < q$, we get $F_2(p) = \frac{1}{q}\int\limits_0^p \frac{1}{q}\,dt = \frac{p}{q^2}$. But if $q<p<2q$, then we integrate from $p - q$ to q. And so in this case $F_2(p) = = \frac{1}{q}\int\limits_{p-q}^{q} \frac{1}{q}\,dt = \frac{2q-p}{q^2}$. Thus

$$F_2(p) = \begin{cases} \dfrac{p}{q^2} & \text{if } 0 < p < q, \\ \dfrac{2q-p}{q^2} & \text{if } q < p < 2q \end{cases}$$

Also note that $F_2(p) \equiv 0$ if $p < 0$ and if $p > 2\,q$ since the weight of a catch cannot be negative and the weight of a catch consisting of two fish cannot exceed $2q$ for the reason that there are no fish in the pond weighing more than q. The graph of function $F_2(p)$ is shown in Fig. 189. We suggest that the reader obtain the function $F_3(p)$ and construct its graph. (See Exercise.)

The following are the two simplest properties of the functions $F_n(p)$.

Fig. 189



1. $\int_0^\infty F_n(p)\,dp = 1$ (for any n). This property is obvious if we proceed from the fact that $F_n(p)$ is the distribution function. The reader will find no difficulty in proving that this relation holds true for the functions $F_2(p)$, $F_3(p)$ of the foregoing example.

2. Denote by $\bar{p}_{n+1}$ the mean weight of a catch consisting of $(n + 1)$ fish. This is to be understood as follows. Suppose we cast the line $(n + 1)$ times in a row and then compute the mean weight of the catch consisting of $(n + 1)$ fish. Then $\bar{p}_{n+1} = \bar{p}_n + \bar{p}_1$, that is, the mean weight of the catch consisting of $(n + 1)$ fish is equal to the sum of the mean weight of the catch consisting of n fish and the mean weight of the catch consisting of a single fish. Therefore $\bar{p}_2 = \bar{p}_1 + \bar{p}_1 = 2\bar{p}_1$, $\bar{p}_3 = \bar{p}_2 + \bar{p}_1 = 3\bar{p}_1$, etc.; that is,

$$\bar{p}_n = n\bar{p}_1 \tag{16}$$

We conclude with a few general properties of *random variables*, which are quantities that assume definite values as a result of a trial. (For example, the weight of a fish caught in a trial — pulling in the fish line — is a random variable.) Suppose we have two random variables ξ and η; denote by p, respectively by q, the possible values of these quantities, and by $f(p)$ and $\varphi(q)$ the corresponding distribution functions. As we have seen, the mean value $\bar{\xi}$ (we can also write $\bar{p}$) of ξ can be computed from the formula $\bar{\xi} = \int_{-\infty}^{\infty} pf(p)\,dp$. The quantity $\bar{\eta}$ is expressed in similar fashion. It is easy to demonstrate that we always have

$$\overline{\xi + \eta} = \bar{\xi} + \bar{\eta},\ \overline{C\xi} = C\bar{\xi} \quad (C = \text{constant}) \tag{17}$$

Indeed, if we denote by p_i and q_i the values of ξ and η in the ith trial, then for an extremely large number N of trials we can write

$$\overline{\xi + \eta} = \frac{1}{N}\sum_{i=1}^{N}(p_i + q_i) = \frac{1}{N}\sum_{i=1}^{N}p_i + \frac{1}{N}\sum_{i=1}^{N}q_i = \bar{\xi} + \bar{\eta}$$

The second equation in (17) can be verified similarly.

If the variables ξ and η are independent, then it can also be demonstrated that $\overline{\xi\eta} = \bar{\xi}\cdot\bar{\eta}$ for the probability that ξ will take on a value lying between p and $p + dp$, and η, a value between q and $q + dq$ is, by virtue of the independence condition, equal to $f(p)\,dp\;\varphi(q)\,dq$. The corresponding value of $\xi\eta$ is equal to pq. Therefore, the mean value $\overline{\xi\eta}$ is obtained from the formula

$$\overline{\xi\eta} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} pq[f(p)\,dp\;\varphi(q)\,dq] = \int_{-\infty}^{\infty} pf(p)\,dp\cdot\int_{-\infty}^{\infty} q\varphi(q)\,dq = \bar{\xi}\cdot\bar{\eta}$$

which is what we set out to prove.

The dispersion of ξ about its mean value is characterized by the mean square of the difference $\Delta_\xi^2 = \overline{(\xi - \bar{\xi})^2} = \int_{-\infty}^{\infty} (p - \bar{\xi})^2 f(p)\,dp$, which is called the *variance* of ξ. (Note that the last integral is indeed a positive number since the integrand is positive.) It is easy to verify that the variance of a sum of independent quantities is equal to the sum of their variances:

$$\Delta_{\xi+\eta}^2 = \overline{[(\xi + \eta) - (\overline{\xi + \eta})]^2} = \overline{[(\xi - \bar{\xi}) + (\eta - \bar{\eta})]^2}$$
$$= \overline{(\xi - \bar{\xi})^2} + 2\overline{(\xi - \bar{\xi})}\;\overline{(\eta - \bar{\eta})} + \overline{(\eta - \bar{\eta})^2} = \Delta_\xi^2 + 2\cdot 0\cdot 0 + \Delta_\eta^2$$

since $\overline{\xi - \bar{\xi}} = \bar{\xi} - \bar{\bar{\xi}} = \bar{\xi} - \bar{\xi} = 0$. This property can be extended directly to the sum of any number of independent random variables.

From this it follows, in particular, that for the sum ξ of n independent values of ξ_1 the variance Δ_n^2 is equal to $\Delta_n^2 = n\Delta_1^2$.

Exercise

Find the function $F_3(p)$ for the case

$$f(p) = \begin{cases} \dfrac{1}{q} & \text{if } 0 < p < q, \\ 0 & \text{if } p < 0 \text{ or } p > q \end{cases}$$

Construct the graph of this function.

13.8 The case of a very large number of trials

In this section we consider the behaviour of the function $F_n(p)$ that was introduced in Sec. 13.7 for very large values of n. For the sake of simplicity, instead of the weight of a fish we first consider a variable ξ_1, the mean value of which is zero. Such a quantity can of course take on values of both signs. $F_n(p)$ is the distribution function of the sum of n independent values of ξ_1.

We start with the formula (15) obtained in Sec. 13.7. But, as was pointed out in the footnote on page 546, the integral in the formula must be taken over the range from $-\infty$ to ∞. Expand $F_n(p-\mu)$ in a series of powers of μ, confining yourself to the term containing μ^2, to get

$$F_n(p-\mu) = F_n(p) - \mu\frac{dF_n(p)}{dp} + \frac{1}{2}\mu^2\frac{d^2F_n(p)}{dp^2} \tag{18}$$

Using this expansion, we get, from (15),

$$F_{n+1}(p) = F_n(p)\int_{-\infty}^{\infty} f(\mu)\,d\mu - \frac{dF_n(p)}{dp}\int_{-\infty}^{\infty} f(\mu)\,\mu\,d\mu + \frac{1}{2}\frac{d^2F_n(p)}{dp^2}\int_{-\infty}^{\infty} f(\mu)\,\mu^2\,d\mu \tag{19}$$

Note that $\int_{-\infty}^{\infty} f(\mu)\,d\mu = 1$ since $f(p)$ is the distribution function. Besides, $\int_{-\infty}^{\infty} f(\mu)\,\mu\,d\mu = 0$ for, by virtue of Sec. 13.7, this integral is equal to the mean value of ξ_1. Finally, we introduce the variance $\Delta_1^2 = \int_{-\infty}^{\infty} \mu^2 f(\mu)\,d\mu$ of ξ_1.

Now formula (19) yields

$$F_{n+1}(p) = F_n(p) + \frac{1}{2}\Delta_1^2\frac{d^2F_n(p)}{dp^2} \tag{20}$$

Now let us consider $F_n(p)$ as a function of two variables p and n and write $F_n(p) = F(p;n)$. We will make use of the partial derivatives with respect to p and also with respect to n, and we will not be disturbed by the fact that n takes on only integral values, since for large n a change of n by one unit may be regarded as an extremely small change in comparison with n. We rewrite (20) thus:

$$F(p;n+1) = F(p;n) + \frac{1}{2}\Delta_1^2\frac{\partial^2 F}{\partial p^2} \tag{21}$$

We now expand the left side in a Taylor series, confining ourselves to the first two terms, which yields

$$F(p;n+1) = F(p;n) + \frac{\partial F}{\partial n}\cdot 1 \tag{22}$$

Equating the right sides of (21) and (22), and then dropping the term $F(p;n)$, we arrive at the basic equation for the function F:

$$\frac{\partial F}{\partial n} = \frac{1}{2}\Delta_1^2\frac{\partial^2 F}{\partial p^2} \tag{23}$$

Before giving the solution of this equation, we carry out a supplementary investigation of the function $F(p;\, n)$, which will do much to clarify the use of Taylor's formula. We consider the graph of the function $F(p;\, n) = F_n(p)$, which is the distribution function of the random variable ξ. Since $\bar{\xi} = n\bar{\xi}_1 = 0$, the centre of gravity of the graph, for all n, is located at the origin. For the width of the graph we can take the *root-mean-square deviation* $\Delta_n = \sqrt{\overline{\Delta_n^2}}$, which has the same dimensions as ξ. However, at the end of Sec. 13.7 we showed that $\Delta_n^2 = n\Delta_1^2$ or $\Delta_n = \sqrt{n}\Delta_1$; the width of the graph is of the order of $\sqrt{n}$. But then from the normalization condition $\int F\, dp = 1$ it follows that $F \propto n^{-1/2}$, whence $\frac{\partial F}{\partial n} \propto n^{-3/2}$, $\frac{\partial^2 F}{\partial n^2} \propto n^{-5/2}$, etc. And from the estimate of the width of the graph it follows that

$$\frac{\partial F}{\partial p} \propto n^{-1/2}\colon\ \sqrt{n} = n^{-1},\ \ \frac{\partial^2 F}{\partial p^2} \propto n^{-1}\colon\ \sqrt{n} = n^{-3/2},\ \text{etc.}$$

Thus, the expansions (21) and (22) are carried to terms of the same asymptotic order and for this reason (23) is asymptotically exact.

It is curious that the asymptotic order of the width of the graph of F_n could have been obtained directly from (23) by differentiating with respect to the parameter under the integral sign (Sec. 3.6) and integrating by parts:

$$\frac{d}{dn}\,\Delta_n^2 = \frac{d}{dn}\int p^2 F_n(p)\, dp = \int p^2 \frac{\partial F}{\partial n}\, dp = \frac{1}{2}\,\Delta_1^2 \int p^2 \frac{\partial^2 F}{\partial p^2}\, dp$$

$$= \frac{1}{2}\,\Delta_1^2 \left[p^2\,\frac{\partial F}{\partial p}\,\Big|_{-\infty}^{\propto} - \int 2p\,\frac{\partial F}{\partial p}\, dp\right] = -\,\Delta_1^2 \left[pF\,\Big|_{-\infty}^{\infty} - \int F\, dp\right] = \Delta_1^2$$

(Here we assume that $F \to 0$ sufficiently fast as $p \to \pm\infty$.) From this $\Delta_n^2 \propto n\Delta_1^2$.

One should not grudge the time spent on such an analysis. On the one hand, even without the solution an important general property of F was obtained in the process, namely the expression $\sqrt{n}\Delta_1$ of the width. What is more — and this is quite important — your vigilance was enhanced. You have a better grasp of the following general rule: in order to introduce approximations (i.e. to retain certain terms and reject others in the Taylor series), one has to learn as much as possible about the function.

Let us examine the solution of equation (23). It can be shown that asymptotically, for large n, the solution of (23) is the function $F(p;\, n) = A\frac{1}{\sqrt{n}}e^{-p^2/2n\Delta_1^2}$, where A is any constant. We give this without any derivation. The fact that this function satisfies (23) can readily be verified by setting up $\partial F/\partial n$, $\partial^2 F/\partial p^2$ and substituting the expressions obtained into (23). We choose the value of the constant A from the condition that the requirement $\int F dp = 1$ be fulfilled:

$$\frac{A}{\sqrt{n}} \int_{-\infty}^{\infty} e^{-p^2/2n\Delta_1^2} dp = 1 \tag{24}$$

Make the change of variable $p/\Delta_1 \sqrt{2n} = t$, $dp/\Delta_1 \sqrt{2n} = dt$. Then

$$\frac{A}{\sqrt{n}} \int_{-\infty}^{\infty} e^{-p^2/2n\Delta_1^2} dp = A\Delta_1 \sqrt{2} \int_{-\infty}^{+\infty} e^{-t^2} dt = A \cdot \Delta_1 \sqrt{2} \cdot \sqrt{\pi} = 1$$

Thus, $A\Delta_1 \sqrt{2\pi} = 1$, whence $A = 1/\Delta_1 \sqrt{2\pi}$ and, finally,

$$F(p\,;\,n) = \frac{1}{\Delta_1 \sqrt{2\pi n}} e^{-p^2/2n\Delta_1^2} \tag{25}$$

Now let us consider a case where the mean value $\bar{\xi}_1$ is not necessarily zero. We then put $\xi_1 - \bar{\xi}_1 = \xi_1'$, whence $\xi_1 = \bar{\xi}_1 + \xi_1'$, where $\bar{\xi}_1' = 0$. Therefore the sum ξ of n independent values of ξ_1 is equal to the result of adding the constant $n\bar{\xi}_1$ and the sum ξ' made up of n independent values of ξ_1'. For n large, the sum ξ' has the distribution law (25). But if we add a constant to a random variable, the corresponding distribution function is merely shifted by the amount of that constant, which is to say that as a result we obtain the distribution function

$$F(p\,;\,n) = \frac{1}{\Delta_1 \sqrt{2\pi n}} e^{-(p-n\bar{p}_1)^2/2n\Delta_1^2} \tag{26}$$

where we have reverted to the notation of Sec. 13.7: $\bar{p}_1$ instead of $\bar{\xi}_1$.

We could have obtained the solution (26) directly by skipping over the special case (25). To do this, it is necessary to obtain, with the aid of the expansions (19) and (22), the equation for the function F, no longer assuming that $\bar{p}_1 = 0$. This equation is of the form

$$\frac{\partial F}{\partial n} = -\bar{p}_1 \frac{\partial F}{\partial p} + \frac{1}{2} \Delta_1^2 \frac{\partial^2 F}{\partial p^2} \tag{27}$$

(when $\bar{p}_1 = 0$ it passes into the equation (23)). After that it is easy to verify by direct substitution that the function (26) satisfies equation (27).

In Fig. 190 we have the graph of $F(p\,;\,n)$ for the case $n = 4$, $\Delta_1 = 1$, $\bar{p}_1 = 1$; for the sake of pictorialness, the scales on the axes are different.

For each specific n, the function $F(p\,;\,n)$ is a bell-shaped curve symmetric about the vertical straight line passing through the point of maximum. As is evident from (26), the maximum results when

Fig. 190

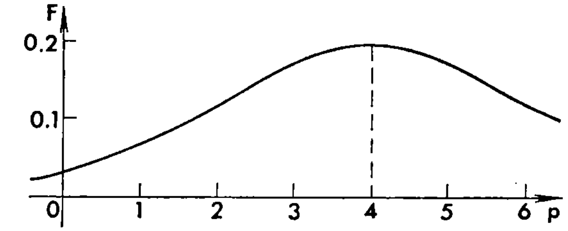
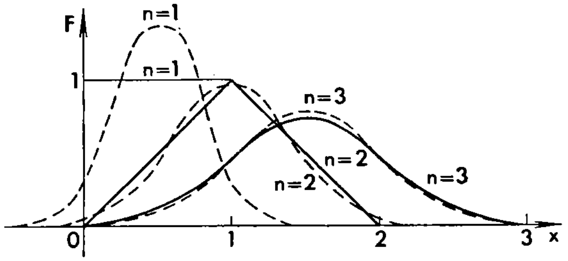


Fig. 191



$p = n\bar{p}_1$, which coincides with the mean value $\bar{p}_n$ found in Sec. 13.7. The height of the maximum is $1/\Delta_1\sqrt{2\pi n}$, i.e. it is proportional to $1/\sqrt{n}$, as has already been stated. Thus, the maximum is shifted rightwards as n increases.

Let us determine the width of the curve, i.e. let us find out how much we have to depart from $p_{\max} = n\bar{p}_1$ for the height of the curve to diminish e-fold compared to the maximum. For this purpose, we have to determine p from the condition

$$\frac{1}{\Delta_1\sqrt{2\pi n}}\, e^{-\frac{(p-n\bar{p}_1)^2}{2n\Delta_1^2}} = \frac{1}{\Delta_1\sqrt{2\pi n}}\,\frac{1}{e}$$

whence $\frac{(p-n\bar{p}_1)^2}{2n\Delta_1^2} = 1$, or $p - n\bar{p}_1 = \pm\,\Delta_1\sqrt{2n}$.

Thus, $p - p_{\max} = \pm\,\Delta_1\sqrt{2n}$, or the width of the curve is proportional to $\sqrt{n}$, as has already been pointed out. Naturally, the height of the maximum is inversely proportional to the width of the curve, as it should be when the area under the curve is preserved.

Note that even if, by its meaning, the variable under study assumes only positive values, then the function (26) produces nonzero values when $p < 0$, which is of course at variance with reality. However, when $p < 0$, $F(p;\, n)$ is so small for rather large n that this drawback is of no practical value.

Fig. 191 shows how the exact curves $F(p,\, n)$, which are obtained from formula (15) and are shown by the solid lines, approach the approximate curves, which are obtained from formula (26) for $n = 1$, $n = 2$, $n = 3$ and are shown as dashed lines.[*] These curves correspond to the case

$$f(p) = \begin{cases} 1 & \text{if } 0 < p < 1, \\ 0 & \text{if } p > 1 \text{ (see example in Sec. 13.7)} \end{cases}$$

Now let us take an example in which the random variable ξ_1 can assume only two values: 1 with a probability α $(0 < \alpha < 1)$ and 0 with a probability $\beta = 1 - \alpha$. Then the sum ξ made up of n independent values of ξ_1 may be interpreted as the number of occurrences of a certain event for n independent trials if the probability of its occurrence in each trial is equal to α. The problem of computing the probability of distinct values for ξ was solved in Secs. 13.2 and 13.3. In Sec. 13.3, with the aid of the Stirling formula, we demonstrated (formula (8)) that for large n the probability that ξ would take on a certain integer value $n\alpha + \delta$ between 0 and n is equal to

$$w(n\alpha + \delta) = \frac{1}{\sqrt{2\pi\alpha\beta n}} e^{-\delta^2/2\alpha\beta n}$$

Since for large n the distances between adjacent possible values of ξ are small in comparison with the interval of all its values and even with the interval of its expected values (see Sec. 13.3), it follows that for such n we can regard ξ as a continuous random variable with distribution density $F(p; n)$. Then the probability that ξ will assume the value $n\alpha + \delta$ may be computed approximately from the formula

$$w(n\alpha + \delta) = \int\limits_{n\alpha+\delta-\frac{1}{2}}^{n\alpha+\delta+\frac{1}{2}} F(p; n)\, dp = F(n\alpha + \delta; n)$$

Comparing the last two formulas and setting $n\alpha + \delta = p$, we get

$$F(p; n) = \frac{1}{\sqrt{2\pi\alpha\beta n}} e^{-(p-n\alpha)^2/2\alpha\beta n}$$

However, for the random variable at hand it will be true that

$$\bar{p}_1 = \bar{\xi}_1 = 1 \cdot \alpha + 0 \cdot \beta = \alpha,$$

$$\Delta_1^2 = \overline{(\xi_1 - \bar{\xi}_1)^2} = (1 - \alpha)^2 \alpha + (0 - \alpha)^2 \beta = \alpha\beta$$

If in the last expression for $F(p; n)$ we replace $\alpha\beta$ by Δ_1^2 and $n\alpha$ by $n\bar{p}_1$, we again arrive at the formula (26), thus completely proving it for random variables of the special type we are considering.

[*] In Fig. 191 it was hard to show that $F(p; 1)$ — a discontinuous function — is depicted as a step with a vertical decline at $p = 1$. The function $F(p; 2)$ is represented as a right triangle lying on the hypotenuse. The points $p = 1$, $F = 1$ on $F(p; 1)$ and $F(p; 2)$ coincide.

Suppose the variables $\xi_1, \xi_2, \ldots, \xi_n$ independently assume random values in the interval $-\infty, \infty$, each of these variables having its own probability distribution:

$$f_1(x),\ f_2(x),\ \ldots, f_n(x)$$

Their sum $\xi = \xi_1 + \xi_2 + \ldots + \xi_n$ will also take on certain random values. Let its probability distribution be $F(x)$.

Proof is given in probability theory that in this case, if n is great, we have

$$F(x) = \frac{1}{\Delta\sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2\Delta^2}} \tag{28}$$

Such a probability distribution law is called a *normal law*. To specify the formula (28), we have to indicate two quantities: $\bar{x}$ and Δ.

The factor in front of $e^{-\frac{(x-\bar{x})^2}{2\Delta^2}}$ is determined from the condition $\int\limits_{-\infty}^{+\infty} F(x)\,dx = 1$. Indeed, let $F(x) = Ae^{-\frac{(x-\bar{x})^2}{2\Delta^2}}$, then $\int\limits_{-\infty}^{+\infty} F(x)\,dx = $ $= A\int\limits_{-\infty}^{+\infty} e^{-\frac{(x-\bar{x})^2}{2\Delta^2}} dx$. Let us find the last integral. Assuming $\frac{x-\bar{x}}{\Delta\sqrt{2}} = t$, $dx = \Delta\sqrt{2}\,dt$, we get $\int\limits_{-\infty}^{+\infty} e^{-\frac{(x-\bar{x})^2}{2\Delta^2}} dx = \Delta\sqrt{2}\int\limits_{-\infty}^{+\infty} e^{-t^2} dt = \Delta\sqrt{2\pi}$. From the condition $\int\limits_{-\infty}^{+\infty} F(x)\,dx = 1$ we find $A \cdot \Delta\sqrt{2\pi} = 1$, whence $A = \frac{1}{\Delta\sqrt{2\pi}}$.

We suggest that the reader verify (see Exercises) that for $F(x)$ determined by the formula (28) the relations

$$\int\limits_{-\infty}^{+\infty} xF(x)\,dx = \bar{x}, \quad \int\limits_{-\infty}^{+\infty} (x-\bar{x})^2 F(x)\,dx = \Delta^2$$

hold true, that is, $\bar{x}$ and Δ^2 are the mean value and the variance of the random variable ξ.

Earlier, all these relations were derived for the particular case of the sum of n variables with the same probability distribution $f(p)$.

Exercises

1. Find $\int\limits_{-\infty}^{+\infty} xF(x)\,dx$ and $\int\limits_{-\infty}^{+\infty} (x-\bar{x})^2 F(x)dx$, where $F(x)$ is given by the formula (28).

2. $$f(p) = \begin{cases} \frac{1}{q} & \text{if } 0 < p < q, \\ 0 & \text{if } p > q \text{ and if } p < 0 \end{cases}$$

Use formula (26) to find the probability distribution $F(p;\ n)$ for the case $n = 10$, $n = 20$.

Hint. First find the quantities $\bar{p}_1$ and Δ_1^2.

3. A fisherman is catching fish in a pond containing fish weighing not more than 2 kg. Suppose there is an equal probability of catching a fish weighing between 0 and 2 kg for each cast of the line.

(a) What is the mean weight of a catch consisting of casting the line 20 times in succession?

(b) What is the probability that in casting the line 20 times the total catch will not exceed 20 kg? 22 kg? 25 kg?

13.9 Correlational dependence

The theory of probability has important applications to the study of relationships between quantities. Up to this point we have considered only functional relationships, that is, such that specification of one set of values completely determines another set of values. However, let us consider, say, such quantities as the length l and the weight p of a fish that has been caught. It is natural to expect that the longer the length l, the greater p is; that is, that there is a relationship here. But this relationship is accomplished only in the mean, since the length l of a fish does not uniquely determine the weight p, for fish of the same length may have different weights: a long fish may turn out to be so skinny that its weight will be less than that of a short fish, and so on.

This type of "flexible" dependence between quantities that is accomplished in the mean is called *correlational dependence,* to distinguish it from the "rigid" or "inflexible" functional dependence. Instances of correlational dependence are the relationship between the age of a person and his height, between the knowledge of a student and the mark he gets at an examination, and the like. One also hears of a correlation between luck at cards and luck in love.

A correlation obtains when the effect is felt of factors that cannot be taken into account due, say, to the involved nature of their influence. However, such a dependence may be the result of necessity, irrespective of the complexity of the factors. The point is that if, say, two quantities x and y are functionally dependent on a single parameter, $x = x(t)$ and $y = y(t)$, then these two relations define a rigid functional relationship between x and y (see page 121). But if there are two or more such parameters, that is

$$x = x(t_1, t_2, \ldots, t_n),\quad y = y(t_1, t_2, \ldots, t_n)$$

then these relations *fundamentally* can no longer determine a functional relationship between x and y. And only if we take into account the frequency of realization of the different combinations of values of the parameters $t_1, t_2, \ldots, t_n$ can we speak of a relationship between x and y, which, however, will only be a correlation. Such precisely is the situation for the quantities l and p above, where, incidentally, it would be difficult to indicate the whole set of essential parameters.

An important instance of correlational dependence is found in curve fitting (Secs. 2.3, 2.4). Even if the true relationship $y = f(x)$ between the physical quantities x and y is a functional relationship, the relationship between the measured values of x and y, given errors of measurement, is only a correlation.

Let us consider a general case. Suppose we have two random variables ξ and η that are related in some fashion. As we have demonstrated above, if these variables are independent, then $\overline{\xi\eta} = \overline{\xi}\,\overline{\eta}$, that is, $\overline{\xi\eta} - \overline{\xi}\,\overline{\eta} = 0$. For this reason, the latter difference may be taken for a crude measure of the relationship between ξ and η. However, one more often passes to a nondimensional *correlation coefficient*:

$$r = r(\xi, \eta) = \frac{\overline{\xi\eta} - \overline{\xi}\,\overline{\eta}}{\Delta_\xi \Delta_\eta}$$

where the denominator consists of the mean square deviations of ξ and η.

By what has already been stated, this coefficient is equal to zero for independent variables. (Generally, the converse is not true.) Now let us consider another extreme case where the relationship between ξ and η is functional and also linear, i.e., $\eta = a\xi + b$, where a and b are constants. Then it is easy to verify that $\Delta_\eta^2 = a^2\Delta_\xi^2$ or $\Delta_\eta = |a|\Delta_\xi$. From the equation $\Delta_\xi^2 = \overline{(\xi - \overline{\xi})^2} = \overline{\xi^2} - 2\overline{\xi\overline{\xi}} + \overline{\xi}^2 = \overline{\xi^2} - \overline{\xi}^2$ it follows that

$$\frac{\overline{\xi\eta} - \overline{\xi}\,\overline{\eta}}{\Delta_\xi \Delta_\eta} = \frac{(\overline{a\xi^2 - b\xi}) - \overline{\xi}(\overline{a\xi + b})}{\Delta_\xi\,|a|\,\Delta_\xi} = \frac{a\overline{\xi^2} + b\overline{\xi} - a\overline{\xi}^2 - b\overline{\xi}}{|a|\,\Delta_\xi^2} = \frac{a}{|a|}$$

That is, $r(\xi, \eta) = 1$ if $a > 0$, and $r(\xi, \eta) = -1$ if $a < 0$.

It can be demonstrated (though we will not do so here) that in all cases except that of a linear functional relationship, the coefficient of correlation lies *strictly* between -1 and 1, $-1 < r(\xi, \eta) < 1$. This coefficient characterizes the degree to which the relationship between ξ and η departs from a linear functional relationship.

Let us take an example. Suppose the physical quantities x and y are related linearly, $y = ax + b$, but the values of the coefficients a and b are unknown and are obtained by an experiment in which we specify values of x and measure values of y. For the sake of simplicity we will assume that the values of x are known with utmost precision and the degree of accuracy in determining the corresponding value of y is the same for all x. Denoting by ξ the measured value of x and by η the measured value of y (ξ, η are random variables), we assume that with equal probability ξ takes on all values between $-l$ and l, which means the appropriate distribution function is of the form

$$\varphi(x) = \begin{cases} \dfrac{1}{2l} & (-l \leqslant x \leqslant l), \\ 0 & (|x| > l) \end{cases} \tag{29}$$

We also assume that for every value $\xi = x$ the variable η is normally distributed about the value $ax + b$ with variance Δ^2 that does not depend on x. (It is natural to assume the normal distribution law if the error in determining y is made up of a large number of independent errors due to a variety of reasons.) We denote by $\psi_x(y)$ the distribution function of η for a given value $\xi = x$ so that

$$\psi_x(y) = \frac{1}{\Delta\sqrt{2\pi}} e^{-[y-(ax+b)]^2/2\Delta^2}$$

Furthermore, denote by $f(x, y)$ the *joint distribution function* of the variables ξ, η; this means that the probability of a simultaneous occurrence of ξ lying between x and $x + dx$ and η lying between y and $y + dy$ is equal to $f(x, y)\, dx\, dy$. To calculate the function $f(x, y)$ note that if a large number N of trials has been carried out, then the number of simultaneous occurrences of ξ lying between x and $x + dx$ and of η lying between y and $y + dy$ is equal to $[f(x, y)\, dx\, dy]N$. On the other hand, we can reason as follows: for N trials the number of cases where ξ lies between x and $x + dx$ is equal to $N_1 = [\varphi(x)\, dx]\, N$. But of these N_1 cases, the number of cases where η lies between y and $y + dy$ is equal to

$$[\psi_x(y)\, dy]\, N_1 = \psi_x(y)\, \varphi(x)\, dx\, dy \cdot N$$

Equating both results we find that $f(x, y) = \varphi(x)\, \psi_x(y)$. That is, under the assumptions we have made,

$$f(x, y) = \begin{cases} \dfrac{1}{2l\Delta\sqrt{2\pi}} e^{-[y-(ax+b)]^2/2\Delta^2} & (|x| \leqslant l), \\ 0 & (|x| > l) \end{cases} \tag{30}$$

Let us compute the correlation coefficient $r(\xi, \eta)$. For reasons of symmetry it is clear that $\bar{\xi} = 0$, $\bar{\eta} = b$. The mean value $\overline{\xi\eta}$ is computed by the general rule (the sum of the products of the values of the random variable into their probabilities), that is,

$$\overline{\xi\eta} = \iint xyf(x, y)\, dx\, dy, \text{ or}$$

$$\overline{\xi\eta} = \int_{-l}^{l} dx \int_{-\infty}^{\infty} \frac{xy}{2l\Delta\sqrt{2\pi}} e^{-[y-(ax+b)]^2/2\Delta^2} dy = \frac{1}{2l}\int_{-l}^{l} x(ax + b)\, dx = \frac{l^2}{3}\, a$$

From the formulas (29) we get $\Delta_\xi^2 = \int_{-l}^{l} x^2 \frac{1}{2l}\, dx = \frac{l^2}{3}$. In order to compute Δ_η we first find the distribution function $\psi(y)$ of η (this is not the same thing as $\psi_x(y)$!). Since $\psi(y)\, dy$ is the probability that η will lie between y and $y + dy$ and ξ will assume an arbitrary value, it follows that $\psi(y)\, dy = \int_{-\infty}^{\infty} [f(x, y)\, dy]\, dx$. And so

$$\psi(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_{-l}^{l} \frac{1}{2l\Delta\sqrt{2\pi}} e^{-[y-(ax+b)]^2/2\Delta^2} dx$$

whence

$$\Delta_\eta^2 = \int_{-\infty}^{\infty} (y - b)^2\, \psi(y)\, dy = \int_{-\infty}^{\infty} dy \int_{-l}^{l} (y - b)^2 \frac{1}{2l\Delta\sqrt{2\pi}} e^{-[y-(ax+b)]^2/2\Delta^2} dx$$

After inverting the order of integration and carrying out the computations, which we leave to the reader, we get $\Delta_\eta^2 = \Delta^2 + \frac{l^2a^2}{3}$. From this we find the desired coefficient

$$r(\xi, \eta) = \frac{l^2a}{3\sqrt{\frac{l^2}{3}\left(\Delta^2 + \frac{l^2a^2}{3}\right)}} = \frac{\frac{al}{\Delta}}{\sqrt{\left(\frac{al}{\Delta}\right)^2 + 3}}$$

It is quite evident that $|r| < 1$; as $\frac{al}{\Delta} \to \infty$ (for instance, when $a =$ constant $\neq 0$, $l =$ constant, $\Delta \to 0$), r will tend to ± 1, which means the correlation is extremal; when $\frac{al}{\Delta} \to 0$ (for example, for $a =$ constant, $b =$ constant, $\Delta \to \infty$), then $r \to 0$, and the correlation is lost.

Suppose an experiment is carried out with N measurements made, then for the values $x = x_1, x_2, \ldots, x_N$ we obtain the corresponding values $y = y_1, y_2, \ldots, y_N$. If the measurements are independent, then such a set of values has a joint distribution function equal, by virtue of (30), to

$$f(x_1, y_1) \cdot f(x_2, y_2) \ldots f(x_N, y_N) = \frac{1}{(2l\Delta\sqrt{2\pi})^N} e^{-\Sigma[y_i-(ax_i+b)]^2/2\Delta^2} \quad (31)$$

(Here the summation Σ is taken with respect to i from 1 to N). If we know the result of the experiment and we proceed from the linear relationship $y = ax + b$, but with coefficients a and b unknown, then it is natural to choose them so that this result should be in the range of the greatest possible distribution density, that is, that it should be the most probable, in a certain sense. This means that the coefficients a and b are chosen from the condition of maximizing the right member of (31), that is to say, minimizing the sum $\sum[y_i - (ax_i + b)]^2$. We arrive at the method of least squares that has already been discussed in Sec. 2.3. Thus, the method of least squares is substantiated by arguments of probability theory.

After the foregoing experiment we can compute the correlation coefficient from the formula (note that $\overline{\xi\eta} - \bar{\xi}\bar{\eta} = \overline{(\xi - \bar{\xi})(\eta - \bar{\eta})}$)

$$r = \frac{\overline{(\xi - \bar{\xi})(\eta - \bar{\eta})}}{\sqrt{\overline{(\xi - \bar{\xi})^2}\,\overline{(\eta - \bar{\eta})^2}}} \approx \frac{\sum(x_i - \bar{x})(y_i - \bar{y})/N}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2/N^2}}$$

where $x = \sum x_i/N$, $\overline{y} = \sum y_i/N$. Multiplying the numerator and denominator by N, we get

$$r \approx r_N = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

If $|r_N|$ turns out to be small, the suspicion may arise that indeed $r = 0$ and the fact that r_N differs from zero arose due to the natural spread of experimental values. It can be demonstrated (though we will not give the proof) that if the true value of $r = 0$, then for large N the observed value of r_N approximately obeys the normal law with mean value $\bar{r}_N = 0$ and variance $\Delta_N^2 = \frac{1}{N-1}$. Arguing as in Sec. 13.3, we find that the probability of the inequality $|r_N| < \varepsilon$ is equal to

$$\int_{-\varepsilon}^{\varepsilon} \frac{1}{\Delta_N\sqrt{2\pi}} e^{-r^2/2\Delta_N^2}\, dr = \Phi\left(\frac{\varepsilon}{\Delta_N}\right) = \Phi(\varepsilon\sqrt{N-1})$$

Equating, for example, the right side to 0.95, from the table we find that with the probability 0.95 we have $|r_N| < \varepsilon$, where $\varepsilon\sqrt{N-1} = 1.96 \approx 2$. Thus, if experiment yields $|r_N| \geqslant \frac{2}{\sqrt{N-1}}$, then with probability at least 0.95 we can assert that the correlation coefficient between the physical quantities under study is different from zero, that is, these quantities are connected by a correlation.

Note in conclusion that if the correlation coefficient turns out to be zero, then this does not yet imply the absence of a relationship, for this relationship may turn out to be nonlinear. Therefore, any empirical eliciting of the relationship $y(x)$ should always begin with a plotting of the experimental results on a coordinate plane xy. It may turn out that the empirical points will closely enough approximate a parabola or some other simple curve, which will then determine the functional relationship $y(x)$. Only if the empirical points form a "cloud" is it necessary to treat them statistically, in which case it is advisable to apply the notion of correlation if the "cloud" resembles a linear relationship, like that shown in Fig. 12. A non-linear relationship can be investigated over small ranges of x by approximately assuming it to be linear, or linearity can be sought with the aid of a preliminary transformation of the variables by the methods given in Sec. 2.4.

13.10 On the distribution of primes

Here is an interesting example of the application of probability theory to number theory. We consider the question of the distribution of prime numbers among the natural numbers. This is a very complicated question and so our reasoning will not be rigorous, though it will give a good general picture of the subject.

In ancient times it had already been proved that there does not exist a greatest prime number: hence, there is an infinity of primes. Even a cursory glance at the first several hundred prime numbers of the positive integers convinces us that the primes are distributed in a highly irregular fashion. For example, following the prime 113 come 13 composite numbers, the fourteenth number (127) being prime. There are only three composite numbers between the prime 127 and the prime 131. There are five numbers between the primes 131 and 137, and then the next prime is 139.

And so we have the interesting question of how many primes there are between the numbers n and $n + \Delta$. We will use the term distribution density of primes for the ratio of the number of primes between n and $n + \Delta$ to the number Δ. This section will be devoted to determining the dependence of the distribution density of primes on n.

We consider Δ to be small in comparison with n but substantially greater than 1. (We thus assume that n is a sufficiently large number.) Under this assumption there are many primes between n and $n + \Delta$. We denote the distribution density of primes by $f(n)$. Then there will be $f(n)\Delta$ primes in the interval between n and $n + \Delta$.

We can make a list of all prime numbers not exceeding n in the following manner.

Write down all the natural numbers from 2 to n. Then cross out all numbers divisible by 2 (not counting the number 2 itself), then cross out all numbers divisible by 3, except 3 itself, then all numbers divisible by 5, except 5, and so forth. In this way we cross out all composite numbers and leave only primes. This method goes by the name of the *sieve of Eratosthenes.*

Clearly, 1/2 of all the numbers from 2 to n are divisible by 2, only 1/3 of these numbers are divisible by 3, 1/5 by 5, etc. And so if we are considering the natural numbers from 2 to n, then $\left(\frac{1}{p}\right)$ th part of them is divisible by the prime number p and $\left(1 - \frac{1}{p}\right)$ th part is not divisible by p. Here we assume that $p \ll n$.

Suppose we have two primes, p_1 and p_2. Take any natural number n. It may or may not be divisible by p_1. The same goes for p_2. We assume that the two events — the first being that n is not divisible by p_1 and the second that n is not divisible by p_2 — are independent.

Of the numbers from 2 to n, the portion of numbers that are not divisible by 2 is equal to $\left(1 - \frac{1}{2}\right)$, the portion of numbers that are not divisible by 3 is equal to $\left(1 - \frac{1}{3}\right)$, the portion of numbers that are not divisible by 5 is equal to $\left(1 - \frac{1}{5}\right)$, and, generally, the portion of numbers that are not divisible by the prime p is $\left(1 - \frac{1}{p}\right)\cdot$ Since we assume as independent the events consisting in the fact that a number is not divisible by various smaller primes, then by the law of compound probabilities the portion of primes located between 2 and n is equal to

$$f(n) = \left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{3}\right)\left(1 - \frac{1}{5}\right)\cdots\left(1 - \frac{1}{p}\right) \qquad (32)$$

(We equated this portion to the distribution density of primes $f(n)$ on the basis of the approximate equality $F(n)/n \approx F'(n)$ for large $F(n)$ and n.)

If $p > 1$, then $e^{-\frac{1}{p}} \approx 1 - \frac{1}{p}$ (we disregard the terms $1/p^2$, $1/p^3$, ...). Therefore

$$\left.\begin{array}{l} 1 - \frac{1}{2} \approx e^{-\frac{1}{2}}, \\ 1 - \frac{1}{3} \approx e^{-\frac{1}{3}}, \\ \cdots\cdots\cdots\cdots \\ 1 - \frac{1}{p} \approx e^{-\frac{1}{p}} \end{array}\right\}$$

The formula (32) takes the form $f(n) = e^{-\frac{1}{2}} \cdot e^{-\frac{1}{3}} \dots e^{-\frac{1}{p}}$, whence

$$\ln f(n) = \left(-\frac{1}{2}\right) + \left(-\frac{1}{3}\right) + \dots + \left(-\frac{1}{p}\right) = -\sum_i \frac{1}{p_i} \tag{33}$$

In the interval between ν and $\nu + d\nu$ (where $d\nu$ is small compared with ν) there are $f(\nu)d\nu$ prime numbers. In view of the fact that the interval is short in length, we can assume that all primes in this interval are approximately equal to ν and therefore $\frac{1}{p_i}$ is $\frac{1}{\nu}$ for this interval. Then for numbers in the interval from ν to $\nu + d\nu$, $\sum \frac{1}{p_i}$ takes the form $\frac{1}{\nu} f(\nu)\, d\nu$. Consequently, for the sum $\sum \frac{1}{p_i}$ that corresponds to the interval from 2 to n we get

$$\sum_i \frac{1}{p_i} = \int_a^b \frac{f(\nu)}{\nu}\, d\nu$$

What are the limits of integration in this integral? We are interested in all primes not exceeding n, and so we take n for the upper limit of integration. Note that the equation $1 - \frac{1}{p_i} = e^{-\frac{1}{p_i}}$ is good only for large p_i. If p_i is small, then the last formula is noticeably in error. For example, for $p = 2$ we get $1 - \frac{1}{2} = 0.5$, $e^{-\frac{1}{2}} = 0.61$ (the error exceeds 20%). Replacing the sum by an integral is also good only for intervals where p_i is sufficiently great. It is therefore impossible to indicate the lower limit of integration. We will leave the lower limit of integration as being indefinite. Formula (33) takes the form

$$\ln f(n) = -\int_a^n \frac{f(\nu)}{\nu}\, d\nu \tag{34}$$

Taking the derivative of both sides of this equation, we get $\frac{1}{f(n)} \frac{df(n)}{dn} = -\frac{f(n)}{n}$. We have an equation with variables separable that is easy to solve. Rewrite it thus:

$$\frac{df(n)}{f^2(n)} = -\frac{dn}{n}$$

Integrate both sides from n_0 to n to get

$$\int_{n_0}^{n} \frac{df}{f^2} = -\int_{n_0}^{n} \frac{dn}{n} \quad \text{or} \quad -\frac{1}{f(n)} + \frac{1}{f(n_0)} = -\ln n + \ln n_0$$

whence

$$\frac{1}{f(n)}=\frac{1}{f(n_0)}-\ln n_0+\ln n=C+\ln n$$

where $C=\frac{1}{f(n_0)}-\ln n_0$

We finally get

$$f(n)=\frac{1}{C+\ln n} \tag{35}$$

In the case of large n, the constant C may be neglected compared with $\ln n$. And so for very large n we have

$$f(n)=\frac{1}{\ln n} \tag{36}$$

Our derivation was extremely crude and inaccuracies in it can easily be pointed out. For example, was it necessary to test all primes less than n for divisibility? Of course not. Actually, it would suffice to test only those primes that do not exceed $\sqrt{n}$. Suppose n is not divisible by any prime less than $\sqrt{n}$. Suppose n is divisible by the prime $p_1 > \sqrt{n}$. Then $n/p_1 = k$ and since $p_1 > \sqrt{n}$, it follows that $k < \sqrt{n}$. Note that $n = p_1k$, whence $n/k = p_1$. Consequently, n is divisible by $k < \sqrt{n}$, which runs counter to the hypothesis.

It turns out (if we reject all terms except the principal one) that (36) also satisfies the modified relation (34) in which the upper limit n of the integral is replaced by $\sqrt{n}$. On the basis of entirely different ideas, it has been demonstrated (in a much more involved manner) that the accuracy of formula (36) for large n is very great and better than could have followed from our reasoning. In particular, it turned out that (35) offers the best asymptotic representation of $f(n)$ precisely for $C = 0$.

Very often the question investigated is not the density of primes but the number of primes $A(n)$ not exceeding a given number n. It is clear that $A(n)=\int_a^n f(\nu)\,d\nu$, where the lower limit of integration a is not known. Thus

$$A(n)=\int_a^n \frac{d\nu}{\ln \nu}$$

Put $\nu = ny$, $d\nu = n\,dy$ here; then

$$A(n)=n\int_{\frac{a}{n}}^{1}\frac{dy}{\ln n+\ln y}$$

Note that for large n

$$\frac{1}{\ln n + \ln y} = \frac{1}{\ln n} \cdot \frac{1}{1 + \dfrac{\ln y}{\ln n}} = \frac{1}{\ln n}\left[1 - \frac{\ln y}{\ln n} + \frac{\ln^2 y}{\ln^2 n} - \dots\right]$$

And so

$$A(n) = \frac{n}{\ln n}\int\limits_{\frac{a}{n}}^{1}\left[1 - \frac{\ln y}{\ln n} + \frac{\ln^2 y}{\ln^2 n} - \dots\right] dy$$

Confining ourselves to the first term in the expansion and replacing $\frac{a}{n}$ by 0 when n is large, we get

$$A(n) = \frac{n}{\ln n} \tag{37}$$

If we take a large number of terms, the formula becomes more exact. For example, confining ourselves to three terms of the expansion, we get

$$A(n) = \frac{n}{\ln n}\left[1 + \frac{1}{\ln n} + \frac{2}{\ln^2 n}\right]^{*} \tag{38}$$

It turns out that the error of any one of such formulas is asymptotically small compared with the last "exact" term.

Exercises

1. Approximate the number of primes lying in the interval between 3000 and 3100, between 3000 and 3200, and between 3000 and 3500. Compare the result with the true figure by counting the number of primes in a table of prime numbers.
2. Determine the number of primes less than 4000 by the formula (37). Refine the result by retaining in (38) the term on the right containing $\frac{1}{\ln n}$, and then the term containing $\frac{1}{(\ln n)^2}$. Compare the result with the exact figure obtained by counting the number of primes less than 4000 in a table of primes.
3. Calculate the number of primes in the interval between 2000 and 5000. Perform the calculations in different ways:
(a) by assuming that $n = 2000$, $\Delta = 3000$;
(b) by finding the difference between $A(5000)$ and $A(2000)$. Compute these quantities using (37) and then, more exactly, by retaining the term containing $\frac{1}{\ln n}$.

[*] Here we had to find $\int\limits_0^1 \ln y\, dy$ and $\int\limits_0^1 \ln^2 y\, dy$. Both integrals can readily be calculated by integration by parts. Note that $y \ln^k y = 0$ for $y = 0$ for any k.

Table of the Probability Integral

$$\Phi(x) = \frac{2}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}}\, dt$$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.00	0.000						
01	0.008	0.36	0.281	0.71	0.522	1.06	0.711
02	0.016	37	0.289	72	0.528	07	0.715
03	0.024	38	0.296	73	0.535	08	0.720
04	0.032	39	0.303	74	0.541	09	0.724
05	0.040	40	0.311	75	0.547	10	0.729
06	0.048	41	0.318	76	0.553	11	0.733
07	0.056	42	0.326	77	0.559	12	0.737
08	0.064	43	0.333	78	0.565	13	0.742
09	0.072	44	0.340	79	0.570	14	0.746
10	0.080	45	0.347	80	0.576	15	0.750
11	0.088	46	0.354	81	0.582	16	0.754
12	0.096	47	0.362	82	0.588	17	0.758
13	0.103	48	0.369	83	0.593	18	0.762
14	0.111	49	0.376	84	0.599	19	0.766
15	0.119	50	0.383	85	0.605	20	0.770
16	0.127	51	0.390	86	0.610	21	0.774
17	0.135	52	0.397	87	0.616	22	0.778
18	0.143	53	0.404	88	0.621	23	0.781
19	0.151	54	0.411	89	0.627	24	0.785
20	0.159	55	0.418	90	0.632	25	0.789
21	0.166	56	0.425	91	0.637	26	0.792
22	0.174	57	0.431	92	0.642	27	0.796
23	0.182	58	0.438	93	0.648	28	0.799
24	0.190	59	0.445	94	0.653	29	0.803
25	0.197	60	0.451	95	0.658	30	0.806
26	0.205	61	0.458	96	0.663	31	0.810
27	0.213	62	0.465	97	0.668	32	0.813
28	0.221	63	0.471	98	0.673	33	0.816
29	0.228	64	0.478	99	0.678	34	0.820
30	0.236	65	0.484	1.00	0.683	35	0.823
31	0.243	66	0.491	01	0.687	36	0.826
32	0.251	67	0.497	02	0.692	37	0.829
33	0.259	68	0.503	03	0.697	38	0.832
34	0.266	69	0.510	04	0.702	39	0.835
35	0.274	70	0.516	05	0.706	40	0.838

Continued

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1.41	0.841	1.66	0.903	1.91	0.944	2.55	0.989
42	0.844	67	0.905	92	0.945	60	0.991
43	0.847	68	0.907	93	0.946	65	0.992
44	0.850	69	0.909	94	0.948	70	0.993
45	0.853	70	0.911	95	0.949	80	0.995
46	0.856	71	0.913	96	0.950	90	0.996
47	0.858	72	0.915	97	0.951	3.00	0.997
48	0.861	73	0.916	98	0.952	10	0.998
49	0.864	74	0.918	99	0.953	20	0.999
50	0.866	75	0.920	2.00	0.954	30	0.999
51	0.869	76	0.922	05	0.960	40	0.999
52	0.871	77	0.923	10	0.964	3.50	$1-0.5\cdot 10^{-3}$
53	0.874	78	0.925	15	0.968	3.9	$1-10^{-4}$
54	0.876	79	0.927	20	0.972	4.4	$1-10^{-5}$
55	0.879	80	0.928	25	0.976	4.9	$1-10^{-6}$
						5.3	$1-10^{-7}$
56	0.881	81	0.930	30	0.979		
57	0.884	82	0.931	35	0.981		
58	0.886	83	0.933	40	0.984		
59	0.888	84	0.934	45	0.986		
60	0.890	85	0.936	50	0.988		
61	0.893	86	0.937				
62	0.895	87	0.939				
63	0.897	88	0.940				
64	0.899	89	0.941				
65	0.901	90	0.943				

ANSWERS AND SOLUTIONS

Sec. 13.1

$$\frac{9}{36}=\frac{1}{4},\quad \frac{4}{36}=\frac{1}{9},\quad \frac{1}{36},\quad \frac{8}{35}.$$

Sec. 13.2

1. $\frac{1}{4}$. 2. $\frac{1}{8}$, $\frac{3}{8}$.

3. The probability of black is equal to

$$w_{\mathrm{b}} = \frac{4}{6} = \frac{2}{3}$$

and the probability of white is

$$w_{\mathrm{w}} = \frac{1}{3}$$

Since $w_{\mathrm{ww}} = w_{\mathrm{w}} \cdot w_{\mathrm{w}}$, it follows that $w_{\mathrm{ww}} = \frac{1}{9}$. Similarly

$$w_{\mathrm{bb}} = w_{\mathrm{b}} \cdot w_{\mathrm{b}} = \frac{4}{9}$$

The probability that white turns up once and black the next time (the order in which they appear is immaterial) can be computed from the general formula $\frac{n!}{m!k!}\alpha^m\beta^k$. We get $\frac{4}{9}$.

4. Using the general formula we find $\frac{2}{9}$, $\frac{4}{9}$.

5. 0.18. **6.** 0.243, 0.027. **7.** 0.29, 0.33.

Sec. 13.3

1. Using formula (3) and setting $n = 1000$, $\delta = 0$, and then $n = 1000$, $\delta = 10$, we obtain 0.025, 0.020.

2. The probability of a certain number of heads not exceeding 500 is equal to

$$w = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} e^{-\frac{t^2}{2}}\,dt = \frac{1}{\sqrt{2\pi}}\int_{0}^{+\infty} e^{-\frac{t^2}{2}}\,dt = \frac{1}{2}\,\frac{2}{\sqrt{2\pi}}\int_{0}^{\infty} e^{-\frac{t^2}{2}}\,dt = \frac{1}{2}$$

The probability of obtaining at least 500 heads is

$$1 - \frac{1}{2} = \frac{1}{2}$$

The probability of obtaining not more than 510 heads ($\delta = 10$) is

$$w = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{2}{\sqrt{10}}} e^{-\frac{t^2}{2}}dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0.63} e^{-\frac{t^2}{2}}\,dt = \frac{1}{\sqrt{2\pi}}\left(\int_{-\infty}^{0} e^{-\frac{t^2}{2}}\,dt + \int_{0}^{0.63} e^{-\frac{t^2}{2}}\,dt\right)$$

$$= \frac{1}{2}\,\frac{2}{\sqrt{2\pi}}\int_{0}^{\infty} e^{-\frac{t^2}{2}}\,dt + \frac{1}{2}\,\frac{2}{\sqrt{2\pi}}\int_{0}^{0.63} e^{-\frac{t^2}{2}}\,dt = 0.5 + \frac{1}{2}\cdot 0.471 = 0.74$$

Therefore the probability of obtaining at least 510 heads is $1 - 0.74 = 0.26$.

3. Putting in (8) $\delta = 0$ (10 hits), we get $w = 0.13$, putting $\delta = -2$ (8 hits), we get $w = 0.11$.

4. Denote by δ the deviation of the number of hits from the most probable number. Then the probability that one of the events for which $\delta \leqslant \delta_0$ is realized is equal to

$$w = \frac{1}{\sqrt{2\pi n\alpha\beta}} \int\limits_{-\infty}^{\delta_0} e^{-\frac{\delta^2}{2n\alpha\beta}}\, d\delta$$

Putting $\frac{\delta}{\sqrt{n\alpha\beta}} = t$, $d\delta = \sqrt{n\alpha\beta}\, dt$ here, we get $w = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{t_0} e^{-\frac{t^2}{2}}\, dt,$

where $t_0 = \frac{\delta_0}{\sqrt{n\alpha\beta}}$. In this case, $n = 100$, $\alpha = 0.1$, $\beta = 0.9$. The most probable number of hits is $n\alpha = 100 \cdot 0.1 = 10$. Compute the probability that the number of hits does not exceed 8 ($\delta = -2$). We get

$$w = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{-0.67} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int\limits_{0.67}^{+\infty} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \left[\int\limits_{0}^{+\infty} e^{-\frac{t^2}{2}} dt - \int\limits_{0}^{0.67} e^{-\frac{t^2}{2}} dt \right]$$

$$= 0.5 - 0.5\, \Phi(0.67) = 0.25$$

Hence the probability that the number of hits is not less than 8 is equal to $1 - 0.25 = 0.75$.

For the other two cases the probabilities are respectively 0.5 and 0.25.

5. 0.73, 0.37.

Sec. 13.5

1. $\frac{w_\mu(m)}{w_\mu(m-1)} = \frac{\mu^m e^{-\mu}(m-1)!}{m!\, \mu^{m-1}\, e^{-\mu}} = \frac{\mu}{m}$, whence $w_\mu(m) > w_\mu(m-1)$ as

long as $m < \mu$, and $w_\mu(m) < w_\mu(m-1)$ when $m > \mu$. Hence, if μ is nonintegral, then $w_\mu(m)$ is the largest when m is equal to the integral part of μ. But if $\mu = 1, 2, \ldots$, then the largest are $w_\mu(\mu - 1) = w_\mu(\mu)$.

2. If a large number N of trials is carried out, the variable x assumes the value x_1 a total of $p_1 N$ times, the value x_2, $p_2 N$ times, ..., and the value x_n, $p_n N$ times. Therefore the mean value (per trial) is

$$\bar{x} = \frac{p_1 N \cdot x_1 + p_2 N \cdot x_2 + \ldots + p_n N \cdot x_n}{N} = p_1 x_1 + p_2 x_2 + \ldots + p_n x_n$$

For the Poisson law this yields

$$\overline{m} = \sum_{m=0}^{\infty} \frac{\mu^m}{m!} e^{-\mu} m = e^{-\mu} \sum_{m=1}^{\infty} \frac{\mu^m}{(m-1)!} = e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^{k+1}}{k!} = e^{-\mu} \mu e^{\mu} = \mu$$

where $m - 1 = k$.

Sec. 13.6

The system of equations is of the form

$$\frac{dp_i}{dt} = -(\alpha + \omega) p_i + \omega p_{i-1} + \alpha p_{i+1} \quad (i = \ldots, -2, -1, 0, 1, 2, \ldots)$$

This system has to be solved for the initial condition $p_0(0) = 1$, $p_i(0) = 0 (i \neq 0)$. When using the method of succesive approximation it is convenient first to make a change of variable, $p_i = e^{-(\alpha+\omega)t} q_i$, whence

$$\frac{dq_i}{dt} = \omega q_{i-1} + \alpha q_{i+1} \quad (i = \ldots, -2, -1, 0, 1, 2, \ldots)$$

The zero-order approximation is

$q_0 = 1$, and the others are $q_i = 0$.

The first approximation is

$q_{-1} = \alpha t$, $q_0 = 1$, $q_1 = \omega t$, and the others are $q_i = 0$.

The second approximation is

$q_{-2} = \frac{\alpha^2 t^2}{2}$, $q_{-1} = \alpha t$, $q_0 = 1 + \alpha\omega t^2$, $q_1 = \omega t$, $q_2 = \frac{\omega^2 t^2}{2}$, and the others are $q_i = 0$, and so forth. The mean value of the number of a state i $= (\omega - \alpha) t$. The problem discussed in the text is obtained for $\alpha = 0$, $\omega = 1$.

Sec. 13.7

$$F_3(p) = \begin{cases} \frac{p^2}{2q^3} & \text{if } 0 < p < q, \\ \frac{1}{2q^3} (-2p^2 + 6pq - 3q^2) & \text{if } q < p < 2q, \\ \frac{1}{2q^3} (p^2 - 6pq + 9q^2) & \text{if } 2q < p < 3q \end{cases}$$

If $p < 0$ and if $p > 3q$, then $F_3(p) \equiv 0$. The graph of the function $F_3(p)$ for $q = 1$ is shown in Fig. 191.

Sec. 13.8

1. $$\int\limits_{-\infty}^{+\infty} xF(x)\,dx = \frac{1}{\Delta\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} x\cdot e^{\frac{(x-\bar{x})^2}{2\Delta^2}}\,dx.$$

Put $\frac{x-\bar{x}}{\Delta\sqrt{2}} = t$, $dx = \Delta\sqrt{2}\,dt$. Then the last integral becomes

$$\frac{1}{\sqrt{\pi}}\int\limits_{-\infty}^{+\infty}(\bar{x}+\Delta\sqrt{2}t)\,e^{-t^2}\,dt = \bar{x} + \frac{\Delta\sqrt{2}}{\sqrt{\pi}}\int\limits_{-\infty}^{+\infty} te^{-t^2}\,dt$$

The integral $\int\limits_{-\infty}^{+\infty} te^{-t^2}\,dt = 0$ because the integrand is odd so that its graph is symmetric about the origin.

To find $\int\limits_{-\infty}^{+\infty}(x-\bar{x})^2\frac{1}{\Delta\sqrt{2\pi}}e^{-\frac{(x-\bar{x})^2}{2\Delta^2}}\,dx$ make the change of variable $\frac{x-\bar{x}}{\Delta\sqrt{2}} = t$, $dx = \Delta\sqrt{2}\,dt$. Then the original integral becomes $\frac{2\Delta^2}{\sqrt{\pi}}\int\limits_{-\infty}^{+\infty} t^2e^{-t^2}\,dt$. Integrate by parts and put $te^{-t^2}\,dt = dg$, $t = f$.

Then $g = -\frac{1}{2}e^{-t^2}$, $df = dt$. We get

$$\frac{2\Delta^2}{\sqrt{\pi}}\int\limits_{-\infty}^{+\infty} t^2e^{-t^2}\,dt = \frac{2\Delta^2}{\sqrt{\pi}}\left[-\frac{1}{2}\,te^{-t^2}\Big|_{-\infty}^{+\infty} + \frac{1}{2}\int\limits_{-\infty}^{+\infty} e^{-t^2}\,dt\right] = \Delta^2$$

2. We find $\bar{p}_1 = \frac{q}{2}$; $\Delta_1^2 = \frac{q^2}{12}$ and so

$$F(p,\,10) = \frac{1}{q}\sqrt{\frac{3}{5\pi}}\,e^{-\frac{3(p-5q)^2}{5q^2}},$$

$$F(p;\,20) = \frac{1}{q}\sqrt{\frac{3}{10\pi}}\,e^{-\frac{3(p-10q)^2}{10q^2}}.$$

3. (a) Since $\bar{p}_1 = 1$ kg and $\bar{p}_n = n\bar{p}_1$, it follows that $\bar{p}_{20} = 20\cdot 1\text{ kg} = 20\text{ kg}$.
(b) Take advantages of the results of (a) setting q equal to 2 in the expression for F $(p;\,20)$ to get

$$F(p;\,20) = \frac{1}{2}\sqrt{\frac{3}{10\pi}}\,e^{-\frac{3(p-20)^2}{40}}$$

or

$$F(p;\,20) = 0.154e^{-0.075(p-20)^2}$$

The probability that the catch will not exceed 20 kg is $w = 0.154\int\limits_0^{20} e^{-0.075(p-20)^2}\,dp$. In order to reduce this integral to the tabulated function $\Phi(x)$ put $\frac{1}{2}x^2 = 0.075(p-20)^2$, whence $x = 0.387(p-20)$, $dx = 0.387dp$.
Therefore

$$w = \frac{0.154}{0.387}\int\limits_{-7.7}^{0} e^{-\frac{x^2}{2}}\,dx = 0.4\int\limits_0^{7.7} e^{-\frac{x^2}{2}}\,dx = 0.4\frac{\sqrt{2\pi}}{2}\Phi(7.7) = 0.5$$

The probability that the catch does not exceed 22 kg is

$$w = 0.4\int\limits_{-7.7}^{+0.77} e^{-\frac{x^2}{2}}\,dx = 0.4\left(\int\limits_{-7.7}^{0} e^{-\frac{x^2}{2}}dx + \int\limits_0^{0.77} e^{-\frac{x^2}{2}}dx\right)$$
$$= 0.4\frac{\sqrt{2\pi}}{2}[\Phi(7.7) + \Phi(0.77)] = 0.78$$

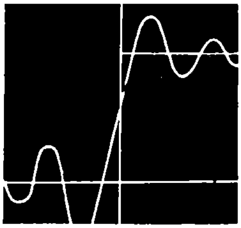
The probability that the catch does not exceed 25 kg is 0.98.

Sec. 13.10

1. Taking $n = 3000$, $\Delta = 100$, 200, 500, we get the approximate values 12, 25, 62. The exact values are equal to 12, 22, 59 respectively.
2. By formula (37) $A(4000) = 482$. By the condensed formula (38), $A = 540$. By the complete formula (38), $A = 554$. The exact value is $A(4000) = 550$.
3. Taking $n = 2000$, $\Delta = 3000$, we get $N = 395$. By formula (37) $N = 322$. By the condensed formula (38), $N = 358$. The exact value is $N = 366$.

Chapter 14

FOURIER TRANSFORMATION[*]



14.1 Introduction

In our discussion, in Chapters 7 and 8, of linear differential equations with constant coefficients and also of systems of such equations (in other words, in the consideration of linear processes homogeneous in time with a finite number of degrees of freedom) we saw how important was the role of the function e^{pt}. The point is that if we put this function in the left member of the equation and perform all operations, the result will be the very same function multiplied by a constant factor. For this reason the solution of a homogeneous equation was expressed in terms of functions of the form e^{pt}, so also was the Green's function expressed in terms of these functions; what is more, the solution of a nonhomogeneous equation was of the simplest form when the right member contained the function e^{pt}. We also include here, as a special case, functions of the form $\cos \omega t$ or $e^{\gamma t} \sin \omega t$, which by the Euler formula are expressed directly in terms of the exponential e^{pt} with complex p (see formulas (3)).

In many problems the independent variable t — ordinarily the time — can assume all values, that is, $-\infty < t < \infty$, and the solutions must, according to the meaning of the problem, remain finite both for finite t and for $t \to \pm\infty$. Since for $p = \gamma + i\omega$ $e^{pt} = e^{\gamma t}(\cos \omega t + i \sin \omega t)$, $|\cos \omega t + i \sin \omega t| = 1$, it follows that for $\gamma > 0$ we will have $|e^{pt}|_{t\to\infty} \to \infty$ and for $\gamma < 0$ it will be true that $|e^{pt}|_{t\to-\infty} \to \infty$. Hence, if we require the exponential e^{pt} to be bounded as $t \to \pm\infty$, then it must be true that $\gamma = 0$, that is to say, we must make use only of "harmonics", or the functions

$$e^{i\omega t} = \cos \omega t + i \sin \omega t \tag{1}$$

When harmonics with different amplitudes and frequencies are superimposed on one another, that is, when we consider sums of the

$$f(t) = \sum_k a_k e^{i\omega_k t} \tag{2}$$

[*] In this chapter we will have need of complex numbers (Secs. 5.1-5.3) and the delta function (Secs. 6.1-6.3). In Sec. 14.7 use is made of the concept of a multidimensional vector space (Sec. 9.6).

form, we can get functions of a much more complicated nature. The set of frequencies ω_k considered here is called the *spectrum* of the function $f(t)$; in this case we have a *discrete spectrum.*[*] We can also make use of sums of sines or cosines with distinct frequencies and amplitudes, for by formula (1) we can pass from exponentials to cosines and sines, and, by the formulas

$$\cos \omega t = \frac{e^{i\omega t} + e^{-i\omega t}}{2}, \quad \sin \omega t = \frac{e^{i\omega t} - e^{-i\omega t}}{2i} \tag{3}$$

from cosines and sines to exponentials. For the sake of uniformity we will mainly use exponentials (cf. Sec. 5.5).

A still broader class of functions $f(t)$ is obtained if in place of a sum over the individual frequencies we take advantage of the integral over all frequencies, which is to say, if we make use of a representation like

$$f(t) = \int_{-\infty}^{\infty} F(\omega)\, e^{i\omega t}\, d\omega \tag{4}$$

Here we have a *continuous spectrum,* which may occupy the entire ω-axis or some interval J on that axis if the function $F(\omega)$ is zero outside the interval; which means that actually the integration (4) is carried out only over J (then J is termed the *carrier* of the function $F(\omega)$). Incidentally, in Sec. 6.1 we saw that if the function $F(\omega)$ is a sum of delta-like terms, then the entire integral (4) passes into the sum (2), which means the spectrum will be discrete and will consist of the set of those values of ω at which the delta-like terms have singularities. In the general case, the spectrum can have both a continuous and a discrete part.

It is not difficult to grasp the meaning of the function $F(\omega)$ in the representation (4). Since in this representation we have the term $F(\omega)\, e^{i\omega t}\, d\omega = [F(\omega)\, d\omega]\, e^{i\omega t}$ on the small frequency interval from ω to $\omega + d\omega$, it follows, by comparison with (2), that $F(\omega)\, d\omega$ is the amplitude corresponding to the indicated interval of frequencies. Hence $F(\omega)$ may be regarded as the "density of the amplitude" corresponding to a small frequency interval and calculated per unit length of this interval. That is why the function $F(\omega)$ is called the *spectral density* of the function $f(t)$. Transition from the sum (2) to the integral (4) is similar to the transition (cf. Sec. 12.1) from the discrete model of a string in the form of elastically connected beads to the continuous model, in which the mass is spread out over the entire length of the string with a definite density. So also in the representation (4), the amplitude of the harmonics is spread out over the entire frequency spectrum with density $F(\omega)$.

[*] Take care to distinguish between the spectrum of a function which we deal with in this chapter and the spectrum of a boundary-value problem (page 300).

In physical problems we mostly deal with real functions $f(t)$, for example, when $f(t)$ is a force f acting on a system as a function of the time t.

The reality condition of $f(t)$ may be written thus:

$$f(t) \equiv f^*(t), \text{ that is } \int\limits_{-\infty}^{+\infty} F(\omega)\, e^{i\omega t}\, d\omega = \int\limits_{-\infty}^{+\infty} F^*(\omega)\, e^{-i\omega t}\, d\omega$$

We do not put the conjugacy symbol on ω because it is assumed that ω is real. For the sums of the harmonics to be equal for any values of t, it is necessary that the corresponding amplitudes be equal. Let us find the factor in front of $e^{i\omega_0 t}$; in the left-hand integral this is $F(\omega_0)$. In the right-hand integral, from the condition $e^{-i\omega t} = e^{i\omega_0 t}$ we find $\omega = -\omega_0$; hence, the factor in front of $e^{i\omega_0 t}$ in the right-hand integral is equal to $F^*(\omega) = F^*(-\omega_0)$. Thus, the condition of reality of the function $f(t)$ yields

$$F^*(-\omega_0) = F(\omega_0)$$

This equality refers to any ω_0 and so we can drop the subscript 0 and write

$$F^*(-\omega) = F(\omega) \tag{5}$$

The function $F(\omega)$ is a complex function. Let us write out explicitly the real and imaginary parts with the aid of two real functions $A(\omega)$ and $B(\omega)$:

$$F(\omega) = A(\omega) + iB(\omega)$$

The reality condition of the original function $f(t)$ that produced formula (5) leads to

$$A(-\omega) - iB(-\omega) = A(\omega) + iB(\omega)$$

that is

$$A(-\omega) = A(\omega), \quad B(-\omega) = -B(\omega) \tag{6}$$

Thus, the real part of F is an even function of ω and the imaginary part of F is an odd function of ω.

From the representation with the aid of $e^{i\omega t}$ let us pass to the representation with the aid of $\cos \omega t$ and $\sin \omega t$:

$$\begin{aligned} f(t) &= \int\limits_{-\infty}^{+\infty} F(\omega)\, e^{i\omega t}\, d\omega = \int\limits_{-\infty}^{+\infty} [A(\omega) + iB(\omega)]\, [\cos \omega t + i \sin \omega t]\, d\omega \\ &= \int\limits_{-\infty}^{+\infty} A(\omega) \cos \omega t\, d\omega - \int\limits_{-\infty}^{+\infty} B(\omega) \sin \omega t\, d\omega \\ &\qquad + i \int\limits_{-\infty}^{+\infty} A(\omega) \sin \omega t\, d\omega + i \int\limits_{-\infty}^{+\infty} B(\omega) \cos \omega t\, d\omega \end{aligned}$$

Let us consider a special but important case of the real function $f(t)$. Then the conditions (6) will make the last two integrals vanish. They contain the product of an even function by an odd function so that $\int\limits_{-\infty}^{0} k\,d\omega = -\int\limits_{0}^{\infty} k\,d\omega$, $\int\limits_{-\infty}^{+\infty} k\,d\omega = 0$, where k is the integrand of the third or fourth integral.

Contrariwise, the integrand in the first two integrals is symmetric:

$$n(\omega) = n(-\omega), \quad \int\limits_{-\infty}^{0} n(\omega)\,d\omega = \int\limits_{0}^{\infty} n(\omega)\,d\omega,$$

$$\int\limits_{-\infty}^{+\infty} n(\omega)\,d\omega = 2\int\limits_{0}^{\infty} n(\omega)\,d\omega$$

so that finally

$$f(t) = 2\int\limits_{0}^{\infty} A(\omega)\cos\omega t\,d\omega - 2\int\limits_{0}^{\infty} B(\omega)\sin\omega t\,d\omega \qquad (7)$$

To summarize, the real function $f(t)$ is represented in the form of an integral of the real functions $\cos\omega t$ and $\sin\omega t$, the corresponding spectral density of $A(\omega)$ and $B(\omega)$ is also real. In this case the integration is performed only over the positive frequencies with ω varying from 0 to ∞.

It is a remarkable fact that it is possible to represent as (4) almost any function $f(t)$ that remains finite as $t \to \pm\infty$. This will be shown in Sec. 14.2 where we will also answer the basic question of how to find the spectral density $F(\omega)$ of the given function $f(t)$. We will see right now that this question goes far beyond being of merely theoretical interest.

Recall the simple facts of the theory of oscillations (Secs. 7.3, 7.5; also see HM, Chs. 6 and 8). Suppose we have an oscillator with slight damping, i.e., little friction (if we are dealing with a mechanical oscillatory system) or with low resistance (if it is an electric oscillatory circuit), and so on. If such an oscillator experiences a harmonic external force with frequency ω_k, then forced harmonic oscillations with the same frequency will build up in the oscillator. The amplitude of these oscillations is the greater, the closer ω is to the frequency ω_0 of the natural oscillations of the oscillator. This "selectivity" of the oscillator relative to the frequency of the external action is expressed the more sharply, the less the damping. In the limit, when we consider an oscillator without damping, for $\omega = \omega_0$ resonance sets in, which means the amplitude of the forced oscillations increases without limit.

Now suppose the harmonic external action is applied to a system of oscillators with small damping that have distinct natural frequencies. Then the oscillator whose natural frequency is equal to the frequency of the external action will respond most strongly to this action. Finally, if such a system of oscillators is acted upon by a mixture of harmonics, that is, by a function of the form (2), then those oscillators will respond whose natural frequency is the same as one of the external frequencies ω_k. Here the amplitude of the forced oscillations of the oscillator with natural frequency $\tilde{\omega}_0$ is proportional to the amplitude of the action, that is, it is proportional to the a_k for which $\omega_k = \tilde{\omega}_0$. Thus, such a system of oscillators realizes a *harmonic analysis* (the term *Fourier analysis* is also used; this is a kind of *spectral analysis*) of the external action, which amounts to breaking the action up into its component harmonics. A similar situation results in the case of imposing an external action of the type (4). The amplitude of an oscillator with natural frequency $\tilde{\omega}_0$ is proportional to $F(\tilde{\omega}_0)$, where F is the spectral density (see formula (4)). An exact statement of the conditions necessary for this will be given below. With a set of oscillators having all possible values $\tilde{\omega}_0$, it is possible to "sense", i.e. determine, the entire course of the function $F(\omega)$.

This kind of harmonic analysis can actually be accomplished. For example, in acoustics use is made of a system of resonators, each one of which is tuned to a specific frequency. If this system is acted upon by an acoustic (sound) oscillation, which can always be represented as a mixture of "pure sounds", or harmonic oscillations, then those resonators will respond whose natural frequency corresponds to the spectrum of the action. By determining the amplitude of oscillation of the resonators we thus accomplish the harmonic analysis of the external action.

Exercises

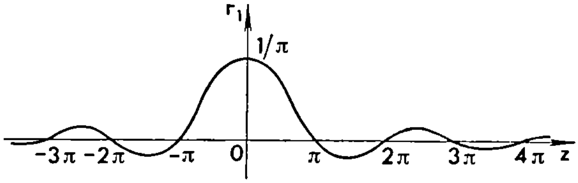
1. What must the function $F(\omega)$ be equal to for the integral (4) to pass into the sum (2)?
2. Under what condition will the sum (2) be a periodic function of t?

14.2 Formulas of the Fourier transformation

We begin with a question that will be needed later on: What will happen if the amplitude of oscillation is uniformly spread over the whole frequency axis, in other words, what function $f(t)$ in the formula (4) has the spectral density $F(\omega) \equiv 1$? This means we have to investigate the integral

$$I(t) = \int_{-\infty}^{\infty} e^{i\omega t}\, d\omega$$

Fig. 192



Although this is a divergent integral (Sec. 3.1), it can be represented as the limit of an integral over a finite interval:

$$I(t) = \lim_{N\to\infty} I_N(t)$$

where

$$I_N(t) = \int_{-N}^{N} e^{i\omega t}\,d\omega = \frac{e^{itN} - e^{-itN}}{it} = 2\,\frac{\sin tN}{t}$$

(in passing to the limit we made use of the second formula in (3)). This result can be represented as

$$I_N(t) = 2\pi N \frac{\sin tN}{\pi t N} = 2\pi N F_1(tN) \tag{8}$$

where $F_1(t) = \dfrac{\sin t}{\pi t}$.

The graph of the function $F_1(t)$ is shown in Fig. 192. It can be demonstrated (see the solution of Exercise 2, Sec. 3.6) that

$$\int_{-\infty}^{\infty} \frac{\sin t}{t}\,dt = \pi, \quad \text{or} \quad \int_{-\infty}^{\infty} F_1(t)\,dt = \int_{-\infty}^{\infty} \frac{\sin t}{\pi t}\,dt = 1$$

The function F_1 has a principal maximum at $t = 0$, $F_1(0) = \dfrac{1}{\pi}$; to the left and right it decreases in oscillatory fashion (see Fig. 192).

From (8) we see that, up to the constant factor 2π, $I_N(t)$ results from the function $F_1(t)$ via the very same transformation which, in Sec. 6.1, led to the concept of the delta function. Thus, passing to the limit, we get

$$\int_{-\infty}^{\infty} e^{i\omega t}\,d\omega = I(t) = 2\pi\delta(t) \tag{9}$$

In this derivation there are two points that require amplification. First, when defining the delta function in Sec. 6.1 we used as illustrations only functions that do not assume negative values. However

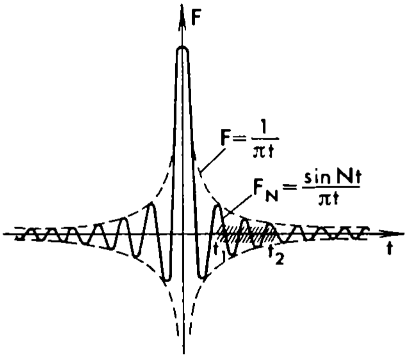


Fig. 193

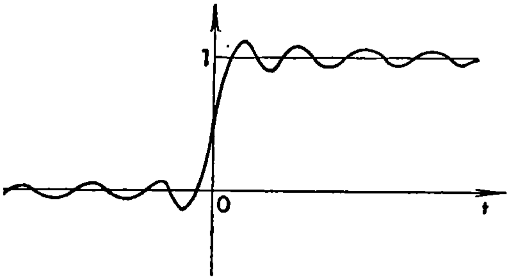


Fig. 194

this condition actually was not made use of so that negative values may be admitted. Second, and this is more important, the function $F_1(t)$ is not a rapidly decreasing function. Indeed, the graph of $F_N(t) = NF_1(Nt)$ for large N is shown in Fig. 193. We see that it oscillates frequently on any finite fixed interval t_1, t_2 not containing the point $t = 0$, but the amplitude of oscillations is not small, i.e. the function $F_N(t)$ is not close to zero, as in the examples of Sec. 6.1. If we took $\int\limits_{-\infty}^{+\infty} |F_N|\, dt$, then such an integral would diverge. But this is not essential, for with increasing N the integral of the function $F_N(t)$ behaves in the following manner:

$$\int\limits_{-\infty}^{t} F_N(t)\, dt = \int\limits_{-\infty}^{t} N\frac{\sin Nt}{\pi Nt}\, dt = \int\limits_{-\infty}^{Nt} \frac{\sin s}{\pi s}\, ds \underset{N\to\infty}{\to} \begin{cases} 0 & (t < 0), \\ 1 & (t > 0) \end{cases}$$

The graph of this integral is shown in Fig. 194. Thus, for large N we have

$$\int\limits_{-\infty}^{t} F_N(t)\, dt \approx e(t) \text{ (see Sec. 6.3), that is, } F_N(t) \approx e'(t) = \delta(t)$$

Thus, the delta-like nature of the function $F_N(t)$ for large N is due not to its rapid decrease but to its frequent oscillations, as a result of which at a finite distance from $t = 0$ the function when integrated is to all intents and purpose identically zero.

Arguing in exactly the same way, we could show that $\int\limits_{-\infty}^{+\infty} Q(t)\,F_N(t - t_0)\,dt \to Q(t_0)$ as $N \to \infty$ in accordance with the properties of the delta function. In such an integral, the major contribution is made by the principal maximum of F_N at $t = t_0$, where $F_N = N/\pi$; the "wings", that is, the regions $t < t_0 - \varepsilon$ and $t > t_0 + \varepsilon$, make slight contributions because the integral of an alternating function is small.

With the aid of (9) it is easy to consider the case of spectral density given by $F(\omega) = Ae^{i\omega\tau}$ (A = constant, τ = constant): substituting into (4), we get

$$\int\limits_{-\infty}^{\infty} Ae^{i\omega\tau}e^{i\omega t}\,d\omega = A\int\limits_{-\infty}^{\infty} e^{i\omega(t+\tau)}\,d\omega = 2\pi A\,\delta(t + \tau)$$

Conversely, it follows from this that to the function $f(t) = B\delta(t - \tau)$ corresponds the density $F(\omega) = \frac{B}{2\pi}\,e^{-i\omega\tau}$.

We can now pass to the case of an arbitrary function $f(t)$ in the integral (4). As we saw in Sec. 6.2, every function $f(t)$ can be represented in the form of a sum (to be more precise, an integral) of delta-like functions

$$f(t) = \sum_{\tau} [f(\tau)\;d\tau]\;\delta(t - \tau)$$

By virtue of what has just been demonstrated, to every term there corresponds a spectral density $\frac{f(\tau)\,d\tau}{2\pi}\,e^{-i\omega\tau}$, which means that to the whole sum — the function $f(t)$ — there corresponds the spectral density

$$F(\omega) = \sum_{\tau} \frac{f(\tau)\,d\tau}{2\pi}\,e^{-i\omega\tau}$$

If one recalls that actually this is not a sum but an integral and, besides, if we denote the variable of integration by t, then we get

$$F(\omega) = \frac{1}{2\pi}\int\limits_{-\infty}^{\infty} f(t)\,e^{-i\omega t}\,dt \qquad (10)$$

The formulas (4) and (10) are the Fourier transformation formulas. Using them, we can pass from any function $f(t)$, which is finite as

$t \to \pm \infty$ (this is necessary for the integral (10) to be meaningful), to its spectral density and, conversely, given the spectral density, we can restore the function. These formulas are remarkably symmetric to within the constant factor $\frac{1}{2\pi}$ and the sign in the exponent.

If $f(t)$ tends to zero as $t \to -\infty$ and as $t \to +\infty$, then the integral (10) can, in principle, always be computed at least numerically.

Let us consider the special case of a real function $f(t)$. We write

$$e^{-i\omega t} = \cos \omega t - i \sin \omega t$$

and obtain

$$F(\omega) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} f(t) \cos \omega t \, dt - \frac{i}{2\pi} \int\limits_{-\infty}^{\infty} f(t) \sin \omega t \, dt$$

Each of the integrals is a real function. Recall how we wrote $F(\omega) = A(\omega) + iB(\omega)$. Clearly,

$$\left.\begin{aligned} A(\omega) &= \frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} f(t) \cos \omega t \, dt, \\ B(\omega) &= -\frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} f(t) \sin \omega t \, dt \end{aligned}\right\} \qquad (11)$$

From this it is easy to verify the properties (5) and (6) of the density $F(\omega)$ for the real function $f(t)$.

Comparing (7) and (11), we are convinced that the coefficient of $\cos \omega t$, that is, $A(\omega)$, is in turn expressed in terms of the integral of the function $f(t)$ multiplied by $\cos \omega t$, similarly for $B(\omega)$.

Note two special cases of the integral (10). Let $f(t) = C =$ constant. From the formula

$$f(t) = C = \int F(\omega) \, e^{i\omega t} \, d\omega$$

it is clear that to obtain C on the left, that is, $Ce^0 = Ce^{i0t}$, we have to take $F(\omega) = C\delta(\omega)$. By the general rule,

$$\int \delta(\omega) \, \psi(\omega) \, d\omega = \psi(0),$$

$$\int \delta(\omega) \, e^{i\omega t} \, d\omega = e^{i0t} \equiv 1$$

Let us analyze $f(t) = De^{i\omega_0 t}$ in similar fashion. In this case $F(\omega) = D\delta(\omega - \omega_0)$, which can also readily be verified by substitution into (4). The same results can also be obtained with the aid of (9):

$$\frac{C}{2\pi}\int\limits_{-\infty}^{+\infty} e^{-i\omega t}\,dt = C\delta(-\omega) = C\delta(\omega),$$

$$\frac{D}{2\pi}\int\limits_{-\infty}^{+\infty} e^{i\omega_0 t}e^{-i\omega t}\,dt = D\delta(\omega - \omega_0)$$

However, such integrals cannot be evaluated directly and one has to consider the nondecaying function C or $De^{i\omega_0 t}$ as the limit of a decaying function, for instance, $Ce^{-\alpha t^2}$ or $De^{i\omega_0 t - \alpha|t|}$ as $\alpha \to 0$. The result is a rather long and complicated chain of reasoning. It is better to obtain these formulas from (9), as was done above, and remember them.

Suppose an undamped oscillator with natural frequency ω_0 is acted upon by a force $f(t)$ so that the equation of motion is of the form

$$m\frac{d^2x}{dt^2} = -m\omega_0^2 x + f(t)$$

We assume that $f(t) = 0$ for $t = -\infty$ and the oscillator was at rest in the equilibrium position, $x = \frac{dx}{dt} = 0$. It is easy to construct a solution of the problem (cf. Sec. 7.5):

$$x = \frac{1}{m\omega_0}\int\limits_{-\infty}^{t} \sin\omega_0(t - \tau)\,f(\tau)\,d\tau$$

This solution is constructed with the aid of the Green's function of the problem: an impulse, i.e. a force $\delta(t - \tau)$, acting on the oscillator at rest, imparts a velocity $\frac{1}{m}$ and free oscillations set in: $x(t > \tau) = \frac{1}{m\omega_0}\sin\omega_0(t - \tau)$.

Suppose the force $f(t)$ acts during a limited period of time, $f = 0$ for $t > T$. How will the oscillator oscillate when the force ceases to act? We expand

$$\sin\omega_0(t - \tau) = \sin\omega_0 t\cos\omega_0\tau - \cos\omega_0 t\sin\omega_0\tau$$

to get

$$x(t) = \frac{1}{m\omega_0}\left[\sin\omega_0 t \int_{-\infty}^{T} f(\tau)\cos\omega_0\tau\, d\tau - \cos\omega_0 t \int_{-\infty}^{T} f(\tau)\sin\omega_0\tau\, d\tau\right]$$

Since $f(\tau) = 0$ for $t > T$, we can write $\int_{-\infty}^{T} f(\tau)\cos\omega\tau\, d\tau$ instead of $\int_{-\infty}^{\infty} f(\tau)\cos\omega\tau\, d\tau$; the same goes for the integral with the sine. Recalling the formulas (11), we get, for oscillations after the force ceases to act,

$$x(t) = \frac{2\pi}{m\omega_0}[A(\omega_0)\sin\omega_0 t + B(\omega_0)\cos\omega_0 t] \text{ for } t > T \qquad (12)$$

The amplitude of continuous oscillations of an oscillator with natural frequency ω_0 after the force ceases to act is

$$x_m = \frac{2\pi}{m\omega_0}\sqrt{A^2(\omega_0) + B^2(\omega_0)} = \frac{2\pi}{m\omega_0}|F(\omega_0)| \qquad (13)$$

Consequently, the amplitude depends only on the spectral density of the acting force $F(\omega)$ at the resonance frequency, or $\omega = \omega_0$. This is also demonstrated by the assertions made in Sec. 14.1.

Physicists often regard $|F(\omega)|^2$ and not $F(\omega)$ as the *spectral density*. For example, if the force $f(t)$ is the pressure in a sound wave or an electric field in an electromagnetic wave, then $|F(\omega)|^2\, d\omega$ is a quantity proportional to the energy of the sound or electromagnetic oscillations over the portion of the spectrum from ω to $\omega + d\omega$.

After some simple manipulations, we can rewrite (12) as

$$\left.\begin{aligned} x(t) &= \frac{\pi}{im\omega_0}[F(\omega_0)e^{i\omega_0 t} - F(-\omega_0)e^{-i\omega_0 t}], \\ m\frac{dx}{dt} &= \pi[F(\omega_0)e^{i\omega_0 t} + F(-\omega_0)e^{-i\omega_0 t}] \end{aligned}\right\} (t > T) \qquad (14)$$

Now let us try to solve the problem of the oscillations of an oscillator directly with the aid of the Fourier integral. To do this we write

$$x(t) = \int_{-\infty}^{+\infty} X(\omega)e^{i\omega t}\, d\omega \qquad (15)$$

and substitute this expression into the oscillation equation. So as not to deal with the condition of rest at $t = -\infty$, we consider a damped oscillator. It is clear that the motion of such an oscillator does not depend on what occurred in the distant past.

Thus we consider the equation

$$m\frac{d^2x}{dt^2} = -m\omega_0^2 x - h\frac{dx}{dt} + f(t) \tag{16}$$

Substituting x and f in the form of Fourier integrals, we get

$$\int_{-\infty}^{+\infty} X(\omega)\,(-m\omega^2 + m\omega_0^2 + i\omega h)\, e^{i\omega t}\, d\omega = \int_{-\infty}^{+\infty} F(\omega)\, e^{i\omega t}\, d\omega$$

From this we immediately get

$$X(\omega) = \frac{F(\omega)}{m(\omega_0^2 - \omega^2) + i\omega h} \tag{17}$$

If the damping is slight, then $X(\omega)$ is particularly great for those values of ω for which only the small quantity $i\omega h$ remains in the denominator. This occurs when $\omega_0^2 - \omega^2 = 0$, i.e., when $\omega = \pm\omega_0$. Hence $X(\omega_0)$ and $X(-\omega_0)$ are particularly great in the Fourier transform, or image (see Sec. 14.4), of the function $x(t)$. In the limit, in the case of continuous oscillations, that is, as $h \to 0$, the expression (14) can be obtained from (15) and (17) by methods of complex variable theory. We will not go into that any further here.

The last remark concerning harmonic analysis is that instruments ordinarily record only the amplitude of oscillations of various frequency. By (13) this means that only the modulus (absolute value) * of the Fourier transform of the force $|F(\omega)|$ being studied is determined. This holds true for the ear as a recorder of sound vibrations, and also for the eye and the spectroscope (the spectroscope only determines the intensity of light of various frequencies).

In this procedure we lose information concerning the phase of $F(\omega)$. Imagine for instance two functions $F_1(\omega)$ and $F_2(\omega)$ such that $|F_1(\omega)| = |F_2(\omega)|$ but F_1 and F_2 themselves are distinct. This means that $F_2(\omega) = F_1(\omega)\, e^{i\varphi(\omega)}$, where φ is any real function. To these two Fourier transforms there correspond distinct $f_1(t)$ and $f_2(t)$:

$$f_1(t) = \int F_1(\omega)\, e^{i\omega t}\, d\omega, \quad f_2(t) = \int F_2(\omega)\, e^{i\omega t}\, d\omega$$

But if $f(t)$ is the air pressure in a sound wave, then our ear perceives the same sounds and is not capable of distinguishing between $f_1(t)$ and $f_2(t)$. In order to distinguish between f_1 and f_2 the curve $f(t)$ has to be recorded by means of a high-speed pressure gauge. Only recently, persistent attempts have been made to find ways of studying the phase $F(\omega)$ in the case of light waves.

We will come back to the question of phase in Sec. 14.8.

* Recall that for a real force $F(-\omega) = F^*(\omega)$ so that $|F(-\omega)| \equiv |F(\omega)|$; the sign of ω is inessential, it suffices to consider $\omega > 0$.

Exercises

1. Find the spectral density of the function $f(t) = e^{-\alpha|t|} (\alpha > 0)$. Going to the limit as $\alpha \to +0$, obtain the spectral density of the function $f(t) \equiv 1$. For this example, derive formula (1) of Ch. 1 from formula (4).
2. Obtain in explicit form the solution, bounded as $t \to \pm\infty$, of equation (16).

14.3 Causality and dispersion relations

In Sec. 7.5 we saw that the solution of equation (16) is obtained from the solution $G(t, \tau)$ (the so-called Green's function) of a similar equation with a special nonhomogeneous term

$$m\frac{d^2x}{dt^2} + h\frac{dx}{dt} + m\omega_0^2 x = \delta(t - \tau)$$

via the formula

$$x(t) = \int_{-\infty}^{\infty} G(t, \tau) f(\tau)\, d\tau$$

On the basis of Sec. 6.2, the solution of any linear problem is of a similar aspect; the role of the input signal being played by $f(t)$, that of the response of the system at hand to this signal by $x(t)$. Then the function $G(t, \tau)$ is the response to a unit instantaneous impulse acting on the system at time τ.

In Sec. 6.2 we saw that the result of action on a linear system is of a similar form also in the case of a space coordinate serving as the independent variable (it is precisely this case that was discussed in detail there). It turns out that if for the independent variable we take the time with its specific single directionality, then the corresponding Green's function has important properties of an extremely general nature, which we will now discuss.

If the parameters of the linear system at hand (say, an oscillator), which is acted upon by signals, do not change in the course of time, then all actions and the corresponding responses admit of an arbitrary time shift. This means then that $G = G(t - \tau)$ and so

$$x(t) = \int_{-\infty}^{\infty} G(t - \tau) f(\tau)\, d\tau \tag{18}$$

We introduce the spectral densities of the functions $x(t)$, $f(t)$ and $2\pi G(t)$ (the factor 2π is put here to simplify subsequent formulas):

$$x(t) = \int X(\omega)\, e^{i\omega t}\, d\omega,$$

$$f(t) = \int F(\omega)\, e^{i\omega t}\, d\omega, \quad 2\pi G(t) = \int L(\omega)\, e^{i\omega t}\, d\omega \tag{19}$$

all integrals are taken from $-\infty$ to ∞. Quite naturally, we will then regard all the participating functions of time as being finite as $t \to \pm\infty$, for we consider the Fourier representation only for such functions. (This finiteness requirement was also implicit in the construction of the solution (17); it was that requirement that separated the solution we needed of (16) from the two-parameter family of all its solutions.) Substituting (18) into (19) and changing the order of integration, we get

$$\begin{aligned}\int X(\omega)\, e^{i\omega t}\, d\omega &= \int\left[G(t-\tau)\int F(\omega)\, e^{i\omega\tau}\, d\omega\right] d\tau\\ &= \iint G(t-\tau)\, F(\omega)\, e^{i\omega\tau}\, d\omega\, d\tau\\ &= \int F(\omega)\left[\int G(t-\tau)\, e^{i\omega\tau}\, d\tau\right] d\omega\end{aligned}$$

Putting $t - \tau = \tau_1$, we get

$$\begin{aligned}&\int F(\omega)\left[\int G(\tau_1)\, e^{i\omega(t-\tau_1)}\, d\tau_1\right] d\omega\\ &\qquad= \int F(\omega)\, e^{i\omega t}\left\{\frac{1}{2\pi}\int [2\pi G(\tau_1)]\, e^{-i\omega\tau_1}\, d\tau_1\right\} d\omega\\ &\qquad= \int F(\omega)\, L(\omega)\, e^{i\omega t}\, d\omega\end{aligned}$$

Comparing the last integral with the original one, we conclude that

$$X(\omega) = L(\omega)\, F(\omega)$$

To summarize, then, the Fourier transforms (images) of the input and output signals are connected in a very simple way: the latter is obtained from the former by multiplying by the *transfer function* $L(\omega)$. For example, from (17) it is evident that the transfer function for an elementary linear oscillator is of the form

$$L(\omega) = \frac{1}{m(\omega_0^2 - \omega^2) + ih\omega} \tag{20}$$

Conversely, given the transfer function $L(\omega)$ and computing the Green's function $G(t)$, we can pass to (18) for transformation of signals. There arises an interesting question. Let $L(\omega)$ be the transfer function of the signal converter. Then the *principle of causality* must hold true: it states that a response cannot arise before a signal is delivered. Mathematically, this principle is equivalent to the requirement that $G(t-\tau) \equiv 0$ for $t < \tau$, i.e., $G(t) \equiv 0$ for $t < 0$. (For this reason, the upper limit of the integral (18) is, for real systems, not equal to ∞ but to t, which is what we stated in Sec. 7.5.) What requirements must the function $L(\omega)$ satisfy for the principle of causality to be fulfilled? It is quite natural that such requirements must be placed in the foundation of the theory even prior to deriving appropriate equations.

For the sake of simplicity, let us confine ourselves to the case where $L(\omega)$ is of the form of a ratio of polynomials: $L(\omega) = \frac{P(\omega)}{Q(\omega)}$ with the degree of the denominator exceeding that of the numerator at least by 2. (Actually, similar investigations can be carried out for an appreciably broader class of functions $L(\omega)$.) Besides, we assume that $Q(\omega)$ does not have any real zeros. Then in order to compute the integral

$$G(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} L(\omega)\, e^{i\omega t}\, d\omega$$

for $t > 0$ we can make use of that very same device that we employed in evaluating the integral (36) of Ch. 5, the only difference being that instead of the complex plane z we have to consider the complex plane ω. Here, as in Fig. 67, the "closing" semicircle is drawn in the upper half-plane of ω, since for $t > 0$ we have $|e^{i\omega t}| \leqslant 1$ in this half-plane (whereas in the lower half-plane $|e^{i\omega t}|$ assumes arbitrarily large values). On the basis of the general theorem (31) of Ch. 5 on residues we find that for $t > 0$

$$G(t) = i \sum_{\operatorname{Im} \omega_k > 0} \operatorname{Res}_{\omega = \omega_k} \{L(\omega)\, e^{i\omega t}\} \quad (t > 0) \tag{21}$$

In the right-hand member the summation is carried out over all singular points (poles) of the function $L(\omega)$, that is, over the zeros of the polynomial $Q(\omega)$ lying in the upper half-plane, $\operatorname{Im} \omega > 0$.

The case $t < 0$ is considered in similar fashion. Then in Fig. 67 we have to use the lower semicircle instead of the upper one; this yields

$$G(t) = -i \sum_{\operatorname{Im} \omega_k < 0} \operatorname{Res}_{\omega = \omega_k} \{L(\omega)\, e^{i\omega t}\} \quad (t < 0) \tag{22}$$

From this, since t is arbitrary, we conclude that the condition for the principle of causality to hold (that is, $G(t) \equiv 0$ for $t < 0$) is the absence of singular points in the transfer function $L(\omega)$ in the lower half-plane: $\operatorname{Im} \omega < 0$. It is easy to verify that the transfer function (20) satisfies this condition.

The resulting condition can also be interpreted as follows. Since the influence function $G(t)$ for $t \neq 0$ satisfies a homogeneous equation, from formulas (21) and (22) it is evident that to every singular point $\omega_k = \alpha_k + i\beta_k$ of the transfer function there corresponds a term of the form $e^{i\omega_k t} = e^{-\beta_k t} e^{i\alpha_k t}$ in the solutions of the homogeneous equation. If all ω_k lie in the upper half-plane, then all $\beta_k > 0$, and for this reason all these summands decay as $t \to \infty$. This is as it

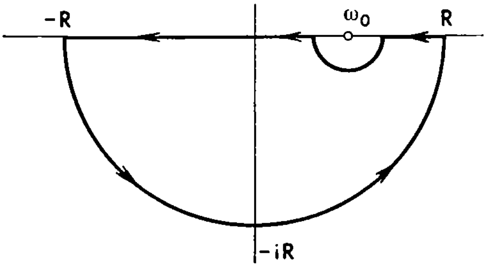


Fig. 195

should be if the system only converts signals and does not have any energy sources of its own. Thus, the physical equivalent of the condition concerning the location of singular points of the transfer function consists in the internal energy stability of the system, that is, in the damping of the energy induced in it. (Note that one could consider linear systems with internal energy sources, that is, unstable systems; then for the principle of causality to hold we would have to give up the requirement that the solutions be bounded as $t \to \infty$.)

We can indicate yet another important condition tantamount to the principle of causality. To derive it, consider the integral

$$I = \oint\limits_{(L)} \frac{L(\omega)}{\omega - \omega_0}\, d\omega \tag{23}$$

extended around the contour in Fig. 195, where ω_0 is a fixed real value of ω. Fulfilment of the principle of causality is equivalent to the absence of singular points inside (L) and for this reason, due to the arbitrary nature of ω_0, to the condition $I = 0$. Now let $R \to \infty$ and let the radius of the small circle $\varepsilon \to 0$. Then, reasoning as in Sec. 5.9, we find that the integral around the large circle tends to zero. The integral over the horizontal segments tends to

$$-\,\mathrm{P}\int\limits_{-\infty}^{\infty} \frac{L(\omega)}{\omega - \omega_0}\, d\omega \tag{24}$$

where P stands for *Gauchy's principal value.* By definition we have

$$\mathrm{P}\int\limits_{-\infty}^{\infty} \frac{L(\omega)}{\omega - \omega_0}\, d\omega = \lim_{\varepsilon \to 0}\left[\int\limits_{-\infty}^{\omega_0-\varepsilon} + \int\limits_{\omega_0+\varepsilon}^{\infty}\right]$$

(This complication is due to the divergence of the integral (24) for $\omega = \omega_0$ in the ordinary sense of Sec. 3.1.) Finally, the integral around the small semicircle tends to

$$L(\omega_0)\int \frac{d\omega}{\omega - \omega_0} = -\pi i L(\omega_0)$$

When passing to the limit in (23) as $R \to \infty$, $\varepsilon \to 0$, we get $-\mathrm{P}\int\limits_{-\infty}^{\infty} \frac{L(\omega)}{\omega - \omega_0}\,d\omega - \pi i L(\omega_0) = 0$, or

$$L(\omega_0) = \frac{i}{\pi}\,\mathrm{P}\int\limits_{-\infty}^{\infty} \frac{L(\omega)}{\omega - \omega_0}\,d\omega$$

Representing $L(\omega)$ as $\operatorname{Re} L(\omega) + i \operatorname{Im} L(\omega)$ and separating the real part from the imaginary part in the last equation, we get the so-called *dispersion relations* for the transfer function:

$$\operatorname{Re} L(\omega_0) = -\frac{1}{\pi}\,\mathrm{P}\int\limits_{-\infty}^{\infty} \frac{\operatorname{Im} L(\omega)}{\omega - \omega_0}\,d\omega,$$

$$\operatorname{Im} L(\omega_0) = \frac{1}{\pi}\,\mathrm{P}\int\limits_{-\infty}^{\infty} \frac{\operatorname{Re} L(\omega)}{\omega - \omega_0}\,d\omega$$

which, as we see, are equivalent to the principle of causality. These relations play an important part in quantum mechanics. They permit obtaining important propositions even for systems for which no appropriate theory has yet been developed and so the type of equations is not known.

The principle of causality is the underlying mechanism of an unworkable device mentioned in an article by R. Hagedorn [7] for taking photographs in complete darkness prior to switching on the light. Suppose a flash of light occurs in complete darkness. It is described by a delta function whose spectral representation contains harmonics of the form $e^{i\omega t}$ that are present at all times. It is further suggested a filter be set up to screen all harmonics except one specific one, with the result that we get monochromatic light that acts not only after but even prior to the light flash! But this is absurd, for the impossibility of such filtering follows from the principle of causality. We can also use dispersion relations to show that a filter that passes only one frequency is impossible. The law of transmission of any filter is such that the sum of the harmonics from any flash of light must be identically equal to zero prior to the flash.

14.4 Properties of the Fourier transformation

We will call the function $f(t)$ the *original* (*preimage*) and the corresponding function $F(\omega)$ the *image* (*Fourier transform*) of $f(t)$ under the Fourier transformation. The *Fourier transformation* itself (otherwise called the *Fourier operator*, cf. Sec. 6.2) is defined by formula (10) and we use it to pass from the original to the image. Formula (4), where we passed from the image to the original, defines the *inverse Fourier transformation.*

Let us consider some properties of the Fourier transformation. First of all, it is linear, which means that when the originals are added, so also are the images, when an original is multiplied by a constant, the image is multiplied by that constant too. This property follows immediately from similar properties of the integral and has already been used by us in Sec. 14.2 where we made use of the principle of superposition in deriving (10). (Physically, this follows from the linearity of the oscillators under consideration.) Naturally, the inverse transformation is also linear.

When the original is shifted by a constant a, its image is multiplied by $e^{-ia\omega}$. True enough, since after a shift we get the function $f(t - a)$ and its spectral density is

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} f(t-a)\, e^{-i\omega t}\, dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(\tau)\, e^{-i\omega(\tau+a)}\, d\tau$$

$$= e^{-i\omega a}\frac{1}{2\pi}\int_{-\infty}^{\infty} f(\tau)\, e^{-i\omega\tau}\, d\tau = e^{-ia\omega}F(\omega)$$

(we made the substitution $t - a = \tau$).

This property is particularly evident in the case of separate harmonics: if we shift the harmonic $e^{i\omega_0 t}$ by a, we get $e^{i\omega_0(t-a)} = e^{-ia\omega_0}e^{i\omega_0 t}$, that is, it is multiplied by $e^{-ia\omega_0}$ and so its image, $\delta(\omega - \omega_0)$, is multiplied by $e^{-ia\omega_0}$ or, what is the same thing for this function, by $e^{-ia\omega}$. And since all these images receive the same factor, their sum, or $F(\omega)$, has that factor too.

In similar fashion it can be verified that if the image is shifted by an amount a, the original is multiplied by e^{iat}. It is precisely this property of invariance, to within a multiplying constant, of the function $e^{i\omega t}$ relative to shifts, together with the possibility of superposition, that makes this function play such a role in the Fourier transformation. It is quite natural that when studying processes that obey linear equations that are homogeneous in time, we have to rely on functions which themselves admit a time shift. *

We will consider in more detail the transformations of a family of solutions of the problem just posed. The term *transformation group*

* Here the equations themselves may not be only differential equations. A typical integral equation homogeneous in time is

$$f(t) = \int_{-\infty}^{\infty} k(t-\tau)\, f(\tau)\, d\tau$$

is used for a set of transformations having two properties: if together with any two transformations it contains the transformation resulting from their successive application and if every transformation has an inverse. For example, a transformation group is formed by any collection of multiplications by nonzero constants (this is necessary for the existence of the inverse transformation). Under this transformation, every function $f(t)$ goes into $Cf(t)$. Another important instance is the group of shifts under which every function $f(t)$ goes into $f(t + \tau)$ (the constant $\tau \gtrless 0$ determines the shift). If a study is being made of a linear problem that is homogeneous in time, then its general solution, (i.e., the family of all solutions), must be invariant with respect to both these transformation groups. Proof is given in linear algebra that this general solution, which, generally, includes many parameters, may be represented in the form of a sum of one-parameter families of solutions, these families also being invariant under the indicated transformation groups (exceptions will be given below).

However, it is easy to verify that a one-parameter family of functions possessing such invariance must necessarily be of the form Ce^{pt} (C is arbitrary and serves as a parameter inside the family, p is fixed and serves as a parameter of the family itself). Indeed, since the family is a one-parameter family and is invariant under multiplication by constants, it is exhausted by functions of the form $Cf(t)$, where C is arbitrary and $f(t)$ is some function of the family. But from the invariance with respect to shifts it follows that $f(t + \tau) = C_\tau f(t)$, whence

$$f'(t) = \lim_{\tau\to 0} \frac{f(t+\tau) - f(t)}{\tau} = \lim_{\tau\to 0} \frac{C_\tau f(t) - f(t)}{\tau} = \left[\lim_{\tau\to 0} \frac{C_\tau - 1}{\tau}\right] f(t)$$

Denoting the limit in the square brackets by p, we arrive at the differential equation $f'(t) = pf(t)$, whence $f(t) = e^{pt}$ to within a multiplying constant.

From this we find that if a family of functions is invariant with respect to both indicated transformation groups and has, say, two degrees of freedom, then it is of the form $C_1e^{p_1t} + C_2e^{p_2t}$, where C_1 and C_2 are arbitrary constants. However, if the problem at hand contains parameters and for some values of these parameters we get $p_1 = p_2$, then the aspect of the two-parameter family of solutions changes. In that case it may be shown, as in Sec. 7.3, that the indicated family takes on the aspect $C_1e^{p_1t} + C_2te^{p_1t}$. (Here, the one-parameter family $C_2te^{p_1t}$ is not by itself invariant to shifts.) If three values coincide, $p_1 = p_2 = p_3$, then the three-parameter family of solutions becomes $C_1e^{p_1t} + C_2te^{p_1t} + C_3t^2e^{p_1t}$, etc.

From the properties that have been demonstrated it follows that the Fourier transform (image) is multiplied by $i\omega$ when differentiating the function $f(t)$. Indeed, for the image of the function $\frac{f(t+h)-f(t)}{h}$ we have

$$\frac{1}{h}e^{ih\omega}F(\omega) - \frac{1}{h}F(\omega) = \frac{e^{ih\omega}-1}{h}F(\omega) = \left[i\omega + \frac{(i\omega)^2}{2}h + \dots\right]F(\omega)$$

Passing to the limit as $h \to 0$, we get our assertion. It can be verified in similar fashion that when differentiating the image, the preimage is multiplied by $-it$. These properties can be seen directly if we form the derivative of the integral with respect to the parameter,

$$\frac{df}{dt} = \frac{d}{dt}\int F(\omega)\, e^{i\omega t} d\omega = \int i\omega F(\omega)\, e^{i\omega t}\, d\omega$$

and also for the inverse transformation

$$F(\omega) = \frac{1}{2\pi}\int f(t)\, e^{-i\omega t}\, dt, \quad \frac{dF}{d\omega} = -\frac{i}{2\pi}\int tf(t)\, e^{-i\omega t}\, dt$$

The foregoing properties of the Fourier transformation find remarkable application to the solution of linear differential equations with constant coefficients. We made use of them in Sec. 14.2 when we considered the problem of forced oscillations of an oscillator with friction.

It is interesting to consider the equation of free oscillations, that is, equation (16) for $f(t) \equiv 0$. Performing the Fourier transformation, we get

$$(-m\omega^2 + h\omega i + m\omega_0^2)\, X(\omega) = 0 \tag{25}$$

whence $X(\omega) \equiv 0$ and $x(t) \equiv 0$. We only get the trivial (zero) solution. The solutions found in Sec. 7.3 were unbounded (exponentially large) as $t \to -\infty$ and therefore they are not obtainable via the Fourier transformation. That is why, in Sec. 14.2, in solving equation (16) we only got one solution which did not include arbitrary constants. Now suppose there is no friction, i.e. $h = 0$. Then in place of (25) we get

$$(-m\omega^2 + m\omega_0^2)\, X(\omega) = 0 \quad \text{or} \quad (\omega^2 - \omega_0^2)\, X(\omega) = 0 \tag{26}$$

It might appear that from this it also follows that $X(\omega) \equiv 0$. But actually (26) is satisfied by the function

$$X(\omega) = C_1\delta(\omega - \omega_0) + C_2\delta(\omega + \omega_0) \tag{27}$$

where C_1 and C_2 are arbitrary constants. Indeed (see Sec. 6.1),

$$\begin{aligned}(\omega^2 - \omega_0^2)\, &[C_1\delta(\omega - \omega_0) + C_2\delta(\omega + \omega_0)] \\ &= C_1(\omega^2 - \omega_0^2)\,\delta(\omega - \omega_0) + C_2(\omega^2 - \omega_0^2)\,\delta(\omega + \omega_0) \\ &= C_1(\omega_0^2 - \omega_0^2)\,\delta(\omega - \omega_0) + C_2[(-\omega_0)^2 - \omega_0^2]\,\delta(\omega + \omega_0) = 0\end{aligned}$$

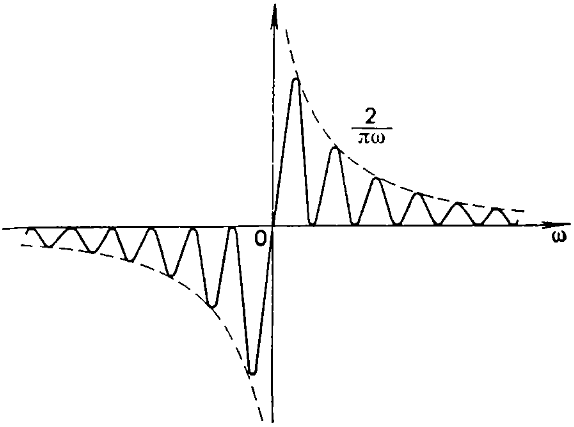


Fig. 196

From (27), by the inversion formula (4), we get

$$x(t) = \int [C_1\delta(\omega - \omega_0) + C_2\delta(\omega + \omega_0)]\, e^{i\omega t}\, d\omega = C_1 e^{i\omega_0 t} + C_2 e^{-i\omega_0 t}$$

We arrive at a solution that we know from Sec. 7.3.

Now let us consider the Fourier transforms (images) of certain functions mentioned in Sec. 6.3. Here, the Fourier operator will be denoted by $\mathfrak{F}$. Formula (9) signifies that

$$\mathfrak{F}[\delta(t)] = \frac{1}{2\pi} \tag{28}$$

It is easy to verify that, conversely $\mathfrak{F}[1] = \delta(\omega)$.

In order to compute $\mathfrak{F}$ [sgn t] (see Sec. 6.3), first replace the limits of integration in the right member of (10) by $\pm N$, where N is great, to get

$$\frac{1}{2\pi}\int_{-N}^{N} \operatorname{sgn} t \cdot e^{-i\omega t}\, dt = \frac{1}{2\pi}\left[-\int_{-N}^{0} e^{-i\omega t}\, dt + \int_{0}^{N} e^{-i\omega t}\, dt\right] = \frac{1 - \cos \omega N}{\pi i \omega}$$

The graph of this function multiplied by i is shown in Fig. 196; for large N it oscillates frequently about the graph $\frac{1}{\pi\omega}\cdot$ Therefore, on the basis of the same reasoning as in Sec. 14.2, it is natural to assume that

$$\mathfrak{F}[\operatorname{sgn} t] = \frac{1}{\pi i \omega} \tag{29}$$

or

$$\int_{-\infty}^{\infty} \frac{1}{\pi i \omega} e^{i\omega t}\, d\omega = \operatorname{sgn} t$$

True enough, for if we regard the integral in the sense of the principal value, then we can see that the last formula holds true, just as in Sec. 14.3.

From (29) there follows the Fourier transform of the unit function:

$$\mathcal{F}[e(t)] = \mathcal{F}\left[\frac{1}{2}+\frac{1}{2}\operatorname{sgn} t\right]=\frac{1}{2}\,\delta(\omega)+\frac{1}{2\pi i\omega} \qquad (30)$$

This formula is in agreement with (28), since it follows from (30) that

$$\mathcal{F}[\delta(t)] = \mathcal{F}[e'(t)] = i\omega\mathcal{F}[e(t)] = \frac{1}{2\pi}$$

Of the other properties of the Fourier transformation, let us examine the asymptotic behaviour of the Fourier transform as $\omega \to \pm\infty$. For the sake of simplicity we assume that the function $f(t)$ is nonzero only on the finite interval t or at least tends to zero as $t \to \pm\infty$ together with its derivatives. Then one should not think that $F(\omega)$ too must necessarily have the same properties as $\omega \to \pm\infty$: the asymptotic behaviour of $F(\omega)$ depends on something quite different.

From (10) it follows immediately that if $f(t) = \delta(t - t_0)$, then $F\omega = \frac{1}{2\pi}\, e^{-i\omega t_0}$, whence $F(\omega)$ as $\omega \to \pm\infty$ remains bounded but does not tend to zero. It can be shown that this will always happen if $f(t)$ contains delta terms.

Now suppose $f(t)$ does not have any delta terms, but has so-called *discontinuities of the first kind*, that is, points t_0 for which the limits $f(t_0 - 0)$ and $f(t_0 + 0)$ are finite but are not equal. (A typical function with a discontinuity of the first kind is the unit function shown in Fig. 72.) Then, as was shown in Sec. 6.3, $f'(t)$ will have delta terms. But by virtue of page 592, the function $f'(t)$ has a spectral density $i\omega F(\omega)$ and so, by the foregoing, we get that $i\omega F(\omega)$ is bounded as $\omega \to \pm\infty$ but it does not tend to zero. Hence, under these assumption $F(\omega)$ tends to zero at the rate of $\frac{1}{\omega}$ as $\omega \to \pm\infty$.

If the function itself, $f(t)$, is continuous but its derivative has discontinuities of the first kind (this means that the graph of $f(t)$ has corners, as in Fig. 76), then in similar fashion we find that $F(\omega)$ is of the order of $\frac{1}{\omega^2}$ as $\omega \to \pm\infty$, and so forth. Finally, if derivatives of $f(t)$ of any order are continuous, then $F(\omega)$, as $\omega \to \pm\infty$, tends to zero faster than any negative power of $|\omega|$; such, for example, will be the case for $f(t) = \frac{1}{t^2+1}$ (see formula (40) of Ch. 5 in the right member of which we have substitute $|\omega|$ for ω when $\omega < 0$), e^{-at^2} (see Sec. 14.5), and so on.

For large ω, the Fourier integral is an integral of a rapidly oscillating function. Such integrals were considered in Sec. 3.4, and now we can better understand why in the asymptotic formulas there the values of the integrand and its derivatives participated only at the endpoints of the interval of integration. Indeed, there we considered integrals of the type

$$\int_a^b f(t)\, e^{i\omega t}\, dt$$

as $\omega \to \pm\infty$. But such an integral can be rewritten as

$$\int_{-\infty}^{\infty} \tilde{f}(t)\, e^{i\omega t}\, dt$$

where the function $\tilde{f}(t)$ is obtained from $f(t)$ by continuing $f(t)$ outside the interval $a \leqslant t \leqslant b$ by identical zero. Thus we obtain (up to the inessential sign of ω and the coefficient $\frac{1}{2\pi}$) the Fourier transform (image) of the function $\tilde{f}(t)$. But this function and its derivatives have discontinuities at $t = a$ and $t = b$, where $f(t)$ vanishes identically. By virtue of what has been said, these singularities are what determine the coefficients in the asymptotic formulas if the function $f(t)$ itself is sufficiently smooth. It is also clear that if $f(t)$ and its derivatives have discontinuities between a and b, then these discontinuities too will have their effect on the asymptotic formulas, as was mentioned in Sec. 3.4.

On the other hand, the procedure of integration by parts — Sec. 3.4 was based on this — is useful in the theory of Fourier transformation since it permits not only determining the order of the Fourier transform as $\omega \to \pm\infty$, but also obtaining the appropriate asymptotic formulas.

Let us consider, for example, the Fourier transform of the function $f(t) = \dfrac{1}{1 + |t|}$:

$$F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{1 + |t|}\, e^{-i\omega t}\, d\omega$$

Here, $f(t)$ has a singularity (corner) at $t = 0$. Integration by parts gives

$$F(\omega) = \frac{1}{2\pi}\left[\frac{1}{1 + |t|}\, \frac{e^{-i\omega t}}{-i\omega}\Bigg|_{t=-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{e^{-i\omega t}}{-i\omega}\, \frac{-1}{(1 + |t|)^2}\, \mathrm{sg}\ t\, dt\right]$$

where $\operatorname{sgn} t = \frac{d|t|}{dt} = \frac{|t|}{t}$. Hence

$$F(\omega) = \frac{1}{2\pi}\,\frac{i}{\omega}\int\limits_{-\infty}^{\infty} \frac{\operatorname{sgn} t}{(1+|t|)^2}\, e^{-i\omega t}\, dt$$

Here the integrand already has a jump at $t = 0$ equal to 2 and so a second integration by parts yields a delta term, which generates the principal term of the asymptotic formula:

$$F\omega = \frac{1}{2\pi}\,\frac{i}{\omega}\left\{-\int\limits_{-\infty}^{\infty} \frac{e^{-i\omega t}}{-i\omega}\left[2\delta(t) + \frac{-2}{(1+|t|)^3}\right] dt\right\}$$

$$= \frac{1}{\pi}\left[\frac{1}{\omega^2} - \frac{2}{2\omega^2}\int\limits_{-\infty}^{\infty} \frac{1}{(1+|t|)^3}\, e^{-i\omega t}\, dt\right] \tag{31}$$

Since the last integral tends to zero as $\omega \to \pm\infty$, we get the asymptotic expression

$$F(\omega) \approx \frac{1}{\pi\omega^2} \tag{32}$$

which is more exact than the general assertion that $F(\omega)$ must be of the order of $\frac{1}{\omega^2}$ as $\omega \to \pm\infty$.

Expression (32) may be refined by a second integration by parts of the last integral in the right member of (31). The computation, which we leave to the reader, yields

$$F(\omega) = \frac{1}{\pi}\left[\frac{1}{\omega^2} - \frac{6}{\omega^4} + \frac{12}{\omega^4}\int\limits_{-\infty}^{\infty} \frac{1}{(1+|t|)^5}\, e^{-i\omega t}\, dt\right]$$

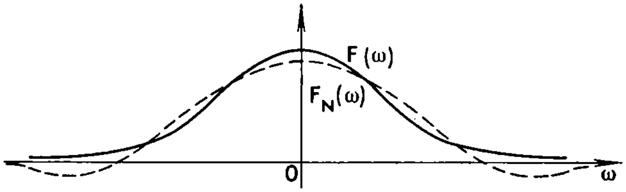
which leads to the asymptotic expression

$$F(\omega) \approx \frac{1}{\pi}\left(\frac{1}{\omega^2} - \frac{6}{\omega^4}\right)$$

that is more exact than (32). This process can be continued.

Let us take a look at another question. In Sec. 14.5 it was shown that the function $F(\omega) = \frac{1}{2\sqrt{\pi}}\, e^{-\omega^2/4}$ serves as the Fourier transform of the function $f(t) = e^{-t^2}$. Here, $f(t)$ and its derivatives do not have any singularities at all and that is why it turned out that $F(\omega)$ tends to zero faster than any negative power of $|\omega|$ as $\omega \to \pm\infty$. At

Fig. 197



the same time, by virtue of the foregoing, for an arbitrarily large fixed N the function

$$F_N(\omega) = \frac{1}{2\pi} \int_{-N}^{N} e^{-t^2} e^{-i\omega t}\, dt$$

that approximates $F(\omega) = F_\infty(\omega)$ is of the order of $|\omega|^{-1}$ as $\omega \to \pm\infty$.

How do we fit these facts?

The point is that for large N the coefficients of the expansion of the function $F_N(\omega)$ in negative powers of ω becomes extremely small so that terms involving ω^{-1}, ω^{-2}, and so on disappear in the limit. This expansion is easily obtained by integration by parts and begins with the terms

$$F_N(\omega) \sim \frac{1}{\pi} e^{-N^2} \frac{\sin N\omega}{\omega} - \frac{2}{\pi} N e^{-N^2} \frac{\cos N\omega}{\omega^2} + \ldots$$

For example, already for $N = 5$ the coefficient of ω^{-1} is of the order of 10^{-12}, which is extremely small. The graphs of the functions $F(\omega)$ and $F_N(\omega)$ are shown schematically in Fig. 197. The difference in the nature of the "wings" for large N is of hardly any significance since the values themselves are very small.

Exercises

1. Prove that if $F(\omega)$ is the image of $f(t)$, then $\frac{1}{|a|} F\left(\frac{\omega}{a}\right)$ serves as the image of the function $f(at)$ $(a = \text{constant})$.
2. Use Exercise 1 of Sec. 14.2 to find the Fourier transforms of the functions $e^{i\beta t - \alpha|t - t_0|}$, $\frac{te^{-\alpha|t|}}{|t|}$ $(\alpha > 0)$.

14.5 Bell-shaped transformation and the uncertainty principle

An interesting example of the Fourier transformation is given by the function

$$f(t) = Ce^{-at^2} \quad (C, a = \text{constant} > 0) \tag{33}$$

(its graph is bell-shaped, see Fig. 187). The formula (10) gives us

$$F(\omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty}Ce^{-at^2}e^{-i\omega t}\,dt=\frac{C}{2\pi}\int_{-\infty}^{\infty}e^{-at^2-i\omega t}\,dt$$

Transforming the exponent by the formula

$$-at^2-i\omega t=-a\left(t^2+\frac{i\omega}{a}\,t\right)=-a\left(t+\frac{i\omega}{2a}\right)^2-\frac{\omega^2}{4a}$$

and setting $\sqrt{a}\left(t+\frac{i\omega}{2a}\right)=z$, we get

$$F(\omega)=\frac{C}{2\pi}\int_{-\infty}^{\infty}e^{-a\left(t+\frac{i\omega}{2a}\right)^2}e^{-\frac{\omega^2}{4a}}\,dt=\frac{C}{2\pi\sqrt{a}}e^{-\frac{\omega^2}{4a}}\int_{(L)}e^{-z^2}\,dz \tag{34}$$

where (L) is a straight line in the complex plane z parallel to the real axis and passing through the point $z=\frac{\omega}{2\sqrt{a}}i=yi$.

We will now show that the integral

$$\int_{(L)}e^{-z^2}\,dz=\int_{-\infty}^{\infty}e^{-(x+iy)^2}\,dx=I_y \tag{35}$$

actually does not depend on y. To do this, differentiate I_y with respect to the parameter y (see Sec. 3.6) to get

$$\frac{dI_y}{dy}=\int_{-\infty}^{\infty}\frac{\partial}{\partial y}\,e^{-(x+iy)^2}\,dx=-2i\int_{-\infty}^{\infty}e^{-(x+iy)^2}(x+iy)\,dx$$

But the last integral can be evaluated:

$$\frac{dI_y}{dy}=ie^{-(x+iy)^2}\Big|_{x=-\infty}^{\infty}=ie^{-x^2+y^2}e^{-2ixy}\Big|_{x=-\infty}^{\infty}=0$$

(the function obtained after integration vanishes at both limits). Thus, $I_y=\text{constant}=I_y|_{y=0}$. Setting $y=0$ in (35), we arrive at the familiar integral (Sec. 4.7)

$$\int_{(L)}e^{-z^2}\,dz=\int_{-\infty}^{\infty}e^{-x^2}\,dx=\sqrt{\pi}$$

Substituting the value found into (34), we get the spectral density of the bell-shaped transformation (33):

$$F(\omega)=\frac{C}{2\sqrt{\pi a}}e^{-\frac{\omega^2}{4a}} \tag{36}$$

We see that the relationship of density and ω is similar, bell-shaped, but with different coefficients.

An important consequence follows from the formulas (33) and (36). A bell-shaped function does not completely vanish for any value of the variable. Still, we can reasonably determine the width of the bell (cf. Sec. 3.2). For example, if for the width we take that portion over which the height diminishes e times from the maximum value, then the width of the bell (33) is equal to $\Delta t = 2/\sqrt{a}$, whereas the width of the transformed bell (36) is equal to $\Delta\omega = 4\sqrt{a}$. Thus, if we vary a, then one of the bells becomes narrower and the other just as many times wider so that

$$\Delta t \cdot \Delta\omega = 8 = \text{constant} \qquad (37)$$

It can be shown that an analogous rule holds for any shape of the functions being transformed. This result is of fundamental importance. Suppose, as in Sec. 14.2, we are considering the action of a force $f(t)$ on a system of oscillators with distinct natural frequencies. We see that the more localized (compressed) the outer force is in time, the more "spread" is its spectrum; that is, the greater the number of oscillators with distinct frequencies that this force will excite with roughly the same amplitude. Conversely, by increasing the selectivity, i.e., compressing the spectrum, we are forced to spread out the external action in time. This impossibility of simultaneously localizing an external action in time and enhancing the selectivity of that action is one of the manifestations of the so-called *uncertainty principle*, which plays a fundamental role in modern physics.

From the uncertainty principle it follows, for one thing, that if a certain system is in oscillation with a variable frequency, then it is meaningless to speak of the value of the frequency at a given time: for instance, in acoustics one cannot speak of the exact pitch of sound at a given instant of time, and so forth. The time interval over which this frequency is determined cannot be taken substantially less than the oscillation period; sound of a definite pitch cannot last too short a time!

In quantum mechanics, the energy of a particle is connected with the frequency of the wave function. The wave function ψ of a particle with a definite energy E is proportional to $e^{-iEt/\hbar}$, where $\hbar = 1.05 \times \times 10^{-27}$ erg-s is Planck's constant (Planck introduced the quantity $h = 6.62 \cdot 10^{-27}$ erg-s; it is more convenient to write formulas with $\hbar = h/2\pi$). If a particle is observed during a short time interval Δt, then, as follows from the foregoing, the frequency of its wave function, i.e., the quantity $E/\hbar$, can only be found with a low accuracy. We get the relation: $\Delta E \cdot \Delta t$ is of the order of $\hbar$. Similarly, the wave function of a particle with a definite momentum along the x-axis, $p = mv_x$, is proportional to $e^{ipx/\hbar}$. If it is known that the particle resides in a definite interval between x and $x + \Delta x$, then the wave function of the particle is nonzero only within this interval. Expanding the wave

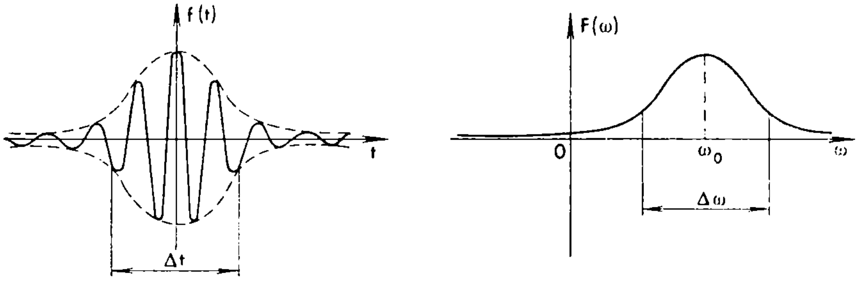


Fig. 198

function in a Fourier integral, we conclude that the momentum of the particle is known only with a definite accuracy Δp; the Fourier integral will contain large terms involving $e^{ipx/\hbar}$, where $p_1 < p <$ $< p_1 + \Delta p$. From this it follows that $\Delta p \cdot \Delta x$ is of the order of $\hbar$.

From (33) and (36), on the basis of the properties of the Fourier transformation described in Sec. 14.4, it follows that

$$\text{if } f(t) = Ce^{-at^2+i\omega_0 t}, \text{ then } F(\omega) = \frac{C}{2\sqrt{\pi a}}\, e^{-\frac{(\omega-\omega_0)^2}{4a}}$$

The real part of $f(t)$ and the function $F(\omega)$ are shown in Fig. 198. The width Δt of the "wave packet" (which is defined in similar fashion to the function (33)) and the width $\Delta\omega$ of the corresponding spectrum are connected by the uncertainty relation (37), irrespective of the frequency ω_0 of oscillations. It is on this frequency that the shift of the spectrum along the axis of frequencies depends.

Note the remarkable application of the bell-shaped transformation to the derivation of the normal law of probability theory (Sec. 13.8). Let the random variables $\xi_1, \xi_2, \ldots, \xi_n$ be independent but distributed by the same law of density distribution $\varphi(x)$; for the sake of simplicity we assume that the mean value $\overline{\xi_1} = 0$. Consider the random variable $\eta_k = \frac{1}{2\pi} e^{-i\omega\xi_k}$ where ω is a specified real constant. Since η_k takes on the values $\frac{1}{2\pi} e^{-i\omega x}$ with probability density $\varphi(x)$, its mean value is

$$\overline{\eta_k} = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} e^{-i\omega x} \varphi(x)\, dx = \Phi(\omega)$$

which is just the Fourier transform of the function $\varphi(x)$.

Denoting by η the product of the variables η_k we get

$$\eta = \eta_1\eta_2, \ldots \eta_n = \frac{1}{(2\pi)^n} e^{-i\omega(\xi_1+\xi_2+\ldots+\xi_n)} = \frac{1}{(2\pi)^n} e^{-i\omega\xi}$$

where $\xi = \xi_1 + \xi_2 + \dots + \xi_n$, whence

$$(2\pi)^{n-1}\eta = \frac{1}{2\pi} e^{-i\omega\xi}$$

That is, by virtue of the preceding section the mean value of $(2\pi)^{n-1}\eta$ serves as the Fourier transform of the distribution function $\varphi_n(x)$ of the random variable ξ.

In Sec. 13.7 we saw that when independent random variables are multiplied, so also are their mean values.
Hence,

$$\overline{(2\pi)^{n-1}\eta} = (2\pi)^{n-1}\bar{\eta}_1\bar{\eta}_2 \dots \bar{\eta}_n = (2\pi)^{n-1}[\Phi(\omega)]^n$$

It is easy to verify that the function $\Phi(\omega)$ for $\omega = 0$ has a maximum. Indeed,

$$\Phi(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi(x)\,dx = \frac{1}{2\pi}$$

whereas for other ω,

$$|\Phi(\omega)| = \left|\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\omega x}\varphi(x)\,dx\right| \leqslant \frac{1}{2\pi}\int_{-\infty}^{\infty} |e^{-i\omega x}\varphi(x)|\,dx$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi(x)\,dx = \frac{1}{2\pi}$$

But on pages 81-82 we saw that in this case, for large n, the function $[\Phi(\omega)]^n$ may be replaced approximately by $Ae^{-c\omega^2}$, where $A = e^{n\psi(0)}$, $c = -\frac{1}{2}n\psi''(0)$, $\psi(\omega) = \ln\Phi(\omega)$.

In our example we get

$$\Phi(0) = \frac{1}{2\pi}, \quad \Phi'(0) = 0,$$

$$\Phi''(0) = -\frac{1}{2\pi}\int_{-\infty}^{\infty} x^2\varphi(x)\,dx = -\frac{\Delta^2}{2\pi}$$

(Δ^2 is the variance of the variable ξ_1, see Sec. 13.8), whence

$$\psi(0) = \ln\frac{1}{2\pi}, \quad \psi''(0) = \frac{\Phi''(0)\,\Phi(0) - [\Phi'(0)]^2}{\Phi^2(0)} = -\Delta^2$$

and so

$$(2\pi)^{n-1}[\Phi(\omega)]^n \approx (2\pi)^{n-1} e^{n\ln\frac{1}{2\pi}} e^{-\frac{1}{2}n\Delta^2\omega^2} = \frac{1}{2\pi} e^{-\frac{1}{2}n\Delta^2\omega^2}$$

By the foregoing we have obtained the Fourier transform of the function $\varphi_n(x)$. As we see, it is bell-shaped. By formulas (33) and (36) $\left(\text{in which } \frac{C}{2\sqrt{\pi a}} = \frac{1}{2\pi}, \frac{1}{4a} = \frac{n\Delta^2}{2}, \text{ whence } a = \frac{1}{2n\Delta^2}, C = \frac{1}{\sqrt{2\pi n}\,\Delta}\right)$ we get the distribution function of the random variable ξ:

$$\varphi_n(x) = \frac{1}{\Delta\sqrt{2\pi n}} e^{-x^2/2n\Delta^2}$$

We have arrived at the normal law (equation (25) of Ch. 13).

Exercises

1. Prove the uncertainty principle for a function $f(t)$ that is constant and nonzero only on a finite interval.
2. Establish a relationship between the uncertainty principle and the result of Exercise 1 of Sec. 14.4.

14.6 Harmonic analysis of a periodic function

Suppose an external action $f(t)$ is a periodic function with a certain period $T > 0$, that is,

$$f(t+T) \equiv f(t) \tag{38}$$

Then only harmonics $e^{i\omega t}$ having the same property (38) will take part in the expansion (4) into harmonics. From this we get

$$e^{i\omega(t+T)} \equiv e^{i\omega t} \text{ that is, } e^{i\omega T} = 1, \quad i\omega T = 2k\pi i.$$

(see Sec. 5.4) and we see that the frequency ω can take on only the values

$$\omega = \omega_k = \frac{2k\pi}{T} (k = ..., -2, -1, 0, 1, 2 ...) \tag{39}$$

Thus, a period function has a discrete spectrum and for this reason a delta-like spectral density

$$F(\omega) = \sum_{k=-\infty}^{\infty} a_k \delta\left(\omega - \frac{2k\pi}{T}\right)$$

Substituting into (4), we get

$$f(t) = \sum_{k=-\infty}^{\infty} a_k e^{\frac{2k\pi i}{T}t} \tag{40}$$

This is the so-called *Fourier series* into which the periodic function is expanded.

To compute the coefficients of (40), take the mean values of both sides of (40). By virtue of the periodicity of $f(t)$ the mean value of the left-hand side is

$$\bar{f} = \frac{1}{T}\int_0^T f(t)\,dt$$

In the right-hand member, the mean value of each harmonic is equal to zero since for $\omega \neq 0$ as well

$$\frac{1}{2N}\int\limits_{-N}^{N} e^{i\omega t}\,dt = \frac{1}{2N}\left.\frac{e^{i\omega t}}{i\omega}\right|_{t=-N}^{N} = \frac{\sin\omega N}{N\omega} \xrightarrow[N\to\infty]{} 0$$

However, one term on the right of (40) is, for $k = 0$, simply equal to the constant a_0, and its mean value is of course a_0. For this reason, from (40) we get

$$\overline{f} = \frac{1}{T}\int\limits_0^T f(t)\,dt = a_0$$

To find the coefficient a_k, multiply both sides of (40) by $e^{-\frac{2k\pi i}{T}t}$ (for a certain fixed k) and then take the mean values of both sides. Arguing as in the preceding paragraph, we get

$$a_k = \overline{f(t)\,e^{-\frac{2k\pi i}{T}t}} = \frac{1}{T}\int\limits_0^T f(t)\,e^{-\frac{2k\pi i}{T}t}\,dt \qquad (41)$$

If the function $f(t)$ is real, then we often use a different, real, form of the Fourier series. For this purpose, in the expansion (40) we combine the symmetric terms

$$\begin{aligned} f(t) &= a_0 + \sum_{k=1}^{\infty}\left(a_k e^{-\frac{2k\pi i}{T}t} + a_{-k}e^{-\frac{2k\pi i}{T}t}\right) \\ &= a_0 + \sum_{k=1}^{\infty}\left[(a_k + a_{-k})\cos\frac{2k\pi}{T}t + i(a_k - a_{-k})\,\sin\frac{2k\pi}{T}t\right] \\ &= a_0 + \sum_{k=1}^{\infty}\left(A_k\cos\frac{2k\pi}{T}t + B_k\sin\frac{2k\pi}{T}t\right) \qquad (42) \end{aligned}$$

Here, the coefficients are, by virtue of (41),

$$a_0 = \frac{1}{T}\int\limits_0^T f(t)\,dt,$$

$$A_k = a_k + a_{-k} = \frac{1}{T}\int\limits_0^T f(t)\left(e^{-\frac{2k\pi i}{T}t} + e^{\frac{2k\pi i}{T}t}\right)dt = \frac{2}{T}\int\limits_0^T f(t)\cos\frac{2k\pi}{T}t\,dt$$

$$(k = 1, 2, \ldots),$$

$$B_k = i(a_k - a_{-k}) = \frac{i}{T}\int\limits_0^T f(t)\left(e^{-\frac{2k\pi i}{T}t} - e^{\frac{2k\pi i}{T}t}\right)dt = \frac{2}{T}\int\limits_0^T f(t)\sin\frac{2k\pi}{T}t\,dt$$

$$(k = 1, 2, \ldots)$$

All quantities in the right members of (42) are real.

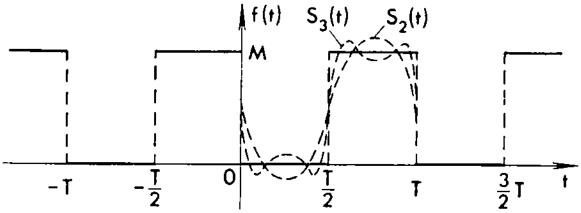


Fig. 199

To illustrate, take the series expansion of (42). Consider the "periodic step" shown in Fig. 199 by the heavy lines. Here $a_0 = \overline{f} = \frac{M}{2}$. By formulas (43)

$$A_k = \frac{2}{T}\int\limits_{\frac{T}{2}}^{T} M \cos\frac{2k\pi}{T} t\, dt = \frac{2M}{T}\frac{T}{2k\pi}\sin\frac{2k\pi}{T} t\Big|_{t=\frac{T}{2}}^{T} = 0,$$

$$B_k = \frac{2}{T}\int\limits_{\frac{T}{2}}^{T} M \sin\frac{2k\pi}{T} t\, dt = \frac{2M}{T}\frac{T}{2k\pi}\left(-\cos\frac{2k\pi}{T} t\Big|_{t=\frac{T}{2}}^{T}\right)$$

$$= -\frac{M}{k\pi}(1 - \cos k\pi)$$

From this,

$$B_1 = -\frac{2M}{1\pi},\ B_2 = 0,\ B_3 = -\frac{2M}{3\pi},\ B_4 = 0,\ B_5 = -\frac{2M}{5\pi}, \ldots$$

and we get the expansion

$$f(t) = \frac{M}{2} - \frac{2M}{\pi}\left(\frac{1}{1}\sin\frac{2\pi t}{T} + \frac{1}{3}\sin\frac{6\pi t}{T} + \frac{1}{5}\sin\frac{10\pi t}{T} + \ldots\right)$$

The second partial sum $S_2(t)$ and the third partial sum $S_3(t)$ of this series are shown in Fig. 199 by the dashed lines (for the sake of simplicity they are shown only over a period, they continue periodically in both directions). It is interesting to note that at the points of discontinuity the sum of the series is equal to the arithmetic mean (the half-sum, that is) of the limiting values of the function $f(t)$ on the left and on the right.*

Since to the oscillation period T there corresponds the frequency $\omega = \frac{2\pi}{T} = \omega_1$ (see (39)), we see from (39) and (40) (or (42)) that the general nonharmonic oscillation with frequency $\omega = \omega_1$ results from the

* Note that the partial sums $S_n(t)$ on the function $f(t)$ (Fig. 199) appreciably overshoot the step boundaries. This common property of $S_n(t)$ is termed the *Gibbs phenomenon.*

combination of harmonic oscillations with frequencies ω_1, $\omega_2 = 2\omega_1$, $\omega_3 = 3\omega_1$, and so on. The first frequency, ω_1, which is the frequency of the oscillation itself, is called the *fundamental frequency* and the corresponding harmonic is also termed the *fundamental harmonic*, the other frequencies and harmonics are called *higher*.

These results are very important, for example, in acoustics. We are familiar with the fact that the pitch of a sound corresponds to the frequency of vibrations of the air: for example, the sound "la" (or A) of the first octave has a frequency of 440 hertz (440 vibrations per second). But we also know that one and the same note sounded on different instruments exhibits different timbre (distinct tones). What does the timbre depend on? It turns out that the timbre is due to the higher harmonics superimposed on the fundamental harmonic. Each of them enters into the oscillation with a specific amplitude, and if we vary proportions between the amplitudes, this will change the timbre.

Another corollary of the results obtained is this. We know that if an oscillator with a natural frequency of ω_0 is excited by an external force with frequency ω, then for $\omega = \omega_0$ we have resonance.

But in practice we know that resonance sets in when $\omega = \frac{\omega_0}{2}$, $\frac{\omega_0}{3}$, $\frac{\omega_0}{4}$, and so on. The important thing here is that the exciting force should be periodic but harmonic! A force $f = \cos\frac{\omega_0 t}{2}$ does not excite an oscillator with natural frequency ω_0. Let us take another typical example. Suppose an oscillator with frequency ω_0 is excited by separate impulses: the best way is to provide one impulse per oscillation. The best time is when $x = 0$, $\frac{dx}{dt} > 0$, $\frac{dx}{dt}$ is maximal; the work of the force will then also be a maximum (we assume that the force acts in the direction of increasing x). Clearly if the force acts once each period, then the frequency of the force coincides with the frequency ω_0 of the oscillations.

Now change the action of the force. Suppose the force acts ever other time, say every odd oscillation. It is clear that the oscillations will build up indefinitely this way too and we will have resonance, although $\omega = \omega_0/2$. The crux of the matter here is that the force $f = \sum\delta(t - nT)$ (n an integer) consisting of separate impulses determined by the interval T is particularly rich in overtones. In the first case $T = \frac{2\pi}{\omega_0}$, in the second case of infrequent impulses $T = 2\cdot\frac{2\pi}{\omega_0} = \frac{2\pi}{\omega}$ the frequency of the force $\omega = \omega_0/2$, but in the second case as well we have a term involving $e^{i\omega_0 t} = e^{i2\omega t}$ in the expansion of $f(t)$. This is due to the presence of higher frequencies. Now if a force is acting with frequency $\frac{\omega_0}{3}$, then it consists of a mixture of harmonics with frequencies $\frac{\omega_0}{3}$, $2\frac{\omega_0}{3}$, $3\frac{\omega_0}{3}$, $4\frac{\omega_0}{3}$, and so on. Of these, it is precisely the third with frequency $3\frac{\omega_0}{3} = \omega_0$ that produces resonance.

On the contrary, a periodic external force with frequency $2\omega_0$, $3\omega_0, \ldots$ does not produce resonance but only generates higher harmonics in the oscillator under consideration.

Exercise

Expand in a Fourier series: (a) the function $A\,|\sin \alpha t|$; (b) a periodic function with period T and equal to ht for $0 < t < T$; (c) $\sum\limits_{n=-\infty}^{\infty} A\delta(t - 2nT)$.

14.7 Hilbert space

The expansions given in Sec. 14.6 have a marvellous geometrical interpretation. We have already pointed out (in Sec. 9.6) that any collection of entities in which linear operations can be performed (these operations are: addition of the entities and multiplication by scalars) may be interpreted as a multidimensional vector space. Now such operations can be performed on functions $f(t)$ considered over a certain interval of the t-axis (this interval may be finite or infinite and, in particular, it may coincide with the entire axis, but it must be the same for all functions considered)! What this means is that these functions may be regarded as vectors: they form a vector space. However, characteristic of a function space is the fact that it is infinite-dimensional. This is because when choosing an arbitrary function f there is an infinity of degrees of freedom because if, say, we indicate any finite number of values $f(t_k)$, then this function can also be chosen in many ways for the remaining values of t.

Let us first assume that the functions at hand take on real values. We can introduce into a function space the *scalar product*

$$(f, \varphi) = \int_a^b f(t)\, \varphi(t)\, dt \tag{44}$$

where a and b are the endpoints of the interval of the t-axis over which the functions are considered. The natural character of this formula is a consequence of the following reasoning. Choose n values of the independent variable

$$t = t_1,\ t_2, \ldots, t_n$$

and consider the values of any function f solely at these points $f(t_1)$, $f(t_2)$, ..., $f(t_n)$. We will then have n degrees of freedom, that is, these functions form an n-dimensional vector space, in which (by Sec. 9.6) a scalar product can be introduced via the formula

$$(f, \varphi) = \sum_{k=1}^{n} f(t_k)\, \varphi(t_k) \tag{45}$$

As the sample values t_k become denser, the n-dimensional space becomes an infinite-dimensional space, and the sum (45) naturally turns into the integral (44).

By formula (44) and according to the rules of vector algebra we can introduce the "modulus of the vector f", which is called the *norm* of the function f and is denoted by $\|f\|$:

$$\|f\|^2 = (f, f) = \int_a^b [f(t)]^2\, dt \tag{46}$$

The space of functions with scalar product (44) and, hence, with norm (46) is called a *Hilbert space* (more exactly, a real Hilbert space of functions on a specified interval with endpoints a, b). It includes continuous, discontinuous, and even unbounded functions. But since every vector must have a finite modulus, it follows that the space includes only those functions for which the integral (46) is either a proper integral or an improper convergent integral (see Sec. 3.1), that is to say, such as of necessity has a finite value. In particular, the delta function is not an element of Hilbert space because the integral of its square is equal to infinity (why?). In what follows we will confine ourselves to functions with a finite norm.

The idea of *orthogonality* is naturally introduced into Hilbert space: two functions $g_1(t)$ and $g_2(t)$ are orthogonal to each other (over an interval with endpoints a, b) if their scalar product is equal to zero, i.e. if

$$\int_a^b g_1(t)\, g_2(t)\, dt = 0 \tag{47}$$

Two orthogonal functions are similar to two perpendicular vectors. If there is an *orthogonal system of functions*, that is to say, a collection of pairwise orthogonal functions

$$g_1(t),\ g_2(t), \ldots,\ g_n(t), \ldots \tag{48}$$

then the problem often arises of expanding any given function $f(t)$ in terms of these functions, that is, an expansion of $f(t)$ in a series like

$$f(t) = \sum_{k=1}^{\infty} a_k g_k(t) \tag{49}$$

In Ch. 9 we considered the problem of resolving a vector in ordinary (three-dimensional) space into orthogonal (i.e. perpendicular) vectors. If there are three such orthogonal vectors, then any vector can be resolved in terms of them: such a triad of vectors is said to be complete and we can take it for the basis of that space. Now if there are two orthogonal vectors, then only vectors lying in the plane of these two vectors can be resolved in terms of the two vectors. In three-dimensional space, such a pair of vectors is not complete, it becomes complete only after adjoining a third vector.

Similarly, the orthogonal system of functions (48) is said to be *complete* if any function $f(t)$ can be expanded in a series of the form (49) in terms of that system; any such system of functions forms a basis in Hilbert space. If the system (48) is incomplete, then not all functions can be resolved in terms of the system, but only those functions that satisfy definite relations. It turns out that any incomplete orthogonal system of functions can be extended to form a complete orthogonal system by adjoining to the original system a certain number (which may even be infinite!) of functions.

In a finite-dimensional vector space it is very easy to determine the completeness or incompleteness of a system of orthogonal vectors: if the number of vectors in the system is equal to the dimensionality of the space, then the system is complete, but if the number is less than the dimensionality of the space, then the system is incomplete. In contrast, in an infinite-dimensional space, even an incomplete system may contain an infinitude of vectors so that the completeness of the system cannot be determined by that number alone. Ordinarily, it is not at all simple to establish the completeness of the system (48).

But if the completeness of the system of functions (48) has been established in some way, then it is very easy to find the coefficients of the expansion of the given function $f(x)$ in the series (49). To do this, form the scalar product of both sides of

$$f(t) = a_1 g_1(t) + a_2 g_2(t) + a_3 g_3(t) + \dots \tag{50}$$

by one of the functions $g_k(t)$. Then by virtue of the orthogonality relation, all terms in the right member of (50) vanish except one, in which the function is multiplied into itself. We then get the equality $(f, g_k) = a_k(g_k, g_k)$, or

$$a_k = \frac{(f, g_k)}{(g_k, g_k)} = \frac{\int\limits_a^b f(t)\, g_k(t)\, dt}{\int\limits_a^b g_k^2(t)\, dt} \tag{51}$$

(cf. formula (28) of Ch. 9).

An important example of a complete orthogonal system of functions on a finite interval $0 \leqslant t \leqslant T$ is given by

$$1,\ \cos\frac{2\pi}{T}t,\ \sin\frac{2\pi}{T}t,\ \cos\frac{4\pi}{T}t,\ \sin\frac{4\pi}{T}t,\ \cos\frac{6\pi}{T}t,\ \sin\frac{6\pi}{T}t,\ \ldots \tag{52}$$

The orthogonality of this system can be obtained by direct computation of integrals of the type (47) (see Exercise 1). The completeness of this system was actually demonstrated in Sec. 14.5, since the expansion in terms of the system (52) is nothing but the expansion (42), and we have seen that it is possible for any function $f(t)$ (since a periodic function with period T can assume absolutely arbitrary values for $0 < t < T$). If we compute the coefficients of the expansion in terms of the system (52) with the aid of the general formula (51) and the easy-to-verify equations

$$\int_0^T 1^2\,dt = T,\ \int_0^T \cos^2\frac{2k\pi}{T}t\,dt = \frac{T}{2},\ \int_0^T \sin^2\frac{2k\pi}{T}t\,dt = \frac{T}{2}$$
$$(k = 1, 2, \ldots)$$

we get the formulas (43).

An interesting generalization of the Pythagorean theorem to Hilbert space results if we take the scalar product of the left and also the right member of (50) into itself. Then in the right-hand member all the pairwise scalar products are equal to zero by virtue of the orthogonality relation, leaving only the scalar squares of all terms, and we get

$$(f, f) = a_1^2(g_1, g_1) + a_2^2(g_2, g_2) + \ldots$$

or

$$\|f\|^2 = a_1^2\|g_1\|^2 + a_2^2\|g_2\|^2 + a_3^2\|g_3\|^2 + \ldots \tag{53}$$

On the left we have the square of the modulus of the vector f, on the right, the sum of the squares of its projections on the basis vectors $g_1, g_2, \ldots$

If we consider a complex Hilbert space, which means the functions of a real argument can assume complex values, then the formula for the scalar product will, instead of (44), have the form

$$(f, \varphi) = \int_a^b f(t)\,[\varphi(t)]^*\,dt$$

where the asterisk denotes a complex conjugate quantity (see Secs. 5.2 and 9.6). For this reason the formula for the norm will also change:

$$\|f\|^2 = (f, f) = \int_a^b f(t)\,[f(t)]^*\,dt = \int_a^b |f(t)|^2\,dt$$

Similar changes will occur in other formulas as well. An important instance of a complete orthogonal system of functions in complex Hilbert space over the interval $0 \leqslant t \leqslant T$ is the set of functions

$$\ldots, e^{-\frac{4\pi i}{T}t}, e^{-\frac{2\pi i}{T}t}, 1, e^{\frac{2\pi i}{T}t}, e^{\frac{4\pi i}{T}t}, \ldots, e^{\frac{2k\pi i}{T}t}, \ldots \qquad (54)$$

The expansion in terms of this system of functions is precisely the expansion (40) of Sec. 14.6.

Such a geometric approach to the set of functions enables us to obtain many important and interesting consequences that lie beyond the scope of this book.

We conclude this section with a few words on the development of the concept of function. In the 18th century, when the concept first appeared, a function was conceived as of necessity being specified by a formula. For this reason, the case of a Fourier expansion of a function that is specified by distinct formulas over different ranges of the argument (see Sec. 14.6) was a puzzle to mathematicians for some time: was this one function (it was a single series and it was formed via a very definite law) or was it several? An analysis of similar instances led, in the 19th century, to the presently accepted definition of a function as an arbitrary law of correspondence between dependent and independent variables. This approach turned out to be useful for the logical substantiation of mathematics as a whole.

However, from the viewpoint of applications this definition is too amorphous and hazy. Such functions had the right to exist as, say, the *Dirichlet function* that is equal to 0 for irrational values and to 1 for rational values of the independent variable (try to imagine the graph of such a function!) and other similar functions that appear to have meaning only in a formally logical sense. In applications, a function is called upon to constitute a working organism and not a disorganized hodgepodge of values. Today we have a particularly clear view of the role of functions specified by formulas, to be more exact, analytic functions (Ch. 5).

But the logical analysis of the function concept that was carried out in the 19th century was also fertile as far as applications were concerned. Thus, functions specified by several formulas (piecewise analytic functions) are frequently encountered in applications and no longer cause discomfort (the most elementary case being the unit function $e(x)$ in Sec. 6.3). Their position was still more clarified after the discovery, in the 20th century, of generalized functions: the Dirac delta function and associated functions (a mathematically rigorous theory of generalized functions was developed in 1936 by the Soviet mathematician S.L. Sobolev). For instance, the function $f(x)$ that is equal to $f_1(x)$ for $x < a$ and to $f_2(x)$ for $x > a$ may be written down as a single formula:

$$f(x) = f_1(x)\, e(a - x) + f_2(x)\, e(x - a)$$

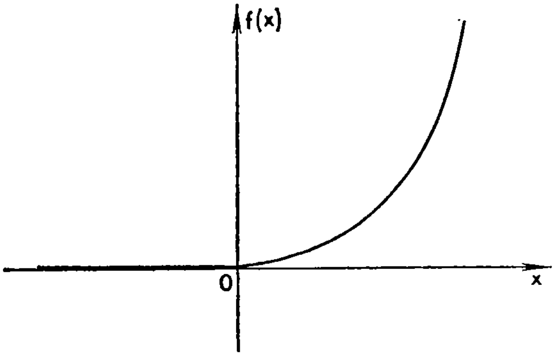


Fig. 200

Apparently, in applications today only the following functions have an individual ("personal") value: analytic, piecewise analytic, and simple generalized functions. (In the theory of random processes of the Brownian motion type, an important role is played by continuous nowhere differentiable — and hence nonanalytic — functions that describe particle trajectories; but these functions cannot be reproduced and so only have a probabilistic value and not an individual value).

The role played by singularities of a function — in particular those that arise when a function is patched together with different formulas — comes to the fore when passing to complex values of the independent variable. Consider, for example, the function

$$f(x) = x^2 e(x) = \begin{cases} 0 & (-\infty < x < 0), \\ x^2 & (0 < x < \infty) \end{cases}$$

shown in Fig. 200. The two portions of $f(x)$ have been patched together at $x = 0$ with the continuity of the derivative observed. This function can be approximately represented in the form

$$f(x) = \frac{x^2}{1 + e^{-\alpha x}}$$

where α is very great. But if we allow for the independent variable assuming complex values, then the right-hand side will have poles at $1 + e^{-\alpha x} = 0$, or

$$x = \frac{\pi}{2\alpha}(2k + 1)\,i \quad (k = 0,\ \pm 1,\ \pm 2, \ldots)$$

as $\alpha \to \infty$, these poles fill the entire imaginary axis, which separates the ranges of the two formulas for the function.

As we have seen in Sec. 14.4, "patching" is also seen under harmonic analysis of a function in the asymptotic behaviour of its Fourier transform.

Characteristic of the 20th century is yet another approach to the notion of function, namely as an element of a function space, for example, of Hilbert space, that is, as a member of a functional assemblage. Such an approach has certain diverse theoretical and applied advantages in many problems, but they lie outside the scope of this text.

Exercises

1. Prove that the system of functions (52) is orthogonal on the interval $0 \leqslant t \leqslant T$, that the same system is nonorthogonal on the interval $0 \leqslant t \leqslant \frac{T}{2}$; that the same system on the interval $0 \leqslant t \leqslant 2T$ is orthogonal but incomplete.
2. Prove that the system of functions (54) is orthogonal on the interval $0 \leqslant t \leqslant T$.

14.8 Modulus and phase of spectral density

In Sec. 14.2 we pointed out that the simplest receptors of oscillations (the ear, the eye, a photographic plate) record only the absolute amplitude; the readings of these receptors do not depend on the phase of the oscillations. This approach to oscillations is characteristic of the 19th century with its interest in energetics, since the energy of oscillations is determined solely by the modulus of the amplitude, to be more precise, by the modulus of the spectral density. If the oscillation is described by a real function $f(t)$, the energy flux for distinct types of oscillations is proportional to $[f(t)]^2$ or $[f'(t)]^2$ so that the total oscillation energy during time $-\infty < t < \infty$ is expressed in terms of the integrals I_0 or I_1:

$$I_0 = \int\limits_{-\infty}^{\infty} [f(t)]^2\, dt, \quad I_1 = \int\limits_{-\infty}^{\infty} [f'(t)]^2\, dt$$

But these integrals are readily expressible in terms of the square of the modulus of the spectral density. To obtain this expression for I_0, square both sides of (4) and then combine both integrals into one in the righthand member:

$$[f(t)]^2 = \int F(\omega) e^{i\omega t}\, d\omega \int F(\omega) e^{i\omega t}\, d\omega = \iint F(\omega_1)\, F(\omega_2)\, e^{i(\omega_1+\omega_2)t}\, d\omega_1\, d\omega_2$$

Now integrate both sides with respect to t and take advantage of the formulas (5), (9), and also of (4) of Ch. 6:

$$\begin{aligned} I_0 = \int [f(t)]^2\, dt &= \iiint F(\omega_1)\, F(\omega_2)\, e^{i(\omega_1+\omega_2)t}\, dt\, d\omega_1\, d\omega_2 \\ &= 2\pi \iint F(\omega_1)\, F(\omega_2)\, \delta(\omega_1 + \omega_2)\, d\omega_1\, d\omega_2 \\ &= 2\pi \int F(-\omega_2)\, F(\omega_2)\, d\omega_2 = 2\pi \int |F(\omega)|^2\, d\omega \end{aligned}$$

Similarly, since $i\omega F$ serves as the Fourier transform of the function $f'(t)$, it follows that

$$I_1 = \int [f'(t)]^2 \, dt = 2\pi \int | i\omega F(\omega) |^2 \, d\omega = 2\pi \int \omega^2 \, | F(\omega) |^2 \, d\omega$$

The corresponding formulas for periodic functions with period T were actually derived in Sec. 14.7 (see (53)):

$$I_0 = \int_0^T [f(t)]^2 \, dt = T \sum_{k=-\infty}^{\infty} | a_k |^2$$

where a_k are the coefficients of the expansion of the function $f(t)$ in the Fourier series (40). Here it suffices to find the energy during a single period, and in place of the integral with respect to the frequency we have a sum. The expression for I_1 is similar.

The two expressions of I — in terms of the integral with respect to time or the integral (sum) with respect to the frequencies — can be visualized as the expression of the square of the modulus of a vector by the Pythagorean theorem as the sum of the squares of its component (cf. the interpretation of formula (53) in Sec. 14.7). Here, there are different expressions because we make use of different systems of coordinates: in one, the coordinates are values of the function at distinct points, i.e. $f(t_1), f(t_1 + \Delta t), \ldots,$ in the other, the coordinates are Fourier coefficients. The equality of the two expressions for I shows that each of them is complete, not a single one of the components of the vector is forgotten.

Thus, the expression for energy does not depend on the phase of the oscillations: replacing $F(\omega)$ by $F(\omega) \, e^{i\alpha(\omega)}$, where $\alpha(\omega)$ is a real function, does not alter the integral I. Not only the total energy remains unchanged, but also the energy obtained by an oscillator tuned to one or another frequency. In the energy approach, the phase of oscillations is not essential.

In this 20th century, particularly in the latter half, special importance is attached to the transmission of information — from radio and television to cybernetic systems of control and research problems. It is clear that we lose information when we fail to record the phase: if different $f(t)$ correspond to the same spectrum, with respect to $| F |^2$, with distinct phases, then when the phase is unknown we cannot reproduce $f(t)$ if all we know is $| F |^2$. Writing $| F |$ and the phase, we would obtain, to put it crudely, twice as much information from the given wave. What is more, it appears possible to transmit information via changes in the phase of the oscillations at a fixed amplitude.

There are two modes of transmitting information in radio engineering: by means of *amplitude modulation,* (Fig. 201),

$$f(t) = a(t) \cos \omega_0 t$$

Fig. 201

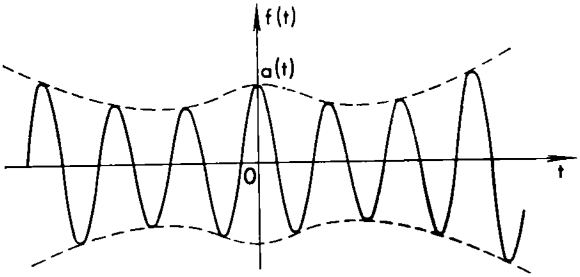
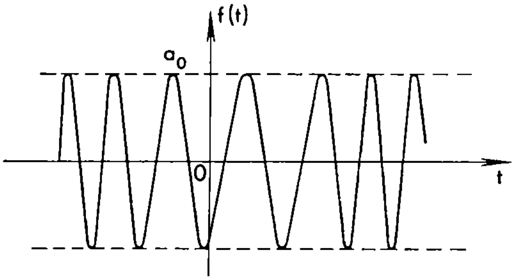


Fig. 202



and by means of *frequency modulation*, (Fig. 202),

$$f(t) = a_0 \cos\left(\omega_0 t + c\int^t b(t)\,dt\right)$$

The information to be transmitted is contained in the function $a(t)$ in the former case and in the function $b(t)$ in the latter. From the viewpoint of harmonic analysis, that is, the expansion of $f(t)$ into the Fourier integral, in both cases we have to do with a spectrum in which $F(\omega)$ is nonzero in the neighbourhood of ω_0, which is a quantity called the *carrier frequency*.

In the case of amplitude modulation, to determine the spectrum we expand $a(t)$:

$$a(t) = \int A(\omega)\, e^{i\omega t}\, d\omega$$

Clearly, if $A(t)$ is nonzero only in the band $-\Delta < \omega < \Delta$, the expansion of $f(t)$ will be concentrated in the band of frequencies of the same width, i.e. $F(\omega) \neq 0$ for $\omega_0 - \Delta < \omega < \omega_0 + \Delta$ (and of course for $-\omega_0 - \Delta < \omega < -\omega_0 + \Delta$ which, by virtue of (5), does not yield anything new for the frequencies). This width does not depend on the absolute value of the amplitude.

In the case of frequency modulation the instantaneous value of the frequency is equal to the derivative of the phase:

$$\omega(t) = \frac{d}{dt}\left[\omega_0 t + c\int^t b(t)\,dt\right] = \omega_0 + cb(t)$$

However, by the uncertainty principle (Sec. 14.5), we must have a sufficient time interval τ for the frequency to change substantially and for this change to be recorded:

$$\tau > (c\overline{b}(t))^{-1}$$

where $\overline{b}$ is to be understood as $(\overline{b^2})^{1/2}$. The corresponding time of variation of the function $b(t)$ itself depends on the frequency of the signal: this time is equal to $\tau = 1/\Delta$. Thus, if $c\overline{b}(t)/\Delta > 1$, then the width of the band used for transmission, that is, the width of the spectrum $F(\omega)$ of the function $f(t)$ is equal to $c\overline{b}(t)$ and is greater than the width of the frequency signal Δ.

Under certain conditions, a large width of the radio signal being transmitted has an advantage and permits reducing interference.

Figs. 201 and 202 show, in particular, that the two functions $f(t)$ with similar spectrum and similar relationship $|F(\omega)|^2$ may appear to be quite different. This difference is due to the difference of the phases, i.e., of the function $\varphi(\omega)$ in the expression $F(\omega) = \sqrt{|F(\omega)|^2} e^{i\varphi(\omega)}$.

ANSWERS AND SOLUTIONS

Sec. 14.1

1. $F(\omega) = \sum\limits_k a_k \delta(\omega - \omega_k)$.

2. This occurs if all ω_k are commensurable, i.e., $\omega_k = n_k\alpha$, where α does not depend on k and all n_k are integers. Then all terms, and so also the sum, have period $T = \frac{2\pi}{\alpha}$.

Sec. 14.2

1. $$F(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\alpha|t|}\, e^{-i\omega t}\,dt = \frac{1}{2\pi}\left(\int_{-\infty}^{0} e^{\alpha t - i\omega t}\,dt + \int_0^{\infty} e^{-\alpha t - i\omega t}\,dt\right)$$
$$= \frac{1}{2\pi}\left(\frac{1}{\alpha - i\omega} - \frac{1}{-\alpha - i\omega}\right) = \frac{1}{\pi}\,\frac{\alpha}{\alpha^2 + \omega^2}.$$

If we put $\alpha = \frac{1}{m}$, we get precisely one of the examples from which we obtained the delta function in Sec. 6.1 as $m \to \infty$

(i.e. $\alpha \to +0$). Hence, for $f(t) \equiv 1$ it will be true that $F(\omega) = \delta(\omega)$. The formula (4) yields

$$\int_{-\infty}^{\infty} \frac{1}{\pi} \frac{\alpha}{\alpha^2 + \omega^2} e^{it\omega}\, d\omega = e^{-\alpha|t|}$$

That is,

$$\int_{-\infty}^{\infty} \frac{\alpha \cos t\omega}{\alpha^2 + \omega^2}\, d\omega = \pi e^{-\alpha|t|}$$

When $\alpha = 1$ we get formula (1) of Ch. 1 in different notation.

2. From (17) we get

$$x(t) = \int_{-\infty}^{\infty} \frac{F(\omega)\, e^{i\omega t}\, d\omega}{m(\omega_0^2 - \omega^2) + i\omega h}$$

where

$$F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)\, e^{-i\omega t}\, dt$$

Replacing the variable of integration t by τ in the last integral, then substituting this integral into the first one, and finally changing the order of integration, we get, after some manipulation,

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(\tau)\, d\tau \int_{-\infty}^{\infty} \frac{e^{i\omega(t-\tau)}}{m(\omega_0^2 - \omega^2) + i\omega h}\, d\omega$$

The inner integral can be evaluated by the methods of Sec. 5.9. The integrand has simple poles for

$$\omega = \omega_{1,2} = \frac{ih}{2m} \pm \sqrt{-\frac{h^2}{4m^2} + \omega_0^2} = i\gamma \pm \omega^0$$

Both of these values are found in the upper half-plane.
For $t > \tau$ we find, using the upper semicircle $|\omega| = R \to \infty$, that the inner integral is equal to

$$2\pi i \left[\frac{e^{i\omega_1(t-\tau)}}{-2m\omega_1 + ih} + \frac{e^{i\omega_2(t-\tau)}}{-2\mathrm{m}\omega_2 + ih}\right] = \frac{2\pi}{m\omega^0} e^{-\gamma(t-\tau)} \sin \omega^0(t - \tau)$$

For $t < \tau$ we have to use the lower semicircle, which leads to the integral vanishing. Thus,

$$x(t) = \frac{1}{m\omega_0} \int_{-\infty}^{t} e^{-\gamma(t-\tau)} \sin \omega^0(t - \tau)\, f(\tau)\, d\tau$$

Sec. 14.4

1. The image of the function $f(at)$ is

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} f(at)\, e^{-i\omega t}\, dt = \frac{1}{a}\int_{-a\infty}^{a\infty} f(t_1)\, e^{-i\frac{\omega}{a}t_1}\, dt_1 = \frac{1}{a}F\left(\frac{\omega}{a}\right)$$

with the substitution ($at = t_1$) for $a > 0$; for $a < 0$, we have to invert the limits of integration, which yields

$$-\frac{1}{a}F\left(\frac{\omega}{a}\right) = \frac{1}{|a|}F\left(\frac{\omega}{a}\right)\cdot$$

2. $$e^{-it_0(\omega-\beta)}\frac{1}{\pi}\,\frac{\alpha}{\alpha^2+(\omega-\beta)^2}, \quad \frac{1}{\alpha}\cdot i\omega\,\frac{\alpha}{\alpha^2+\omega^2} = \frac{i\omega}{\alpha^2+\omega^2}\cdot$$

Sec. 14.5

1. Since a shift along the t-axis does not affect the uncertainty relation, we put $f(t) = C$ ($|x| < h$), $f(t) = 0$ ($|x| > h$) so that $\Delta t = 2h$. Then by (10), $F(\omega) = C\frac{\sin \omega h}{\omega h}\cdot$ For the width of the function obtained we take the width of the middle "hump", i.e. the distance between adjacent zeros. The $\Delta\omega = \frac{\pi}{h}$, whence

$$\Delta t\cdot\Delta\omega = 2h\,\frac{\pi}{h} = 2\pi = \text{constant.}$$

2. If in some way we have determined the width Δt of the function $f(t)$ and the width $\Delta\omega$ of its density $F(\omega)$, then the width of the function $f(at)$ is equal to $\frac{\Delta t}{|a|}$ and the width of its density $\frac{1}{|a|}F\left(\frac{\omega}{a}\right)$ is equal to $|a|\,\Delta\omega$. Here, $\frac{\Delta t}{|a|}\cdot|a|\,\Delta\omega = \Delta t\cdot\Delta\omega$, that is to say, it does not depend on a.

Sec. 14.6

(a) $\frac{2A}{\pi} - \sum_{k=1}^{\infty}\frac{4A}{\pi}\,\frac{\cos 2k\alpha t}{4k^2-1}$

(b) $\frac{hT}{2} - \frac{hT}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^k}{k}\sin\frac{2k\pi t}{T}$

(c) $\frac{A}{2T} + \frac{A}{T}\sum_{k=1}^{\infty}\cos\frac{k\pi t}{T}\cdot$ The period of the kth harmonic is equal to $\frac{2T}{k}\cdot$

Sec. 14.7

1. For $m \neq n$ we have

$$\int_0^T \cos\frac{2m\pi}{T}t\cdot\cos\frac{2n\pi}{T}t\,dt$$

$$=\frac{1}{2}\int_0^T\left[\cos\frac{2(m+n)\pi}{T}t+\cos\frac{2(m-n)\pi}{T}t\right]dt$$

$$=\frac{1}{2}\left[\frac{T}{2(m+n)\pi}\sin\frac{2(m+n)\pi}{T}t\right.$$

$$\left.+\frac{T}{2(m-n)\pi}\sin\frac{2(m-n)\pi}{T}t\right]_{t=0}^{T}=0$$

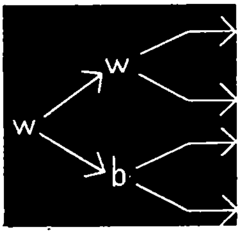
Here, $m = 0$ is possible; then $\cos\frac{2m\pi}{T}t \equiv 1$. In similar fashion we can verify the orthogonality to each other of the sines, and also of the sine and cosine. If the integral is taken from 0 to $\frac{T}{2}$, then the sines are no longer orthogonal to the cosines (the integral turns out to be nonzero). For the interval $0 \leqslant t \leqslant 2T$ the integrals are also zero, but since all functions (52) are periodic with period T, only that function can be expanded (in terms of these functions) whose values on the portion $T \leqslant t \leqslant 2T$ are an exact repetition of its values on the portion $0 \leqslant t \leqslant T$.

2. For $m \neq n$ it will be true that

$$\int_0^T e^{\frac{2m\pi i}{T}t}\left(e^{\frac{2n\pi i}{T}t}\right)^* dt=\int_0^T e^{\frac{2(m-n)\pi i}{T}t}dt=\frac{T}{2(m-n)\pi i}e^{\frac{2(m-n)\pi i}{T}t}\Big|_{t=0}^{T}=0$$

Chapter 15

DIGITAL COMPUTERS



In conclusion at least a few words are in order concerning the electronic digital computers that have wrought a revolution in modern applied mathematics, and even beyond.

The most elementary computing devices like the slide rule, the abacus, the desk calculator, and tables have been in use for a long time. But they fall far short of the demands made by modern engineering, science and economics. Many important problems are solvable in principle but require such an enormous amount of computation that their solution by the abovementioned devices cannot be obtained in reasonable periods of time. In order to obtain solutions, one had to forego many essential factors and this led to quantitative and ofttimes fundamental errors.

The search for more effective computational tools and the attainments of modern technology led to the construction of electronic digital computers whose speed exceeds that of the classical tools by many orders of magnitude. Although the underlying idea of such machines had been advanced in the 19th century, only the development of modern electronics made their realization a possibility. The first electronic digital computer was constructed in the U.S.A. in 1943, and in 1946 the American mathematician J. von Neumann (1903-1957) formulated the basic ideas and principles for constructing such machines. The widespread introduction of computers made it possible to solve a series of important problems and extended appreciably the range of fields in which mathematics and the allied science of cybernetics achieved useful results. There can be no question that the further development and, particularly, the expanded use of calculating machines will in this generation result in a fundamental reorganization of scientific research, technology, economic calculations, management, and so forth.

15.1 Analogue computers

Suppose a certain problem has been reduced to the solution of a mathematical equation. For short we will call the quantities that enter into the equation mathematical quantities (although they may have a definite physical meaning, which was why the equation was under consideration in the first place).

There are two basic modes of representing mathematical quantities that are to be operated on computationally. In one mode, these quantities are represented directly by the physical quantities of length, angle, electric voltage, and so forth, irrespective of the meaning that the original physical problem had. Then a physical scheme is outlined in which the physical quantities are transformed via the same law as the mathematical quantities are to be. A computing machine based on such a principle is an *analogue computer*, the simplest instance of which is the slide rule, where addition of mathematical quantities (the logarithms of the factors) are simulated by the addition of lengths. Here is another example: to compute an integral we can simulate the integrand by the law of opening a water faucet; then the integral itself will correspond to the total volume of water, which can be measured with ease. The most common analogue machines are those based on electrical analogies.

In the second mode, a device is made to represent in digital notation the mathematical quantities involved; the operations on these quantities are then replaced by arithmetical operations on the digits (numbers). Computing machines based on this principle are called *digital computers*. They include the abacus and the desk calculator. The biggest advances have been made by digital machines, but for the sake of completeness we will take a brief glance at analogue computers as well.

First of all, the *input parameters*, that is, the *operands*, or quantities that are to be operated on, ordinarily are capable of varying continuously over certain ranges; hence, this class of machines is termed *continuous-mode machines*. True, the input parameters could possibly be in the form of electric resistance that is varied discretely with the aid of a resistance box; such discreteness would not be of a fundamental nature, but only due to the design of the machine, whereas digital machines are fundamentally machines of discrete operation (*discrete-mode machines*).

Furthermore it is clear that the precision of input of the parameters and that of the results is not great in the case of analogue machines. Ordinarily it is of the order of percents, at best, tenths of a percent. This naturally restricts the possibilities of simulating involved computations. Besides, analogue machines are usually rather specialized, adapted to the solution of a definite narrow class of problems. But if such specialized problems have to be solved over and over again and high accuracy is not required, then the use of analogue machines and devices proves to be extremely effective. For example, electrical *integrators* for the solution of systems of ordinary differential equations are in wide use.

It often happens that one and the same relationship of quantities can be effected via different physical schemes. This justifies the *simulating* of physical processes. Suppose we have a physical system in which it is difficult to measure or compute some quantity S directly. We construct a new system of a different physical nature in which the quantities involved exhibit the very same functional relationship, and then the quantity corresponding to S is measured in the new system. What is more, if the mathematical equivalence of the two physical systems has been established, then the mathematical solution of the problem is unnecessary and so no calculations need be made at all, unless they are needed for some other purpose. Actually, what is utilized here is the similarity of phenomena based on specific characteristics (Sec. 8.10) with a dimensional proportionality (similarity) factor. In recent years, the *analogue solution* of problems based on electromechanical, optical-mechanical electrodiffusion, etc. analogies has seen extensive development.

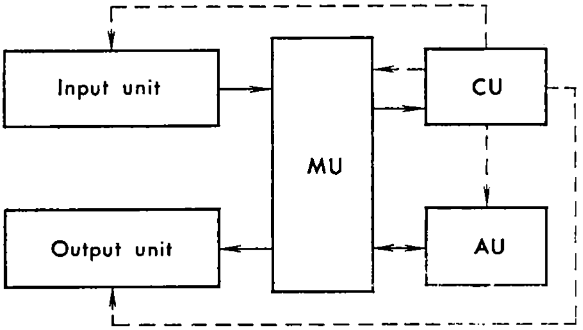
If a process is being studied and by the physical meaning of the problem the independent variable is the time t, then it often happens that the solution on an analogue machine is obtained as a function of t. We then say that the problem is solved in real time and the solution can be transferred directly (without human intervention) to the system under study for further use; this is the underlying principle of many automatic control devices. The use of real time also makes it possible to replace expensive equipment with computing machines when testing complicated engineering designs. For instance, an automatic pilot need not be tested in flight, which is expensive and dangerous, but in a test bed with the aircraft replaced by an analogue machine, which reacts to the automatic pilot as if it were the aircraft.

15.2 Digital computers

Let us now discuss digital computing machines. We will not consider the more common desk calculators that perform the four operations of arithmetic. These are very useful devices but they are rather slow because of the low speed of the computation itself and also because the input data for each arithmetic operation are entered manually by the operator. Both of these drawbacks are overcome in the electronic digital computer.

Fig. 203 is a block diagram of a digital computer showing the basic units and their interrelationships; the information paths are denoted by solid arrows, the control paths by dashed arrows. The *memory unit* (MU) has a definite number, say, 512, of *storage elements* (*locations*) each of which serves to record a single number or instruction. As we shall see in Sec. 15.3, the notation of an instruction does not differ in any way from that of a number. Some of these elements are

Fig. 203



filled from the input unit, whereas others are filled in the process of calculation (the contents of the storage elements may change many times during a computation or some may not be used at all). Guided by programmed instructions, the *control unit* (CU) sends numbers (or instructions) from the memory unit to the *arithmetic unit* (AU), which transforms them in accordance with the instructions and returns them to memory. On signals from the control unit, required results are automatically printed, and when the computation is completed, the control unit stops the machine.

By means of accessory equipment, a digital computer can give the result in the form of a graph approximated by computed coordinates of points or even a moving-picture display showing the variation of such a graph.

A *general-purpose digital computer* is capable of solving any mathematical problem that has an *algorithm* (a clear-cut sequence of the procedures determining the computational process such that starting with the input data leads of necessity to the required result). However, nonmathematical problems for which a similar algorithm can be indicated are also solvable on these machines. For example, in metalworking a digital computer can direct the machining of a workpiece with contours of any degree of complexity. The input data describe the shape and the results of computation are transformed into signals delivered to the control unit that operates the machine tool. The operation scheme is similar for controlling the flight of an aircraft or a technological process. In the case of deviations from the originally specified program, the machine can make optimal decisions by comparing possible variants, and it can also check the result. Digital computers have found extensive use in weather forecasting, transport problems, and the like. Incidentally, *special-purpose computers* are more effective in such cases than general-purpose machines because they are specially designed for the solution of a narrow range of problems.

15.3 Representation of numbers and instructions in digital computers

The decimal (denary, or base-10) number system that we studied in school and use in everyday life is not convenient when working with digital computers. The *binary system* (base-2) is used for number notation in such machines; here only two digits, 0 and 1, are needed (whereas in the base-10 system we use 10 digits) and all numbers are represented as combinations of these digits. For instance, the numeral 10100 indicates powers of two (not ten), i.e., the numbers two, four, etc. This can be written as follows:

$$1_{10} = 1_2,\ 2_{10} = 10_2,\ 3_{10} = 11_2,\ 4_{10} = 100_2,\ 5_{10} = 101_2$$

where the subscript indicates the number system.

Any integer can be written in the base-2 system merely by isolating powers of two beginning with the highest one. For example, in base-10 we have.

$$1971 = 1\cdot 1024 + 1\cdot 512 + 1\cdot 256 + 1\cdot 128 + 0\cdot 64 + \\ + 1\cdot 32 + 1\cdot 16 + 0\cdot 8 + 0\cdot 4 + 1\cdot 2 + 1\cdot 1$$

or $1971_{10} = 11110110011_2$.

In similar fashion using binary numbers, we can write bicimal fractions in place of ordinary decimal fractions. For example, take the binary number 10.1011_2. It is equivalent, in base-10, to $2 + \frac{1}{2} + + \frac{1}{8} + \frac{1}{16} = \frac{43}{16} = 2.6875_{10}$. Any nonintegral number can be written in the form of a finite or infinite bicimal; fractions are naturally rounded off in actual computations.

Tables of addition and multiplication in the base-2 system are very simple:

$$0 + 0 = 0,\quad 1 + 0 = 0 + 1 = 0,\quad 1 + 1 = 10, \qquad (1)$$

$$0\cdot 0 = 0,\quad 1\cdot 0 = 0\cdot 1 = 0,\quad 1\cdot 1 = 1 \qquad (2)$$

Using these tables, we can perform arithmetic operations on numbers written in the binary (base-2) system in exactly the same way we do in the denary (base-10) system.

Of course, don't think that a computing expert uses base-2 when working with a computer. Programs are written in base-10 notation and the conversion to base-2 is handled inside the computer by a special instruction, which we dispense with here so as to simplify our discussion. Neither will we go into the base-8 (octonary) and binary-coded decimal systems, which are used in such cases. The computation result is printed by the machine in the ordinary base-10 notation.

Numbers are entered in the memory unit (Sec. 15.2) and each one is placed in a storage element (location); all storage elements are of the same length, that is to say, they have the same number of *orders* (*digit positions*), each of which contains 0 or 1. Soviet digital computers have *floating-point representation*, which means that we write the number located between — 1 and 1, and the power of two into which (power) the number has to be multiplied. A certain number of digit positions (at the end of the storage element, for instance) are allotted to the exponent and the symbol code of the exponent.

For example, suppose there are six orders for these digits in a location accommodating 30 orders, the + sign represented as 0, the — sign as 1. Then the set of digits in the storage location

symbol code of number ↓ — fractional part — symbol code of exponent ↓ — exponential part

1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1

(3)

denotes the number

$$-\ 0.101011_2 \cdot 2^{+101_2} = -\left[\left(\frac{1}{2}+\frac{1}{8}+\frac{1}{32}+\frac{1}{64}\right)\cdot 2^{4+1}\right]_{10} = -21.5_{10}$$

The largest number that can be entered in this fashion is of the form

0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1

(4)

and is equal to $(1 - 2^{-23})\, 2^{31} = 2^{31} - 2^{8} \approx 2^{31}$, whereas the smallest positive number has the form

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1

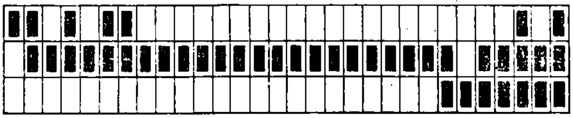
(5)

and is equal to $2^{-23} \cdot 2^{-31} = 2^{-54}$ (think this through!).

Actually, the recording of digits is done in different ways depending on the design of the machine. So that the numbers can be entered in the input unit of the machine (Sec. 15.2), they are punched on special tape (*punched tape*) or on a set of special cards (*punched cards*). For the sake of definiteness we will discuss only cards. A row on a card corresponds to a storage location, the length of the row being equal to the number of orders in the storage element; a punched hole corresponds to the digit 1, the absence of a hole to the digit 0. For example, if the numbers indicated in (3), (4) and (5) followed in a sequence in an input program, then the corresponding punched card would have a portion like that shown in Fig. 204 (check this). The holes are punched beforehand on a special perforator that is not connected with the computer and resembles somewhat the ordinary cash register in a shop.

A large program is written on a whole stack of cards, so if an error is detected in the program or if the input data are changed, then any card can be replaced with ease.

Fig. 204



Every storage location in the memory unit of the machine is a set of code elements, each of which can be in one of two states. This is for storing numbers and processing them as the computations proceed. The number of elements in a set is equal to the number of orders (digit positions) in the storage location, and each of the states of an element corresponds to the value 0 or 1 in the appropriate order. Such code elements come in different types, but they all have to possess two properties: *inertialessness*, which means the instantaneous transition from one state to the other, and *stability*, which means the capability of remaining in a given state for an indefinitely long time in the absence of a transition signal. For the code elements, use is made of magnetic ferrite cores that change the direction of the magnetic field when a current flows in the winding, magnetized and unmagnetized portions on a magnetic drum or magnetic tape (their operation is much like that of an ordinary tape recorder), and so forth.

Numbers are transferred from memory to the arithmetic unit and back again through a system of channels (there may be as many channels as there are digit positions in a location, in which case we have a *parallel-mode machine*).

The more locations there are in the memory unit, the more information that can be entered in the computer and the greater the flexibility of operation. For this reason, many computers have their memory capacity extended by an *external memory unit* which is usually in the form of a magnetic tape; a special device acting on instructions provided for in advance can extract the numbers from the external memory unit and transfer them to the *internal memory unit* or vice versa, naturally. Whereas the number of locations in the internal memory unit is in the thousands, the number of locations on magnetic tape can run into millions. Reading numbers from the tape and writing them back on the tape is slower than the same processes in the internal memory because winding the tape takes time.

If the machine is processing information oth r than numbers (say, words), then it has to be coded by means of binary digits. Then a device has to be provided for that can *decode* the result, that is, return it to the natural form of the original information.

The computer performs operations on the numb rs stored in the memory unit by means of *instructions* that are e tered in the memory unit prior to operation and have the same type of notation as numbers. We will consider only the so-called *three-address instructions*. Each such instruction is a storage location that we can imagine to be divided into four parts of definite length, say 3, 9, 9, and 9 digits each. These parts accommodate:

(1) the code of the operation that is to be performed;

(2) the address (number of the storage location) of the first operand (number to be operated on);

(3) the address of the second operand;

(4) the address to which the result is to be sent.

All locations are numbered serially. For example, if the memory unit has 512 locations, then nine binary digits is just enough to indicate the address (why?). Suppose the operation code for addition is 1. Now if it is required to add the numbers located in locations 271 and 59 and the result to be put in 422, then with our storage location divided as indicated above the appropriate instruction will be in the following form (verify this):

$$\text{parts:}\quad \underbrace{001}_{\text{1st}}\quad \underbrace{100001111}_{\text{2nd}}\quad \underbrace{000111011}_{\text{3rd}}\quad \underbrace{110100110}_{\text{4th}} \qquad (6)$$

After this instruction has been carried out, all storage locations, including 271 and 59, will remain unchanged (as they were before), with the exception of 422, in which the sum of the numbers at the addresses 271 and 59 appear in place of what location 422 had before.

The instruction (6) is punched on a card and is stored in one of the storage locations of the memory unit just exactly like the number $0.0110000111100011101111_2 \cdot 2^{-110_2}$ would be. The difference is evident only when the machine is in operation, since if the program has been properly written, signals in the form of instructions enter the control unit only from those locations containing instructions.

Electronic circuitry provides for the operation of the computer, including the operations designated by the instructions. These circuits transmit electric signals according to specific rules. In the early days of computers, these circuits were based on the use of electron tubes (triodes and pentodes), then later on semiconductors (transistors), and integrated circuits. The basic types of signal converters are shown in Fig. 205. The first generates a signal (positive voltage of a sufficient level) at the output B if and only if the signal at the input A is absent. The second generates a signal at the output C if and only if signals have been fed to both inputs A and B, and the third generates a signal when a signal has been fed to at least one input. More complicated converters are obtained by combining these elementary ones. For example, check to see that the circuit shown in Fig. 206 realizes the addition table (1) if a signal represents unity and absence of a signal, zero. The multiplication table (2) is realized by the elementary "and" circuit.

We now give a rough outline of how a machine operates. At the start, the *program* (which is a definite sequence of instructions)

Fig. 205

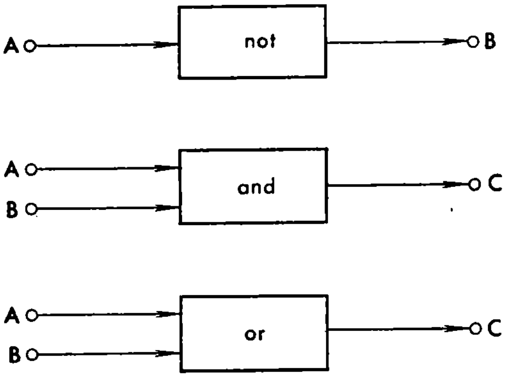
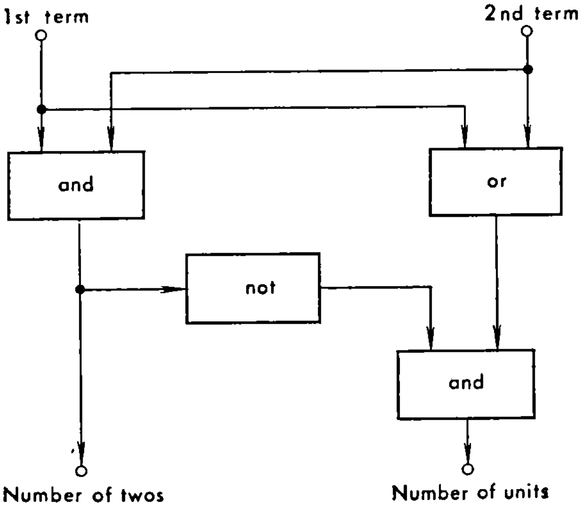


Fig. 206



and the *initial information* (the operands, or numbers to be operated on) are entered in the memory unit by means of punched cards. Then the control unit accepts the contents of the first location as an instruction and, in accordance with this instruction, the arithmetic unit performs the indicated operations. Then the control unit accepts the contents of the second location as an instruction, and so forth, with the exception of special instructions called "transfer of control", which we will discuss in Sec. 15.4, and after the performance of which the control unit does not proceed to the next location but to the address indicated in the instruction. Certain instructions provide for printing the contents of a location, but after the execution of such an instruction the control unit then passes to the next location.

(True, if a lot of printing is required, the work is slow, since printing takes more time than arithmetic operations.) This continues until the control unit reaches an instruction to *stop* the computer, which brings the computation to an end. There is also an *emergency stop* which stops the machine if the operation provided for by the program cannot be performed; for example, if in the course of a computation, a number is obtained that is too big for the storage element (this is called *overflow*). A special device makes it possible at any time, from the control console, to check the contents of any location and also to enter accessory information.

Exercises

1. Write the following base-10 numbers in binary (base-2): 9999, $-1/3$, 2.75.
2. Write the same numbers, using formula (3) as a pattern.

15.4 Programming

We now give a few simple examples to illustrate the basic principles of *programming* a computer. It should be noted that writing a serious program is a very responsible undertaking and often requires a large amount of time and skill. For the sake of simplicity we will write the sign of the operation instead of the operation code (for instance, the + sign instead of the add code).

Suppose it is required to find the solution of a system of equations of the first degree (system (6) of Ch. 8), assuming the values of the numbers a_i, b_i, d_i to be known. The solution, as we know, can be found by formulas (7) of Ch. 8.

We will place the instructions in storage locations labelled 1,2, ... ; we do not yet know how many instructions will be needed. The six input parameters a_1, b_1, a_2, b_2, d_1, d_2 will be located, respectively, in storage elements labelled $\alpha + 1$, $\alpha + 2$, $\alpha + 3$, $\alpha + 4$, $\alpha + 5$, $\alpha + 6$. The value of α will be determined later on. Several more locations will be used for storing intermediate results. It does not matter what the state of these locations was prior to starting our computation because any previous writing in a location is automatically erased when a new entry is made. The remaining storage elements of the memory unit are not used in the writing and performance of this program. We begin by calculating the value of $d_1b_2 - b_1d_2$. To compute d_1b_2 we need the instruction

$$(1) \quad \times \quad \alpha + 5 \quad \alpha + 4 \quad \alpha + 7$$

Here we indicate the number of the instruction though actually it is never punched in a card. When this instruction is executed, the number d_1b_2 appears in $(\alpha + 7)$. The next instruction is

$$(2) \quad \times \quad \alpha + 2 \quad \alpha + 6 \quad \alpha + 8$$

and after its execution the number b_1d_2 appears in $(\alpha + 8)$. Next, subtract from the number d_1b_2 (location $(\alpha + 7)$) the number b_1d_2 (location $(\alpha + 8)$). Since these numbers will not be needed any more, the result can be put in $(\alpha + 7)$ via the instruction

$$(3) \quad - \quad \alpha + 7 \quad \alpha + 8 \quad \alpha + 7$$

Of course we could have made use of the location $(\alpha + 9)$, but in more complicated programs one has to save on storage space.

In similar fashion we find the denominator $a_1b_2 - b_1a_2$

$$\begin{array}{lllll}
(4) & \times & \alpha + 1 & \alpha + 4 & \alpha + 8, \\
(5) & \times & \alpha + 2 & \alpha + 3 & \alpha + 9, \\
(6) & - & \alpha + 8 & \alpha + 9 & \alpha + 8
\end{array}$$

Now take the numerator located at $(\alpha + 7)$ and divide it by the denominator located at $(\alpha + 8)$, and then print the result:

$$\begin{array}{lllll}
(7) & : & \alpha + 7 & \alpha + 8 & \alpha + 7 \\
(8) & \text{print } \alpha + 7 & & &
\end{array}$$

On the last instruction, the machine prints the contents of the storage element, i.e., the value of x, and then proceeds to the next instruction.

We compute y in similar fashion, having in view that the denominator for y has already been computed in $(\alpha + 8)$:

$$\begin{array}{lllll}
(9) & \times & \alpha + 1 & \alpha + 6 & \alpha + 7, \\
(10) & \times & \alpha + 5 & \alpha + 3 & \alpha + 9, \\
(11) & - & \alpha + 7 & \alpha + 9 & \alpha + 7, \\
(12) & : & \alpha + 7 & \alpha + 8 & \alpha + 7, \\
(13) & \text{print } \alpha + 7, & & & \\
(14) & \text{stop} & & &
\end{array}$$

The machine is stopped on the 14th instruction. Thus, our program contains 14 instructions and so we can take $\alpha = 14$. Then the entire program will occupy 20 storage locations and will look like this:

$$\begin{array}{lllll}
(1) & \times & 19 & 18 & 21, \\
(2) & \times & 16 & 20 & 22, \\
(3) & - & 21 & 22 & 21, \\
(4) & \times & 15 & 18 & 22, \\
(5) & \times & 16 & 17 & 23, \\
(6) & - & 22 & 23 & 22, \\
(7) & : & 21 & 22 & 21, \\
(8) & \text{print } 21, & & & \\
(9) & \times & 15 & 20 & 21,
\end{array}$$

(10) $\times$ 19 17 23,

(11) — 21 23 21,

(12) : 21 22 21,

(13) print 21,

(14) stop,

(15) a_1,

(16) b_1,

(17) a_2,

(18) b_2,

(19) d_1,

(20) d_2

Thus, all the intermediate results in our program require three intermediate-result locations (Nos. 21, 22, and 23). The only thing left to do now is to punch this program and start the machine. (Actually, a few more instructions are required that are of no particular importance to us; for instance, the instruction to enter the program the memory unit from the cards, and so on.)

In this example, the number of instructions in the program was equal to the number of operations. But modern electronic digital computers were created mainly for computations requiring millions of operations. It is clearly impossible to write a separate instruction for every operation in that case. Luckily, in such a huge volume of computation many intermediate computations are ordinarily carried out several times according to the same scheme. Then in the program we can provide for the formation of *loops*, in which the control unit passes through one and the same portion of the program several times. Loops are formed with the aid of the *conditional transfer of control instruction (jump)*, which looks like this:

$$\text{jump } N_1 \quad N_2 \quad N_3$$

We wrote "jump" but actually, of course, one has to indicate the code of the operation. There are different versions for realizing this instruction. For the sake of definiteness, we will assume that on this instruction the machine compares the contents $(N_1)'$ of location N_1 and the contents $(N_2)'$ of location N_2, if $(N_1)' < (N_2)'$, then the control unit executes the next instruction, but if $(N_1)' \geqslant (N_2)'$, then it executes the instruction of location N_3, in both cases the contents of the locations remain unchanged. Say, the instruction

$$\text{jump } 1 \quad 1 \quad N_3$$

means that after reading it the control unit executes the instruction contained in location N_3. We can regard this instruction as an *unconditional transfer of control instruction.*

As an example let us write a program for printing the table of reciprocals of the 500 natural numbers between 2001 and 2500. Without the conditional transfer instruction the program would be very large, whereas actually it is very small. We put the number 2000 in location $(\alpha + 1)$, the number 1 in $(\alpha + 2)$, and the number 2499 in $(\alpha + 3)$ (it will soon become clear why this is done). Suppose the first instruction is of the form

$$(1) \ + \ \alpha + 1 \quad \alpha + 2 \quad \alpha + 1$$

On execution of this instruction, the number 2001 appears in location $(\alpha + 1)$ instead of 2000, we compute the reciprocal of 2001 and print the result:

$$(2) \ : \quad \alpha + 2 \quad \alpha + 1 \quad \alpha + 4,$$
$$(3) \ \text{print} \ \alpha + 4$$

We compare the number in $(\alpha + 1)$ with 2499 and, since it has not yet exceeded 2499, we pass again to the first instructions:

$$(4) \ \text{jump} \ \alpha + 3 \quad \alpha + 1 \quad 1$$

Since in this case $(\alpha + 3)' = 2499$ and $(\alpha + 1)' = 2001$, on reading the fourth instruction, the control unit goes to the first instruction. Upon executing this instruction, the number 2002 instead of 2001 appears in $(\alpha + 1)$. The next two instructions have to do with computing and printing the reciprocal of 2002. On the fourth instruction the machine compares 2499 with 2002 and again goes back to the first instruction, and so forth, and only when, after adding unity for the last time, the number 2500 appears in location $(\alpha + 1)$ and the reciprocal of 2500 has been printed does the fourth instruction let the control unit take over, for then we have $(\alpha + 3)' = 2499$, $(\alpha + 1)' = 2500$. The machine is then stopped:

(5) stop

All the results have been printed out.

Thus, we take $\alpha = 5$ and the complete program looks like this:

```
(1) +      6  7  6,
(2) :      7  6  9,
(3) print  9
(4) jump   8  6  1,
(5) stop,
(6) 2000,
(7) 1,
(8) 2499
```

We needed only one storage location (No. 9) for storing intermediate results, but in the process of operation the contents of the sixth location varied from 2000 to 2500 at intervals of 1.

Let us examine a variant in which at the end of the operation period the reciprocals are not printed out but are located in storage elements of the memory unit with number labels from 10 to 509 (these quantities may be needed for subsequent computations):

```
(1) +     6 7 6,
(2) +'    3 9 3,
(3) :     7 6 9,
(4) jump  8 6 1,
(5) stop,
(6) 2000,
(7) 1,
(8) 2499
(9) (1 in the last digit position of the stor-
    age element, the other positions be-
    ing 0)
```

Here, an *auxiliary number* (without a quantitative value) is delivered to the ninth location of the program; it serves to *convert instructions*, which is a fundamentally new type of operation. After the first execution of the second instruction the third instruction accepted by the arithmetic unit as a number takes the form: 7 6 10. (Note that addition via instructions (1) and (2) is performed by different rules; this was indicated by a prime. Think this over!) For this reason, after the third instruction the number 1: 2001 is sent to the 10th location. After a second execution of the second instruction the third instruction becomes 7 6 11, and so after the second execution of the third instruction the number 1: 2002 is sent to the 11th location, and so on. This operation of changing the address in an instruction is called *address substitution*.

Thus, instructions can be automatically converted in the process of operation of the machine. This opens up fresh opportunities in the use of electronic digital computers.

The conditional transfer of control instruction is also used without loops, when it is required to organize *program splitting*, which is the performing of distinct sequences of operations depending on circumstances that are not known beforehand. Suppose that in a computation a number a is to appear in a location labelled N and we are to leave it unchanged if $a \geqslant 0$ or square it and leave the result in location N if $a < 0$. Since ordinarily the number 0 is placed in the location labelled 0, the required operation can be performed by appropriately placing the following instructions:

....

(k)	jump	N	0	$k+2$,
$(k+1)$	$\times$	N	N	N

The program splitting is handled automatically so that we do not even know which variant has been accomplished if such information has not been provided for.

Finally, let us examine a program in which the total number of operations has not been provided for beforehand. Suppose we have to solve the cubic equation

$$x = 0.1\,x^3 + 1$$

by the iterative method (Sec. 1.3) accurate to 0.001, beginning with the value $x_0 = 0$. To do this, place the number 0 in $(\alpha + 1)$ (the appropriate row of the card is simply not punched), the number 0.1 in the location $(\alpha + 2)$, the number 1 in $(\alpha + 3)$ and the number 0.001 in $(\alpha + 4)$. We will place the successive approximations in storage location $(\alpha + 1)$. Computing the next approximation from the previous one in this manner is accomplished via the instructions:

$$\begin{array}{llllll}
(1) & \times & \alpha+1 & \alpha+1 & \alpha+5, \\
(2) & \times & \alpha+5 & \alpha+1 & \alpha+5, \\
(3) & \times & \alpha+2 & \alpha+5 & \alpha+5, \\
(4) & + & \alpha+5 & \alpha+3 & \alpha+5
\end{array}$$

Thus, the next approximation is placed in location $(\alpha + 5)$. It must be compared with the preceding one located at $(\alpha + 1)$ and if they coincide to within 0.001, then we place the next approximation in location $(\alpha + 1)$ and repeat the iteration. If the approximations differ by less than 0.001, the result is printed. This can be done via the instructions (which should be checked)

$$(5)\quad |-|\quad \alpha+1\quad \alpha+5\quad \alpha+1$$

(this instruction sends the number $|\,(\alpha + 1)' - (\alpha + 5)'\,|$ to location $(\alpha + 1)$),

$$\begin{array}{lllll}
(6) & \text{jump} & 0 & \alpha+1 & 9, \\
(7) & + & \alpha+5 & 0 & \alpha+1, \\
(8) & \text{jump} & 1 & 1 & 1, \\
(9) & \text{print} & \alpha+5 & & \\
(10) & \text{stop} & & &
\end{array}$$

Thus we can set $\alpha = 10$ and write out the entire program in 14 storage locations. By this program, the computer will perform iterations until the last two approximations coincide to within the accuracy indicated; it will then print the result and stop.

Note that if the iteration process does not converge, the computer will overflow (that is, it will exceed the capacity of the number representation) or it will lock into a loop, in which case the machine is not able to stop by itself and requires interference from the control panel.

Programs for more complicated problems may be quite extensive, but they frequently include simpler (and often repeated) problems like, say, computing the sine of the relevant quantities, and the like.

These simpler problems are solved by means of ***standard subprograms***, which are merely a set of instructions compiled beforehand and residing in specific storage locations of the internal memory of the computer or in the external memory section. In the writing of a program of considerable complexity, such standard subprograms are brought into the computation on a single instruction.

In recent years, the work of programming a computer has been simplified greatly by the development of several *universal machine languages*, which make it possible to write the program in more conventional mathematical terms than what we have just described. A program, written in such a language does not depend on the type of computer and is printed on a device very much like a typewriter. Then a special *translator* (a system having a number of inputs and outputs) automatically converts the given program to that of the machine to be used in the computation.

Two languages of this kind are in wide use: ALGOL (from ALGOrithmic Language) and FORTRAN (meaning FORmula TRANslator). They are of considerable aid in making electronic digital computers more accessible to the scientific community.

Exercises

1. Write a program for computing a table of the squares of the natural numbers from 1 to 1000 inclusive, with the values of n and n^2 alternating in the sequence: 1, 1, 2, 4, 3, 9, ...
2. Write a program for computing the sum $\sqrt{5} + \sqrt{6} + \ldots + \sqrt{100}$ (this sum was discussed in Sec. 1.2). Let the extraction of the square root be a separate operation indicated by the code $\sqrt{\ }$.
3. Program the computation of the integral $\int\limits_0^1 \frac{dx}{1+x}$ by Simpson's formula (Sec. 1.1) with the interval of integration partitioned into 100 parts; into 1000 parts.
4. Write a program for integrating the differential equation $y' = x^2 - y^2$ over the interval $-1 \leqslant x \leqslant 0$ with the initial condition $y(-1) = 0$ by the recalculation method described in Sec. 8.7 (cf. Table 5) with the integration step 0.001, and print the results at intervals of 0.1. Use the operation $\{\ \ \}$ to indicate the fractional part of a number ($\{5.3\} = 0.3$, $\{-5.3\} = 0.7$, $\{5\} = 0$).

15.5 Use computers!

From Sec. 15.4 it is clear that programming a digital computer is not difficult in simple cases. Programming is even being taught in secondary school. In many cases computers have made it possible to increase accuracy and speed of computations by several orders of magnitude; certain complicated problems have for the first time entered the realm of solvability. (To get an idea of the range of these operations, recall that small-scale digital computers perform several thousand arithmetic operations per second, medium-sized computers, tens of thousands per second, and large ones, like the Soviet BESM-6, do a million operation per second.)

The only barrier left to the widespread use of computers would seem to be psychological. To put it crudely, many research workers are simply frightened: they fear computers somewhat along the same lines that past generations of mathematicians were overawed by integrals that could not be evaluated, transcendental finite equations, differential equations not solvable by quadratures, and the like. Instead of fundamentally altering the approach to such integrals and equations, scientists persistently sought to find new cases of integrability. All this appreciably restricted the possibilities of mathematical applications. Today the situation is similar with regard to electronic digital computers.

One should not of course veer to the other extreme and think that computers make superfluous all analytic (exact, approximate, asymptotic) formulas and methods, "hand" computation aided by the electric desk calculator and the slide rule, and pencil-and-paper calculations. Analytic solutions, when they are possible, often have the inestimable advantage of being compact, particularly if the problem includes parameters or the solution is obtained as a function of several independent variables. Asymptotic formulas are effective in cases where the use of numerical methods becomes involved. Hand calculation is the most mobile, so to say, and is particularly well adapted to rough estimates, which should be resorted to as often as possible, even when preparing for very extensive computations. Computers are not meant to replace other fruitful mathematical methods, but to combine with them and thus substantially expand the range of application of mathematics.

When a computer is used to solve a problem that has already been stated in mathematical terms, the point of greatest responsibility is usually that of preparing the problem for the programming process. It is often quite a job to choose a method capable of yielding a reliable result and yet one that is within the capacity of the computer, which is powerful but not all-powerful. In this preparatory stage, one frequently has to recast one's "mathematical thinking" that was reared on hand methods.

For example, in the premachine era it was taken for granted that if it is possible in a nonlinear differential equation to lower the order of the equation by means of a substitution, then this should be done. Say, to solve the equation

$$y'' = f(y, y') \; (y = y(x)) \tag{7}$$

the advised procedure was: consider the relationship $y' = p$ as a function of y, whence $y'' = \frac{dy'}{dx} = \frac{dp}{dy}\frac{dy}{dx} = p\frac{dp}{dy}$ and equation (7) takes the form

$$p\frac{dp}{dy} = f(y, p) \quad (p = p(y)) \tag{8}$$

which is an equation of the first order. If we are able to integrate it, that is, to find the general solution $p = \varphi(y, C_1)$, then we write

$$p = \frac{dy}{dx} = \varphi(y, C_1)$$

whence

$$\frac{dy}{\varphi(y, C_1)} = dx \text{ and } \int\frac{dy}{\varphi(y, C_1)} = x + C_2 \tag{9}$$

This procedure sometimes achieves its aim, particularly in an analytic investigation of the solution, but for a numerical solution, especially using a computer, it ordinarily proves inadvisable, for we have to solve (8) numerically (and if it is solved in quadratures, then we have to compute the appropriate integrals), and then compute the integral (9), and, finally, invert the resulting relationship $x(y)$. But it is much simpler — both for compiling tables of the values of a particular solution and for complete tables of the general solution — to integrate equation (7) directly numerically, without lowering the order of the equation. That is exactly what should be done on a computer.

Which means that one should be capable, at least in a crude manner, to estimate the volume of computation required to bring the solution of the problem to completion.

Here is another example. The very statement of the problem in Secs. 1.2 and 3.4 on approximating sums by means of integrals was of course due to the requirements of hand computation. In hand calculation, the indicated approximate formulas and methods for their refinement are extremely useful. But when computers are used, direct summing of the terms turns out to be more effective in most cases.

But here too one should not act unthinkingly. Suppose we want to find the sum of the infinite series

$$S = \frac{1}{1^2} + \frac{1}{2^2} + \dots + \frac{1}{n^2} + \dots$$

Regarding the computer as all-powerful, we might attempt to calculate and add terms of the series until we get a machine zero, that is, until they become zero after the rounding that is necessary to represent the number in the storage location, after which the partial sums of the series will cease to increase. But for the parameters of the computer given in Sec. 15.3 (see (4), (5)), this will occur when

$$\frac{1}{n^2} < 2^{-55}, \text{ or } n > 2^{27.5} \approx 2 \cdot 10^8$$

As may be seen from Sec. 3.4, the error is then close to $1/n$, or eight orders of magnitude higher than the last terms. Since adding a term of a series to a partial sum requires several operations, for a computer operating at 20 000 operations per second it will take 5 hours of time, 300 roubles of government money and will incur the fiercest criticism of one's colleague, especially when they find out what the computing was all about. (Now suppose the length of a storage location to be 40 instead of 30 digits, which is closer to actual numbers, the operation period and cost of the computation will rise by a factor of 32.)

Of course computations should not be handled that way. The result can be obtained much faster and more accurately if, say, we sum the first 1000 terms of the series (which takes less than a second) and then replace the remainder by an integral via the method of Sec. 3.4. The precision of the result can be found by repeating the computation for 2000 terms. A still better way is to take a reference book (say [6], where you will find that $S = \pi^2/6$), where the value of π^2 can be found very precisely in tables. (True, it was simply our luck in this case because numerical series rarely develop into finite expressions in terms of known constants.)

Computations done on electronic digital computers are fundamentally discrete. For this reason, when one has a digital computer in mind, he must formulate the problem as a differential or integral equation, all the time bearing in mind how the problem and the method of its solution will appear in discrete terms. For equations involving more than one independent variable this gives rise to a diversity of complications, many of which have not yet been overcome.

Problems involving parameters require special attention. For example, suppose we want to compile a table that could be used to solve the complete cubic equation

$$ax^3 + bx^2 + cx + d = 0 \tag{10}$$

If we assume that each of the parameters a, b, c, d can take on 50 values, which isn't many, then we get a total of $50^4 \approx 6 \cdot 10^6$ combinations of these values. An average-size computer can handle the job in about one month of continuous operation with most of the time devoted to printing. This will leave you with 200 kilometres of tape weighing two tons, and you will rue the day you got the idea of such a problem.

Generally, most people newly introduced to computers and amazed at what they can do try to obtain as many numerical results as possible, they are guided by the naive principle that "the more figures, the more information and, hence, the greater utility". But these people are then flooded by a tide of numbers, and the fresh task of extricating something of value often proves more complicated than the original problem. All of which is reminiscent of the medieval legend of the magician and his apprentice, who in the absence of his mentor called forth a jinn and caused him to haul water but was unable to stop him and almost drowned. Hence so important and timely the frequently repeated statement by the prominent computer expert R. Hamming [8] "... it is a good general rule to begin a problem in computation with a searching examination of 'What are we going to do with the answers?'"

Let us return to equation (10). Actually the situation is not so hopeless after all. First divide both sides by a:

$$x^3 + \frac{b}{a} x^2 + \frac{c}{a} x + \frac{d}{a} = 0$$

Then make the substitution $x = y + \alpha$ choosing α so as to eliminate the term with the unknown squared. This gives us $\alpha = -\frac{b}{3a}$ and we get the equation (check this)

$$y^3 + py + q = 0, \quad \text{where } p = \frac{c}{a} - \frac{b^2}{3a^2}, \quad q = \frac{d}{a} - \frac{bc}{3a^2} + \frac{2b^3}{27a^3}$$

Finally, make the substitution $y = \beta z$ to get

$$\beta^3 z^3 + p\beta z + q = 0, \quad \text{or} \quad z^3 + \frac{p}{\beta^2} z + \frac{q}{\beta^3} = 0$$

Choosing $\beta = \sqrt[3]{q}$, we arrive at the equation

$$z^3 + rz + 1 = 0, \quad \text{where } r = pq^{-2}/^3 \tag{11}$$

The idea behind these manipulations is clear: we successively reduced the number of parameters till we reached (11) with only one, r, which is a combination of the initial parameters. To compute r from these parameters requires simple arithmetic operations and a table of cube roots. Using a computer, it is now easy to set up a table of values of the solutions of equation (11) depending on the single parameter r, even if we assign 5000 values to r instead of 50. Using the solution z, we easily find the solution x:

$$x = \beta z + \alpha = [(27a^3d - 9abc + 2b^3)^{1/3} z - b]/3a$$

The solution, as we see, comes out in the form of two tables of one entry each (the table of cube roots and $z(r)$), which of course is incomparably simpler than one table of four entries.

The problem of reducing the number of parameters can also arise in the solution of a differential equation containing parameters. A case of this kind was considered in Sec. 8.10 where a problem containing five parameters (if one takes into account that the solution itself is a function of t, we find a table with six entries (!) is required to represent the solution) was reduced to the problem (71) of Ch. 8 with two parameters ((72) of Ch .8) with the aid of similarity transformations.

By way of another illustration, consider the solution of equation (7) without parameters for arbitrary initial conditions:

$$y(x_0) = y_0, \quad y'(x_0) = y'_0$$

It would appear at first glance that the solution requires a table of four entries for its representation, since the parameters x_0, y_0, y'_0 and the independent variable x may assume arbitrary values. But the number of entries can easily be reduced to three. Indeed, equation (7) does not contain x, which means it is invariant under the transformation $x \to x +$ constant. For this reason, along with the solution $y = g(x)$, (7) has the solution $y = g(x + C)$ for any C. But this means that the solution y does not depend on x and x_0 separately but on the combination $x - x_0$:

$$y = \psi(x - x_0, y_0, y'_0) \tag{12}$$

And so it suffices to compile a table of three entries $y = \psi(x, y_0, y'_0)$ to solve equation (7) for $x_0 = 0$, then for any x_0 we can find the solution from (12).

In this last example, the number of entries in the complete table of the solution was obtained by indicating the group of transformations $x \to x + C$ under which the equation at hand remains invariant (transformation groups are discussed in Sec. 14.4). This also means that the set of graphs of all solutions is invariant under translations along the x-axis: under a translation, each such graph passes into the graph of some other solution. The interesting thing here is that we arrive at this conclusion by analyzing the equation itself and not its solutions!

It turns out that a knowledge of the continuous group of transformations (i.e., transformations dependent on one or several continuous parameters) that leave the given differential equation invariant always permits reducing the number of entries in the complete table of its solution. Unfortunately, it is not always possible by far to detect such groups, but sometimes they are suggested by physical reasoning (as the time-shift group for an autonomous system).

Do not grudge the time taken to reduce the number of parameters in a problem! First of all, it cuts the overall amount of computation and makes the results more surveyable: for example, if a parameter takes on 20 values, getting rid of it will cut computations and the extent of the final table by a factor of 20. Secondly, the combinations of parameters obtained after transforming the problem are often profoundly meaningful from a physical point of view.

At times, the use of computers suggests the advisability of fundamentally changing the conventional procedure for problem solving. For example, in computing the integral

$$I = \int_{(\Omega)} f(M)\, d\Omega$$

over a region (Ω) of high dimensionality and of complex shape it is better to replace quadrature formulas of the Simpson type (Sec. 1.1) by a method based on approaching the integral as the arithmetic mean of the function value. Say, in the region (Ω) we choose at random points M_1, M_2, ..., M_N (computers have randomizing devices that generate random numbers which are taken for the coordinates of these points) and then put

$$I \approx \frac{1}{N}[f(M_1) + f(M_2) + \dots + f(M_N)]$$

For large N — and the computer can see to it that N is large enough — the accuracy of the formula is quite acceptable. This procedure is one of the simplest instances of a general, specifically machine, method called the *Monte Carlo method* (named after the city of gambling fame, the business of which is completely based on independent random trials) in which the desired quantity is represented as the mean value of a certain random variable (Sec. 13.7), and this mean value is replaced by the arithmetic mean of realization of this variable in large numbers of independent trials.

A fundamentally important problem when working with computers is the effect of rounding off errors. When we have long chains of computations and succeeding steps rely all the time on preceding results, rounding errors can build up to a point where we will be dealing solely with errors.

Here is a striking example of this effect (it is given in [1]). In evaluating the integral

$$I_n = \frac{1}{e}\int_0^1 x^n e^x\, dx \quad (n = 0,\ 1,\ 2,\ \dots) \qquad (13)$$

it is easy to establish by integration by parts that

$$I_n = 1 - nI_{n-1} \quad (n = 1, 2, 3, \ldots) \tag{14}$$

Besides,

$$I_0 = \frac{1}{e}\int_0^1 e^x\,dx = \frac{1}{e}(e-1) = 1 - \frac{1}{e} = 0.632 \tag{15}$$

Formulas (14) and (15) permit computing successively $I_1 = 1 - 1I_0$, $I_2 = 1 - 2I_1$, and so on. The results I_n^{I} obtained in computing these values to three decimal places are:

n	0	1	2	3	4	5	6	7	8
I_n^{I}	0.632	0.368	0.264	0.208	0.168	0.160	0.040	0.720	−4.760

The results are clearly absurd, since from (13) it is at once apparent that $I_0 > I_1 > I_2 > \ldots > 0$.

The root of the trouble is clear: in the computation of I_n^{I} the original rounding error of I_0 is multiplied by $1 \cdot 2 \cdot 3 \ldots n$, and since the exact value of I_n is bounded (and even tends to zero as $n \to \infty$), the relative error increases swiftly. Computations with large numbers of digits, which is typical of computers, helped matters somewhat but not for long. Working to 9 significant digits, we reach nonsense results from $n = 14$ on. The benefit of a repeated computation with a different number of decimal digits is that we learn how reliable our computations were.

Now it is not hard to rearrange our material so that the errors fall off instead of building up. Replace n by $n + 1$ and rewrite (14) as

$$I_n = \frac{1}{n+1}(1 - I_{n+1})$$

Then put a "starting" I_n equal to zero and compute I_n from large n to small n. For example, put $I_{10} = 0$ and we get

n	9	8	7	6	5	4	3	2	1	0
I_n^{II}	0.100	0.100	0.112	0.127	0.146	0.171	0.207	0.264	0.368	0.632

More exact computations show that $I_9 = 0.092$, $I_8 = 0.101$, whereas all the other figures are correct.

When computing integrals and solving differential equations one often gets into a paradoxical situation due to the effects of rounding errors: to improve accuracy we refine the partition, but if the method used is unstable, then because of the increased number of operations the rounding errors come to the fore and the overall error increases. This must be kept in mind when working with computers.

Yes, definitely get into the habit of using computers. Do the computations yourself, don't entrust this important work to others. A computing expert who is not sufficiently acquainted with the specific nature of the problem (provided, of course, that he is not a co-author) will do a lot of extra work, may find it difficult to reorganize himself or back out of a dead-end alley, or resort to intuition based on the physics of the situation or to rough estimates — and all this will unavoidably affect the result.

Finally, it is well worth remembering the basic proposition of Hamming that is often repeated in his book [8]:

"The purpose of computing is insight, not numbers."

All results must be physically meaningful, they must exhibit trends, influences, and the crowning achievement of the job must be the creation of an approximate interpolation theory with coefficients obtained from precise calculations. This ideal may not always be attainable, but let us at least strive to attain it!

Exercise

Indicate the group of transformations that leave invariant the equation $y''' = f(x, y'')$; also indicate the number of entries needed to form a table of its general solution.

ANSWERS AND SOLUTIONS

Sec. 15.3

1. 10011100001111, $-1/11 = -0.010101...$, 10.11

2.

0	1	0	0	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	0	0	1
0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Sec. 15.4

1. Here and henceforth only one of several possible variants is indicated:

```
(1) +     9  7  9,
(2) print 9,
(3) ×     9  9  10,
(4) print 10,
(5) jump  8  9  1,
(6) stop,
(1) 1,
(8) 999,
(9) 0
```

2.

```
 (1) +      9  7   9,
 (2) √      9 11,
 (3) +     10 11  10,
 (4) jump   8  9   1,
 (5) print 10,
 (6) stop,
 (7)  1,
 (8) 99,
 (9)  4,
(10)  0
```

3.

```
 (1) +     18 23 18,
 (2) :     19 18 26,
 (3) +     20 26 20,
 (4) +     18 23 18,
 (5) :     19 18 26,
 (6) +     21 26 21,
 (7) jump  24 18 1,
 (8) —     21 26 21,
 (9) ×     21 16 21,
(10) ×     20 17 20,
(11) +     22 21 22,
(12) +     22 20 22,
(13) :     22 25 22,
(14) print 22,
(15) stop,
(16) 2,
(17) 4,
(18) 1,
(19) 1,
(20) 0,
(21) 0,
(22) 1.5,
(23) 0.01,
(24) 1.99,
(25) 300
```

In the second variant, the last three initial parameters have to be replaced, respectively, by 0.001, 1.999, and 3000; the computer handles the rest! It is even easy to write a program in which replacing the step would only require replacing a single initial parameter.

4.

```
(1) print 22,
(2) print 23,
(3) ×     23  23  29,
(4) —     24  29  29,
(5) ×     29  25  30,
 (6) +      23  30  30,
 (7) +      22  25  22,
 (8) ×      22  22  24,
 (9) —      30  30  30,
(10) —      24  30  30,
(11) +      29  30  29,
(12) ×      29  26  29,
(13) +      23  29  23,
(14) ×      22  27  29,
(15) { }    29  29,
(16) jump    0  29  18,
(17) jump    0   0   3,
(18) print  22,
(19) print  23,
(20) jump   28  22   3,
(21) stop,
(22) —1,
(23) 0,
(24) 1,
(25) 0.001,
(26) 0.0005,
(27) 10,
(28) —0.1
```

Sec. 15.5

The two-parameter group $y \to y + C_1x + C_2$. If $z = \varphi(x, x_0, z_0)$ is a solution of the equation $z' = f(x, z)$ with the initial condition $z(x_0) = z_0$, then the solution of the given equation for the initial conditions $y(x_0) = y_0$, $y'(x_0) = y_0'$, $y''(x_0) = y_0''$ is of the form

$$y = y_0 + y_0'(x - x_0) + \int_{x_0}^{x} ds \int_{x_0}^{s} \varphi(\sigma, x_0, y_0'')\, d\sigma$$

$$= y_0 + y_0'(x - x_0) + \psi(x, x_0, y_0'')$$

which means a table of three entries suffices for ψ.

REFERENCES

1. Babuška, I., Práger, M., and Vitásek, E. *Numerical Processes in Differential Equations*, New York, Interscience (1966).
2. Bridgman, P. W. *Dimensional Analysis*, New Haven, Yale University Press (1932).
3. Bronstein, I. N. and Semendyayew, K.A. *A Guide-Book to Mathematics*, Frankfurt, Deutsch (1971).
4. Einstein, A. *The Meaning of Relativity*, 5th ed. Princeton, N. J., Princeton University Press (1955).
5. Frank-Kamenetskii, D. A. *Diffusion and Heat Transfer in Chemical Kinetics*, New York, Plenum Press (1969).
6. Gradshteyn, I. S. and Ryzhik, I. M. *Tables of Integrals, Series and Products*, New York. Academic Press (1965).
7. Hagedorn, R. "Causality and Dispersion Relations" in *Preludes in Theoretical Physics in Honour of V. F. Weisskopf*, Amsterdam, North-Holland (1966).
8. Hamming, R. W. *Numerical Methods for Scientists and Engineers*, New York, McGraw-Hill (1962).
9. Jahnke, E., Emde, F., and Lösch, F. *Tables of Higher Functions*, 6th ed. New York, McGraw-Hill (1960).
10. Kamke, E. *Differentialgleichungen, Lösungsmethoden und Lösungen*, Vols. I, II. Leipzig, Akademische Verlagsgesellshaft (1944--1950).
11. Migdal, A. B. and Krainov, V. P. *Approximate Methods in Quantum Mechanics*, New York, W. Benjamin (1969).
12. Paradoksov, P. *Usp. Fiz. Nauk 89*, 707 (1966) [Eng. translation *Sov. Phys.-Usp. 9*, 618 (1967)].
13. Planck, M. "Das Kultur der kleinsten Wirkung" in *Die Kultur der Gegenwart*, vol 1, Physik (1915).
14. Sedov, L.I. *Similarity and Dimensional Methods in Mechanics*, New York, Academic Press (1959).

INDEX