MIR PUBLISHERS

А. Г. КУРОШ

КУРС ВЫСШЕЙ АЛГЕБРЫ

ИЗДАТЕЛЬСТВО «НАУКА»

МОСКВА

A. KUROSH

HIGHER ALGEBRA

Translated from the Russian
by
George Yankovsky

Moscow
MIR PUBLISHERS

UDC 512.8—20

На английском языке

CONTENTS

INTRODUCTION

The education of the mathematics major begins with the study of three basic disciplines: mathematical analysis, analytic geometry and higher algebra. These disciplines have a number of points of contact, some of which overlap; together they constitute the foundation upon which rests the whole edifice of modern mathematical science.

Higher algebra—the subject of this text—is a far-reaching and natural generalization of the basic school course of elementary algebra. Central to elementary algebra is without doubt the problem of solving equations. The study of equations begins with the very simple case of one equation of the first degree in one unknown. From there on, the development proceeds in two directions: to systems of two and three equations of the first degree in two and, respectively, three unknowns, and to a single quadratic equation in one unknown and also to a few special types of higher-degree equations which readily reduce to quadratic equations (quartic equations, for example).

Both trends are further developed in the course of higher algebra, thus determining its two large areas of study. One—the foundations of linear algebra—starts with the study of arbitrary systems of equations of the first degree (linear equations). When the number of equations equals the number of unknowns, solutions of such systems are obtained by means of the theory of determinants. However, the theory proves insufficient when studying systems of linear equations in which the number of equations is not equal to the number of unknowns. This is a novel feature from the standpoint of elementary algebra, but it is very important in practical applications. This stimulated the development of the theory of matrices, which are systems of numbers arranged in square or rectangular arrays made up of rows and columns. Matrix theory proved to be very deep and has found application far beyond the limits of the theory of systems of linear equations. On the other hand, investigations into systems of linear equations gave rise to multidimensional (so-called vector or linear) spaces. To the nonmathematician, multidimensional space (four-dimensional, to begin with) is a nebulous and often confusing concept. Actually, however, the notion is a strictly mathematical one, mainly algebraic, and serves as an important tool in a variety of mathematical investigations and also in physics.

The second half of the course of higher algebra, called the algebra of polynomials, is devoted to the study of a single equation in one unknown but of arbitrary degree. Since there is a formula for solving quadratic equations, it was natural to seek similar formulas for

higher-degree equations. That is precisely how this division of algebra developed historically. Formulas for solving equations of third and fourth degree were found in the sixteenth century. The search was then on for formulas capable of expressing the roots of equations of fifth and higher degree in terms of the coefficients of the equations by means of radicals, even radicals within radicals. It was futile, though it continued up to the beginning of the nineteenth century, when it was proved that no such formulas exist and that for all degrees beyond the fourth there even exist specific examples of equations with integral coefficients whose roots cannot be written down by means of radicals.

One should not be saddened by this absence of formulas for solving equations of higher degrees, for even in the case of third and fourth degree equations, where such formulas exist, computations are extremely involved and, in a practical sense, almost useless. On the other hand, the coefficients of equations one encounters in physics and engineering are usually quantities obtained in measurements. These are approximations and therefore the roots need only be known approximately, to within a specified accuracy. This led to the elaboration of a variety of methods of approximate solution of equations; only the most elementary methods are given in the course of higher algebra.

However, in the algebra of polynomials the main thing is not the problem of finding the roots of equations, but the problem of their existence. For example, we even know of quadratic equations with real coefficients that do not have real-valued roots. By extending the range of numbers to include the collection of complex numbers, we find that quadratic equations do have roots and that this holds true for equations of the third and fourth degree as well, as follows from the existence of formulas for their solution. But perhaps there are equations of the fifth and higher degree without a single root even in the class of complex numbers. Will it not be necessary, when seeking the roots of such equations, to pass from complex numbers to a still bigger class of numbers? The answer to this question is contained in an important theorem which asserts that any equation with numerical coefficients, whether real or complex, has complex-valued (real-valued, as a special case) roots; and, generally speaking, the number of roots is equal to the degree of the equation.

Such, in brief, is the basic content of the course of higher algebra. It must be stressed that higher algebra is only the starting point of the vast science of algebra which is very rich, extremely ramified and constantly expanding. Let us attempt, even more sketchily, to survey the various branches of algebra which, in the main, lie beyond the scope of the course of higher algebra.

Linear algebra, which is a broad field devoted mainly to the theory of matrices and the associated theory of linear transforma-

tions of vector spaces, includes also the theory of forms, the theory of invariants and tensor algebra, which plays an important role in differential geometry. The theory of vector spaces is further developed outside the scope of algebra, in functional analysis (infinite-dimensional spaces). Linear algebra continues, so far, to occupy first place among the numerous branches of algebra as to diversity and significance of its applications in mathematics, physics and the engineering sciences.

The algebra of polynomials, which over many decades has been growing as a science concerned with one equation of arbitrary degree in one unknown, has now in the main completed its development. It was further developed in part in certain divisions of the theory of functions of a complex variable, but basically grew into the theory of fields, which we will speak of later on. Now the very difficult problem of systems of equations of arbitrary degree (not linear) in several unknowns—it embraces both divisions of the course of higher algebra and is hardly touched on in this text—actually has to do with a special branch of mathematics called algebraic geometry.

An exhaustive treatment of the problem of the conditions under which an equation can be solved in terms of radicals was given by the French mathematician Galois (1811-1832). His investigations pointed out new vistas in the development of algebra and led, in the twentieth century, after the work of the German woman-algebraist E. Noether (1882-1935), to the establishment of a fresh viewpoint on the problems of algebraic science. There is no doubt now that the central problem of algebra is not the study of equations. The true subject of algebraic study is algebraic operations, like those of addition and multiplication of numbers, but possibly involving entities other than numbers.

In school physics one deals with the operation of composition of forces. The mathematical disciplines studied in the junior courses of universities and teachers' colleges provide numerous examples of algebraic operations: the addition and multiplication of matrices and functions, operations involving vectors, transformations of space, etc. These operations are usually similar to those involving numbers and bear the same names, but occasionally some of the properties which are customary in the case of numbers are lost. Thus, very often and in very important instances, the operations prove to be noncommutative (a product is dependent on the order of the factors), at times even nonassociative (a product of three factors depends on the placing of parentheses).

A very systematic study has been made of a few of the most important types of algebraic systems (or structures), that is, sets composed of entities of a certain nature for which certain algebraic operations have been defined. Such, for example, are fields. These

are algebraic systems in which (like in the systems of real and complex numbers) are defined the operations of addition and multiplication, both commutative and associative, connected by the distributive law (the ordinary rule of removing brackets holds) and possessing the inverse operations of subtraction and division. The theory of fields was a natural area for the further development of the theory of equations, while its principal branches—the theory of fields of algebraic numbers and the theory of fields of algebraic functions—linked it up with the theory of numbers and the theory of functions of a complex variable, respectively. The present course of higher algebra includes an elementary introduction to the theory of fields, and some portions of the course—polynomials in several unknowns, the normal form of a matrix—are presented directly for the case of an arbitrary base field.

Broader than a field is the concept of a ring. Unlike the field, division is not required here and, besides, multiplication may be noncommutative and even nonassociative. The simplest instances of rings are the set of all integers (including negative numbers), the set of polynomials in one unknown and the set of real-valued functions of a real variable. The theory of rings includes such old branches of algebra as the theory of hypercomplex numbers and the theory of ideals. It is related to a number of mathematical sciences (functional analysis being one) and has already made inroads into physics. The course of higher algebra actually contains only the definition of a ring.

Still greater in its range of applications is the theory of groups. A group is an algebraic system with one basic operation, which must be associative but not necessarily commutative, and must possess an inverse operation (division if the basic operation is multiplication). Such, for example, is the set of integers with respect to the operation of addition and also the set of positive real numbers with respect to the operation of multiplication. Groups were already important in the theory of Galois, in the problem of the solvability of equations in terms of radicals; today groups are a powerful tool in the theory of fields, in many divisions of geometry, in topology, and also outside mathematics (in crystallography and theoretical physics). Generally speaking, within the sphere of algebra, group theory takes second place after linear algebra as to its range of applications. Our course of higher algebra contains a chapter on the fundamentals of the theory of groups.

In recent decades an entirely new branch of algebra—lattice theory—has come to the fore. A lattice is an algebraic system with two operations—addition and multiplication. These operations must be commutative and associative and must also satisfy the following requirements: both the sum and the product of an element with itself must be equal to the element; if the sum of two elements

is equal to one of them, then the product is equal to the other, and conversely. An example of a lattice is the system of natural numbers relative to the operations of taking the least common multiple and the greatest common divisor. Lattice theory has interesting ties with the theory of groups and the theory of rings, and also with the theory of sets; one old branch of geometry (projective geometry) actually proved to be a part of the theory of lattice. It is also worth mentioning the expansion of lattice theory into the theory of electric circuits.

Certain similarities between parts of the theories of groups, rings and lattices led to the development of a general theory of algebraic systems (or universal algebras). The theory has only taken a few steps but its general outlines are evident and certain links with mathematical logic that have been perceived point to a rich future in this area.

The foregoing scheme does not of course embrace the whole range of algebraic science. For one thing, there are a number of divisions of algebra bordering on other areas of mathematics, such as topological algebra, which deals with algebraic systems in which the operations are continuous relative to some convergence defined for the elements of the systems. An example is the system of real numbers. Closely related to topological algebra is the theory of continuous (or Lie) groups, which has found numerous applications in a broad range of geometrical problems, in theoretical physics and hydrodynamics. Incidentally, the theory of Lie groups is characterized by such an interweaving of algebraic, topological, geometric and function-theoretic methods as to be more properly considered a special branch of mathematics altogether. Next we have the theory of ordered algebraic systems which arose out of investigations into the fundamentals of geometry and has found applications in functional analysis. Finally, there is differential algebra which has established fresh relationships between algebra and the theory of differential equations.

Quite naturally, the flowering of algebraic science so evident today is not accidental, but is an organic part of the general advance of mathematics and is due, in large measure, to the demands made upon algebra by the other mathematical sciences. On the other hand, the development of algebra itself has exerted a far-reaching influence on the elaboration of allied branches of science; this influence has been particularly enhanced by the spread of applications so characteristic of modern algebra. One is often tempted to speak of an "algebraization" of mathematics.

We conclude this rather sketchy survey of algebra with a general historical background.

Babylonian and, later, ancient Greek mathematicians studied certain problems of algebra, in particular the solution of simple

equations. The peak of algebraic investigations during this period was reached in the works of the Greek mathematician Diophantos of Alexandria (third century). These studies were then extended by mathematicians of India: Aryabhata (sixth century), Brahmagupta (seventh century), and Bhaskara (twelfth century). In China, algebraic problems got an early start: Ch'ang Ts'ang (second century B.C.), Ching Chou-chan (first century A.D.). An outstanding Chinese algebraist was Ch'in Chiu-shao (thirteenth century).

A major contribution to the development of algebra was made by scholars of the Middle East whose writings were in Arabic, particularly the Uzbek scholar Muhammad al-Khowârizmî (ninth century) and the Tajik mathematician and poet Omar Khayyam (1040-1123). In particular, the very term "algebra" came from the title of al-Khowârizmî's treatise *Hisâb al-jabr w'al-muqâ-balah.*

The above-mentioned studies of Babylonian, Greek, Indian, Chinese, and Central-Asian algebraists have to do with those problems of algebra which constitute the present school course of elementary algebra and only occasionally touch on equations of the third degree. That, in the main, was the range of problems that interested medieval European algebraists and those of the Renaissance, such as the Italian mathematician Leonardo of Pisa (Fibonacci) (twelfth century) and the founder of present-day algebraic symbolism, the Frenchman Vieta (or Viète) (1540-1603). We have already mentioned that in the sixteenth century methods were found for solving equations of the third and fourth degree; here we must mention the names of the Italians Ferro (1465-1526), Tartaglia (1500-1557), Cardano (1501-1576) and Ferrari (1522-1565).

The seventeenth and eighteenth centuries saw an intensive elaboration of the general theory of equations (or the algebra of polynomials) in which outstanding scholars of the time participated: Descartes (1596-1650), Sir Isaac Newton (1643-1727), d'Alembert (1717-1783) and Lagrange (1736-1813). In the eighteenth century, the Swiss mathematician Cramer (1704-1752) and Laplace (1749-1827) of France, laid the foundation of the theory of determinants. At the turn of the century, the great German mathematician Gauss (1777-1855) proved the earlier mentioned fundamental theorem on the existence of roots of equations with numerical coefficients.

The first third of the nineteenth century stands out in the history of algebra as the time when the problem of the solvability of equations by radicals was resolved. Proof of the impossibility of obtaining formulas for the solution of equations of degree five or higher was obtained by the Italian mathematician Ruffini (1765-1822) and in more rigorous form by the Norwegian Abel (1802-1829). As already mentioned, an exhaustive treatment of the problem of the conditions under which an equation admits of solution in terms of radicals was given by Galois.

Galois' theory spurred the advance of algebra in the latter half of the nineteenth century. There appeared the theory of fields of algebraic numbers and of fields of algebraic functions and the associated theory of ideals. Here, mention should be made of the German mathematicians Kummer (1810-1893), Kronecker (1823-1891), and Dedekind (1831-1916), and the Russian mathematicians E. I. Zolotarev (1847-1878) and G. F. Voronoi (1868-1908). Particular advances were made in the theory of finite groups which grew out of the research of Lagrange and Galois; this work was carried out by the French mathematicians Cauchy (1789-1857) and Jordan (1838-1922), the Norwegian Sylow (1832-1918), the German algebraists Frobenius (1849-1918) and Hölder (1859-1937). The investigations of the Norwegian S. Lie (1842-1899) initiated the theory of continuous groups.

The works of Hamilton (1805-1865) and the German mathematician Grassmann (1809-1877) laid the foundations for the theory of hypercomplex systems or, as we now say, the theory of algebras. A prominent role in the development of this branch of algebra was played (at the end of the century) by the Russian mathematician F. E. Molin (1861-1941).

Linear algebra attained great heights in the nineteenth century primarily due to the work of the English mathematicians Sylvester (1814-1897) and Cayley (1821-1895). Work continued on the algebra of polynomials; we note only the method of approximate solution of equations found by the Russian geometer N. I. Lobachevsky (1792-1856) and the work of the German Hurwitz (1859-1919). Algebraic geometry was begun in the latter part of the nineteenth century, particularly in the works of the German mathematician M. Noether (1844-1922).

In the twentieth century, algebraic studies expanded considerably and algebra, as we already know, occupies a very special place of honour in mathematics. New divisions of algebra have sprung up, including the general theory of fields (in the 1910's), the theory of rings and the general theory of groups (1920's), topological algebra and lattice theory (1930's), the theory of semigroups and the theory of quasigroups, the theory of universal algebras, homological algebra, the theory of categories (all in the 1940's and 1950's). Prominent mathematicians are presently engaged in all spheres of algebra, and in a number of countries (in the Soviet Union, for example) whole schools of algebra are in evidence.

Among the prerevolutionary Russian algebraists, noteworthy contributions to algebra were also made by S.O. Shatunovsky (1859-1929) and D. A. Grave (1863-1939). However, it was only after the Great October Revolution of 1917 that algebraic investigations in the Soviet Union reached high peaks. These studies now embrace practically all divisions of modern algebraic science and in some the work of Soviet algebraists is of a leading nature. Suffice

it to name only two algebraists: N. G. Chebotarev (1894-1947), who worked in the theory of fields and Lie groups, and O. Yu. Schmidt (1891-1956), the famous polar explorer who was also a noted algebraist and founded the Soviet school of group theory.

We conclude this brief survey of the historical background and modern state of algebra with the remark that most of the fields of research mentioned here lie beyond the scope of the present course of higher algebra. The aim of the survey was to help the reader to find the proper place for this text in algebraic science as a whole within the edifice of mathematics.

CHAPTER 1

SYSTEMS OF LINEAR EQUATIONS. DETERMINANTS

1. The Method of Successive Elimination of Unknowns

We begin the course of higher algebra with a study of systems of first-degree equations in several unknowns or, to use the more common term, *systems of linear equations.**

The theory of systems of linear equations serves as the foundation for a vast and important division of algebra—linear algebra—to which a good portion of this book is devoted (the first three chapters in particular). The coefficients of the equations considered in these three chapters, the values of the unknowns and, generally, all numbers that will be encountered are to be considered real. Incidentally, all the material of these three chapters is readily extendable to the case of arbitrary complex numbers which are familiar from elementary mathematics.

In contrast to elementary algebra, we will study systems with an arbitrary number of equations and unknowns; at times, the number of equations of a system will not even be assumed to coincide with the number of unknowns. Suppose we have a system of s linear equations in n unknowns. Let us agree to use the following symbolism: the unknowns will be denoted by x and subscripts: $x_1, x_2, \ldots, x_n$; we will consider the equations to be enumerated thus: first, second, . . ., sth; the coefficient of x_j in the ith equation will be given as a_{ij}**. Finally, the constant term of the ith equation will be indicated as b_i.

---

* The term "linear" stems from analytic geometry, where a first-degree equation in two unknowns defines a straight line in a plane.

** We thus use two subscripts, the first indicates the position number of the equation, the second the position number of the unknown. They are to be read: a_{11} "a sub one one" and not "a eleven"; a_{34} "a sub three four" and not "a thirty-four", and are not separated by a comma.

Our system of equations will now be written as follows:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{s1}x_1 + a_{s2}x_2 + \ldots + a_{sn}x_n &= b_s \end{aligned}\right\} \tag{1}$$

The coefficients of the unknowns form a rectangular array:

$$\begin{pmatrix} a_{11}a_{12} \ldots a_{1n} \\ a_{21}a_{22} \ldots a_{2n} \\ \ldots\ldots\ldots \\ a_{s1}a_{s2} \ldots a_{sn} \end{pmatrix} \tag{2}$$

called a *matrix* of *s rows* and *n columns*; the numbers a_{ij} are termed *elements* of the matrix.* If $s=n$ (which means the number of rows is equal to the number of columns), then the matrix is called *a square matrix of order n*. The diagonal of the matrix from upper left corner to lower right corner (i.e., composed of the elements $a_{11}, a_{22}, \ldots, a_{sn}$) is called the *principal diagonal*. We call a square matrix of order n a *unit matrix of order n* if all the elements of its principal diagonal are equal to unity and all other elements are zero.

The *solution* of the system of linear equations (1) is a set of n numbers $k_1, k_2, \ldots, k_n$ such that each of the equations (1) becomes an identity upon substitution of the corresponding numbers k_i, $i = 1, 2, \ldots, n$ for the unknowns x_i.**

A system of linear equations may not have any solutions; it is then called *inconsistent*. Such, for example, is the system

$$x_1 + 5x_2 = 1,$$
$$x_1 + 5x_2 = 7$$

The left members of these equations coincide, but the right members are different and so no set of values of the unknowns can satisfy both equations simultaneously.

If a system of linear equations has solutions, it is termed *consistent*. A consistent system is called *determinate* if it has a unique solution—only such are considered in elementary algebra—and *indeterminate* if there are more solutions than one. As we shall learn later on, there may even be an infinity of solutions. For instance,

---

* Thus, if the matrix (2) is regarded by itself (not connected with the system (1)), then the first subscript of element a_{ij} indicates the number of the row, the second the number of the column at the intersection of which the element is positioned.

** We stress the fact that the numbers $k_1, k_2, \ldots, k_n$ constitute a *single* solution of the system and not n solutions.

the system

$$\left.\begin{aligned} x_1 + 2x_2 &= 7, \\ x_1 + x_2 &= 4 \end{aligned}\right\}$$

is determinate: it has the solution $x_1 = 1$, $x_2 = 3$ and, as may readily be verified by the method of elimination, this solution is unique. On the other hand, the system

$$\left.\begin{aligned} 3x_1 - x_2 &= 1, \\ 6x_1 - 2x_2 &= 2 \end{aligned}\right\}$$

is indeterminate since it has infinitely many solutions of the form

$$x_1 = k, \qquad x_2 = 3k - 1 \tag{3}$$

where k is an arbitrary number; the solutions obtained using formulas (3) exhaust the solutions of the system.

The problem of the theory of systems of linear equations consists in elaborating methods to determine whether a given system of equations is consistent or not and, in the case of consistency, to establish the number of solutions and also to indicate a procedure for finding the solutions.

We begin with the most convenient practical method for finding solutions to systems with numerical coefficients, namely, *the method of successive elimination of unknowns*, or *Gauss' method*.

First, a preliminary remark. In future we will manipulate systems of equations in the following manner: both members of one of the equations of the system multiplied by one and the same number will be subtracted from the corresponding members of some other equation of the system. For the sake of definiteness, let us subtract both members of the first equation of system (1), multiplied by a number c, from the corresponding members of the second equation. We obtain a new system of linear equations:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a'_{21}x_1 + a'_{22}x_2 + \ldots + a'_{2n}x_n &= b'_2, \\ a_{31}x_1 + a_{32}x_2 + \ldots + a_{3n}x_n &= b_3, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{s1}x_1 + a_{s2}x_2 + \ldots + a_{sn}x_n &= b_s \end{aligned}\right\} \tag{4}$$

where

$$a'_{2j} = a_{2j} - ca_{1j} \quad \text{for} \quad j = 1, 2, \ldots, n, \; b'_2 = b_2 - cb_1$$

The systems (1) and (4) are equivalent, which is to say they are either both inconsistent or they are both consistent and have the same solutions. Indeed, let $k_1, k_2, \ldots, k_n$ be an arbitrary solution of system (1). Obviously, these numbers satisfy all the equations of (4) except the second. However, they likewise satisfy the second equa-

tion of the system (4). It will suffice to recall how this equation is expressed in terms of the second and first equations of system (1). Conversely, any solution of (4) will also satisfy (1). Indeed, the second equation of (1) is obtained by subtracting, from both members of the second equation of (4), the corresponding members of the first equation of the system multiplied by the number $-c$.

Quite naturally, *if manipulations of this kind are applied several times to system* (1), *the newly obtained system of equations will remain equivalent to the original system* (1).

It may happen that as a result of such manipulations, there will appear in our system an equation whose coefficients in the left-hand member are equal to zero. Now if the constant term of this equation is zero, then the equation is satisfied for any values of the unknowns and so *by discarding this equation we arrive at a system of equations equivalent to the original system.* But if the constant term of the equation at hand is nonzero, then the equation cannot be satisfied for any values of the unknowns and for this reason *the system obtained* (*and the equivalent original system as well*) *will be inconsistent.*

Let us now examine Gauss' method.

We are given an arbitrary system of linear equations (1). To be specific, suppose that the coefficient $a_{11} \neq 0$, though in reality it may of course be equal to zero and then we would have to start with some other, nonzero, coefficient of the first equation of the system.

Let us now transform system (1) by eliminating the unknown x_1 from all equations except the first. To do this, multiply both members of the first equation by the number $\frac{a_{21}}{a_{11}}$ and subtract from the corresponding members of the second equation; then subtract both members of the first equation, multiplied by $\frac{a_{31}}{a_{11}}$, from the corresponding members of the third equation, and so on.

We thus arrive at a new system made up of s linear equations in n unknowns:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n &= b_1, \\ a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n &= b'_2, \\ a'_{32}x_2 + a'_{33}x_3 + \ldots + a'_{3n}x_n &= b'_3, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a'_{s2}x_2 + a'_{s3}x_3 + \ldots + a'_{sn}x_n &= b'_s \end{aligned}\right\} \qquad (5)$$

We do not need to write out explicitly the expressions of the new coefficients a'_{ij} and the new constant terms b'_i via the coefficients and constant terms of the original system (1).

As we know, the system of equations (5) is equivalent to (1). Now transform (5). We no longer involve the first equation and

manipulate only that portion of (5) consisting of all equations except the first. We of course assume that there are no equations with all coefficients of the left members zero (such would have been rejected if their constant terms were likewise zero, and if that were not so, we would have proved the inconsistency of our system). Thus, among the coefficients a'_{ij} there are some different from zero; for definiteness, we put $a'_{22} \neq 0$. Now transform (5) by subtracting from both members of the third and of each of the succeeding equations both members of the second equation multiplied respectively by the numbers

$$\frac{a'_{32}}{a'_{22}}, \quad \frac{a'_{42}}{a'_{22}}, \quad \ldots, \quad \frac{a'_{s2}}{a'_{22}}$$

In this way we eliminate the unknown x_2 from all equations, except the first and second, and arrive at the following system of equations which is equivalent to (5) and hence to (1):

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n &= b_1, \\ a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n &= b'_2, \\ a''_{33}x_3 + \ldots + a''_{3n}x_n &= b''_3, \\ \ldots\ldots\ldots\ldots\ldots & \\ a''_{t3}x_3 + \ldots + a''_{tn}x_n &= b''_t \end{aligned}\right\}$$

Our system now contains t equations, $t \leqslant s$, since some of the equations were possibly discarded. Naturally the number of equations of the system could already have diminished after eliminating the unknown x_1. Subsequently, only a portion of the system obtained (that containing all equations except the first two) will be subject to transformations.

The question arises as to when this process of successive elimination of unknowns will stop.

If we arrive at a system in which one of the equations has a non-zero constant term and all the coefficients of the left member are equal to zero, then, as we know, our original system was inconsistent.

If that is not the case, then we obtain the following system of equations which is equivalent to system (1):

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1,\,k-1}x_{k-1} + a_{1k}x_k + \ldots + a_{1n}x_n &= b_1, \\ a'_{22}x_2 + \ldots + a'_{2,\,k-1}x_{k-1} + a'_{2k}x_k + \ldots + a'_{2n}x_n &= b'_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a^{(k-2)}_{k-1,\,k-1}x_{k-1} + a^{(k-2)}_{k-1,\,k}x_k + \ldots + a^{(k-2)}_{k-1,\,n}x_n &= b^{(k-2)}_{k-1}, \\ a^{(k-1)}_{kk}x_k + \ldots + a^{(k-1)}_{kn}x_n &= b^{(k-1)}_k \end{aligned}\right\} \quad (6)$$

Here $a_{11} \neq 0$, $a'_{22} \neq 0$, . . ., $a^{(k-2)}_{k-1, k-1} \neq 0$, $a^{(k-1)}_{kk} \neq 0$. Note also that $k \leqslant s$, and, obviously, $k \leqslant n$.

In this case system (1) *is consistent. It will be determinate for* $k = n$ *and indeterminate for* $k < n$.

Indeed, if $k = n$, then system (6) has the form

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a'_{22}x_2 + \ldots + a'_{2n}x_n &= b'_2, \\ \ldots\ldots\ldots\ldots\ldots & \\ a^{(n-1)}_{nn}x_n &= b^{(n-1)}_n \end{aligned}\right\} \qquad (7)$$

From the last equation we obtain a quite definite value for the unknown x_n. Substituting it into the next to the last equation, we find a uniquely defined value for the unknown x_{n-1}. Continuing in similar fashion, we find that system (7) and, for this reason, system (1) as well have a unique solution, that is to say, they are consistent and determinate.

But if $k < n$, for the "free" unknowns x_{k+1}, . . ., x_n we take arbitrary numerical values, then, moving, in system (6) from bottom to top, we find quite definite values for the unknowns x_k, x_{k-1}, . . ., x_2, x_1 (as above). Since the values for the free unknowns may be chosen in an infinity of ways, our system (6) and, hence, (1) as well are consistent but indeterminate. It is easy to verify that by using the foregoing method (given all possible choices of values for the free unknowns) we can find all the solutions of system (1).

At first glance, yet another form to which a system of linear equations may be reduced by the Gaussian method would appear possible, namely, the form obtained by adjoining to system (7) a number of equations containing only the unknown x_n. Actually, however, in this case the transformations have simply not been completed: since $a^{(n-1)}_{nn} \neq 0$, the unknown x_n may be eliminated in all equations from the $(n + 1)$th on.

Note that the "triangular" form of the system of equations (7) or the "trapezoidal" form of system (6) (for $k < n$) resulted from the assumption that the coefficients a_{11}, a'_{22}, etc. are different from zero. In the general case, the system of equations which we arrive at after completing the process of elimination of unknowns takes on a triangular or trapezoidal form only after an appropriate alteration in the numbering of the unknowns.

To summarize, then, we find that *the Gaussian method is applicable to any system of linear equations. The system is inconsistent if after the transformations we obtain an equation in which the coefficients of all unknowns are zero and the constant term is nonzero; but if no such equation is encountered, the system is consistent. A consistent system of equations is determinate if it reduces to the triangular form* (7) *and indeterminate if it reduces to the trapezoidal form* (6) *for* $k < n$.

Let us apply what has been said to the case of a system of *homogeneous* linear equations, that is, equations whose constant terms are zero. Such a system is always consistent since it has a *zero solution* (0, 0, . . ., 0). Suppose that in the system at hand the number of equations is *less* than the number of unknowns. Then our system cannot reduce to the triangular form since in the Gaussian elimination process the number of equations of the system can diminish but not increase; hence, it reduces to the trapezoidal form and so is indeterminate.

To put it otherwise, *if in a system of homogeneous linear equations the number of equations is less than the number of unknowns, then this system has, in addition to the zero solution, nonzero solutions*, that is, solutions in which the values of some (or even all) unknowns are nonzero. *There is an infinity of such solutions.*

In practical solutions of a system of linear equations by the Gaussian method, one should write down the matrix of the coefficients of the system and adjoin a column made up of the constant terms, which, for the sake of convenience, are separated by a vertical line, and then perform all the manipulations on the rows of this "augmented" matrix.

Example 1. Solve the system

$$\left.\begin{aligned} x_1 + 2x_2 + 5x_3 &= -9, \\ x_1 - x_2 + 3x_3 &= 2, \\ 3x_1 - 6x_2 - x_3 &= 25 \end{aligned}\right\}$$

Transform the augmented matrix of the system:

$$\left(\begin{array}{ccc|c} 1 & 2 & 5 & -9 \\ 1 & -1 & 3 & 2 \\ 3 & -6 & -1 & 25 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & 2 & 5 & -9 \\ 0 & -3 & -2 & 11 \\ 0 & -12 & -16 & 52 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & 2 & 5 & -9 \\ 0 & -3 & -2 & 11 \\ 0 & 0 & -8 & 8 \end{array}\right)$$

We thus arrive at the following system of equations:

$$\left.\begin{aligned} x_1 + 2x_2 + 5x_3 &= -9, \\ -3x_2 - 2x_3 &= 11, \\ -8x_3 &= 8 \end{aligned}\right\}$$

which has the unique solution

$$x_1 = 2, \quad x_2 = -3, \quad x_3 = -1$$

The original system proved to be determinate.

Example 2. Solve the system

$$\left.\begin{aligned} x_1 - 5x_2 - 8x_3 + x_4 &= 3, \\ 3x_1 + x_2 - 3x_3 - 5x_4 &= 1, \\ x_1 - 7x_3 + 2x_4 &= -5, \\ 11x_2 + 20x_3 - 9x_4 &= 2 \end{aligned}\right\}$$

We transform the augmented matrix of the system:

$$\left(\begin{array}{cccc|c} 1 & -5 & -8 & 1 & 3 \\ 3 & 1 & -3 & -5 & 1 \\ 1 & 0 & -7 & 2 & -5 \\ 0 & 11 & 20 & -9 & 2 \end{array}\right) \rightarrow \left(\begin{array}{cccc|c} 1 & -5 & -8 & 1 & 3 \\ 0 & 16 & 21 & -8 & -8 \\ 0 & 5 & 1 & 1 & -8 \\ 0 & 11 & 20 & -9 & 2 \end{array}\right)$$

$$\rightarrow \left(\begin{array}{cccc|c} 1 & -5 & -8 & 1 & 3 \\ 0 & -89 & 0 & -29 & 160 \\ 0 & 5 & 1 & 1 & -8 \\ 0 & -89 & 0 & -29 & 162 \end{array}\right) \rightarrow \left(\begin{array}{cccc|c} 1 & -5 & -8 & 1 & 3 \\ 0 & -89 & 0 & -29 & 160 \\ 0 & 5 & 1 & 1 & -8 \\ 0 & 0 & 0 & 0 & 2 \end{array}\right)$$

We arrive at a system containing the equation $0 = 2$. Consequently, the original system is inconsistent.

Example 3. Solve the system

$$\left.\begin{aligned} 4x_1 + x_2 - 3x_3 - x_4 &= 0, \\ 2x_1 + 3x_2 + x_3 - 5x_4 &= 0, \\ x_1 - 2x_2 - 2x_3 + 3x_4 &= 0 \end{aligned}\right\}$$

This is a system of homogeneous equations, and the number of equations is less than the number of unknowns; it must therefore be indeterminate. Since all the constant terms are zero, we perform manipulations solely with the matrix of the coefficients of the system:

$$\begin{pmatrix} 4 & 1 & -3 & -1 \\ 2 & 3 & 1 & -5 \\ 1 & -2 & -2 & 3 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & 9 & 5 & -13 \\ 0 & 7 & 5 & -11 \\ 1 & -2 & -2 & 3 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & 2 & 0 & -2 \\ 0 & 7 & 5 & -11 \\ 1 & -2 & -2 & 3 \end{pmatrix}$$

We arrive at the following system of equations:

$$\left.\begin{aligned} 2x_2 - 2x_4 &= 0, \\ 7x_2 + 5x_3 - 11x_4 &= 0, \\ x_1 - 2x_2 - 2x_3 + 3x_4 &= 0 \end{aligned}\right\}$$

We can take either one of the unknowns x_2 or x_4 for the free unknown. Let $x_4 = \alpha$. Then from the first equation it follows that $x_2 = \alpha$, and from the second equation we get $x_3 = \frac{4}{5}\alpha$ and, finally, from the third equation $x_1 = \frac{3}{5}\alpha$. Thus,

$$\frac{3}{5}\alpha,\ \alpha,\ \frac{4}{5}\alpha,\ \alpha$$

is the general form of the solutions of the given system of equations.

2. Determinants of Second and Third Order

The method of solving systems of linear equations given in Sec. 1 is extremely simple and requires the performance of the same kind of computations, which are readily carried out on computing machines. Its drawback, however, is that it does not enable us to

state the conditions of consistency or determinacy of the system by means of coefficients and constant terms of the system. On the other hand, even in the case of a determinate system, this method does not permit finding formulas that express the solution of the system in terms of its coefficients and constant terms. However, all this proves to be necessary in theoretical problems, in particular, in geometrical investigations; for this reason, the theory of systems of linear equations has to be elaborated by different and more profound methods. The general case will be pursued in the next chapter; for the present, we consider determinate systems having an equal number of equations and unknowns. We begin with the systems in two and three unknowns of elementary algebra.

Let there be given *a system of two linear equations in two unknowns*:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1, \\ a_{21}x_1 + a_{22}x_2 &= b_2 \end{aligned}\right\} \tag{1}$$

whose coefficients form the second-order square matrix

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \tag{2}$$

Applying to system (1) the method of equalizing the coefficients, we obtain

$$\begin{aligned} (a_{11}a_{22} - a_{12}a_{21})\, x_1 &= b_1a_{22} - a_{12}b_2, \\ (a_{11}a_{22} - a_{12}a_{21})\, x_2 &= a_{11}b_2 - b_1a_{21} \end{aligned}$$

Suppose that $a_{11}a_{22} - a_{12}a_{21} \neq 0$. Then

$$x_1 = \frac{b_1a_{22} - a_{12}b_2}{a_{11}a_{22} - a_{12}a_{21}}, \quad x_2 = \frac{a_{11}b_2 - b_1a_{21}}{a_{11}a_{22} - a_{12}a_{21}} \tag{3}$$

It is easy to show, by substituting the values of the unknowns into (1), that (3) is a solution of system (1). The question of the uniqueness of this solution will be considered in Sec. 7.

The common denominator of the values of the unknowns (3) is very simply expressed in terms of the elements of matrix (2): it is equal to the product of the elements of the principal diagonal minus the product of the elements of the secondary diagonal. This number is called the *determinant* of the matrix (2); we call it a *second-order determinant* since the matrix (2) is a second-order matrix. To symbolize a determinant, we use vertical lines in place of parentheses:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} \tag{4}$$

Examples.

(1) $$\begin{vmatrix} 3 & 7 \\ 1 & 4 \end{vmatrix} = 3\cdot 4 - 7\cdot 1 = 5,$$

(2) $$\begin{vmatrix} 1 & -2 \\ 3 & 5 \end{vmatrix} = 1\cdot 5 - (-2)\cdot 3 = 11$$

It is worth stressing once again, that while a matrix is an array of numbers, a determinant is a number associated in a definite way with a square matrix. The products $a_{11}a_{22}$ and $a_{12}a_{21}$ are called the *terms* of a second-order determinant.

The numerators of expressions (3) have the same form as the denominators, that is, they are also determinants of second order; the numerator of the expression for x_1 is the determinant of the matrix obtained from matrix (2) by replacing its first column by the column of constant terms of system (1), the numerator of the expression for x_2 is the determinant of the matrix obtained from matrix (2) by replacing its second column. We can now write formula (3) as follows:

$$x_1 = \frac{\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}, \qquad x_2 = \frac{\begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}} \tag{5}$$

This rule for the solution of a system of two linear equations in two unknowns (called *Cramer's rule*) is formulated as follows.

If the determinant, (4), *of the coefficients of a system of equations,* (1), *is different from zero, we obtain the solution of system* (1) *by taking for the values of the unknowns the fractions whose common denominator is determinant* (4) *and whose numerator for the unknown* x_i $(i = 1, 2)$ *is a determinant obtained by replacing in determinant* (4) *the ith column (that is, the column of coefficients of the desired unknown) by the column of the constant terms of system* (1).*

Example. Solve the system

$$\left.\begin{aligned} 2x_1 + x_2 &= 7, \\ x_1 - 3x_2 &= -2 \end{aligned}\right\}$$

The determinant of the coefficients is

$$d = \begin{vmatrix} 2 & 1 \\ 1 & -3 \end{vmatrix} = -7$$

It is different from zero and, for this reason, Cramer's rule is applicable. The determinants

$$d_1 = \begin{vmatrix} 7 & 1 \\ -2 & -3 \end{vmatrix} = -19, \qquad d_2 = \begin{vmatrix} 2 & 7 \\ 1 & -2 \end{vmatrix} = -11$$

---

* For brevity we speak here of replacing columns "in the determinant". In the same way, we will in future, if it is more convenient, speak of rows and columns of a determinant, of its elements and diagonals, etc.

are the numerators for the unknowns. Thus, the following set of numbers is the solution of our system:

$$x_1=\frac{d_1}{d}=\frac{19}{7}, \qquad x_2=\frac{d_2}{d}=\frac{11}{7}$$

The introduction of second-order determinants does not substantially simplify the solution of a system of two linear equations in two unknowns, which does not present any difficulties as it is. However, for the case of *systems of three linear equations in three unknowns*, similar methods are of practical utility. Suppose we have a system

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned}\right\} \tag{6}$$

with the coefficient matrix

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \tag{7}$$

If we multiply both sides of the first equation of (6) by the number $a_{22}a_{33} - a_{23}a_{32}$, both sides of the second equation by $a_{13}a_{32} - a_{12}a_{33}$, both sides of the third equation by $a_{12}a_{23} - a_{13}a_{22}$, and then add all three equations, it is easy to verify that the coefficients of x_2 and x_3 will turn out to be zero, that is, these unknowns are eliminated simultaneously and we obtain the equation

$$\begin{aligned} &(a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} \\ &- a_{11}a_{23}a_{32})\,x_1 = b_1a_{22}a_{33} + a_{12}a_{23}b_3 + a_{13}b_2a_{32} - a_{13}a_{22}b_3 \\ &\qquad - a_{12}b_2a_{33} - b_1a_{23}a_{32} \end{aligned} \tag{8}$$

Here, the coefficient of x_1 is called a *third-order determinant* corresponding to matrix (7). The symbolism is the same as in the case of second-order determinants; thus,

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{aligned} &a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ &- a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32} \end{aligned} \tag{9}$$

The expression for a third-order determinant is rather involved, but the rule for its formation from the elements of matrix (7) is extremely simple, as witness: one of the three terms (of the determinant) in (9) with the plus sign is the product of the elements of the principal diagonal, each of the other two is a product of the elements lying parallel to this diagonal, with the third factor added from the opposite corner of the matrix. The terms with the minus sign

in (9) are constructed in a similar manner but relative to the secondary diagonal. We obtain a technique for computing determinants of the third order that produces quick results (after a certain amount

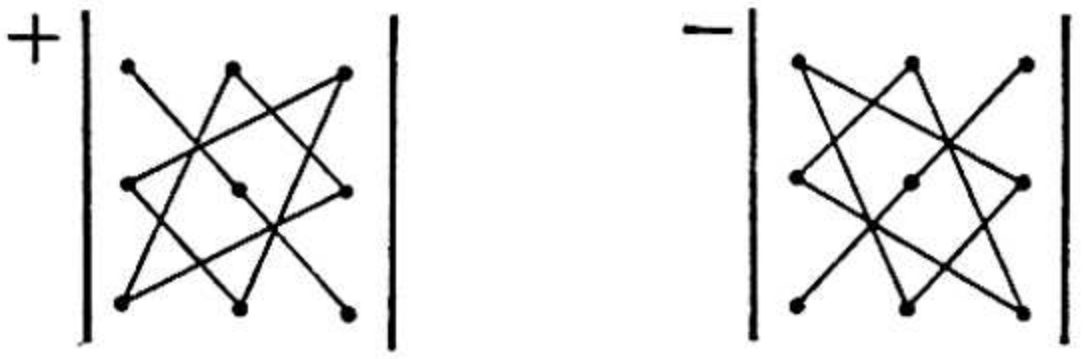

Fig. 1

of practice). Fig. 1 gives a schematic view of computing the positive terms (left) and the negative terms (right) of a third-order determinant.

Examples.

(1)
$$\begin{vmatrix} 2 & 1 & 2 \\ -4 & 3 & 1 \\ 2 & 3 & 5 \end{vmatrix} = 2\cdot 3\cdot 5 + 1\cdot 1\cdot 2 + 2\cdot(-4)\cdot 3 - 2\cdot 3\cdot 2 - 1\cdot(-4)\cdot 5 - 2\cdot 1\cdot 3 = 30 + 2 - 24 - 12 + 20 - 6 = 10$$

(2)
$$\begin{vmatrix} 1 & 0 & -5 \\ -2 & 3 & 2 \\ 1 & -2 & 0 \end{vmatrix} = 1\cdot 3\cdot 0 + 0\cdot 2\cdot 1 + (-5)\cdot(-2)\cdot(-2) - (-5)\cdot 3\cdot 1 - 0\cdot(-2)\cdot 0 - 1\cdot 2\cdot(-2) = -20 + 15 + 4 = -1$$

The right-hand side of (8) is also a third-order determinant, namely, the determinant of the matrix obtained from matrix (7) by replacing its first column by the column of constant terms of system (6). If we denote determinant (9) by the letter d and the determinant obtained by replacing its jth column ($j = 1, 2, 3$) by the column of constant terms of system (6) by the symbol d_j, then equation (8) becomes $dx_1 = d_1$, whence, for $d \neq 0$, it follows that

$$x_1 = \frac{d_1}{d} \tag{10}$$

In exactly the same way, by multiplying equation (6) by the numbers $a_{23}a_{31} - a_{21}a_{33}$, $a_{11}a_{33} - a_{13}a_{31}$, $a_{13}a_{21} - a_{11}a_{23}$, respectively, we obtain for x_2 the following expression (again for $d \neq 0$):

$$x_2 = \frac{d_2}{d} \tag{11}$$

Finally, multiplying these equations, respectively, by $a_{21}a_{32} - a_{22}a_{31}$, $a_{12}a_{31} - a_{11}a_{32}$, $a_{11}a_{22} - a_{12}a_{21}$, we arrive at the expression for x_3:

$$x_3 = \frac{d_3}{d} \tag{12}$$

Substituting expressions (10) to (12) into equation (6) (it is of course assumed that the determinants d and all d_j are written in expanded form), we would find—after cumbersome computations, all, however, well within the grasp of the reader—that all these equations are satisfied, that is, that the numbers (10)-(12) constitute the solution of system (6). Thus, *if the determinant of the coefficients of a system of three linear equations in three unknowns is nonzero, then the solution of this system may be found by Cramer's rule as stated for the case of a system of two equations.* In Sec. 7 the reader will find a different proof of this assertion (one that does not rely on the calculations we have omitted here) and also a proof of the uniqueness of the solution (10)-(12) of system (6) for the more general case.

Example. Solve the system of equations

$$\left.\begin{aligned} 2x_1 - x_2 + x_3 &= 0, \\ 3x_1 + 2x_2 - 5x_3 &= 1, \\ x_1 + 3x_2 - 2x_3 &= 4 \end{aligned}\right\}$$

The determinant of the coefficients is nonzero:

$$d = \begin{vmatrix} 2 & -1 & 1 \\ 3 & 2 & -5 \\ 1 & 3 & -2 \end{vmatrix} = 28$$

so the Cramer rule is applicable. The numerators for the unknowns are

$$d_1 = \begin{vmatrix} 0 & -1 & 1 \\ 1 & 2 & -5 \\ 4 & 3 & -2 \end{vmatrix} = 13, \quad d_2 = \begin{vmatrix} 2 & 0 & 1 \\ 3 & 1 & -5 \\ 1 & 4 & -2 \end{vmatrix} = 47,$$

$$d_3 = \begin{vmatrix} 2 & -1 & 0 \\ 3 & 2 & 1 \\ 1 & 3 & 4 \end{vmatrix} = 21$$

Hence, the following numbers constitute the solution of the system:

$$x_1 = \frac{13}{28}, \quad x_2 = \frac{47}{28}, \quad x_3 = \frac{21}{28} = \frac{3}{4}$$

3. Arrangements and Permutations

In the study of determinants of order n we will need certain concepts and facts relating to finite sets. Suppose we have a certain finite set M consisting of n elements, which may be enumerated by using the natural numbers 1, 2, . . ., n; since the properties of the elements of the set M will not play any role whatsoever, we simply say that the elements of M are the numbers 1, 2, . . ., n.

Besides the natural order of 1, 2, . . ., n, we can arrange the numbers in many other ways. Thus, we can arrange the numbers 1, 2, 3, 4 as 3, 1, 2, 4 or 2, 4, 1, 3 and so on. Every rearrangement

of the numbers 1, 2, . . ., n in any definite order is called a *permutation* (or *arrangement*)* of n numbers (or n symbols).

The number of distinct arrangements of n symbols is equal to the product $1 \cdot 2 \ldots n$, denoted by $n!$ (read "n factorial"). Indeed, the general form of an arrangement of n symbols is $i_1, i_2, \ldots, i_n$, where each of the i_s is one of the numbers 1, 2, . . ., n, without repetitions. Use any one of the numbers 1, 2, . . ., n for i_1; this yields n distinct possibilities. But if i_1 has been chosen, then for i_2 we can only take one of the remaining $n - 1$ numbers; that is, the number of different ways of choosing the symbols i_1 and i_2 is equal to the product $n(n - 1)$ and so on.

Thus, the number of arrangements of n symbols for $n = 2$ is $2! = 2$ (the arrangements 12 and 21; in examples where $n \leqslant 9$, we do not separate the symbols by commas); for $n = 3$ this number is $3! = 6$, for $n = 4$ it is $4! = 24$. As n increases, the number of arrangements increases very fast: for $n = 5$ it is $5! = 120$, and for $n = 10$ it is already 3,628,800.

If in a certain arrangement we interchange any two symbols (not necessarily adjacent) and leave all the remaining ones fixed, we obtain a new arrangement. This operation is called a *transposition.*

All $n!$ *arrangements of* n *symbols may be ordered so that each is obtained from the preceding one via a single transposition; any arrangement can serve as the starting point.*

This assertion holds true for $n = 2$: if it is required to begin with the arrangement 12, the desired order will be 12, 21; if we begin with the arrangement 21, then the order will be 21, 12. Suppose our assertion has already been proved for $n - 1$, and we prove it for n. Let us begin with the arrangement

$$i_1, i_2, \ldots, i_n \tag{1}$$

We consider all arrangements of n symbols starting with i_1. There are $(n - 1)!$ such arrangements and they may be ordered in accord with the requirements of the theorem, beginning with (1), since this actually reduces to an ordering of all arrangements of $n - 1$ symbols; this ordering, by the induction hypothesis, may be initiated from any arrangement, say, $i_2, \ldots, i_n$. In the last of the arrangements of n symbols thus obtained we perform a transposition of i_1 and any other symbol (say i_2) and, again beginning with the arrangement obtained, we appropriately order all the arrangements with i_2 in first place, and so forth. It is thus obviously possible to enumerate all arrangements of n symbols.

---

* *Translator's note*: the term "arrangement" will be used, since permutation is reserved in this text for a different concept.

From this theorem it follows that *it is possible to pass from any arrangement of n symbols to any other arrangement of the same symbols by means of several transpositions.*

We say that in a given arrangement the numbers i and j constitute an *inversion* if $i > j$ but i comes before j in the arrangement. An arrangement is termed *even* if its symbols form an even number of inversions, otherwise it is *odd.* Thus, the arrangement $1, 2, \ldots, n$ is even for any n since the number of inversions here is zero. The arrangement 451362 ($n = 6$) contains 8 inversions and so is even. The arrangement 38524671 ($n = 8$) contains 15 inversions and so is odd.

Every transposition changes the parity of the arrangement.

To prove this important theorem let us first consider the case where the symbols i and j being interchanged are adjacent; in other words, the arrangement is of the form $\ldots, i, j, \ldots,$ where the dots stand for symbols unaltered by the transposition. The transposition converts our arrangement into the arrangement $\ldots, j, i, \ldots,$ it being understood that in both cases each of the symbols i, j constitutes the same set of inversions with the symbols which remain fixed. Whereas earlier i and j did not constitute an inversion, in the new arrangement there is a fresh inversion; hence, the number of inversions has increased by unity; contrariwise, if they originally formed an inversion, then the inversion now vanishes, the number of inversions being diminished by unity. In both cases the parity of the arrangement is altered.

Now let us suppose that there are s symbols, $s > 0$, between i and j; that is, the arrangement is of the form

$$\ldots, i, k_1, k_2, \ldots, k_s, j, \ldots \tag{2}$$

The symbols i and j may be interchanged by means of a succession of $2s + 1$ transpositions of adjacent elements. These are transpositions interchanging the symbols i and k_1, then interchanging i (now in the place of k_1) and k_2, and so on until i occupies the site of symbol k_s. These s transpositions are then followed by a transposition that interchanges the symbols i and j and then s transpositions of the symbol j with all k's; as a result, j occupies the place of i and the symbols k return to their original sites. We have thus changed the parity of the arrangement an odd number of times and for this reason the arrangements (2) and

$$\ldots, j, k_1, k_2, \ldots, k_s, i, \ldots \tag{3}$$

are of different parity.

For $n \geqslant 2$, *the number of even arrangements of n symbols is equal to the number of odd arrangements, i.e.,* $\frac{1}{2}\, n!$

Indeed, proceeding from the foregoing, order all arrangements of n symbols so that each one is obtained from the preceding one by a single transposition. Adjacent arrangements will have opposite parity, that is, the arrangements are ordered so that even and odd arrangements alternate. Our assertion now follows from the obvious remark that for $n \geqslant 2$ the number $n!$ is even.

Let us now define a new concept, that of a *permutation of degree n*. Write down two arrangements of n symbols, one under the other, and place parentheses around them; for example, for $n = 5$,

$$\begin{pmatrix} 3 & 5 & 1 & 4 & 2 \\ 5 & 2 & 3 & 4 & 1 \end{pmatrix} \tag{4}$$

In this example,* 5 stands under 3, 2 under 5, etc. We say that number 3 *goes into* 5, 5 goes into 2, 1 goes into 3, and the number 4 goes into 4 (or *remains fixed*) and, finally, 2 goes into 1. Thus, two arrangements written one under the other in the form shown in (4) define a certain *one-to-one mapping* of the set of the first five natural numbers onto itself, that is, a mapping in which each of the natural numbers 1, 2, 3, 4, 5 is associated with one of these same natural numbers, distinct numbers corresponding to distinct numbers. And since there are only five numbers (a finite set), *each one* corresponds to one of the five numbers 1, 2, 3, 4, 5, namely, that one into which it "goes".

It is clear that the one-to-one mapping of the set of the first five natural numbers which we obtained by means of (4) could be obtained by writing certain other pairs of arrangements of five symbols one under the other. These are obtained from (4) by means of several transpositions of the columns, such as, for instance,

$$\begin{pmatrix} 2 & 1 & 5 & 3 & 4 \\ 1 & 3 & 2 & 5 & 4 \end{pmatrix}, \quad \begin{pmatrix} 1 & 5 & 2 & 4 & 3 \\ 3 & 2 & 1 & 4 & 5 \end{pmatrix}, \quad \begin{pmatrix} 2 & 5 & 1 & 4 & 3 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix} \tag{5}$$

In all these groups, 3 goes into 5, 5 into 2, etc.

Similarly, two arrangements of n symbols written one under the other define a one-to-one mapping of the set of the first n natural numbers onto itself. Any one-to-one mapping A of the set of the first n natural numbers onto itself is termed a *permutation of degree n*. Obviously, any permutation A may be written with the help of two arrangements, written one under the other:

$$A = \begin{pmatrix} i_1, & i_2, & \ldots, & i_n \\ \alpha_{i_1}, & \alpha_{i_2}, & \ldots, & \alpha_{i_n} \end{pmatrix} \tag{6}$$

---

* This array looks like a matrix of two rows and five columns, but its meaning is quite different.

Here, α_i denotes the number into which i ($i = 1, 2, \ldots, n$) goes in the permutation A.

The permutation A possesses many different notations of the form (6). For instance, (4) and (5) are different ways of denoting one and the same permutation of degree 5.

It is possible to pass from one mode of notation of the permutation A to another simply by performing a number of transpositions of the columns. It is then possible to obtain (6) in a mode such that the upper (or lower) row is any preassigned arrangement of n symbols. In particular, any permutation A of degree n may be written as

$$A = \begin{pmatrix} 1 & 2 & \ldots & n \\ \alpha_1 & \alpha_2 & \ldots & \alpha_n \end{pmatrix} \tag{7}$$

that is, with the numbers in the upper row arranged in their natural order. Given this notation, various permutations differ in the arrangements of the lower row, and for this reason *the number of permutations of degree n is equal to the number of arrangements of n symbols, or $n!$.*

An instance of an nth-degree permutation is the *identity permutation*

$$E = \begin{pmatrix} 1 & 2 & \ldots & n \\ 1 & 2 & \ldots & n \end{pmatrix}$$

in which all symbols remain fixed.

It is well to point out that the upper and lower rows of the permutation A in notation (6) play different roles so that if interchanged the result would be a different permutation. Thus, the permutations of degree 4

$$\begin{pmatrix} 2 & 1 & 4 & 3 \\ 4 & 3 & 1 & 2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 4 & 3 & 1 & 2 \\ 2 & 1 & 4 & 3 \end{pmatrix}$$

are different: in the first, 2 goes into 4, in the second it goes into 3.

Let us take some permutation A of degree n in the arbitrary notation (6). The arrangements constituting the upper and lower rows in this mode can have either identical or opposite parities. As we know, we can proceed to any other mode of permutation A by means of successive transpositions in the upper row and corresponding transpositions in the lower row. However, by performing one transposition in the upper row of (6) and one transposition of the corresponding elements in the lower row, we simultaneously alter the parities of both arrangements and therefore preserve the coincident or opposite nature of these parities. From this it follows that *in all modes of notation of the permutation A, the parities of the upper and lower rows either coincide or are opposite.* In the former case, A is

called *even*, in the latter, *odd*. In particular, the identity permutation is even.

If the permutation A is written as (7) (that is, with the even arrangement 1, 2, . . ., n in the upper row), then the parity of permutation A is determined by the parity of the arrangement $\alpha_1, \alpha_2, \ldots, \alpha_n$ of the lower row. Whence it follows that *the number of even permutations of degree n is equal to the number of odd permutations, that is,* $\frac{1}{2}n!$.

The definition of parity of a permutation may be cast in the following modified form. If, when written in mode (6), the parities of both rows coincide, then the number of inversions is either even in both rows or is odd in both, that is, the total number of inversions in both rows of (6) is even; but if the parities of the rows in mode (6) are opposite, then the total number of inversions in these two rows is odd. Thus, *permutation A is even if the total number of inversions in the two rows in any mode of notation is even, it is odd otherwise.*

Example. Let there be given a permutation of degree 5:

$$\begin{pmatrix} 3 & 1 & 4 & 5 & 2 \\ 2 & 5 & 4 & 3 & 1 \end{pmatrix}$$

There are 4 inversions in the upper row, and 7 inversions in the lower row. The total number in the two rows is 11, and so the permutation is odd.

Rewrite this permutation as

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 1 & 2 & 4 & 3 \end{pmatrix}$$

The number of inversions in the upper row is 0, in the lower, 5; that is, the total number is again odd. Though the modes of notation differ, the permutations preserve the parity of the total number of inversions, but not the actual number of them.

We wish to indicate other ways, equivalent to those given above, of defining parities of permutations.* For this purpose we define *multiplication of permutations*, which is of great interest in itself. As we already know, a permutation of degree n is a one-to-one mapping of the set of numbers 1, 2, . . ., n onto itself. The result of a successive execution of two one-to-one mappings of the set 1, 2, . . ., n onto itself will obviously again be a certain one-to-one mapping of the set onto itself, that is to say, a successive execution of two permutations of degree n leads to a certain very definite third permutation of degree n called the *product* of the first by the second. Thus, if we have the permutations of degree four,

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 4 & 2 \end{pmatrix}.$$

* This material may be omitted in a first reading since it will be required only in Chapter 14.

then

$$AB = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \end{pmatrix}$$

In the permutation A, the symbol 1 goes into 3, but in B the symbol 3 goes into 4, and so for AB the symbol 1 goes into 4, etc.

Multiplication is only possible with permutations of the same degree. *Multiplication of permutations of degree n for $n \geqslant 3$ is non-commutative.* Indeed, using A and B, the product BA yields

$$BA = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 2 & 1 \end{pmatrix}$$

which shows that the permutation BA differs from the permutation AB. Such examples may be chosen for all n, $n \geqslant 3$, although for certain pairs of permutations, commutativity may accidentally be valid.

The multiplication of permutations is associative; that is, we can speak of the product of any finite number of permutations of degree n taken (because of noncommutativity) in a definite order. Let there be given permutations A, B and C and let the symbol i_1, $1 \leqslant i_1 \leqslant n$, go to i_2 in A, i_2 to i_3 in B and to i_4 in the permutation C. Then in the permutation AB, i_1 goes to i_3, in BC the symbol i_2 goes to i_4 and therefore the symbol i_1 goes to i_4 whether we perform $(AB)\,C$ or $A\,(BC)$.

It is obvious that *the product of any permutation A by the identity permutation E (and also the product of E by A) is equal to A:*

$$AE = EA = A$$

Let us now define the *inverse* of the permutation A as the permutation A^{-1} of the same degree such that

$$AA^{-1} = A^{-1}A = E$$

It is easy to see that the inverse of

$$A = \begin{pmatrix} 1 & 2 & \dots & n \\ \alpha_1 & \alpha_2 & \dots & \alpha_n \end{pmatrix}$$

is the permutation

$$A^{-1} = \begin{pmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_n \\ 1 & 2 & \dots & n \end{pmatrix}$$

obtained from A by interchanging the upper and lower rows.

Let us now examine permutations of a special kind which are obtained from the identity permutation E by means of a single transposition performed in the lower row. Such permutations are

odd: they are termed *transpositions* and are of the form

$$\begin{pmatrix} \ldots & i & \ldots & j & \ldots \\ \ldots & j & \ldots & i & \ldots \end{pmatrix} \tag{8}$$

where the dots stand for symbols that remain fixed. Let us agree to denote this transposition by the symbol (i, j). Application of the transposition of symbols i, j to the lower row of (7) of an arbitrary permutation A is equivalent to multiplying A on the right by the permutation (8), that is by (i, j). We know that all arrangements of n symbols may be obtained from one of them, say from $1, 2, \ldots, n$, by successive transpositions, and so any permutation may be obtained from the identity permutation by successive transpositions in the lower row, that is, by successive multiplication by permutations of the form (8). It can therefore be asserted (omitting the factor E) that *any permutation can be represented as a product of transpositions.*

Any permutation may be factored into a product of transpositions in many different ways. It is always possible, for example, to add two identical factors of the form (i, j) (i, j), which when multiplied yield E, that is to say, cancel out. Let us take a somewhat less trivial instance:

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 4 & 3 & 1 \end{pmatrix} = (12)\,(15)\,(34) = (14)\,(24)\,(45)\,(34)\,(13)$$

This new way of defining the parity of a permutation is based on the following theorem.

For all factorizations of a permutation into a product of transpositions, the parity of the number of these transpositions is the same and coincides with the parity of the permutation.

Thus, in the example given above, the permutation is odd, as may also be verified by counting the number of inversions.

This theorem will be proved if we demonstrate that *the product of any k transpositions is a permutation whose parity coincides with the parity of the number k.* For $k = 1$ this is true because a transposition is an odd permutation. Let our assertion be proved for the case of $k - 1$ factors. Then its validity for k factors follows from the fact that the numbers $k - 1$ and k are of opposite parity and the multiplication of a permutation (in this case, the product of the first $k - 1$ factors) by a transposition is equivalent to this transposition performed in the lower row of the permutation, which is to say, it changes the parity.

Decomposition into cycles is a convenient way of writing permutations which makes it easy to find their parity. Any permutation of degree n can leave certain symbols $1, 2, \ldots, n$ fixed while moving others. *A cyclic permutation* (or, simply, a *cycle*) is a permu-

tation such that when it is repeated a sufficient number of times any one of the symbols can be transformed into any other symbol. Such, for instance, is the permutation of degree eight

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 8 & 6 & 4 & 5 & 2 & 7 & 3 \end{pmatrix}$$

It transfers the symbols 2, 3, 6, and 8, with 2 going into 8, 8 into 3, 3 into 6, and 6 again into 2.

All transpositions belong to cycles. By analogy with the earlier used abbreviated notation for transpositions, the following notation is used for cycles: the symbols being transferred are enclosed in parentheses in the order in which they go into one another when the permutation is repeated; any transferable symbol can serve as the starting point, and the last one is that which goes into the first. Thus, for the example given above, this notation has the form

$$(2\ 8\ 3\ 6)$$

The number of symbols transferred by a cycle is called the cycle *length*.

Two cycles of degree n are called *disjoint* if they do not have any common symbols subject to transfer. It is clear that in multiplication of disjoint cycles, the order of the factors does not affect the result.

Any permutation can be factored uniquely into a product of pairwise disjoint cycles. The proof is simple and so we omit it. In actual practice, the factorization is accomplished in the following manner: begin with any one of the symbols subject to transfer, write out those symbols into which it goes in a new permutation until you arrive at the original symbol. After thus "closing" the cycle, begin with one of the remaining transferable symbols to obtain the second cycle, and so on.

Examples.

(1) $$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 1 & 2 & 4 \end{pmatrix} = (13)\,(254)$$

(2) $$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 5 & 2 & 8 & 7 & 6 & 1 & 4 & 3 \end{pmatrix} = (156)\,(38)\,(47)$$

Conversely, for any permutation specified by a decomposition into disjoint cycles, it is possible to find a notation in ordinary form, provided that the degree of the permutation is known. For example,

(3) $$(1372)\,(45) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 1 & 7 & 5 & 4 & 6 & 2 \end{pmatrix}$$

if it is known that the permutation is of degree 7.

Let there be given a permutation of degree n and let s be the number of disjoint cycles in its decomposition plus the number of symbols which it holds

fixed*. The difference $n - s$ is called the *decrement* of this permutation. The decrement is obviously equal to the number of actually transferable symbols diminished by the number of disjoint cycles entering into the decomposition of the permutation. For Examples 1, 2, and 3 above, the decrement will be equal to 3, 4, and 4, respectively.

The parity of a permutation coincides with the parity of the decrement of the permutation.

Indeed, any cycle of length k may be represented in the following manner as the product of $k - 1$ transpositions:

$$(i_1, i_2, \ldots, i_k) = (i_1, i_2)(i_1, i_3) \ldots (i_1, i_k).$$

Let us now suppose we have an expansion of permutation A into disjoint cycles. If each one of the cycles is factored by the indicated method into a product of transpositions, we get a representation of permutation A in the form of a product of transpositions. The number of these transpositions will obviously be less than the number of symbols actually transferable by A by a number equal to the number of disjoint cycles in the decomposition of the permutation. Whence it follows that the permutation A may be factored into a product of transpositions whose number is equal to the decrement, and for this reason the parity of the permutation is determined by the parity of the decrement.

4. Determinants of nth Order

We now wish to generalize the results obtained in Sec. 2 for $n = 2$ and $n = 3$ to the case of an arbitrary n. For this purpose, we have to introduce determinants of order n. However, it is not possible to do that the way we introduced determinants of order two and three, that is by solving a system of linear equations in the general form: as n increased, the computations would become progressively more unwieldy, and totally unmanageable for arbitrary n. We choose a different approach. Considering the determinants of order two and three which we are already familiar with, let us attempt to establish a general law expressing these determinants in terms of the elements of the corresponding matrices, and then let us apply that law as a definition for an nth-order determinant. After that we will prove that Cramer's rule holds true under such a definition.

Recall the expressions for determinants of order two and three:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32}$$

We see that any term of a second-order determinant is a product of two elements which lie in different rows and also in different co-

---

* With every symbol which the permutation holds fixed it is possible to associate a "cycle" of length 1, i.e., say, in Example 2 above we could write: (156) (38) (47) (2). But we shall not do that.

lumns, and also that all products of this type that may be formed from the elements of a second-order matrix (two altogether) are utilized as terms of the determinant. Similarly, every term of a third-order determinant is a product of three elements, also taken one in each row and each column; again, all such products are utilized as terms of the determinant.

Let us now take a square matrix of order n:

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \tag{1}$$

We consider all possible products of the n elements of this matrix located in different rows and different columns, that is, products of the form

$$a_{1\alpha_1} \; a_{2\alpha_2} \dots a_{n\alpha_n} \tag{2}$$

where the subscripts $\alpha_1, \alpha_2, \dots, \alpha_n$ constitute an arrangement of the numbers $1, 2, \dots, n$. The number of such products is equal to the number of different arrangements of n symbols, or $n!$. We consider all these products as terms of the future nth-order determinant associated with the matrix (1).

To determine the sign affixed to product (2) in the determinant, note that, using the subscripts of this product, we can form the permutation

$$\begin{pmatrix} 1 & 2 & \dots & n \\ \alpha_1 & \alpha_2 & \dots & \alpha_n \end{pmatrix} \tag{3}$$

where i goes into α_i if an element in the ith row and α_ith column of matrix (1) enters into the product (2). Examining expressions of determinants of second and third order, we note that the plus sign is affixed to the terms whose subscripts constitute an even permutation, and the minus sign to those terms with an odd permutation of subscripts. It is also natural to retain this regularity in the definition of a determinant of order n.

We thus arrive at the following definition: *the nth-order determinant* associated with matrix (1) is the algebraic sum of $n!$ terms which is constructed in the following fashion: the terms are all possible products of the n elements of the matrix taken one in each row and each column, the term having a plus sign if its subscripts form an even permutation, and a minus sign otherwise.

For the notation of the nth-order determinant associated with matrix (1) we will, as in the case of determinants of order two and

three, use the symbol

$$\begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} \tag{4}$$

Determinants of the nth order become determinants of order two and three, for $n = 2$ and $n = 3$; for $n = 1$, that is, for matrices consisting of a single element, the determinant is equal to that element. So far we do not know whether it is possible, for $n > 3$, to use the nth-order determinant for solving systems of linear equations. That will be shown in Sec. 7. It will be necessary first to subject the nth-order determinants to a detailed study and, in particular, it will be necessary to find procedures for evaluating them, since to compute a determinant directly (via its definition), even for n not very large, would be extremely complicated.

For the present let us establish some of the simpler properties of nth-order determinants that refer mainly to one of the two following problems: on the one hand, we are interested in the conditions under which a determinant is equal to zero, on the other, we will indicate certain matrix transformations which leave its determinant unchanged or result in readily perceivable alterations.

The *transpose operation* with respect to matrix (1) is a transformation of the matrix in which its rows become columns with the same subscripts; in other words, it is a transition from matrix (1) to the matrix

$$\begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \dots & \dots & \dots & \dots \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{pmatrix} \tag{5}$$

or we can say that a transposition is obtained by flipping matrix (1) over the principal diagonal. Accordingly, we say that the determinant

$$\begin{vmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \dots & \dots & \dots & \dots \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{vmatrix} \tag{6}$$

is obtained by taking the transpose of the determinant (4).

Property 1. *Taking the transpose does not change the determinant.* Indeed, every term of determinant (4) is of the form

$$a_{1\alpha_1} a_{2\alpha_2} \ldots a_{n\alpha_n} \tag{7}$$

where the second subscripts form an arrangement of the symbols 1, 2, . . ., n. However, all the factors of product (7) remain in different rows and different columns in determinant (6) as well; hence, (7) serves as a term of the transpose of the determinant too. The converse is also obviously true and for this reason the determinants (4) and (6) consist of the same terms. The sign of the term (7) in determinant (4) is determined by the parity of the permutation

$$\begin{pmatrix} 1 & 2 & \ldots & n \\ \alpha_1 & \alpha_2 & \ldots & \alpha_n \end{pmatrix} \tag{8}$$

In determinant (6) the first subscripts of the elements indicate the column, the second subscripts the row, and so term (7) in determinant (6) is associated with the permutation

$$\begin{pmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_n \\ 1 & 2 & \ldots & n \end{pmatrix} \tag{9}$$

In the general case, the permutations (8) and (9) are different but they obviously have the same parity and so term (7) has the same sign in both determinants. Thus the determinants (4) and (6) are sums of the same terms taken with the same signs, that is, they are equal.

From Property 1 it follows that any assertion about rows holds true for the columns of a determinant and conversely; in other words, in contrast to a matrix, *in a determinant the rows and columns are of equal status.* We will therefore formulate and prove Properties 2 to 9 only for the rows of a determinant; analogous properties for columns will not require special proof.

Property 2. *If one of the rows of a determinant consists of zeros, the determinant is zero.*

Indeed, let all the elements of the ith row of a determinant be zeros. Every term of the determinant must have, as a factor, one element of the ith row, and so in our case all the terms of the determinant are zero.

Property 3. *If a determinant is obtained from another one by interchanging two rows, then all terms of the first determinant will be terms of the second but with signs reversed; which means that interchanging two rows of a determinant only changes the sign.*

Suppose, in determinant (4), the ith and jth rows ($i \neq j$) are interchanged and all other rows remain fixed. We get the deter-

minant

$$\begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots \\ a_{j1} & a_{j2} & \dots & a_{jn} \\ \dots & \dots & \dots & \dots \\ a_{i1} & a_{i2} & \dots & a_{in} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} \begin{matrix} \\ \\ (i) \\ \\ (j) \\ \\ \\ \end{matrix} \tag{10}$$

(row numbers indicated on the right). If

$$a_{1\alpha_1} a_{2\alpha_2} \dots a_{n\alpha_n} \tag{11}$$

is a term of (4), then all its factors in (10) as well obviously remain in different rows and different columns. Thus, determinants (4) and (10) consist of the same terms. Term (11) in determinant (4) is associated with the permutation

$$\begin{pmatrix} 1 & 2 & \dots & i & \dots & j & \dots & n \\ \alpha_1 & \alpha_2 & \dots & \alpha_i & \dots & \alpha_j & \dots & \alpha_n \end{pmatrix} \tag{12}$$

and in determinant (10) with the permutation

$$\begin{pmatrix} 1 & 2 & \dots & j & \dots & i & \dots & n \\ \alpha_1 & \alpha_2 & \dots & \alpha_i & \dots & \alpha_j & \dots & \alpha_n \end{pmatrix} \tag{13}$$

since, for example, element $a_{i\alpha_i}$ now lies in the jth row but remains in the old α_ith column. The permutation (13) however is obtained from (12) via a single transposition in the upper row; it thus has opposite parity. Whence it follows that all terms of determinant (4) enter into determinant (10) with opposite signs. Determinants (4) and (10) differ in sign alone.

Property 4. *A determinant containing two identical rows is equal to zero.*

Indeed, let a determinant be equal to the number d and let the corresponding elements of its ith and jth rows ($i \neq j$) be equal. By Property 3, after an interchange of these two rows, the determinant will be equal to the number $-d$. But since identical rows are interchanged, the determinant does not actually change; thus, $d = -d$, whence $d = 0$.

Property 5. *If all the elements of some row of a determinant are multiples of some number k, then the determinant itself is a multiple of k.*

Let all elements of the ith row be multiplied by k. Each term of the determinant contains exactly one element of the ith row,

therefore every term acquires the factor k, which means the determinant itself is a multiple of k.

This property admits of the following formulation as well: *a common factor of all elements of some row of a determinant may be factored out of the determinant.*

Property 6. *A determinant with two proportional rows is equal to zero.*

Let the elements of the jth row of a determinant differ from the corresponding elements of the ith row ($i \neq j$) by one and the same factor k. Factoring this common factor k out of the jth row of the determinant, we obtain a determinant with two identical rows, which by Property 4 is zero.

Property 4 (and also Property 2 for $n > 1$) is obviously a special case of Property 6 (for $k = 1$ and $k = 0$).

Property 7. *If all the elements of the ith row of a determinant of order n are given as a sum of two terms:*

$$a_{ij} = b_j + c_j, \qquad j = 1, \ldots, n$$

then the determinant is equal to the sum of two determinants in which all rows (except the ith*) are the same as in the given determinant and the ith row in one of the summands consists of the elements* b_j *and in the other, of the elements* c_j.

Indeed, any term of the given determinant may be represented in the form

$$a_{1\alpha_1}a_{2\alpha_2}\ldots a_{i\alpha_i}\ldots a_{n\alpha_n} = a_{1\alpha_1}a_{2\alpha_2}\ldots(b_{\alpha_i}+c_{\alpha_i})\ldots a_{n\alpha_n}$$
$$= a_{1\alpha_1}a_{2\alpha_2}\ldots b_{\alpha_i}\ldots a_{n\alpha_n} + a_{1\alpha_1}a_{2\alpha_2}\ldots c_{\alpha_i}\ldots a_{n\alpha_n}$$

Collecting together the first summands of these sums (with the same signs as the corresponding terms had in the given determinant) we evidently obtain an nth-order determinant which differs from the given determinant solely in the fact that the ith row has elements b_j in place of elements a_{ij}. Accordingly, the second summands form a determinant in the ith row of which are the elements c_j. Thus

$$\begin{vmatrix} a_{11} & a_{12}\ldots a_{1n} \\ \ldots\ldots\ldots\ldots \\ b_1+c_1\, b_2+c_2\ldots b_n+c_n \\ \ldots\ldots\ldots\ldots \\ a_{n1} & a_{n2}\ldots a_{nn} \end{vmatrix} = \begin{vmatrix} a_{11}\ a_{12}\ \ldots\ a_{1n} \\ \ldots\ldots\ldots \\ b_1\ \ b_2\ \ \ldots\ \ b_n \\ \ldots\ldots\ldots \\ a_{n1}\ a_{n2}\ \ldots\ a_{nn} \end{vmatrix} + \begin{vmatrix} a_{11}\ a_{12}\ \ldots\ a_{1n} \\ \ldots\ldots\ldots \\ c_1\ \ c_2\ \ \ldots\ \ c_n \\ \ldots\ldots\ldots \\ a_{n1}\ a_{n2}\ \ldots\ a_{nn} \end{vmatrix}$$

Property 7 is readily extended to the case when any element of the ith row is a sum of m summands, not two, $m \geqslant 2$.

We shall say that the ith row of a determinant is a linear combination of the remaining rows if for every row with subscript j,

$j = 1, \ldots, i-1, i+1, \ldots, n$, there exists a number k_j such that when the jth row is multiplied by k_j and then all the rows except the ith are added together (addition of rows is to be understood in the sense that the elements of the row are added in each column separately), we obtain the ith row. Some of the coefficients k_j may be zero, that is the ith row will actually be a linear combination not of all but only of a few of the remaining rows. In particular, if only one of the coefficients k_j is different from zero, we get the case of proportionality of two rows. Finally, if the row consists entirely of zeros, it will always be a linear combination of the remaining rows—the case when all k_j are zero.

Property 8. *If one of the rows of a determinant is a linear combination of the other rows, then the determinant is zero.*

For example, let the ith row be a linear combination of s other rows, $1 \leqslant s \leqslant n-1$. Then every element of the ith row will be a sum of s summands, and for this reason, using Property 7, we can represent our determinant in the form of a sum of determinants in each of which the ith row will be proportional to one of the other rows. By Property 6, all these determinants are zero; hence the given determinant is zero as well.

This property is a generalization of Property 6 and, as will be proved in Sec. 10, it provides the most general case of a zero determinant.

Property 9. *A determinant remains unchanged if to the elements of one of its rows we add corresponding elements of another row multiplied by the same number.*

Suppose to the ith row of determinant d we add the jth row, $j \neq i$, multiplied by the number k; that is, in the new determinant every element of the ith row will be of the form $a_{is} + ka_{js}$, $s = 1, 2, \ldots, n$. Then, by Property 7, this determinant is equal to the sum of two determinants, the first of which is d and the second of which contains two proportional rows and is therefore zero.

Since the number k may also be negative, *the determinant does not change even if we subtract from one of its rows a row multiplied by some number.* Generally, *a determinant remains unchanged if to one of its rows we add any linear combination of the other rows.*

Let us consider an example. A determinant is called *skew-symmetric* if the elements symmetric about the principal diagonal differ in sign alone, that is, if for all i and j it is true that $a_{ji} = -a_{ij}$, whence it follows that for all i it is true that $a_{ii} = -a_{ii} = 0$. Thus, the determinant is of the form

$$d = \begin{vmatrix} 0 & a_{12} & a_{13} & \ldots & a_{1n} \\ -a_{12} & 0 & a_{23} & \ldots & a_{2n} \\ -a_{13} & -a_{23} & 0 & \ldots & a_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ -a_{1n} & -a_{2n} & -a_{3n} & \ldots & 0 \end{vmatrix}$$

Multiplying each row of this determinant by —1, we obtain the transpose of the determinant, which is again equal to d, whence, by Property 5, it follows that

$$(-1)^n d = d$$

It then follows, for odd n, that $-d = d$, or $d = 0$. Thus *any skew-symmetric determinant of odd order is equal to zero.*

5. Minors and Their Cofactors

We have already pointed out that it would be difficult to compute an nth-order determinant by applying the definition directly, that is every time writing out all $n!$ terms, determining their signs, etc. There are simpler methods for evaluating determinants. They are based on the fact that a determinant of order n may be expressed in terms of a determinant of lower order. For this purpose we introduce the following notion.

Let there be a determinant d of order n. Take an integer k which satisfies the condition $1 \leqslant k \leqslant n - 1$, and in the determinant d choose arbitrary k rows and k columns. The elements which lie at the intersection of these rows and columns, that is, which belong to one of the chosen rows and to one of the chosen columns will obviously form a matrix of order k. The determinant of this matrix is called a *minor of order k* of the determinant d. We can also say that the kth-order minor is a determinant obtained by striking out $n - k$ rows and $n - k$ columns in d. In particular, after striking out one row and one column in the determinant we obtain a minor of $(n - 1)$th order; on the other hand, separate elements of determinant d will be minors of the first order.

Let us take a minor M of order k in a determinant d of order n. If we strike out the rows and columns at the intersection of which this minor stands, we obtain the minor M' of order $(n - k)$ which is called the *complementary minor* of the minor M. If, on the contrary, we strike out the rows and columns which contain elements of the minor M', then what remains is obviously minor M. Thus, we can speak of a *pair of complementary minors* of the determinant. In particular, the element a_{ij} and the minor of order $(n - 1)$ obtained by striking out the ith row and the jth column in the determinant will form a pair of complementary minors.

If a kth-order minor M is located in rows with the position numbers (indices) $i_1, i_2, \ldots, i_k$ and in columns with the position numbers $j_1, j_2, \ldots, j_k$, then we use the term *cofactor* of the minor M for the supplementary minor M' taken with a plus or minus sign according as the sum of the position numbers of all rows and columns in which M is located is even or odd, that is, the sum

$$s_M = i_1 + i_2 + \ldots + i_k + j_1 + j_2 + \ldots + j_k \qquad (1)$$

In other words, the cofactor of M is the number $(-1)^{s_M} M'$.

The product of any minor M of order k by its cofactor in a determinant d is the algebraic sum, whose summands, which are obtained by multiplying the terms of the minor M by the terms of the supplementary minor M' taken with the sign $(-1)^{s_M}$, are certain terms of the determinant d; their signs in this sum coincide with the signs they have in the determinant.

We begin the proof of this theorem with the case when the minor M is located in the upper left corner of the determinant:

$$d = \left|\begin{array}{ccc|ccc} a_{11} & \dots & a_{1k} & a_{1,k+1} & \dots & a_{1n} \\ \dots & M & \dots & \dots & \dots & \dots \\ a_{k1} & \dots & a_{kk} & a_{k,k+1} & \dots & a_{kn} \\ \hline a_{k+1,1} & \dots & a_{k+1,k} & a_{k+1,k+1} & \dots & a_{k+1,n} \\ \dots & \dots & \dots & \dots & M' & \dots \\ a_{n1} & \dots & a_{nk} & a_{n,k+1} & \dots & a_{nn} \end{array}\right|$$

that is, in rows with position numbers $1, 2, \dots, k$ and in columns with the same position numbers. Then the minor M' will occupy the lower right corner of the determinant. The number s_M will then be even:

$$s_M = 1+2+\dots+k+1+2+\dots+k = 2(1+2+\dots+k)$$

therefore, the minor M' itself will serve as the cofactor of M.

Take an arbitrary term

$$a_{1\alpha_1} a_{2\alpha_2} \cdots a_{k\alpha_k} \tag{2}$$

of minor M; its sign in M is $(-1)^l$ if l is the number of inversions in the permutation

$$\begin{pmatrix} 1 & 2 & \dots & k \\ \alpha_1 & \alpha_2 & \dots & \alpha_k \end{pmatrix} \tag{3}$$

In this minor, the arbitrary term

$$a_{k+1,\beta_{k+1}} a_{k+2,\beta_{k+2}} \cdots a_{n\beta_n} \tag{4}$$

of minor M' has the sign $(-1)^{l'}$ where l' is the number of inversions in the permutation

$$\begin{pmatrix} k+1 & k+2 & \dots & n \\ \beta_{k+1} & \beta_{k+2} & \dots & \beta_n \end{pmatrix} \tag{5}$$

Multiplying the terms (2) and (4), we obtain a product of n elements

$$a_{1\alpha_1} a_{2\alpha_2} \cdots a_{k\alpha_k} a_{k+1,\beta_{k+1}} a_{k+2,\beta_{k+2}} \cdots a_{n\beta_n} \tag{6}$$

located in different rows and different columns of the determinant. It is therefore a term of determinant d. The sign of term (6) in the

product MM' is a product of the signs of terms (2) and (4), i.e., $(-1)^l \cdot (-1)^{l'} = (-1)^{l+l'}$. However, term (6) has the same sign in the determinant d as well. Indeed, the lower row of the permutation

$$\begin{pmatrix} 1 & 2 & \dots & k & k+1 & k+2 & \dots & n \\ \alpha_1 & \alpha_2 & \dots & \alpha_k & \beta_{k+1} & \beta_{k+2} & \dots & \beta_n \end{pmatrix}$$

made up of the subscripts of this term contains only $l + l'$ inversions, since no α can form an inversion with any one of the β; all α do not exceed k, all β are not less than $k + 1$.

This proves the particular case of the theorem that we have considered. Let us now take up the general case. Suppose that the minor M lies in the rows with position numbers $i_1, i_2, \dots, i_k$ and in the columns with position numbers $j_1, j_2, \dots, j_k$, with the condition that

$$i_1 < i_2 < \dots < i_k, \quad j_1 < j_2 < \dots < j_k$$

Let us attempt, by interchanging rows and columns of the determinant, to move the minor M to the upper left corner and let us try to do this so that the complementary minor is not changed. For this purpose, interchange the i_1th row with the $(i_1 - 1)$th, then with the $(i_1 - 2)$th and so on until the i_1th row occupies the first row; this requires interchanging the rows $i_1 - 1$ times. Then we successively interchange the i_2th row with rows located above it until it lies directly under the i_1th row (that is, in the position of the original second row); this, as can readily be verified, will require interchanging the rows $i_2 - 2$ times. Similarly, we move the i_3th row to the third row, and so on, until the i_kth row takes up the position of the kth row. In all, we will have to perform

$$(i_1 - 1) + (i_2 - 2) + \dots + (i_k - k) = (i_1 + i_2 + \dots + i_k) - (1 + 2 + \dots + k)$$

transpositions of rows.

The minor M is thus located in the first k rows of the new determinant. We will now successively interchange the columns of the determinant, the j_1th column with all preceding ones, until it occupies first place, then the j_2th column until it occupies second place, and so forth. In all, the columns will be interchanged

$$(j_1 + j_2 + \dots + j_k) - (1 + 2 + \dots + k)$$

times.

All these transformations lead us to a new determinant d' in which the minor M occupies the upper left corner. Since each time we interchanged only adjacent rows or columns, the mutual positions of the rows and columns containing the minor M' in the determinant d remain without change, and so the minor M' remains complementary to the minor M in the determinant d'; however, it now occupies the

lower right corner. As was proved above, the product MM' is the sum of some number of terms of the determinant d' taken with the same signs as they had in d'. However, the determinant d' is obtained from the determinant d by means of

$$[(i_1 + i_2 + \ldots + i_k) - (1 + 2 + \ldots + k)] + [(j_1 + j_2 + \ldots + j_k) - (1 + 2 + \ldots + k)] = s_M - 2(1 + 2 + \ldots + k)$$

transpositions of rows and columns, and so, as we know from Sec. 4, the terms of determinant d' differ from the corresponding terms of determinant d in sign alone, $(-1)^{s_M}$ [naturally, the even number $2(1 + 2 + \ldots + k)$ will not affect the sign]. From this it follows that the product $(-1)^{s_M} MM'$ consists of a certain number of terms of the determinant d taken with the same signs as they have in that determinant. The theorem is proved.

Note that if the minors M and M' are complementary, then the numbers s_M and $s_{M'}$ have the same parity. Indeed, the position number of any row and any column enters as a summand in one and only one of these numbers, and therefore the sum $s_M + s_{M'}$ is equal to the total sum of the position numbers of all rows and columns of the determinant, i.e., it is equal to the even number $2(1 + 2 + \ldots + n)$.

6. Evaluating Determinants

The results of the preceding section enable us to reduce computing an nth-order determinant to the computation of several determinants of order $(n - 1)$. Let us first introduce notation: if a_{ij} is an element of determinant d, then M_{ij} denotes the complementary minor, or, simply, *the minor of that element*, that is, the minor of order $(n - 1)$ obtained by striking out the ith row and the jth column of the determinant. A_{ij} will denote the cofactor of the element a_{ij}; thus,

$$A_{ij} = (-1)^{i+j} M_{ij}$$

As was proved in the preceding section, the product $a_{ij}A_{ij}$ is the sum of several terms of the determinant d which enter into this sum with the same signs as they have in the determinant d. It is easy to count these terms: the number is equal to the number of terms in the minor M_{ij}, or $(n - 1)!$.

Let us now choose any ith row of the determinant d and take the product of each element of the row by its cofactor:

$$a_{i1}A_{i1},\ a_{i2}A_{i2},\ \ldots,\ a_{in}A_{in} \qquad (1)$$

No term of the determinant d can be in two different products of those given in (1): all the terms of the determinant which enter into the product $a_{i1}A_{i1}$ contain the element a_{i1} of the ith row and

for this reason differ from the terms which enter into the product $a_{i2}A_{i2}$, that is, those which contain the element a_{i2} of the ith row, and so on.

On the other hand, the total number of terms of determinant d which appear in all the products of (1) is equal to

$$(n-1)!\cdot n = n!$$

Generally, this exhausts all the terms of the determinant d. We have thus proved that there is *an expansion of the determinant d in terms of the ith row*:

$$d = a_{i1}A_{i1} + a_{i2}A_{i2} + \ldots + a_{in}A_{in} \qquad (2)$$

The determinant d is thus equal to the sum of the products of all the elements of an arbitrary row by their cofactors. A similar expansion of the determinant can also be obtained about any column.

By replacing the cofactors in the expansion (2) by corresponding minors with a plus or a minus sign, we reduce computation of an nth-order determinant to the computation of several determinants of order $(n-1)$. Note that if some of the elements of the ith row are zero, then naturally the corresponding minors need not be evaluated. It is therefore useful, first, to transform the determinant, using Property 9 (see Sec. 4), so that a large enough number of elements in one of the rows or in one of the columns are replaced by zeros. Actually, *Property 9 enables us to replace all elements, except one, by zeros in any row or any column.* Indeed, if $a_{ik} \neq 0$, then any element a_{ij}, $j \neq k$, of the ith row will be replaced by a zero after subtracting the kth column multiplied by $\frac{a_{ij}}{a_{ik}}$ from the jth column. Thus, evaluating a determinant of the nth order may be reduced to computing a single determinant of order $(n-1)$.

Example 1. Evaluate the fourth-order determinant

$$d = \begin{vmatrix} 3 & 1 & -1 & 2 \\ -5 & 1 & 3 & -4 \\ 2 & 0 & 1 & -1 \\ 1 & -5 & 3 & -3 \end{vmatrix}$$

Expand it about the third row by using the zero in that row:

$$d = (-1)^{3+1}\cdot 2\cdot \begin{vmatrix} 1 & -1 & 2 \\ 1 & 3 & -4 \\ -5 & 3 & -3 \end{vmatrix} + (-1)^{3+3}\cdot 1\cdot \begin{vmatrix} 3 & 1 & 2 \\ -5 & 1 & -4 \\ 1 & -5 & -3 \end{vmatrix} + (-1)^{3+4}\cdot(-1)\cdot \begin{vmatrix} 3 & 1 & -1 \\ -5 & 1 & 3 \\ 1 & -5 & 3 \end{vmatrix}$$

Evaluating the third-order determinants thus obtained, we get

$$d = 2\cdot 16 - 40 + 48 = 40$$

Example 2. Evaluate the fifth-order determinant

$$d = \begin{vmatrix} -2 & 5 & 0 & -1 & 3 \\ 1 & 0 & 3 & 7 & -2 \\ 3 & -1 & 0 & 5 & -5 \\ 2 & 6 & -4 & 1 & 2 \\ 0 & -3 & -1 & 2 & 3 \end{vmatrix}$$

Adding three times the fifth row to the second and subtracting four times the fifth row from the fourth row, we get

$$d = \begin{vmatrix} -2 & 5 & 0 & -1 & 3 \\ 1 & -9 & 0 & 13 & 7 \\ 3 & -1 & 0 & 5 & -5 \\ 2 & 18 & 0 & -7 & -10 \\ 0 & -3 & -1 & 2 & 3 \end{vmatrix}$$

Expanding this determinant in terms of the third column, which contains only one nonzero element (with the sum of subscripts, $5 + 3$, being even), we get

$$d = (-1) \cdot \begin{vmatrix} -2 & 5 & -1 & 3 \\ 1 & -9 & 13 & 7 \\ 3 & -1 & 5 & -5 \\ 2 & 18 & -7 & -10 \end{vmatrix}$$

We now transform this determinant by adding two times the second row to the first row and subtracting three times the second from the third row, and two times the second from the fourth:

$$d = - \begin{vmatrix} 0 & -13 & 25 & 17 \\ 1 & -9 & 13 & 7 \\ 0 & 26 & -34 & -26 \\ 0 & 36 & -33 & -24 \end{vmatrix}$$

and then expand it in terms of the first column. Noting that the only nonzero element of this column is associated with an odd sum of subscripts, we get

$$d = \begin{vmatrix} -13 & 25 & 17 \\ 26 & -34 & -26 \\ 36 & -33 & -24 \end{vmatrix}$$

Let us compute this third-order determinant after expanding it in terms of the third row:

$$d = 36 \cdot \begin{vmatrix} 25 & 17 \\ -34 & -26 \end{vmatrix} - (-33) \cdot \begin{vmatrix} -13 & 17 \\ 26 & -26 \end{vmatrix} + (-24) \cdot \begin{vmatrix} -13 & 25 \\ 26 & -34 \end{vmatrix}$$

$$= 36 \cdot (-72) - (-33) \cdot (-104) + (-24) \cdot (-208) = -1032$$

Example 3. *If all the elements of a determinant located on one side of the principal diagonal are equal to zero, then the determinant is equal to the product of the elements on the principal diagonal.*

This assertion is obvious for a second-order determinant. We therefore prove it by induction, that is, we assume that for determinants of order $(n - 1)$

it has been proved, and then we consider the nth-order determinant

$$d = \begin{vmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22} & a_{23} & \dots & a_{2n} \\ 0 & 0 & a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & a_{nn} \end{vmatrix}$$

Expanding it in terms of the first column, we get

$$d = a_{11} \cdot \begin{vmatrix} a_{22} & a_{23} & \dots & a_{2n} \\ 0 & a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{vmatrix}$$

But the induction hypothesis is applicable to the minor on the right-hand side: it is equal to the product $a_{22}a_{33} \dots a_{nn}$ and so

$$d = a_{11}a_{22}a_{33} \dots a_{nn}$$

Example 4. The *Vandermonde determinant* is the determinant

$$d = \begin{vmatrix} 1 & 1 & 1 & \dots & 1 \\ a_1 & a_2 & a_3 & \dots & a_n \\ a_1^2 & a_2^2 & a_3^2 & \dots & a_n^2 \\ \dots & \dots & \dots & \dots & \dots \\ a_1^{n-1} & a_2^{n-1} & a_3^{n-1} & \dots & a_n^{n-1} \end{vmatrix}$$

We shall prove that *for any n the Vandermonde determinant is equal to the product of all possible differences* $a_i - a_j$, *where* $1 \leqslant j < i \leqslant n$. Indeed, for $n = 2$ we have

$$\begin{vmatrix} 1 & 1 \\ a_1 & a_2 \end{vmatrix} = a_2 - a_1$$

Suppose our assertion has already been proved for Vandermonde determinants of order $(n - 1)$. We transform determinant d as follows: subtract from the nth (last) row the $(n - 1)$th row multiplied by a_1, then from the $(n - 1)$th row subtract the $(n - 2)$th also multiplied by a_1, etc. Finally, from the second row subtract the first multiplied by a_1. We obtain

$$d = \begin{vmatrix} 1 & 1 & 1 & \dots & 1 \\ 0 & a_2 - a_1 & a_3 - a_1 & \dots & a_n - a_1 \\ 0 & a_2^2 - a_1a_2 & a_3^2 - a_1a_3 & \dots & a_n^2 - a_1a_n \\ \dots & \dots & \dots & \dots & \dots \\ 0 & a_2^{n-1} - a_1a_2^{n-2} & a_3^{n-1} - a_1a_3^{n-2} & \dots & a_n^{n-1} - a_1a_n^{n-2} \end{vmatrix}$$

Expanding this determinant in terms of the first column, we arrive at a determinant of order $(n - 1)$; after factoring out common factors from all columns,

it will take the form

$$d=(a_2-a_1)(a_3-a_1)\dots(a_n-a_1)\cdot\begin{vmatrix} 1 & 1 & \dots & 1 \\ a_2 & a_3 & \dots & a_n \\ a_2^2 & a_3^2 & \dots & a_n^2 \\ \dots & \dots & \dots & \dots \\ a_2^{n-2} & a_3^{n-2} & \dots & a_n^{n-2} \end{vmatrix}$$

The last factor is the Vandermonde determinant of order $(n-1)$, that is, by hypothesis, it is equal to the product of all the differences $a_i - a_j$ for $2 \leqslant j < i \leqslant n$. Using the symbol Π to denote a product, we can write

$$d=(a_2-a_1)(a_3-a_1)\dots(a_n-a_1)\prod_{2\leqslant j<i\leqslant n}(a_i-a_j)=\prod_{1\leqslant j<i\leqslant n}(a_i-a_j)$$

Using the same method, we can prove that the determinant

$$d'=\begin{vmatrix} a_1^{n-1} & a_2^{n-1} & a_3^{n-1} & \dots & a_n^{n-1} \\ \dots & \dots & \dots & \dots & \dots \\ a_1^2 & a_2^2 & a_3^2 & \dots & a_n^2 \\ a_1 & a_2 & a_3 & \dots & a_n \\ 1 & 1 & 1 & \dots & 1 \end{vmatrix}$$

is equal to the product of all possible differences $a_i - a_j$, where $1 \leqslant i < j \leqslant n$, that is,

$$d'=\prod_{1\leqslant i<j\leqslant n}(a_i-a_j)$$

Generalizing the above-obtained expansions of a determinant about a row or a column, we prove the following theorem which has to do with *the expansion of a determinant in terms of several rows or columns.*

Laplace's theorem. Let there be arbitrarily chosen, in a determinant d of order n, k rows (or k columns), $1 \leqslant k \leqslant n-1$. Then the sum of the products of all kth-order minors contained in the chosen rows by their cofactors is equal to the determinant d.

Proof. Suppose, in determinant d, we choose rows with position numbers $i_1, i_2, \dots, i_k$. We know that the product of any minor M of order k located in these rows by its cofactor consists of a certain number of terms of the determinant d taken with the signs they have in the determinant. The theorem will consequently be proved if we demonstrate that by making M run through all kth-order minors located in the chosen rows we obtain all the terms of the determinant, none being repeated.

Let

$$a_{1\alpha_1}a_{2\alpha_2}\dots a_{n\alpha_n} \tag{3}$$

be an arbitrary term of the determinant d. We separately take the product of those elements of the term which belong to the rows we have chosen with position numbers $i_1, i_2, \dots, i_k$. This is the

product

$$a_{i_1\alpha_{i_1}}a_{i_2\alpha_{i_2}}\ldots a_{i_k\alpha_{i_k}} \quad (4)$$

The k factors of this product lie in k distinct columns, namely, in the columns with position numbers $\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_k}$. These position numbers of the columns are consequently determined by specifying the term (3). If by M we denote the kth-order minor lying at the intersection of the columns with these position numbers $\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_k}$ and of the earlier chosen rows with the position numbers $i_1, i_2, \ldots, i_k$, then the product (4) is one of the terms of the minor M, and the product of all the elements of the term (3) not in (4) is a term of its complementary minor. Thus, any term of the determinant enters into the product of a certain (quite definite) minor of order k made up of the chosen rows multiplied by its complementary minor, and is a product of quite definite terms of these two minors. Finally, in order to obtain the term that we took of the determinant with the sign which it has in the determinant, it remains, as we know, to replace the complementary minor by the cofactor. This completes the proof of the theorem.

It is possible to give a slightly different proof, namely, the product of any kth-order minor M located in the chosen rows by its cofactor consists of $k!\ (n-k)!$ terms, since the kth-order minor M consists of $k!$ terms and its cofactor, differing possibly from the minor of order $n-k$ in sign alone, contains $(n-k)!$ terms. On the other hand, the number of kth-order minors contained in the chosen rows is equal to the number of combinations of n taken k at a time, that is, it is equal to the number

$$\frac{n!}{k!\ (n-k)!}$$

Multiplying out, we find that the sum of the products of all kth-order minors of the chosen rows by their cofactors consists of $n!$ summands. Such, however, is the total number of terms of the determinant d. The theorem will thus be proved if we demonstrate that any term of the determinant d appears at least once (and, in that case, exactly once) in the sum at hand of the products of the minors by their cofactors. It is left to the reader to repeat (with slight simplifications) the reasoning given in the first proof.

The Laplace theorem enables one to reduce the computation of an nth-order determinant to the computation of several determinants of orders k and $n-k$. Generally speaking, there are very large number of such new determinants and so it is advisable to apply the Laplace theorem only when it is possible to choose k rows (or columns) in the determinant so that many of the kth-order minors located in these rows are zero.

Example 1. Suppose we have a determinant, all elements of which in the first k rows and the last $n-k$ columns are zero:

$$d=\begin{vmatrix} a_{11} & \dots & a_{1k} & & & \\ \dots & \dots & \dots & & \mathbf{0} & \\ a_{k1} & \dots & a_{kk} & & & \\ a_{k+1,1} & \dots & a_{k+1,k} & a_{k+1,k+1} & \dots & a_{k+1,n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{nk} & a_{n,k+1} & \dots & a_{nn} \end{vmatrix}$$

This determinant is then equal to the product of two of its minors:

$$d=\begin{vmatrix} a_{11} & \dots & a_{1k} \\ \dots & \dots & \dots \\ a_{k1} & \dots & a_{kk} \end{vmatrix} \cdot \begin{vmatrix} a_{k+1,k+1} & \dots & a_{k+1,n} \\ \dots & \dots & \dots \\ a_{n,k+1} & \dots & a_{nn} \end{vmatrix}$$

To prove this, it suffices to expand the determinant about the first k rows.

Example 2. Suppose we have a determinant d of order $2n$, in the upper left corner of which is an nth-order minor composed entirely of zeros. If the nth-order minors lying in the upper right and lower left and lower right corners of the determinant are denoted, respectively, by M, M' and M'', so that the determinant may be written symbolically as $d=\begin{vmatrix} 0 & M \\ M' & M'' \end{vmatrix}$, then $d=$ $=(-1)^n MM'$.

To prove this, expand the determinant in terms of the first n rows and note that

$$s_M=(1+2+\dots+n)+[(n+1)+(n+2)+\dots+2n]=n+2n^2$$

that is, s_M and n are of the same parity.

Example 3. Evaluate the determinant

$$d=\begin{vmatrix} -4 & 1 & 2 & -2 & 1 \\ 0 & 3 & 0 & 1 & -5 \\ 2 & -3 & 1 & -3 & 1 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 4 & 0 & 2 & 5 \end{vmatrix}$$

Expanding it about the first and third columns which contain nicely located zeros, we get

$$d=(-1)^{1+3+1+3}\begin{vmatrix} -4 & 2 \\ 2 & 1 \end{vmatrix}\cdot\begin{vmatrix} 3 & 1 & -5 \\ -1 & -1 & 0 \\ 4 & 2 & 5 \end{vmatrix}$$

$$+(-1)^{1+4+1+3}\begin{vmatrix} -4 & 2 \\ -1 & 3 \end{vmatrix}\cdot\begin{vmatrix} 3 & 1 & -5 \\ -3 & -3 & 1 \\ 4 & 2 & 5 \end{vmatrix}$$

$$+(-1)^{3+4+1+3}\begin{vmatrix} 2 & 1 \\ -1 & 3 \end{vmatrix}\cdot\begin{vmatrix} 1 & -2 & 1 \\ 3 & 1 & -5 \\ 4 & 2 & 5 \end{vmatrix}$$

$$=(-8)\cdot(-20)-(-10)\cdot(-62)-7\cdot 87=-1069$$

7. Cramer's Rule

The foregoing theory of determinants of order n allows us to show that these determinants, which were introduced only by analogy with second- and third-order determinants, may, like the latter, be used to solve systems of linear equations. Let us first make one additional remark regarding expansions of determinants in terms of a row or a column; this remark will often come in handy in the sequel.

Expand the determinant

$$d = \begin{vmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ a_{21} & \dots & a_{2j} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{nj} & \dots & a_{nn} \end{vmatrix}$$

about the jth column:

$$d = a_{1j}A_{1j} + a_{2j}A_{2j} + \dots + a_{nj}A_{nj}$$

Then, in this expansion, replace the elements of the jth column by a set of n arbitrary numbers $b_1, b_2, \dots, b_n$. The expression

$$b_1A_{1j} + b_2A_{2j} + \dots + b_nA_{nj}$$

which you obtain will obviously serve as an expansion about the jth column for the determinant

$$d' = \begin{vmatrix} a_{11} & \dots & b_1 & \dots & a_{1n} \\ a_{21} & \dots & b_2 & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & b_n & \dots & a_{nn} \end{vmatrix}$$

which is obtained from the determinant d by replacing its jth column by a column of the numbers $b_1, b_2, \dots, b_n$. Indeed, replacing the jth column of d does not affect the minors of the elements of the column, and for this reason does not affect their cofactors.

Let us apply this to the case when for the numbers $b_1, b_2, \dots, b_n$ we take elements of the kth column of the determinant d when $k \neq j$. The determinant resulting from such a replacement will contain two identical columns (jth and kth) and therefore will be zero. Hence, the expansion of this determinant about its jth column will also be zero, that is

$$a_{1k}A_{1j} + a_{2k}A_{2j} + \dots + a_{nk}A_{nj} = 0 \quad \text{for} \quad j \neq k$$

Thus, *the sum of the products of all elements of a certain column of a determinant by the cofactors of the corresponding elements of*

another column is zero. The same result of course holds true for the rows of a determinant.

Let us now examine systems of linear equations; we will confine ourselves for the time being to systems in which *the number of equations is equal to the number of unknowns*, i.e., systems of the form

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{aligned}\right\} \qquad (1)$$

We also assume that the determinant d made up of the coefficients of the unknowns of the system (called, for short, the *determinant of the system*) is nonzero. Given these assumptions, we will prove that the system (1) is consistent and even determinate.

In Sec. 2, when we solved a system of three equations in three unknowns, we multiplied each of the equations by a factor, and then added the equations; the coefficients of two of the unknowns proved to be zero. We now see immediately that the factors which we used were cofactors, in the determinant of the system, of the element which was the coefficient of the desired unknown in the given equation. We now use this device to solve system (1).

First suppose that system (1) is consistent and $\alpha_1, \alpha_2, \ldots, \alpha_n$ is one of its solutions. Hence, the following equations hold true:

$$\left.\begin{aligned} a_{11}\alpha_1 + a_{12}\alpha_2 + \ldots + a_{1n}\alpha_n &= b_1, \\ a_{21}\alpha_1 + a_{22}\alpha_2 + \ldots + a_{2n}\alpha_n &= b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n1}\alpha_1 + a_{n2}\alpha_2 + \ldots + a_{nn}\alpha_n &= b_n \end{aligned}\right\} \qquad (2)$$

Let j be any one of the numbers $1, 2, \ldots, n$. Multiply both sides of the first equation of (2) by A_{1j}, that is, by the cofactor of the element a_{1j} in the determinant d of the system. Multiply both sides of the second equation by A_{2j}, and so on. Finally, multiply both sides of the last equation by A_{nj}. Adding together separately the left and right sides of all equations, we arrive at the following equation:

$$\begin{aligned} &(a_{11}A_{1j} + a_{21}A_{2j} + \ldots + a_{n1}A_{nj})\,\alpha_1 \\ &\quad + (a_{12}A_{1j} + a_{22}A_{2j} + \ldots + a_{n2}A_{nj})\,\alpha_2 \\ &\quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\quad + (a_{1j}A_{1j} + a_{2j}A_{2j} + \ldots + a_{nj}A_{nj})\,\alpha_j \\ &\quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\quad + (a_{1n}A_{1j} + a_{2n}A_{2j} + \ldots + a_{nn}A_{nj})\,\alpha_n \\ &\qquad = b_1A_{1j} + b_2A_{2j} + \ldots + b_nA_{nj} \end{aligned}$$

The coefficient of α_j in this equation is d, the coefficients of all other α will, due to the remark made above, be zero, and the constant term will be the determinant obtained from the determinant d after replacing the jth column in it by a column of the constant terms of system (1). If, as in Sec. 2, we denote this latter determinant by d_j, then our equation takes the form

$$d\alpha_j = d_j$$

whence, because $d \neq 0$,

$$\alpha_j = \frac{d_j}{d}$$

This proves that if system (1) is consistent, then it possesses the unique solution

$$\alpha_1 = \frac{d_1}{d}, \quad \alpha_2 = \frac{d_2}{d}, \quad \ldots, \quad \alpha_n = \frac{d_n}{d} \tag{3}$$

We will now show that the set (3) of numbers actually satisfies system (1) of equations, that is, that (1) is consistent. We will make use of the following commonly employed symbolism.

Any sum of the form $a_1 + a_2 + \ldots + a_n$ will be denoted briefly by $\sum_{i=1}^{n} a_i$. But if we consider a sum whose terms a_{ij} are labelled with two subscripts, and $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, then we can first take the sums of the elements with fixed first subscript, that is, the sums $\sum_{j=1}^{m} a_{ij}$, where $i = 1, 2, \ldots, n$, and then add all the sums. We then obtain the following notation for the sum of all elements a_{ij}:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}$$

However, we could first add the summands a_{ij} with fixed second subscript and then combine the resulting sums. Thus

$$\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} = \sum_{j=1}^{m} \sum_{i=1}^{n} a_{ij}$$

i.e., *in a double sum the order of summation may be reversed.*

Now put the values of the unknowns (3) into the ith equation of system (1). Since the left side of the ith equation may be written as $\sum_{j=1}^{n} a_{ij}x_j$ and since $d_j = \sum_{k=1}^{n} b_k A_{kj}$, we get

$$\sum_{j=1}^{n} a_{ij} \cdot \frac{d_j}{d} = \frac{1}{d} \sum_{j=1}^{n} a_{ij} \left(\sum_{k=1}^{n} b_k A_{kj} \right) = \frac{1}{d} \sum_{k=1}^{n} b_k \left(\sum_{j=1}^{n} a_{ij} A_{kj} \right)$$

With regard to these manipulations, note that the number $\frac{1}{d}$ turned out to be a common factor in all summands and was therefore taken outside the summation sign; besides, after changing the order of summation, the factor b_k was factored out of the inner sum since it is not dependent on the subscript j of the inner summation.

We know that the expression $\sum_{j=1}^{n} a_{ij}A_{kj} = a_{i1}A_{k1} + a_{i2}A_{k2} + \ldots$ $\ldots + a_{in}A_{kn}$ will be equal to d for $k = i$ and to 0 for all other k's. Thus, in our outer sum with respect to k there will be only one summand left, namely, $b_i d$; i.e.,

$$\sum_{j=1}^{n} a_{ij} \cdot \frac{d_j}{d} = \frac{1}{d} \cdot b_i d = b_i$$

This is proof that the set (3) of numbers is indeed a solution to the system (1) of equations.

We have obtained the following important result.

A system of n linear equations in n unknowns, the determinant of which is nonzero, has a unique solution. This solution is obtained from formulas (3), that is by means of *Cramer's rule*. The formulation of this rule is the same as in the case of a system of two equations (see Sec. 2).

Example. Solve the system of linear equations

$$\left.\begin{aligned} 2x_1 + x_2 - 5x_3 + x_4 &= 8, \\ x_1 - 3x_2 - 6x_4 &= 9, \\ 2x_2 - x_3 + 2x_4 &= -5, \\ x_1 + 4x_2 - 7x_3 + 6x_4 &= 0 \end{aligned}\right|$$

The determinant of the system is different from zero:

$$d = \begin{vmatrix} 2 & 1 & -5 & 1 \\ 1 & -3 & 0 & -6 \\ 0 & 2 & -1 & 2 \\ 1 & 4 & -7 & 6 \end{vmatrix} = 27$$

and so Cramer's rule is applicable. The values of the unknowns will have as numerators the determinants

$$d_1 = \begin{vmatrix} 8 & 1 & -5 & 1 \\ 9 & -3 & 0 & -6 \\ -5 & 2 & -1 & 2 \\ 0 & 4 & -7 & 6 \end{vmatrix} = 81, \quad d_2 = \begin{vmatrix} 2 & 8 & -5 & 1 \\ 1 & 9 & 0 & -6 \\ 0 & -5 & -1 & 2 \\ 1 & 0 & -7 & 6 \end{vmatrix} = -108,$$

$$d_3 = \begin{vmatrix} 2 & 1 & 8 & 1 \\ 1 & -3 & 9 & -6 \\ 0 & 2 & -5 & 2 \\ 1 & 4 & 0 & 6 \end{vmatrix} = -27, \quad d_4 = \begin{vmatrix} 2 & 1 & -5 & 8 \\ 1 & -3 & 0 & 9 \\ 0 & 2 & -1 & -5 \\ 1 & 4 & -7 & 0 \end{vmatrix} = 27$$

Thus,

$$x_1 = 3, \quad x_2 = -4, \quad x_3 = -1, \quad x_4 = 1$$

will be the unique solution set of our system.

We did not consider the case when the determinant of a system of n linear equations in n unknowns (1) is zero. It will be discussed in Chapter 2, where it will find its place in the general theory of systems involving any number of equations in any number of unknowns.

One more remark is in order with respect to systems of n linear equations in n unknowns. Given a system of n homogeneous linear equations in n unknowns (see Sec. 1):

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= 0, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= 0 \end{aligned}\right\} \tag{4}$$

In this case, all determinants d_j, $j = 1, 2, \ldots, n$, contain a column made up of zeros and are therefore equal to zero. Thus, if the determinant of system (4) is nonzero, that is if Cramer's rule is applicable, then the only solution of system (4) will be the trivial solution

$$x_1 = 0, \quad x_2 = 0, \ \ldots, \ x_n = 0 \tag{5}$$

Whence follows the result:

If a system of n homogeneous linear equations in n unknowns has nontrivial solutions, then the determinant of the system is necessarily zero.

In Sec. 12 it will also be shown that, conversely, if the determinant of such a system is indeed equal to zero, then the system will have solutions other than the trivial solution, the existence of which is obvious for every system of homogeneous equations.

Example. For what values of k can the system of equations

$$\left.\begin{aligned} kx_1 + x_2 &= 0, \\ x_1 + kx_2 &= 0 \end{aligned}\right\}$$

have nontrivial solutions?

The determinant of this system

$$\begin{vmatrix} k & 1 \\ 1 & k \end{vmatrix} = k^2 - 1$$

will be zero only when $k = \pm 1$. It is easy to see that for each one of these two values of k the given system will indeed have nontrivial solutions.

The significance of Cramer's rule lies mainly in the fact that for cases when it is applicable it offers an explicit expression of

the solution of the system in terms of the coefficients of the system. However, Cramer's rule involves very unwieldy computations; in the case of a system of n linear equations in n unknowns, one has to compute $n + 1$ determinants of the nth order. The method of successive elimination of unknowns given in Sec. 1 is much more convenient in this respect since the computations involved here are actually equivalent to those required in the evaluation of a single determinant of the nth order.

In applications, we often encounter systems of linear equations whose coefficients and constant terms are real numbers obtained in measurements of physical quantities and as such are known only approximately, to within a specified accuracy. The foregoing methods are then sometimes rather inconvenient because they lead to results with poor accuracy. A variety of *iterative procedures* have taken their place. These are methods which yield solutions of systems of equations via successive approximations of the unknowns. The interested reader will find such methods described in texts dealing with the theory of approximate calculations.

CHAPTER 2

SYSTEMS OF LINEAR EQUATIONS (GENERAL THEORY)

8. n-Dimensional Vector Space

To construct a general theory of systems of linear equations we will need more than the apparatus that sufficed with such success in the solution of systems to which Cramer's rule was applicable. Besides determinants and matrices we will need a new concept, which, perhaps, is of still greater general mathematical interest—that of *multidimensional vector spaces.*

First a few preliminary remarks. From the course of analytic geometry we know that any point in a plane is determined (for specified coordinate axes) by its two coordinates, which is to say, by an ordered set of two real numbers. Any vector in a plane is determined by its two components, which again is an ordered set of two real numbers. Similarly, a point in three-dimensional space is determined by three coordinates, a vector in space, by three components.

In geometry and also in mechanics and physics we often encounter objects whose specification requires more than three real numbers. For instance, let us consider a collection of spheres in three-dimensional space. To specify a sphere completely we need the coordinates of its centre and the radius; this amounts to an ordered set of four real numbers, of which, incidentally, the radius can only assume positive values. On the other hand, let us consider various positions of a solid in space. The position of a solid will be fully defined if we indicate the coordinates of its centre of gravity (this requires three real numbers), the direction of some fixed axis passing through the centre of gravity (two numbers—two out of three direction cosines), and, finally, the angle of rotation about this axis. Thus, the position of a solid body in space is determined by an ordered set of six real numbers.

These examples suggest considering collections of all possible ordered sets of n real numbers. After introducing the operations of addition and multiplication by a scalar (this will be done later

on by analogy with appropriate operations involving vectors in three-dimensional space expressed in terms of components), we call this collection an n-dimensional vector space. Thus, n-dimensional space is only an algebraic structure which retains certain of the simplest properties of collections of vectors of three-dimensional space emanating from a coordinate origin.

An ordered set of n numbers (an ordered n-tuple)

$$\alpha = (a_1, a_2, \ldots, a_n) \tag{1}$$

is called an *n-dimensional vector*. The numbers a_i, $i = 1, 2, \ldots, n$, will be called the *components* of the vector α. The vectors α and

$$\beta = (b_1, b_2, \ldots, b_n) \tag{2}$$

will be considered *equal* if their components, in the same places, coincide, that is, if $a_i = b_i$, $i = 1, 2, \ldots, n$. Lower-case Greek letters will be used to denote vectors and lower-case Latin letters to denote scalars.

Examples of vectors are: (1) Vector segments (directed line-segments) emanating from the coordinate origin in a plane or in three-dimensional space will, given a fixed system of coordinates, be two- and three-dimensional vectors in the meaning of the definition given above. (2) The coefficients of a linear equation in n unknowns constitute an n-dimensional vector. (3) Any solution of a system of linear equations in n unknowns is an n-dimensional vector. (4) If an s by n matrix is given (s rows and n columns), then its rows are n-dimensional vectors, its columns, s-dimensional vectors. (5) The s by n matrix itself can be regarded as an sn-dimensional vector: all we need to do is read the elements of the matrix one after the other, row by row; in particular, any square matrix of order n may be regarded as an n^2-dimensional vector, and it is quite obvious that any n^2-dimensional vector may be obtained in this way from a matrix of order n.

The *sum* of vectors (1) and (2) is the vector

$$\alpha + \beta = (a_1 + b_1, a_2 + b_2, \ldots, a_n + b_n) \tag{3}$$

whose components are sums of the corresponding components of the vectors being added. Addition of vectors is commutative and associative because of the commutativity and associativity of the addition of numbers.

The role of zero is played by the *zero vector*:

$$0 = (0, 0, \ldots, 0) \tag{4}$$

Indeed,

$$\alpha + 0 = (a_1 + 0, a_2 + 0, \ldots, a_n + 0)$$
$$= (a_1, a_2, \ldots, a_n) = \alpha$$

We use the same symbol 0 for the zero vector as for the number 0. There is never any difficulty in deciding whether it is the number zero or the zero vector we are talking about at any time. However, from now on the reader should bear in mind the possibility of different interpretations of the symbol 0.

We use the term *opposite vector* (negative) of the vector (1) for the vector

$$-\alpha = (-a_1, -a_2, \ldots, -a_n) \tag{5}$$

It is obvious that $\alpha + (-\alpha) = 0$. It is now easy to see that for the addition of vectors there is an inverse operation—subtraction: the *difference* between the vectors (1) and (2) is the vector $\alpha - \beta = \alpha + (-\beta)$, or

$$\alpha - \beta = (a_1 - b_1, a_2 - b_2, \ldots, a_n - b_n) \tag{6}$$

The addition of n-dimensional vectors defined by formula (3) arose out of the geometric addition of vectors in the plane or in three-dimensional space performed by the parallelogram rule. In geometry we have to do with the multiplication of a vector by a real number ("scalar"): the multiplication of a vector α by a scalar k signifies, for $k > 0$, a stretching of α by a factor k (it is compression if $k < 1$), and for $k < 0$ a stretching by a factor $|k|$ and reversal of direction. Expressing this rule in terms of the components of the vector α and passing to the general case at hand, we obtain the following definition.

The product of a vector (1) *by a scalar* k is the vector

$$k\alpha = \alpha k = (ka_1, ka_2, \ldots, ka_n) \tag{7}$$

whose components are equal to the product of the corresponding components of the vector α by k.

From this definition there follow important properties which may be verified by the reader:

$$k(\alpha \pm \beta) = k\alpha \pm k\beta, \tag{8}$$

$$(k \pm l)\alpha = k\alpha \pm l\alpha, \tag{9}$$

$$k(l\alpha) = (kl)\alpha, \tag{10}$$

$$1 \cdot \alpha = \alpha \tag{11}$$

The following properties are just as easy to verify but they may also be obtained as corollaries to Properties (8)-(11):

$$0 \cdot \alpha = 0, \tag{12}$$

$$(-1) \cdot \alpha = -\alpha, \tag{13}$$

$$k \cdot 0 = 0, \tag{14}$$

if $k\alpha = 0$, then either $k = 0$, or $\alpha = 0$. (15)

The collection of all n-dimensional vectors with real components regarded in conjunction with the operations of addition of vectors and multiplication of a vector by a scalar is called an *n-dimensional vector space.*

Note that the definition of an n-dimensional vector space does not include multiplication of a vector by a vector. It would be easy to define multiplication of vectors—assume, say, that the components of a product of vectors are equal to the products of the corresponding components of the factors. However, such multiplication would not find any serious applications. Thus, vector segments emanating from a coordinate origin in the plane or in three-dimensional space constitute (for a fixed system of coordinates) a two-dimensional and, respectively, a three-dimensional vector space. The addition of vectors and the multiplication of a vector by a scalar are, as we have pointed out above, geometrically important, whereas it is impossible to give any reasonable geometrical interpretation to the componentwise multiplication of vectors.

Let us consider another example. The left side of a linear equation in n unknowns, that is, an expression of the form

$$f = a_1x_1 + a_2x_2 + \ldots + a_nx_n$$

is called a *linear form* in the unknowns $x_1, x_2, \ldots, x_n$. The linear form f is obviously defined completely by the vector $(a_1, a_2, \ldots, a_n)$ of its coefficients; conversely, any n-dimensional vector uniquely determines some linear form. The addition of vectors and the multiplication of a vector by a scalar become corresponding operations involving linear forms; these operations were extensively used in Sec. 1. Componentwise multiplication of vectors in this instance is meaningless.

9. Linear Dependence of Vectors

A vector β of n-dimensional vector space is *proportional* to vector α if there exists a number k such that $\beta = k\alpha$ [see formula (7) of the preceding section]. In particular, the zero vector is proportional to any vector α due to the equality $0 = 0 \cdot \alpha$. But if $\beta = k\alpha$ and $\beta \neq 0$, whence $k \neq 0$, then $\alpha = k^{-1}\beta$, that is, for nonzero vectors, proportionality possesses the property of symmetry.

A generalization of the concept of proportionality of vectors is the following concept which we have already (in the case of rows in a matrix) encountered in Sec. 4; a vector β is called a *linear combination* of the vectors $\alpha_1, \alpha_2, \ldots, \alpha_s$ if there exist numbers $l_1, l_2, \ldots, l_s$ such that

$$\beta = l_1\alpha_1 + l_2\alpha_2 + \ldots + l_s\alpha_s$$

Thus the jth component of the vector β, $j = 1, 2, \ldots, n$, is equal (because of the definition of a sum of vectors and a product of a vector by a scalar) to the sum of the products of the jth components of the vectors $\alpha_1, \alpha_2, \ldots, \alpha_s$, by $l_1, l_2, \ldots, l_s$, respectively.

A system of vectors

$$\alpha_1, \alpha_2, \ldots, \alpha_{r-1}, \alpha_r \qquad (r \geqslant 2) \tag{1}$$

is *linearly dependent* if at least one of the vectors is a linear combination of the remaining vectors of the system; it is called *linearly independent* otherwise.

We give another form of this extremely important definition: a system of vectors (1) is linearly dependent if there exist numbers $k_1, k_2, \ldots, k_r$, *at least one of which is nonzero*, such that the equation

$$k_1\alpha_1 + k_2\alpha_2 + \ldots + k_r\alpha_r = 0 \tag{2}$$

holds true.

Proof of the equivalence of these two definitions is not difficult. For example, let the vector α_r of system (1) be a linear combination of the remaining vectors:

$$\alpha_r = l_1\alpha_1 + l_2\alpha_2 + \ldots + l_{r-1}\alpha_{r-1}$$

From this there follows the equation

$$l_1\alpha_1 + l_2\alpha_2 + \ldots + l_{r-1}\alpha_{r-1} - \alpha_r = 0$$

which is like (2), where $k_i = l_i$ for $i = 1, 2, \ldots, r-1$ and $k_r = -1$ that is $k_r \neq 0$. Conversely, let the vectors (1) be connected by the relation (2) in which, say, $k_r \neq 0$. Then

$$\alpha_r = \left(-\frac{k_1}{k_r}\right)\alpha_1 + \left(-\frac{k_2}{k_r}\right)\alpha_2 + \ldots + \left(-\frac{k_{r-1}}{k_r}\right)\alpha_{r-1}$$

Vector α_r has proved to be a linear combination of the vectors $\alpha_1, \alpha_2, \ldots, \alpha_{r-1}$.

Example. The system of vectors

$$\alpha_1 = (5, 2, 1), \quad \alpha_2 = (-1, 3, 3), \quad \alpha_3 = (9, 7, 5), \quad \alpha_4 = (3, 8, 7)$$

is linearly dependent, since the vectors are connected by the relation

$$4\alpha_1 - \alpha_2 - 3\alpha_3 + 2\alpha_4 = 0$$

In this relation all the coefficients are different from zero. However, there are other linear dependences between the vectors, dependences in which some of the coefficients are zero, for instance

$$2\alpha_1 + \alpha_2 - \alpha_3 = 0, \qquad 3\alpha_2 + \alpha_3 - 2\alpha_4 = 0$$

The latter definition of a linear dependence given above is also applicable to the case of $r = 1$, that is, to the case of a system consisting of one vector α: *this system is linearly dependent if and only if* $\alpha = 0$. Indeed, if $\alpha = 0$, then, say, for $k = 1$ we will have $k\alpha = 0$. Conversely, if $k\alpha = 0$ and $k \neq 0$, then $\alpha = 0$.

Note the following property of the concept of linear dependence.

If some subsystem of the system (1) *of vectors is linearly dependent, then the whole system* (1) *is linearly dependent.*

Indeed, let the vectors $\alpha_1, \alpha_2, \ldots, \alpha_s$ of system (1), where $s < r$, be connected by the relation

$$k_1\alpha_1 + k_2\alpha_2 + \ldots + k_s\alpha_s = 0$$

in which not all coefficients are zero. It then follows that the relation

$$k_1\alpha_1 + k_2\alpha_2 + \ldots + k_s\alpha_s + 0\cdot\alpha_{s+1} + \ldots + 0\cdot\alpha_r = 0$$

or system (1) is linearly dependent.

From this property follows the linear dependence of any system of vectors containing two equal or, generally, two proportional vectors and also of any system containing the zero vector. The property we have just proved can also be stated as follows: *if a system* (1) *of vectors is linearly independent, then any subsystem of* (1) *is also linearly independent.*

The question arises as to how many vectors a linearly independent system of n-dimensional vectors can contain and, in particular, whether there exist systems with an arbitrarily large number of vectors. To answer this question, let us consider the following vectors in an n-dimensional vector space:

$$\left.\begin{aligned} \varepsilon_1 &= (1, 0, 0, \ldots, 0), \\ \varepsilon_2 &= (0, 1, 0, \ldots, 0), \\ &\ldots\ldots\ldots\ldots \\ \varepsilon_n &= (0, 0, 0, \ldots, 1) \end{aligned}\right\} \qquad (3)$$

They are called *unit vectors* of that space. *The system of unit vectors will be linearly independent*: let

$$k_1\varepsilon_1 + k_2\varepsilon_2 + \ldots + k_n\varepsilon_n = 0$$

Since the left side of this equation is equal to the vector $(k_1, k_2, \ldots, k_n)$, it follows that

$$(k_1, k_2, \ldots, k_n) = 0$$

or $k_i = 0$, $i = 1, 2, \ldots, n$, since all the components of the zero vector are zero, and equality of vectors is equivalent to equality of their corresponding components.

Thus, in n-dimensional vector space we have found one linearly independent system consisting of n vectors. The reader will learn later on that there actually exist an infinite number of distinct systems of that kind in this space.

On the other hand, let us prove the following theorem.

For $s > n$, any s vectors of an n-dimensional vector space constitute a linearly dependent system.

Let there be given the vectors

$$\begin{aligned} \alpha_1 &= (a_{11}, a_{12}, \ldots, a_{1n}), \\ \alpha_2 &= (a_{21}, a_{22}, \ldots, a_{2n}), \\ &\ldots\ldots\ldots\ldots \\ \alpha_s &= (a_{s1}, a_{s2}, \ldots, a_{sn}) \end{aligned}$$

We have to choose scalars $k_1, k_2, \ldots, k_s$, not all zero, such that

$$k_1\alpha_1 + k_2\alpha_2 + \ldots + k_s\alpha_s = 0 \qquad (4)$$

Passing from (4) to the corresponding equations between the components, we get

$$\left.\begin{aligned} a_{11}k_1 + a_{21}k_2 + \ldots + a_{s1}k_s &= 0, \\ a_{12}k_1 + a_{22}k_2 + \ldots + a_{s2}k_s &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots&\ldots \\ a_{1n}k_1 + a_{2n}k_2 + \ldots + a_{sn}k_s &= 0 \end{aligned}\right\} \qquad (5)$$

However, equations (5) constitute a system of n homogeneous linear equations in s unknowns $k_1, k_2, \ldots, k_s$. The number of equations in this system is less than the number of unknowns, and therefore, as proved at the end of Sec. 1, the system has nontrivial solutions. We can thus choose scalars $k_1, k_2, \ldots, k_s$ (not all zero) which will satisfy requirement (4). The theorem is proved.

Let us call a linearly independent system of n-dimensional vectors

$$\alpha_1, \alpha_2, \ldots, \alpha_r \qquad (6)$$

a *maximal* linearly independent system if by adjoining to this system any n-dimensional vector β we obtain a linearly dependent system. Since in every linear dependence relating the vectors $\alpha_1, \alpha_2, \ldots, \alpha_r, \beta$, the coefficient of β must be nonzero—otherwise system (6) would be linearly dependent—it follows that the vector β is expressed linearly in terms of the vectors (6). Therefore the system (6) of vectors is a maximal linearly independent system if and only if the vectors (6) are linearly independent and any n-dimensional vector β is a linear combination of them.

From the results obtained above it follows that *in an n-dimensional space any linearly independent system consisting of n vectors is maximal* and also that *any maximal linearly independent system of vectors of this space consists of at most n vectors.*

Every linearly independent system of n-dimensional vectors is contained in at least one maximal linearly independent system. Indeed, if a given system of vectors is not maximal, then one vector may be added to it so that the resulting system remains linearly independent. If this new system is still not maximal, then another vector

may be added to it, and so on. However, this process cannot continue endlessly because every system of n-dimensional vectors consisting of $n + 1$ vectors is linearly dependent.

Since every system consisting of one nonzero vector is linearly independent, we find that *any nonzero vector is contained in some maximal linearly independent system*, and for this reason *there are infinitely many different maximal linearly independent systems of vectors in an n-dimensional vector space.*

The question arises: do there exist, in this space, maximal linearly independent systems with a smaller number of vectors than n or is the number of vectors in any such system invariably equal to n? The answer to this important question will be given below after a few preliminary investigations.

If vector β is a linear combination of the vectors

$$\alpha_1, \alpha_2, \ldots, \alpha_r \tag{7}$$

it is often said that β *is expressed linearly in terms of system* (7). Naturally, if vector β is linearly expressed in terms of some subsystem of this system, then it will be linearly expressed in terms of (7) as well—it would be sufficient to take the remaining vectors of the system with coefficients equal to zero. Generalizing this terminology, we say that *the system of vectors*

$$\beta_1, \beta_2, \ldots, \beta_s \tag{8}$$

is expressed linearly in terms of system (7) if every vector β_i, $i = 1, 2, \ldots, s$, is a linear combination of the vectors of (7).

We prove the transitivity of this concept: *if system* (8) *is expressed linearly in terms of* (7), *and the system of vectors*

$$\gamma_1, \gamma_2, \ldots, \gamma_t \tag{9}$$

is expressed linearly in terms of (8), *then* (9) *is expressed linearly in terms of* (7) *as well.*

Indeed,

$$\gamma_j = \sum_{i=1}^{s} l_{ji}\beta_i, \quad j = 1, 2, \ldots, t \tag{10}$$

but $\beta_i = \sum_{m=1}^{r} k_{im}\alpha_m$, $i = 1, 2, \ldots, s$. Substituting these expressions into (10), we get

$$\gamma_j = \sum_{i=1}^{s} l_{ji} \left(\sum_{m=1}^{r} k_{im}\alpha_m \right) = \sum_{m=1}^{r} \left(\sum_{i=1}^{s} l_{ji}k_{im} \right) \alpha_m$$

In other words, every vector γ_j, $j = 1, 2, \ldots, t$, is a linear combination of vectors of system (7).

Two systems of vectors are termed *equivalent* if each one of them can be expressed linearly in terms of the other. From the proof given here of the transitivity of the property of systems of vectors to be expressed linearly in terms of each other there follows the transitivity of the concept of equivalence of systems of vectors and also the following assertion: *if two systems of vectors are equivalent and if some vector is expressed linearly in terms of one of these systems, then it will be expressed linearly in terms of the other too.*

One cannot assert that if one of two equivalent systems of vectors is linearly independent, then the other system also possesses this property. But if both systems are linearly independent, then an important statement can be made with respect to the number of vectors entering into them. First let us prove the following theorem which, because of the role it will play in the future, it will be convenient to term a **fundamental theorem.**

If in an n-dimensional vector space we have two systems of vectors:

$$\text{(I)} \qquad \alpha_1,\ \alpha_2,\ \ldots,\ \alpha_r,$$

$$\text{(II)} \qquad \beta_1,\ \beta_2,\ \ldots,\ \beta_s$$

the first being linearly independent and expressible linearly in terms of the second, then the number of vectors in the first system does not exceed that in the second system, or $r \leqslant s$.

Let $r > s$. By hypothesis, each vector of system (I) can be expressed linearly in terms of system (II):

$$\left.\begin{aligned} \alpha_1 &= a_{11}\beta_1 + a_{12}\beta_2 + \ldots + a_{1s}\beta_s, \\ \alpha_2 &= a_{21}\beta_1 + a_{22}\beta_2 + \ldots + a_{2s}\beta_s, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_r &= a_{r1}\beta_1 + a_{r2}\beta_2 + \ldots + a_{rs}\beta_s \end{aligned}\right\} \qquad (11)$$

The coefficients of these linear expressions constitute a system of r s-dimensional vectors:

$$\begin{aligned} \gamma_1 &= (a_{11},\ a_{12},\ \ldots,\ a_{1s}), \\ \gamma_2 &= (a_{21},\ a_{22},\ \ldots,\ a_{2s}), \\ &\ldots\ldots\ldots\ldots\ldots \\ \gamma_r &= (a_{r1},\ a_{r2},\ \ldots,\ a_{rs}) \end{aligned}$$

Since $r > s$, these vectors are linearly dependent, that is,

$$k_1\gamma_1 + k_2\gamma_2 + \ldots + k_r\gamma_r = 0$$

where not all coefficients $k_1, k_2, \ldots, k_r$ are zero. Whence we arrive at certain equations between the components:

$$\sum_{i=1}^{r} k_i a_{ij} = 0, \qquad j = 1,\ 2,\ \ldots,\ s \qquad (12)$$

Let us now consider the following linear combination of vectors of system (1):

$$k_1\alpha_1 + k_2\alpha_2 + \ldots + k_r\alpha_r$$

or, more compactly, $\sum_{i=1}^{r} k_i\alpha_i$. Utilizing (11) and (12), we get

$$\sum_{i=1}^{r} k_i\alpha_i = \sum_{i=1}^{r} k_i \left(\sum_{j=1}^{s} a_{ij}\beta_j\right) = \sum_{j=1}^{s} \left(\sum_{i=1}^{r} k_i a_{ij}\right) \beta_j = 0$$

But this runs counter to the linear independence of system (1).

From the fundamental theorem just proved we have the following result.

Any two equivalent linearly independent systems of vectors contain an equal number of vectors.

Any two maximal linearly independent systems of n-dimensional vectors are evidently equivalent. They therefore consist of one and the same number of vectors, and since (as we know) there exist systems of that kind consisting of n vectors, we finally get the answer to the earlier posed question: *every maximal linearly independent system of vectors of an n-dimensional vector space consists of n vectors.*

Some corollaries follow.

If in a given linearly dependent system of vectors we take two maximal linearly independent subsystems, that is, subsystems to which no vector of our system can be adjoined without spoiling the linear independence, then these subsystems contain an equal number of vectors.

Indeed, if in the system of vectors

$$\alpha_1, \alpha_2, \ldots, \alpha_r \tag{13}$$

the subsystem

$$\alpha_1, \alpha_2, \ldots, \alpha_s, \qquad s < r \tag{14}$$

is a maximal linearly independent subsystem, then any one of the vectors $\alpha_{s+1}, \ldots, \alpha_r$ is expressible linearly in terms of system (14). On the other hand, any vector α_i of system (13) is linearly expressible in terms of this system: it is only necessary to take the coefficient 1 for the vector α_i, and the coefficient 0 for all the other vectors. It is now easy to see that systems (13) and (14) are equivalent. From this it follows that (13) is equivalent to any one of its maximal linearly independent subsystems, and therefore all the subsystems are equivalent; i.e., being linearly independent, they contain the same number of vectors each.

The number of vectors in any maximal linearly independent subsystem of a given system of vectors is termed the *rank* of the system. Taking advantage of this concept, we derive yet another corollary from the fundamental theorem.

Suppose there are two systems of n-dimensional vectors:

$$\alpha_1, \alpha_2, \ldots, \alpha_r \tag{15}$$

and

$$\beta_1, \beta_2, \ldots, \beta_s \tag{16}$$

which are not necessarily linearly independent; the rank of system (15) *is equal to the number k, the rank of system* (16), *to the number l. If the first system is expressed linearly in terms of the second, then* $k \leqslant l$. *But if these systems are equivalent, then* $k = l$.

In fact, let

$$\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_k} \tag{17}$$

and

$$\beta_{j_1}, \beta_{j_2}, \ldots, \beta_{j_l} \tag{18}$$

be, respectively, any maximal linearly independent subsystems of (15) and (16). Then systems (15) and (17) are equivalent and the same holds true for (16) and (18). From the fact that (15) is linearly expressible in terms of (16) it now follows that (17) is also linearly expressible in terms of (16) and therefore in terms of the equivalent system (18). It then remains, utilizing the linear independence of system (17), to apply the fundamental theorem. The second assertion of the corollary being proved follows directly from the first.

10. Rank of a Matrix

If we are given a system of n-dimensional vectors, it is natural to ask whether this system of vectors is linearly dependent or not. One cannot hope to find that in every specific instance the question will be resolved without difficulty: a superficial examination of the system of vectors

$$\alpha = (2, -5, 1, -1), \ \beta = (1, 3, 6, 5), \ \gamma = (-1, 4, 1, 2)$$

fails to reveal any linear dependences in it, though in reality these vectors are connected by the relation

$$7\alpha - 3\beta + 11\gamma = 0$$

One way of settling this issue is given in Sec. 1. Since the components of the given vectors are known, we consider as unknown the coefficients of the desired linear dependence and obtain a system of homogeneous linear equations, which we solve by the Gaussian method. In this section we suggest a different approach, which will also bring us closer to our principal objective—the solution of arbitrary systems of linear equations.

Suppose we have an s by n matrix (s rows and n columns)

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{s1} & a_{s2} & \dots & a_{sn} \end{pmatrix}$$

the numbers s and n not being related in any way. Regarded as s-dimensional vectors, the columns of this matrix may, generally speaking, be linearly dependent. The rank of the system of columns, that is the maximal number of linearly independent columns of matrix A (more precisely, the number of columns in any maximal linearly independent subsystem of the system of columns) is called the *rank* of the matrix.

Naturally, in the same way the rows of matrix A may be regarded as n-dimensional vectors. It appears that the rank of the system of rows of the matrix is equal to the rank of the system of its columns, that is, it is equal to the rank of the matrix. The proof of this extremely unexpected assertion will be obtained after we point out yet another way of defining the rank of a matrix (which at the same time indicates a practical method of evaluation).

Let us first generalize the concept of a minor to the case of rectangular matrices. In matrix A we choose arbitrary k rows and k columns, $k \leqslant \min(s, n)$. The elements at the intersection of these rows and columns constitute a square matrix of order k, the determinant of which is called the *kth-order minor* of matrix A. We will now be interested in the orders of those minors of A which differ from zero, namely, *the highest one of these orders.* In searching for it, it is well to bear in mind the following: *if all kth-order minors of matrix A are zero, then so also are all minors of higher order.* Indeed, expanding any minor of order $k + j$, $k < k + j \leqslant \min(s, n)$, by the Laplace theorem in terms of any k rows, we represent this minor as a sum of minors of order k multiplied by certain minors of order j, thus proving that it is zero.

Let us now prove the following **theorem on the rank of a matrix.**

The highest order of nonzero minors of matrix A is equal to the rank of the matrix.

Proof. Let the highest order of nonzero minors of matrix A be r. Let us assume—there is no loss of generality— that the rth-order minor D in the upper left corner of the matrix

$$A = \begin{pmatrix} \boxed{\begin{matrix} a_{11} & \dots & a_{1r} \\ \dots & D & \dots \\ a_{r1} & \dots & a_{rr} \end{matrix}} & \begin{matrix} a_{1,\,r+1} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{r,\,r+1} & \dots & a_{rn} \end{matrix} \\ \begin{matrix} a_{r+1,\,1} & \dots & a_{r+1,\,r} \\ \dots & \dots & \dots \\ a_{s1} & \dots & a_{sr} \end{matrix} & \begin{matrix} a_{r+1,\,r+1} & \dots & a_{r+1,\,n} \\ \dots & \dots & \dots \\ a_{s,\,r+1} & \dots & a_{sn} \end{matrix} \end{pmatrix}$$

is different from zero, $D \neq 0$. Then the first r columns of A will be linearly independent: if the dependence were linear, then, since corresponding components are combined in the addition of vectors, this same linear dependence would exist among the columns of minor D and therefore D would be zero.

Now let us prove that each lth column of A, $r < l \leqslant n$, is a linear combination of the first r columns. We take any i, $1 \leqslant i \leqslant s$, and construct an auxiliary determinant of order $(r+1)$:

$$\Delta_i = \begin{vmatrix} a_{11} & \ldots & a_{1r} & a_{1l} \\ \cdot & \cdot & \cdot & \cdot \\ a_{r1} & \ldots & a_{rr} & a_{rl} \\ a_{i1} & \ldots & a_{ir} & a_{il} \end{vmatrix}$$

obtained by "bordering" the minor D by appropriate elements of the lth column and the ith row. Determinant Δ_i is zero for any i. Indeed, if $i > r$, then Δ_i is a minor of order $(r+1)$ of our matrix A and therefore is zero due to the choice of the number r. But if $i \leqslant r$, then Δ_i can no longer be a minor of matrix A since it cannot be obtained by deleting from this matrix certain of its rows and columns; however, determinant Δ_i now has two equal rows and, hence, is again zero.

Let us examine the cofactors of the elements of the last row of determinant Δ_i. Obviously, the cofactor of the element a_{il} is minor D. But if $1 \leqslant j \leqslant r$, then for the cofactor of element a_{ij} in Δ_i we have the number

$$A_j = (-1)^{(r+1)+j} \begin{vmatrix} a_{11} & \ldots & a_{1,\,j-1} & a_{1,\,j+1} & \ldots & a_{1r} & a_{1l} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{r1} & \ldots & a_{r,\,j-1} & a_{r,\,j+1} & \ldots & a_{rr} & a_{rl} \end{vmatrix}$$

It is not dependent on i and therefore is denoted by A_j. Thus, expanding determinant Δ_i about its last row and equating this expansion to zero, since $\Delta_i = 0$, we get

$$a_{i1}A_1 + a_{i2}A_2 + \ldots + a_{ir}A_r + a_{il}D = 0$$

whence, because $D \neq 0$,

$$a_{il} = -\frac{A_1}{D}a_{i1} - \frac{A_2}{D}a_{i2} - \ldots - \frac{A_r}{D}a_{ir}$$

This equation holds true for all i, $i = 1, 2, \ldots, s$, and since its coefficients are not dependent on i, we find that the entire lth column of A is a sum of the first r columns taken, respectively, with the coefficients $-\frac{A_1}{D}, -\frac{A_2}{D}, \ldots, -\frac{A_r}{D}$.

In the system of columns of matrix A we have thus found a maximal linearly independent subsystem consisting of r columns. This is proof that the rank of matrix A is equal to r, and it completes the proof of the rank theorem.

This theorem provides a practical method for computing the rank of a matrix and therefore for settling the question of the existence of linear dependence in a given system of vectors; forming a matrix for which the given vectors serve as columns and computing the rank of the matrix, we find the maximum number of linearly independent vectors of our system.

The method of finding the rank of a matrix based on the rank theorem requires computing a finite but perhaps very large number of minors of the matrix. The following remark suggests a way of substantially simplifying this procedure. If the reader will again look through the proof of the rank theorem, he will notice that in the proof we did not take advantage of the fact that *all* minors of order $(r + 1)$ of matrix A are equal to zero; actually, we used only those minors of order $(r + 1)$ which border the given nonzero rth-order minor D (that is, those which contain it completely within themselves); for this reason, from the fact that only these minors are equal to zero it follows that r is the maximum number of linearly independent columns of matrix A; this implies that all minors of order $(r + 1)$ of this matrix are zero. We arrive at the following **rule for evaluating the rank of a matrix.**

In computing the rank of a matrix, move from minors of smaller order to minors of greater order. If a nonzero kth-order minor D has already been found, then only the $(k + 1)$th-order minors bordering minor D need be computed; if they are all zero, the rank of the matrix is k.

Example 1. Find the rank of the matrix

$$A = \begin{pmatrix} 2 & -4 & 3 & 1 & 0 \\ 1 & -2 & 1 & -4 & 2 \\ 0 & 1 & -1 & 3 & 1 \\ 4 & -7 & 4 & -4 & 5 \end{pmatrix}$$

The second-order minor in the upper left corner of this matrix is zero. However, the matrix also contains nonzero minors of order two, for instance,

$$d = \begin{vmatrix} -4 & 3 \\ -2 & 1 \end{vmatrix} \neq 0$$

The third-order minor

$$d' = \begin{vmatrix} 2 & -4 & 3 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{vmatrix}$$

bordering minor d is different from zero, $d' = 1$, but both fourth-order minors bordering minor d' are zero:

$$\begin{vmatrix} 2 & -4 & 3 & 1 \\ 1 & -2 & 1 & -4 \\ 0 & 1 & -1 & 3 \\ 4 & -7 & 4 & -4 \end{vmatrix} = 0, \quad \begin{vmatrix} 2 & -4 & 3 & 0 \\ 1 & -2 & 1 & 2 \\ 0 & 1 & -1 & 1 \\ 4 & -7 & 4 & 5 \end{vmatrix} = 0$$

Thus, the rank of matrix A is three.

Example 2. Find the maximal linearly independent subsystem in the system of vectors

$$\alpha_1 = (2, -2, -4), \quad \alpha_2 = (1, 9, 3), \quad \alpha_3 = (-2, -4, 1), \quad \alpha_4 = (3, 7, -1)$$

Form the matrix

$$\begin{pmatrix} 2 & 1 & -2 & 3 \\ -2 & 9 & -4 & 7 \\ -4 & 3 & 1 & -1 \end{pmatrix}$$

in which the given vectors are columns. The rank of this matrix is two: the second-order minor in the upper left corner is nonzero, but both third-order minors bordering it are zero. From this it follows that the vectors α_1, α_2 form in the given system one of maximal linearly independent subsystems.

As a corollary to the rank theorem, we now prove an assertion that was stated earlier.

The maximum number of linearly independent rows of any matrix is equal to the maximum number of its linearly independent columns, which means that it is equal to the rank of the matrix.

To prove this, take the transpose of the matrix (that is, interchange rows and columns retaining the subscripts of the elements). In taking the transpose, the maximal order of nonzero minors of the matrix cannot change since taking transposes does not change the determinant, and for any minor of the original matrix the minor obtained from it by taking the transpose is in the new matrix, and conversely. Whence it follows that the rank of the new matrix is equal to the rank of the original matrix; it is also equal to the maximum number of linearly independent columns of the new matrix (or the maximum number of linearly independent *rows* of the original matrix).

Example. In Sec. 8 we introduced the concept of a linear form in n unknowns and defined addition of linear forms and their multiplication by a scalar. This definition permits extending to linear forms the concept of linear dependence with all its properties.

Let there be a system of linear forms

$$\begin{aligned} f_1 &= x_1 + 2x_2 + x_3 + 3x_4, \\ f_2 &= 4x_1 - x_2 - 5x_3 - 6x_4, \\ f_3 &= x_1 - 3x_2 - 4x_3 - 7x_4, \\ f_4 &= 2x_1 + x_2 - x_3 \end{aligned}$$

In it we have to choose a maximal linearly independent subsystem.

Form the matrix of the coefficients of these forms:

$$\begin{pmatrix} 1 & 2 & 1 & 3 \\ 4 & -1 & -5 & -6 \\ 1 & -3 & -4 & -7 \\ 2 & 1 & -1 & 0 \end{pmatrix}$$

and find its rank. The second-order minor in the upper left corner is nonzero, but, as can easily be verified, all four third-order minors bordering it are zero. Whence it follows that the first two rows of our matrix are linearly independent, and the third and fourth are linear combinations of them. Hence, the system f_1, f_2 is the desired subsystem of the given system of linear forms.

There is yet another important consequence of the rank theorem.

An nth-order determinant is equal to zero if and only if there is a linear dependence among its rows.

This assertion has already been proved in one direction in Sec. 4 (Property 8). Now let there be given an nth-order determinant equal to zero; in other words, suppose we have a square matrix of order n whose only minor having maximal order is zero. It then follows that the highest order of the nonzero minors of this matrix is less than n, that is, the rank is less than n, and so, on the basis of the foregoing proof, the rows of this matrix are linearly dependent.

Quite naturally, this corollary can be stated with columns taken instead of rows.

There is yet another way to compute the rank of a matrix which is not connected with the rank theorem and does not require evaluating determinants. Incidentally, it is only applicable when we wish to know only the rank itself and are not interested in precisely which columns (or rows) comprise the maximal linearly independent system. The procedure is this.

We use the term *elementary transformations* of a matrix A for the following transformations:

(a) interchange (transposition) of two rows or two columns;

(b) multiplication of a row (or a column) by an arbitrary nonzero scalar;

(c) addition of a multiple of one row (or column) to another row (column).

Clearly, *elementary transformations do not change the rank of a matrix.* Indeed, if these transformations are applied, say, to the columns of a matrix, the system of columns (regarded as vectors) is replaced by an equivalent system. We prove it for transformation (c) since for (a) and (b) it is obvious. Let the jth column multiplied by a number k be added to the ith column. If, prior to the manipulation, the vectors

$$\alpha_1, \ldots, \alpha_i, \ldots, \alpha_j, \ldots, \alpha_n \qquad (1)$$

served as columns of the matrix, then after the manipulation the vectors

$$\alpha_1, \ldots, \alpha_i' = \alpha_i + k\alpha_j, \ldots, \alpha_j, \ldots, \alpha_n \qquad (2)$$

will form the columns of the matrix. System (2) is expressible linearly in terms of system (1), and the equation

$$\alpha_i = \alpha_i' - k\alpha_j$$

shows that (1), in turn, is linearly expressible in terms of (2). Consequently, these systems are equivalent and for this reason their maximal linearly independent subsystems consist of the same number of vectors.

Thus, when computing the rank of a matrix, the matrix may first be simplified by means of a combination of elementary transformations.

We say that an s by n matrix has *diagonal form* if all its elements are zero except the elements $a_{11}, a_{22}, \ldots, a_{rr}$ [where $0 \leqslant r \leqslant \min(s, n)$], which are equal to unity. The rank of this matrix is obviously r.

Using elementary transformations, it is possible to reduce any matrix to diagonal form.

Indeed, suppose we have a matrix

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \cdot & \cdots & \cdot \\ a_{s1} & \ldots & a_{sn} \end{pmatrix}$$

If all the elements are zero, then it already has diagonal form. But if there are nonzero elements, then an interchange of rows and columns will change element a_{11} to a nonzero element. Then by multiplying the first row by a_{11}^{-1}, we convert element a_{11} to unity. And if we now subtract from the jth column, $j > 1$, the first column multiplied by a_{1j}, then element a_{1j} will be replaced by a zero. Manipulating in similar fashion all columns beyond the first, and also all rows, we arrive at a matrix of the form

$$A' = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & a_{22}' & \ldots & a_{2n}' \\ \cdot & \cdot & \cdots & \cdot \\ 0 & a_{s2}' & \ldots & a_{sn}' \end{pmatrix}$$

Performing the same manipulations with the submatrix that remains in the lower right corner, and so on, we finally—after a finite number of manipulations—arrive at a diagonal matrix with the same rank as the original matrix A.

Thus, *to find the rank of a matrix it is necessary to convert the matrix, by means of elementary transformations, to diagonal form and count the number of units in the principal diagonal.*

Example. Find the rank of the matrix

$$A = \begin{pmatrix} 0 & 2 & -4 \\ -1 & -4 & 5 \\ 3 & 1 & 7 \\ 0 & 5 & -10 \\ 2 & 3 & 0 \end{pmatrix}$$

Interchanging the first and second columns and multiplying the first row by the number $\frac{1}{2}$, we get the matrix

$$\begin{pmatrix} 1 & 0 & -2 \\ -4 & -1 & 5 \\ 1 & 3 & 7 \\ 5 & 0 & -10 \\ 3 & 2 & 0 \end{pmatrix}$$

Adding two times the first column to the third column and then adding some multiple of the new first row to each of the remaining rows, we get the matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & -3 \\ 0 & 3 & 9 \\ 0 & 0 & 0 \\ 0 & 2 & 6 \end{pmatrix}$$

Finally, multiplying the second row by -1, subtracting from the third column three times the second column, and then subtracting from the third and fifth rows certain multiples of the new second row, we arrive at the desired diagonal form

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

The rank of the matrix A is thus two.

In Chapter 13 we will again encounter elementary transformations and diagonal matrices; true, these will be matrices in which the elements are polynomials, not numbers.

11. Systems of Linear Equations

We now begin the study of arbitrary systems of linear equations without any assumptions concerning the number of equations of a system being equal to the number of unknowns. Incidentally, the results we achieve will be applicable to the case (not considered in Sec. 7) when the number of equations is equal to the number of unknowns, but the determinant of the system is zero.

Suppose we have a system of linear equations

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{s1}x_1 + a_{s2}x_2 + \ldots + a_{sn}x_n &= b_s \end{aligned}\right\} \qquad (1)$$

As we know from Sec. 1, the first thing is to decide whether the system is consistent or not. For this purpose, take the coefficient matrix A of the system and the augmented matrix $\bar{A}$ obtained by adjoining to A a column made up of the constant terms,

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{s1} & a_{s2} & \ldots & a_{sn} \end{pmatrix}, \quad \bar{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n}b_1 \\ a_{21} & a_{22} & \ldots & a_{2n}b_2 \\ \ldots & \ldots & \ldots & \ldots \\ a_{s1} & a_{s2} & \ldots & a_{sn}b_s \end{pmatrix}$$

and evaluate the ranks of these matrices. It is easy to see that *the rank of matrix $\bar{A}$ is either equal to the rank of matrix A or exceeds the latter by unity.* Indeed, take a certain maximal linearly independent system of columns of matrix A. It will also be linearly independent in matrix $\bar{A}$. If it also retains the property of maximality, that is, the column of the constant terms is expressible linearly in terms of it, then the ranks of matrices A and $\bar{A}$ are equal; otherwise, adjoining to this system a column made up of constant terms yields a linearly independent system of columns of matrix $\bar{A}$, which is maximal in it.

The question of consistency of a system of linear equations is fully resolved by the following theorem.

Kronecker-Capelli theorem. *A system of linear equations* (1) *is consistent if and only if the rank of the augmented matrix $\bar{A}$ is equal to the rank of the matrix A.*

Proof. 1. Let system (1) be consistent and let $k_1, k_2, \ldots, k_n$ be one of its solutions. Substituting these numbers, in place of the unknowns, into (1), we get s identities, which show that the last column of $\bar{A}$ is the sum of all the remaining columns taken, respectively, with the coefficients $k_1, k_2, \ldots, k_n$. Any other column of $\bar{A}$ is also in A and therefore is expressible linearly in terms of all the columns of this matrix. Conversely, any column of matrix A is a column of $\bar{A}$ as well, that is, it is linearly expressible in terms of the columns of this matrix. From this it follows that the systems of columns of matrices A and $\bar{A}$ are equivalent and therefore, as

proved at the end of Sec. 9, both these systems of s-dimensional vectors have one and the same rank; in other words, the ranks of the matrices A and $\bar{A}$ are equal.

2. Now suppose that the matrices A and $\bar{A}$ have equal ranks. It then follows that any maximal linearly independent system of columns of A remains a maximal linearly independent system in matrix $\bar{A}$ as well. For this reason, the last column of $\bar{A}$ can be expressed linearly in terms of this system and therefore, generally, in terms of the system of columns of matrix A. Consequently, there exists a system of coefficients $k_1, k_2, \ldots, k_n$ such that the sum of the columns of A taken with these coefficients is equal to the column of constant terms, and therefore the numbers $k_1, k_2, \ldots, k_n$ constitute a solution of system (1). Thus, coincidence of the ranks of matrices A and $\bar{A}$ implies that system (1) is consistent.

The proof is complete. In practical situations, it is first necessary to compute the rank of matrix A; to do this, find one of the nonzero minors of the matrix such that all the minors bordering it are zero. Let it be the minor M. Then compute all the minors of matrix $\bar{A}$ bordering M but not contained in A [the so-called *characteristic determinants* of system (1)]. If they are all zero, then the rank of matrix $\bar{A}$ is equal to the rank of matrix A and therefore system (1) is consistent, otherwise it is not consistent. Thus, the Kronecker-Capelli theorem may be stated as follows: *a system of linear equations* (1) *is consistent if and only if all its characteristic determinants are equal to zero.*

Let us now suppose that *system* (1) *is consistent.* The Kronecker-Capelli theorem which we used to establish the consistency of this system states that a solution exists. However, it does not give us any practical method for finding all the solutions of the system. We shall now investigate this problem.

Let matrix A have rank r. As was proved in the preceding section, r is equal to the maximum number of linearly independent rows of matrix A. To be specific let the first r rows of A be linearly independent, and let each of the remaining rows be a linear combination of them. Then the first r rows of $\bar{A}$ will also be linearly independent: any linear dependence between them would also be a linear dependence among the first r rows of A (recall the definition of addition of vectors!). From coincidence of the ranks of matrices A and $\bar{A}$ it follows that the first r rows of $\bar{A}$ constitute, in it, a maximal linearly independent system of rows; in other words, any other row of this matrix is a linear combination of them.

It follows, then, that any equation of system (1) can be represented as a sum of the first r equations taken with certain coefficients and therefore any general solution of the first r equations will satisfy

all the equations of (1). Consequently, it suffices to find all the solutions of the system

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{r1}x_1 + a_{r2}x_2 + \ldots + a_{rn}x_n &= b_r \end{aligned}\right\} \qquad (2)$$

Since the rows of coefficients of the unknowns in equations (2) are linearly independent, that is the matrix of the coefficients has rank r, it follows that $r \leqslant n$ and, besides, that at least one of the minors of order r of this matrix is nonzero. If $r = n$, then (2) is a system with an equal number of equations and unknowns and with a nonzero determinant; that is, it, and for this reason system (1) as well, has a unique solution, namely, that which is calculable by the Cramer rule.

Now let $r < n$ and, for definiteness, let the rth-order minor made up of the coefficients of the first r unknowns be different from zero. In each of the equations of (2), transpose to the right side all terms with the unknowns $x_{r+1}, \ldots, x_n$ and for these unknowns select certain values $c_{r+1}, \ldots, c_n$. We obtain a system of r equations:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1r}x_r &= b_1 - a_{1,\,r+1}c_{r+1} - \ldots - a_{1n}c_n, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2r}x_r &= b_2 - a_{2,\,r+1}c_{r+1} - \ldots - a_{2n}c_n, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{r1}x_1 + a_{r2}x_2 + \ldots + a_{rr}x_r &= b_r - a_{r,\,r+1}c_{r+1} - \ldots - a_{rn}c_n \end{aligned}\right\} \qquad (3)$$

in the r unknowns $x_1, x_2, \ldots, x_r$. Cramer's rule is applicable and therefore the system has a unique solution $c_1, c_2, \ldots, c_r$; it is obvious that the set of numbers $c_1, c_2, \ldots, c_r, c_{r+1}, \ldots, c_n$ will serve as a solution of system (2). Since the values $c_{r+1}, \ldots, c_n$ for the unknowns $x_{r+1}, \ldots, x_n$, called *free unknowns*, can be chosen in arbitrary fashion, we obtain an infinity of distinct solutions of system (2).

On the other hand, any solution of (2) may be obtained in the indicated way: if some solution $c_1, c_2, \ldots, c_n$ of (2) is given, then we take the numbers $c_{r+1}, \ldots, c_n$ for the values of the free unknowns. Then the numbers $c_1, c_2, \ldots, c_r$ will satisfy system (3) and therefore will constitute the only solution of the system, which solution is computed by Cramer's rule.

The foregoing may be combined into a rule for the solution of an arbitrary system of linear equations.

Let there be a consistent system of linear equations (1) *and let the matrix A of the coefficients have rank r. In A we choose r linearly independent rows and leave in* (1) *only those equations whose coefficients*

lie in the chosen rows. In these equations we leave in the left members r unknowns such that the determinant of their coefficients is nonzero, the remaining unknowns are called free and are transposed to the right sides of the equations. Assigning arbitrary numerical values to the free unknowns and computing the values of the remaining unknowns by Cramer's rule, we obtain all the solutions of system (1).

We also state the following result that we have obtained.

A consistent system (1) *has a unique solution if and only if the rank of matrix A is equal to the number of unknowns.*

Example 1. Solve the system

$$\left.\begin{aligned} 5x_1 - x_2 + 2x_3 + x_4 &= 7, \\ 2x_1 + x_2 + 4x_3 - 2x_4 &= 1, \\ x_1 - 3x_2 - 6x_3 + 5x_4 &= 0 \end{aligned}\right\}$$

The rank of the coefficient matrix is two: the second-order minor in the upper left corner of this matrix is nonzero, but both third-order minors bordering it are zero. The rank of the augmented matrix is three, since

$$\begin{vmatrix} 5 & -1 & 7 \\ 2 & 1 & 1 \\ 1 & -3 & 0 \end{vmatrix} = -35 \neq 0$$

The system is thus inconsistent.

Example 2. Solve the system

$$\left.\begin{aligned} 7x_1 + 3x_2 &= 2, \\ x_1 - 2x_2 &= -3, \\ 4x_1 + 9x_2 &= 11 \end{aligned}\right\}$$

The rank of the coefficient matrix is two, i.e., it is equal to the number of unknowns; the rank of the augmented matrix is also two. Thus, the system is consistent and has a unique solution. The left-hand sides of the first two equations are linearly independent; solving the system of these two equations, we get the values

$$x_1 = -\frac{5}{17}, \qquad x_2 = \frac{23}{17}$$

for the unknowns. It is easy to see that this solution also satisfies the third equation.

Example 3. Solve the system

$$\left.\begin{aligned} x_1 + x_2 - 2x_3 - x_4 + x_5 &= 1, \\ 3x_1 - x_2 + x_3 + 4x_4 + 3x_5 &= 4, \\ x_1 + 5x_2 - 9x_3 - 8x_4 + x_5 &= 0 \end{aligned}\right\}$$

The system is consistent since the rank of the augmented matrix (like the rank of the matrix of coefficients) is two. The left members of the first and third equations are linearly independent since the coefficients of the unknowns x_1 and x_2 constitute a nonzero minor of order two. Solve the system of these two equations, the unknowns x_3, x_4, x_5 being considered free; transpose them to the right members of the equations and assume that they have been given

certain numerical values. Using Cramer's rule, we get

$$x_1=\frac{5}{4}+\frac{1}{4}\,x_3-\frac{3}{4}x_4-x_5,$$

$$x_2=-\frac{1}{4}+\frac{7}{4}\,x_3+\frac{7}{4}\,x_4$$

These equations determine the *general solution* of the given system: assigning arbitrary numerical values to the free unknowns, we obtain all the solutions of our system. Thus, for example, the vectors (2, 5, 3, 0, 0), (3, 5, 2, 1, -2), $\left(0, -\frac{1}{4}, -1, 1, \frac{1}{4}\right)$ and so on are solutions of our system. On the other hand, substituting the expressions for x_1 and x_2 from the general solution into any one of the equations of the system, say the second, which was earlier rejected, we obtain an identity.

Example 4. Solve the system

$$\left.\begin{aligned} 4x_1+\ x_2-2x_3+\ x_4&=\ 3,\\ x_1-2x_2-\ x_3+2x_4&=\ 2,\\ 2x_1+5x_2\ \ \ \ -\ x_4&=-1,\\ 3x_1+3x_2-\ x_3-3x_4&=\ 1 \end{aligned}\right\}$$

Although the number of equations is equal to the number of unknowns, the determinant of the system is zero and, therefore. Cramer's rule is not applicable. The rank of the coefficient matrix is equal to three—in the upper right corner of this matrix is a nonzero third-order minor. The rank of the augmented matrix is also three, so the system is consistent. Considering only the first three equations and taking the unknown x_1 as free, we obtain the general solution in the form

$$x_2=-\frac{1}{5}-\frac{2}{5}\,x_1,\quad x_3=-\frac{8}{5}+\frac{9}{5}\,x_1,\quad x_4=0$$

Example 5. Suppose we have a system consisting of $n+1$ equations in n unknowns. The augmented matrix $\overline{A}$ of this system is a square matrix of order $n+1$. If our system is consistent, then, by the Kronecker-Capelli theorem, the determinant of $\overline{A}$ must be zero.

Thus, let there be a system

$$\left.\begin{aligned} x_1-8x_2&=\ 3,\\ 2x_1+\ x_2&=\ 1,\\ 4x_1+7x_2&=-4 \end{aligned}\right\}$$

The determinant of the coefficients and the constant terms of these equations is different from zero:

$$\begin{vmatrix} 1 & -8 & 3\\ 2 & 1 & 1\\ 4 & 7 & -4 \end{vmatrix}=-77$$

The system is therefore inconsistent.

The converse, generally speaking, is not true: from the determinant of matrix $\overline{A}$ being zero it does not follow that the ranks of matrices A and $\overline{A}$ coincide.

12. Systems of Homogeneous Linear Equations

Let us apply the findings of the preceding section to the case of *a system of homogeneous linear equations*:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= 0, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{s1}x_1 + a_{s2}x_2 + \ldots + a_{sn}x_n &= 0 \end{aligned}\right\} \qquad (1)$$

From the Kronecker-Capelli theorem it follows that this system is always consistent, since adding a column of zeros cannot raise the rank of the matrix. This incidentally is evident by a simple inspection—system (1) definitely has a trivial solution (0, 0, . . ., 0).

Let the coefficient matrix A of system (1) have rank r. *If* $r = n$, *then the trivial solution will be the only solution of* (1); *for* $r < n$, *the system has also nontrivial solutions*; to find all these solutions, use the same technique as above in the case of an arbitrary system of equations. In particular, *a system of* n *homogeneous linear equations in* n *unknowns has nontrivial solutions if and only if the determinant of the system is zero.** Indeed, the fact that the determinant is zero is equivalent to the assertion that the rank of matrix A is less than n. On the other hand, *if in a system of homogeneous equations the number of equations is less than the number of unknowns, then the system must definitely have solutions different from zero*, since in that case the rank cannot be equal to the number of unknowns. This was already obtained in Sec. 1 by other reasoning.

Let us, for example, examine the case of a system consisting of $n - 1$ homogeneous equations in n unknowns; assume that the left members of these equations are linearly independent among themselves. Let

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n-1,\,1} & a_{n-1,\,2} & \ldots & a_{n-1,\,n} \end{pmatrix}$$

be the matrix of the coefficients of this system. Denote by M_i the minor of order $n - 1$ obtained by deleting the ith column from A, $i = 1, 2, \ldots, n$. *Then for one of the solutions of our system we have the set of numbers*

$$M_1, \ -M_2, \ M_3, \ -M_4, \ \ldots, \ (-1)^{n-1}M_n \qquad (2)$$

and any other solution is proportional to it.

---

* One half of this assertion was already proved in Sec. 7.

Proof. Since, by hypothesis, the rank of matrix A is $n-1$, one of the minors M_i must be nonzero; let it be M_n. Assume the unknown x_n to be free and transpose it to the right side of each of the equations. We then get

$$\begin{aligned}
a_{11}x_1 + \quad & a_{12}x_2 + \ldots + \quad a_{1,\,n-1}x_{n-1} = \quad -a_{1n}x_n,\\
a_{21}x_1 + \quad & a_{22}x_2 + \ldots + \quad a_{2,\,n-1}x_{n-1} = \quad -a_{2n}x_n,\\
& \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
a_{n-1,\,1}x_1 + {} & a_{n-1,\,2}x_2 + \ldots + a_{n-1,\,n-1}x_{n-1} = -a_{n-1,\,n}x_n
\end{aligned}$$

Applying Cramer's rule, we obtain the general solution of the given system of equations, which, after simple manipulations, becomes

$$x_i = (-1)^{n-i}\frac{M_i}{M_n}x_n, \quad i = 1,\ 2,\ \ldots,\ n-1 \tag{3}$$

Setting $x_n = (-1)^{n-1}M_n$, we obtain: $x_i = (-1)^{2n-i-1}M_i$, $i = 1, 2, \ldots, n-1$, or, since the difference $(2n-i-1)-(i-1) = 2n-2i$ is an even number, $x_i = (-1)^{i-1}M_i$, that is, the set of numbers (2) will indeed be a solution of our system of equations. Any other solution of this system is obtained from formulas (3) for a different numerical value of the unknown x_n and so is proportional to solution (2). This assertion clearly holds true for the case when $M_n = 0$, but one of the minors M_i, $1 \leqslant i \leqslant n-1$, is nonzero.

Solutions of a system of homogeneous linear equations have the following properties. If the vector $\beta = (b_1, b_2, \ldots, b_n)$ is a solution of system (1), then for any scalar k the vector $k\beta = (kb_1, kb_2, \ldots, kb_n)$ is also a solution of the system. This is verified directly by substitution into any one of the equations (1). If the vector $\gamma = (c_1, c_2, \ldots, c_n)$ is another solution of (1) then the vector $\beta + \gamma = (b_1 + c_1, b_2 + c_2, \ldots, b_n + c_n)$ is also a solution of the system:

$$\sum_{j=1}^{n} a_{ij}(b_j + c_j) = \sum_{j=1}^{n} a_{ij}b_j + \sum_{j=1}^{n} a_{ij}c_j = 0 \qquad i = 1,\ 2,\ \ldots,\ s$$

Thus, generally, *any linear combination of solutions of the homogeneous system* (1) *is a solution of the system.* Note that in the case of a *nonhómogeneous* system, that is, a system of linear equations whose constant terms are not all equal to zero, no such assertion is true: neither the sum of two solutions of a system of nonhomogeneous equations nor the product of a solution of the system by a scalar can serve as solutions of the system.

From Sec. 9 we know that any system of n-dimensional vectors consisting of more than n vectors will be linearly dependent. Whence it follows that from a number of solutions of the homogeneous system (1), which solutions, as we know, are n-dimensional vectors, it is

possible to choose a *finite* maximal linearly independent system, that is, maximal in the sense that any other solution of system (1) will be a linear combination of the solutions that enter into the chosen system. Any maximal linearly independent system of solutions of the homogeneous system of equations (1) is called its *fundamental system of solutions.*

Let us once again stress the fact that *an n-dimensional vector is a solution of system* (1) *if and only if it is a linear combination of vectors comprising the given fundamental system.*

Quite naturally, the fundamental system exists only if system (1) has nontrivial solutions, that is, if the rank of its matrix of coefficients is less than the number of unknowns. Then system (1) can have many different fundamental systems of solutions. All these systems are equivalent however, since each vector of any one of the systems is linearly expressible in terms of any other system, and for this reason the systems *consist of one and the same number of solutions.*

The following theorem is valid.

If the rank r of the coefficient matrix of the system of homogeneous linear equations (1) *is less than the number of unknowns n, then any fundamental system of solutions of* (1) *consists of n — r solutions.*

To prove this, note that $n - r$ is the number of free unknowns in system (1); let the unknowns $x_{r+1}, x_{r+2}, \ldots, x_n$ be free. We consider an arbitrary nonzero determinant d of order $n - r$, which we write as follows:

$$d = \begin{vmatrix} c_{1,\,r+1} & c_{1,\,r+2} & \ldots & c_{1n} \\ c_{2,\,r+1} & c_{2,\,r+2} & \ldots & c_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ c_{n-r,\,r+1} & c_{n-r,\,r+2} & \ldots & c_{n-r,\,n} \end{vmatrix}$$

Taking elements of the ith row of this determinant, $1 \leqslant i \leqslant n - r$, for the values of the free unknowns, we get unique values for the unknowns $x_1, x_2, \ldots, x_r$. In other words, we arrive at a quite definite solution of the system (1) of equations. Let us write the solution in the form of a vector:

$$\alpha_i = (c_{i1}, c_{i2}, \ldots, c_{ir}, c_{i,\,r+1}, c_{i,\,r+2}, \ldots, c_{in})$$

The set of vectors $\alpha_1, \alpha_2, \ldots, \alpha_{n-r}$ that we have obtained serves as a fundamental system of solutions for the system (1) of equations. Indeed, this set of vectors is linearly independent since the matrix made up of them (as rows) contains a nonzero minor d of order $n - r$. On the other hand, let

$$\beta = (b_1, b_2, \ldots, b_r, b_{r+1}, b_{r+2}, \ldots, b_n)$$

be an arbitrary solution of system (1). We will prove that the vector β can be expressed linearly in terms of the vectors $\alpha_1, \alpha_2, \ldots, \alpha_{n-r}$.

Denote by α'_i, $i = 1, 2, \ldots, n - r$, the ith row of the determinant d; regard this row as an $(n - r)$-dimensional vector. Then set

$$\beta' = (b_{r+1}, b_{r+2}, \ldots, b_n)$$

The vectors α'_i, $i = 1, 2, \ldots, n - r$, are linearly independent since $d \neq 0$. However, the system of $(n - r)$-dimensional vectors

$$\alpha'_1, \alpha'_2, \ldots, \alpha'_{n-r}, \beta'$$

is linearly dependent since the number of vectors in it is greater than their dimensionality. Hence there are scalars $k_1, k_2, \ldots, k_{n-r}$ such that

$$\beta' = k_1\alpha'_1 + k_2\alpha'_2 + \ldots + k_{n-r}\alpha'_{n-r} \tag{4}$$

Now consider the n-dimensional vector

$$\delta = k_1\alpha_1 + k_2\alpha_2 + \ldots + k_{n-r}\alpha_{n-r} - \beta$$

Since the vector δ is a linear combination of the solutions of the system (1) of homogeneous equations, it will be a solution of the system. From (4) it follows that in the δ solution the values of all the free unknowns are zero. However, the unique solution of system (1) which is obtained for zero values of the free unknowns will be a trivial solution. Thus, $\delta = 0$, that is,

$$\beta = k_1\alpha_1 + k_2\alpha_2 + \ldots + k_{n-r}\alpha_{n-r}$$

which proves the theorem.

Note that the foregoing proof permits us to assert that we will obtain all the fundamental systems of solutions of the system (1) of homogeneous equations by taking for d all possible nonzero determinants of order $n - r$.

Example. Given the following system of homogeneous linear equations:

$$\left.\begin{aligned} 3x_1 + x_2 - 8x_3 + 2x_4 + x_5 &= 0,\\ 2x_1 - 2x_2 - 3x_3 - 7x_4 + 2x_5 &= 0,\\ x_1 + 11x_2 - 12x_3 + 34x_4 - 5x_5 &= 0,\\ x_1 - 5x_2 + 2x_3 - 16x_4 + 3x_5 &= 0 \end{aligned}\right\}$$

The rank of the coefficient matrix is two, the number of unknowns is equal to five; therefore every fundamental system of solutions of this system of equations consists of three solutions. We solve the system confining ourselves to the first two linearly independent equations and considering x_3, x_4, x_5 as free unknowns. We obtain the general solution in the form

$$x_1 = \frac{19}{8}x_3 + \frac{3}{8}x_4 - \frac{1}{2}x_5,$$

$$x_2 = \frac{7}{8}x_3 - \frac{25}{8}x_4 + \frac{1}{2}x_5$$

Then we take the next three linearly independent three-dimensional vectors (1, 0, 0), (0, 1, 0), (0, 0, 1). Substituting the components of each of them

into the general solution as values for the free unknowns and computing the values for x_1 and x_2, we get the following fundamental system of solutions of the given system of equations:

$$\alpha_1=\left(\frac{19}{8},\ \frac{7}{8},\ 1,\ 0,\ 0\right),\quad \alpha_2=\left(\frac{3}{8},\ -\frac{25}{8},\ 0,\ 1,\ 0\right),$$
$$\alpha_3=\left(-\frac{1}{2},\ \frac{1}{2},\ 0,\ 0,\ 1\right)$$

We conclude this section by considering the relationship between the solutions of nonhomogeneous and homogeneous systems. Suppose we have a system of nonhomogeneous linear equations

$$\left.\begin{array}{l} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n=b_1, \\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n=b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{s1}x_1+a_{s2}x_2+\ldots+a_{sn}x_n=b_s \end{array}\right\} \tag{5}$$

The system of homogeneous linear equations

$$\left.\begin{array}{l} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n=0, \\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n=0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{s1}x_1+a_{s2}x_2+\ldots+a_{sn}x_n=0 \end{array}\right\} \tag{6}$$

obtained from (5) by replacing the constant terms by zeros is called the *reduced system* of (5). There is a close connection between the solutions of (5) and (6), as the following two theorems indicate.

I. *The sum of any solution of system* (5) *and any solution of the reduced system* (6) *is again a solution of system* (5).

Indeed, let $c_1, c_2, \ldots, c_n$ be a solution of (5), and $d_1, d_2, \ldots, d_n$ a solution of (6). Take any one of the equations of system (5), say the kth, and substitute into it the numbers $c_1+d_1, c_2+d_2, \ldots, c_n+d_n$ in place of the unknowns. We get

$$\sum_{j=1}^{n} a_{kj}(c_j+d_j)=\sum_{j=1}^{n} a_{kj}c_j+\sum_{j=1}^{n} a_{kj}d_j=b_k+0=b_k$$

II. *The difference between any two solutions of* (5) *is a solution of* (6).

Indeed, let $c_1, c_2, \ldots, c_n$ and $c_1', c_2', \ldots, c_n'$ be solutions of system (5). Take any one of the equations of (6), say the kth, and substitute into it in place of the unknowns the numbers

$$c_1-c_1',\ c_2-c_2',\ \ldots,\ c_n-c_n'$$

This yields

$$\sum_{j=1}^{n} a_{kj}(c_j-c_j')=\sum_{j=1}^{n} a_{kj}c_j-\sum_{j=1}^{n} a_{kj}c_j'=b_k-b_k=0$$

It follows from these theorems that *by finding one solution of the system* (5) *of nonhomogeneous linear equations and adding it to every solution of the reduced system* (6), *we obtain all solutions of* (5).

CHAPTER 3

THE ALGEBRA OF MATRICES

13. Matrix Multiplication

In the preceding chapters the concept of a matrix was utilized as an essential auxiliary tool in the study of systems of linear equations. Numerous other applications have made it the subject of a large independent theory, many branches of which go beyond the limits of this course. We shall now discuss the fundamentals of this theory which starts with the fact that two algebraic operations, addition and multiplication, are defined in the set of all square matrices of a given order in a very peculiar but fully motivated fashion. We begin with the multiplication of matrices; addition will be introduced in Sec. 15.

From the course of analytic geometry we know that when the axes of a rectangular coordinate system in the plane are rotated through an angle α, the coordinates of a point are transformed according to the following formulas:

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

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

where x and y are the old coordinates of the point, and x', y' are the new coordinates. Thus, x and y are expressed linearly in terms of x' and y' with certain numerical coefficients. There are also many other instances of the substitution of unknowns (or variables) in which the old unknowns are linearly expressed in terms of the new ones. Such a substitution of unknowns is ordinarily called a linear transformation (or linear substitution). We thus arrive at the following definition.

A linear transformation of unknowns is a transition from a set of n unknowns $x_1, x_2, \ldots, x_n$ to a set of n unknowns $y_1, y_2, \ldots, y_n$ such that the old unknowns are expressed linearly in terms of the

new ones with certain numerical coefficients:

$$\left.\begin{aligned} x_1 &= a_{11}y_1 + a_{12}y_2 + \ldots + a_{1n}y_n, \\ x_2 &= a_{21}y_1 + a_{22}y_2 + \ldots + a_{2n}y_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n &= a_{n1}y_1 + a_{n2}y_2 + \ldots + a_{nn}y_n \end{aligned}\right\} \qquad (1)$$

The linear transformation (1) is fully determined by its coefficient matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}$$

since two linear transformations of the same matrix can differ only in the letters denoting the unknowns; we take it, however, that the choice of these designations is wholly in our own hands. Conversely, specifying an arbitrary matrix of order n, we can immediately write the linear transformation for which this matrix serves as the coefficient matrix. Thus, there is a one-to-one correspondence between the linear transformations of n unknowns and the square matrices of order n. Therefore, every concept involving linear transformations and every property of these transformations must be associated with a similar concept or property involving matrices.

Let us examine the problem of a successive performance of two linear transformations. Suppose that following the linear transformation (1) there is effected a linear transformation like

$$\left.\begin{aligned} y_1 &= b_{11}z_1 + b_{12}z_2 + \ldots + b_{1n}z_n, \\ y_2 &= b_{21}z_1 + b_{22}z_2 + \ldots + b_{2n}z_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= b_{n1}z_1 + b_{n2}z_2 + \ldots + b_{nn}z_n \end{aligned}\right\} \qquad (2)$$

which takes the set of unknowns $y_1, y_2, \ldots, y_n$ into the set $z_1, z_2, \ldots, z_n$; denote the matrix of this transformation by B. Substituting into (1) the expressions for $y_1, y_2, \ldots, y_n$ from (2), we get linear expressions for the unknowns $x_1, x_2, \ldots, x_n$ in terms of the unknowns $z_1, z_2, \ldots, z_n$. Thus, *the result of a successive execution of two linear transformations of unknowns will again be a linear transformation.*

Example. The result of the successive performance of linear transformations

$$\begin{aligned} x_1 &= 3y_1 - y_2, & y_1 &= z_1 + z_2, \\ x_2 &= y_1 + 5y_2, & y_2 &= 4z_1 + 2z_2 \end{aligned}$$

is the linear transformation

$$\begin{aligned} x_1 &= 3(z_1 + z_2) - (4z_1 + 2z_2) = -z_1 + z_2, \\ x_2 &= (z_1 + z_2) + 5(4z_1 + 2z_2) = 21z_1 + 11z_2 \end{aligned}$$

Denote by C the matrix of the linear transformation which is the result of the successive performance of transformations (1) and (2) and find the law by which its elements c_{ik}, $i, k = 1, 2, \ldots, n$ are expressed in terms of the elements of the matrices A and B. Writing down the transformations (1) and (2) succinctly in the form

$$x_i = \sum_{j=1}^{n} a_{ij} y_j, \quad i = 1, 2, \ldots, n; \quad y_j = \sum_{k=1}^{n} b_{jk} z_k, \quad j = 1, 2, \ldots, n$$

we obtain

$$x_i = \sum_{j=1}^{n} a_{ij} \left(\sum_{k=1}^{n} b_{jk} z_k \right) = \sum_{k=1}^{n} \left(\sum_{j=1}^{n} a_{ij} b_{jk} \right) z_k, \quad i = 1, 2, \ldots, n$$

Thus, the coefficient of z_k in the expression for x_i (that is, the element c_{ik} of matrix C) is of the form

$$c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk} = a_{i1} b_{1k} + a_{i2} b_{2k} + \ldots + a_{in} b_{nk} \tag{3}$$

The element of matrix C in the ith row and kth column is equal to the sum of the products of the corresponding elements of the ith row of matrix A and the kth column of matrix B.

Formula (3), which expresses the elements of matrix C in terms of the elements of matrices A and B, permits us to write down C immediately, given A and B, without having to examine the linear transformations corresponding to the matrices A and B. In this fashion, a one-to-one correspondence is set up between any pair of square matrices of order n and a definite third matrix. We can say that in the set of all square matrices of nth order we have defined an algebraic operation which is called *matrix multiplication*, and matrix C is called the *product* of the matrix A by the matrix B:

$$C = AB$$

Let us once again formulate the relationship between linear transformations and matrix multiplication.

A linear transformation of unknowns obtained as a result of the successive performance of two linear transformations of matrices A and B has as its coefficient matrix the matrix AB.

Examples.

(1) $$\begin{pmatrix} 4 & 9 \\ -1 & 3 \end{pmatrix} \cdot \begin{pmatrix} 1 & -3 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 4 \cdot 1 + 9 \cdot (-2) & 4 \cdot (-3) + 9 \cdot 1 \\ (-1) \cdot 1 + 3 \cdot (-2) & (-1) \cdot (-3) + 3 \cdot 1 \end{pmatrix} = \begin{pmatrix} -14 & -3 \\ -7 & 6 \end{pmatrix}$$

(2) $$\begin{pmatrix} 2 & 0 & 1 \\ -2 & 3 & 2 \\ 4 & -1 & 5 \end{pmatrix} \cdot \begin{pmatrix} -3 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & -1 & 3 \end{pmatrix} = \begin{pmatrix} -6 & 1 & 3 \\ 6 & 2 & 9 \\ -12 & -3 & 14 \end{pmatrix}$$

(3) $$\begin{pmatrix} 7 & 2 \\ 1 & 1 \end{pmatrix}^2 = \begin{pmatrix} 7 & 2 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 7 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 51 & 16 \\ 8 & 3 \end{pmatrix}$$

(4) Find the result of the successive performance of the linear transformations

$$\begin{aligned} x_1 &= 5y_1 - y_2 + 3y_3, \\ x_2 &= y_1 - 2y_2, \\ x_3 &= 7y_2 - y_3 \end{aligned}$$

and

$$\begin{aligned} y_1 &= 2z_1 + z_3, \\ y_2 &= z_2 - 5z_3, \\ y_3 &= 2z_2 \end{aligned}$$

Multiplying the matrices, we obtain

$$\begin{pmatrix} 5 & -1 & 3 \\ 1 & -2 & 0 \\ 0 & 7 & -1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & -5 \\ 0 & 2 & 0 \end{pmatrix} = \begin{pmatrix} 10 & 5 & 10 \\ 2 & -2 & 11 \\ 0 & 5 & -35 \end{pmatrix}$$

The desired linear transformation is therefore of the form

$$\begin{aligned} x_1 &= 10z_1 + 5z_2 + 10z_3, \\ x_2 &= 2z_1 - 2z_2 + 11z_3, \\ x_3 &= 5z_2 - 35z_3 \end{aligned}$$

Take one of the above examples of matrix multiplication, say (2), and find the product of the same matrices, but in reverse order:

$$\begin{pmatrix} -3 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & -1 & 3 \end{pmatrix} \cdot \begin{pmatrix} 2 & 0 & 1 \\ -2 & 3 & 2 \\ 4 & -1 & 5 \end{pmatrix} = \begin{pmatrix} -8 & 3 & -1 \\ 0 & 5 & 9 \\ 14 & -6 & 13 \end{pmatrix}$$

We see that the product of the matrices depends on the order of the factors; in other words, *matrix multiplication is noncommutative*. Actually, this is something we should have expected, if only because the matrices A and B are not of equal status in the definition of matrix C given above by means of formula (3): in A we take the rows and in B the columns.

Examples of noncommutative matrices of order n, that is, matrices whose product changes with an interchange of the factors, may be given for all n beyond $n = 1$ [second-order matrices in Example (1) are noncommutative]. On the other hand, two given matrices may accidentally turn out to be commutative, as witness the following example:

$$\begin{pmatrix} 7 & -12 \\ -4 & 7 \end{pmatrix} \cdot \begin{pmatrix} 26 & 45 \\ 15 & 26 \end{pmatrix} = \begin{pmatrix} 26 & 45 \\ 15 & 26 \end{pmatrix} \cdot \begin{pmatrix} 7 & -12 \\ -4 & 7 \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix}$$

Matrix multiplication is associative; one can therefore speak of a uniquely defined product of any finite number of matrices of order n taken in a definite order (because of the noncommutativity of multiplication).

Proof. Suppose we have three arbitrary matrices of order n, A, B and C. In abbreviated notation (which indicates the general

aspect of their elements) we have $A = (a_{ij})$, $B = (b_{ij})$, $C = (c_{ij})$. We also introduce the following designations:

$$AB = U = (u_{ij}), \qquad BC = V = (v_{ij}),$$
$$(AB)\,C = S = (s_{ij}), \quad A\,(BC) = T = (t_{ij})$$

We have to prove the truth of the equations $(AB)\,C = A\,(BC)$, that is, $S = T$. However

$$u_{il} = \sum_{k=1}^{n} a_{ik}b_{kl}, \quad v_{kj} = \sum_{l=1}^{n} b_{kl}c_{lj}$$

and, therefore, because of the equations $S = UC$, $T = AV$,

$$s_{ij} = \sum_{l=1}^{n} u_{il}c_{lj} = \sum_{l=1}^{n}\sum_{k=1}^{n} a_{ik}b_{kl}c_{lj},$$

$$t_{ij} = \sum_{k=1}^{n} a_{ik}v_{kj} = \sum_{k=1}^{n}\sum_{l=1}^{n} a_{ik}b_{kl}c_{lj}$$

That is to say, $s_{ij} = t_{ij}$ for $i, j = 1, 2, \ldots, n$.

To go deeper into the properties of matrix multiplication we have to study their determinants. For the sake of brevity, we agree to denote the determinant of matrix A by $|A|$. If in each of the above examples the reader will take the pains to count the determinants of the matrices being multiplied and to compare the product of these determinants with the determinant of the product of the given matrices, he will detect an extremely curious regularity which is expressed as the following very important **theorem on the multiplication of determinants.**

The determinant of a product of several matrices of order n is equal to the product of the determinants of these matrices.

It will suffice to prove this theorem for the case of two matrices. Let there be given the nth-order matrices $A = (a_{ij})$ and $B = (b_{ij})$ and let $AB = C = (c_{ij})$. Construct the following auxiliary determinant Δ of order $2n$: in the upper left corner put matrix A, in the lower right corner, matrix B, the entire upper right corner will be occupied by zeros, finally, put the number -1 along the principal diagonal of the lower left corner and zeros elsewhere. Determinant Δ will then look like this:

$$\Delta = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} & 0 & 0 & \ldots & 0 \\ a_{21} & a_{22} & \ldots & a_{2n} & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} & 0 & 0 & \ldots & 0 \\ -1 & 0 & \ldots & 0 & b_{11} & b_{12} & \ldots & b_{1n} \\ 0 & -1 & \ldots & 0 & b_{21} & b_{22} & \ldots & b_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & -1 & b_{n1} & b_{n2} & \ldots & b_{nn} \end{vmatrix}$$

Applying the Laplace theorem to the determinant Δ—expansion about the first n rows—we get the following equation:

$$\Delta = |A| \cdot |B| \tag{4}$$

Now let us attempt to transform the determinant Δ, without changing its value, so that all elements b_{ij}, $i, j = 1, 2, \ldots, n$, are replaced by zeros. To do this, add to the $(n+1)$th column of Δ its first column multiplied by b_{11}, the second multiplied by b_{21} and so on, and finally, the nth column, multiplied by b_{n1}. Then add to the $(n+2)$th column of determinant Δ the first column multiplied by b_{12}, the second multiplied by b_{22}, and so on. Generally, we add to the $(n+j)$th column of the determinant Δ, where $j = 1, 2, \ldots, n$, the sum of the first n columns taken, respectively, with the coefficients $b_{1j}, b_{2j}, \ldots, b_{nj}$.

It is easy to see that these manipulations do not change the determinant and actually result in the replacement of all elements b_{ij} by zeros. At the same time, in place of the zeros in the upper right corner of the determinant there appear the following numbers: at the intersection of the ith row and the $(n+j)$th column of the determinant, $i, j = 1, 2, \ldots, n$, will stand the number $a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{in}b_{nj}$ equal [because of (3)] to the element c_{ij} of matrix $C = AB$. The upper right corner of the determinant is now occupied by matrix C:

$$\Delta = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} & c_{11} & c_{12} & \ldots & c_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} & c_{21} & c_{22} & \ldots & c_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} & c_{n1} & c_{n2} & \ldots & c_{nn} \\ -1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & -1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & -1 & 0 & 0 & \ldots & 0 \end{vmatrix}$$

Apply the Laplace theorem once again, expanding the determinant about the last n columns. The complementary minor of the minor $|C|$ is equal to $(-1)^n$, and since the minor $|C|$ is located in rows with position numbers $1, 2, \ldots, n$ and in columns with position numbers $n+1, n+2, \ldots, 2n$, and

$$1 + 2 + \ldots + n + (n+1) + (n+2) + \ldots + 2n = 2n^2 + n$$

it follows that

$$\Delta = (-1)^{2n^2+n}(-1)^n |C| = (-1)^{2(n^2+n)} |C|$$

or, due to the evenness of the number $2(n^2+n)$,

$$\Delta = |C| \tag{5}$$

Finally, from (4) and (5) follows the equation we set out to prove:

$$|C| = |A| \cdot |B|$$

The multiplication theorem for determinants could be proved without invoking the Laplace theorem. One such proof is given at the end of Sec. 16.

14. Inverse Matrices

A square matrix is called *singular* if its determinant is zero, otherwise it is *nonsingular*. Accordingly, a linear transformation of unknowns is called *singular* or *nonsingular* depending on whether the coefficient determinant of this transformation is zero or not. The following assertion follows from the theorem proved at the end of Sec. 13.

The product of matrices, at least one of which is singular, is a singular matrix.

The product of any nonsingular matrices is a nonsingular matrix.

From this there follows the assertion (because of the relationship existing between matrix multiplication and the successive performance of linear transformations): *the result of a successive performance of several linear transformations is a nonsingular transformation if and only if all the given transformations are nonsingular.*

The role of unity in matrix multiplication is played by the unit (identity) matrix

$$E = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

It commutes with any matrix A of a given order,

$$AE = EA = A \tag{1}$$

These equalities are proved either by direct application of the rule for multiplying matrices or on the basis of the remark that the unit (identity) matrix corresponds to an *identical* linear transformation of unknowns:

$$\begin{aligned} x_1 &= y_1, \\ x_2 &= y_2, \\ &\dots\dots\dots \\ x_n &= y_n \end{aligned}$$

the performance of which, either prior to or following any other linear transformation, obviously does not alter that transformation.

Note that *matrix E is the only matrix which satisfies condition* (1) *for any matrix A*. Indeed, if there were also matrix E' with this property, we would have

$$E'E = E', \quad E'E = E$$

whence $E' = E$.

The question of whether a given matrix A has an *inverse* turns out to be more complicated. Since matrix multiplication is not commutative, we will now speak of the *right* inverse matrix, that is a matrix A^{-1} such that postmultiplication of A by this matrix yields the identity matrix:

$$AA^{-1} = E \tag{2}$$

Suppose matrix A is singular; then if matrix A^{-1} existed, the product on the left of (2) would, as we know, be a singular matrix, whereas in actual fact the matrix E in the right member of this equation is nonsingular since its determinant is equal to unity. Thus a singular matrix cannot have a right inverse matrix. Similar reasoning shows that it cannot have a left inverse matrix either, and for this reason, *a singular matrix has no inverse at all.*

Passing to the case of a nonsingular matrix, let us first introduce the following auxiliary concept. Suppose we have an nth-order matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

The matrix

$$A^* = \begin{pmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \dots & \dots & \dots & \dots \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{pmatrix}$$

which consists of the cofactors of the elements of A (note that the cofactor of element a_{ij} lies at the intersection of the jth row and the ith column) is called the *adjoint* of matrix A.

Let us find the products AA^* and A^*A. Using the familiar formula (see Sec. 6) for the expansion of a determinant about a row or column, and also the theorem (see Sec. 7) on the sum of the products of the elements of any row (column) of a determinant by the cofactors of the corresponding elements of another row (column) and denoting by d the determinant of the matrix A,

$$d = |A|$$

we get the following equations:

$$AA^* = A^*A = \begin{pmatrix} d & 0 & \dots & 0 \\ 0 & d & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & d \end{pmatrix} \tag{3}$$

From this it follows that *if matrix A is nonsingular, then its adjoint A^* will also be nonsingular; note that the determinant d^* of matrix A^* is equal to the $(n-1)$th power of the determinant d of matrix A.*

Indeed, passing from (3) to the equality between the determinants, we get

$$dd^* = d^n$$

whence, because $d \neq 0$,

$$d^* = d^{n-1}$$

(We could prove that if matrix A is singular, then its adjoint A^* is also singular and has rank which does not exceed 1.)

It is now easy to prove the existence of an inverse matrix for any *nonsingular* matrix A and to find its form. Note first that if we consider the product of two matrices AB and if we divide all the elements of one of the factors, say B, by one and the same number d, then this number also divides all elements of the product AB: to prove this all we need to do is recall the definition of matrix multiplication. Thus, if

$$d = |A| \neq 0$$

then from (3) it follows that *the inverse of A is a matrix obtained from the adjoint A^* by means of division of all its elements by the number d:*

$$A^{-1} = \begin{pmatrix} \frac{A_{11}}{d} & \frac{A_{21}}{d} & \dots & \frac{A_{n1}}{d} \\ \frac{A_{12}}{d} & \frac{A_{22}}{d} & \dots & \frac{A_{n2}}{d} \\ \dots & \dots & \dots & \dots \\ \frac{A_{1n}}{d} & \frac{A_{2n}}{d} & \dots & \frac{A_{nn}}{d} \end{pmatrix}$$

Indeed, from (3) follow the equalities

$$AA^{-1} = A^{-1}A = E \tag{4}$$

We stress once again that the ith *row* of matrix A^{-1} contains the cofactors of the elements of the ith *column* of determinant $|A|$ divided by $d = |A|$.

It is easy to prove that *matrix A^{-1} is the only matrix which satisfies condition (4) for a given nonsingular matrix A*. True enough,

if matrix C is such that

$$AC = CA = E$$

then

$$CAA^{-1} = C\,(AA^{-1}) = CE = C,$$
$$CAA^{-1} = (CA)\,A^{-1} = EA^{-1} = A^{-1}$$

whence $C = A^{-1}$.

From (4) and the multiplication theorem for determinants it follows that *the determinant of matrix* A^{-1} *is equal to* $\frac{1}{|A|}$ *so that this matrix is also nonsingular; its inverse is the matrix* A.

Now, if we have square matrices A and B of order n, of which A is nonsingular and B is arbitrary, then we can perform *the right and left divisions of* B *by* A, that is, we can solve the matrix equations

$$AX = B, \quad YA = B \tag{5}$$

To do this, it will suffice (because of the associativity of matrix multiplication) to set

$$X = A^{-1}B, \quad Y = BA^{-1}$$

These solutions of equations (5) will, in the general case (because matrix multiplication is not commutative), be distinct.

Example 1. Given a matrix

$$A = \begin{pmatrix} 3 & -1 & 0 \\ -2 & 1 & 1 \\ 2 & -1 & 4 \end{pmatrix}$$

Its determinant $|A| = 5$, and so the inverse matrix A^{-1} exists:

$$A^{-1} = \begin{pmatrix} 1 & \frac{4}{5} & -\frac{1}{5} \\ 2 & \frac{12}{5} & -\frac{3}{5} \\ 0 & \frac{1}{5} & \frac{1}{5} \end{pmatrix}$$

Example 2. Given the matrices

$$A = \begin{pmatrix} 3 & 2 \\ 4 & 3 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & 7 \\ 3 & 5 \end{pmatrix}$$

The matrix A is nonsingular, and

$$A^{-1} = \begin{pmatrix} 3 & -2 \\ -4 & 3 \end{pmatrix}$$

Therefore the following matrices are solutions of the equations $AX = B$, $YA = B$:

$$X = \begin{pmatrix} 3 & -2 \\ -4 & 3 \end{pmatrix} \cdot \begin{pmatrix} -1 & 7 \\ 3 & 5 \end{pmatrix} = \begin{pmatrix} -9 & 11 \\ 13 & -13 \end{pmatrix},$$

$$Y = \begin{pmatrix} -1 & 7 \\ 3 & 5 \end{pmatrix} \cdot \begin{pmatrix} 3 & -2 \\ -4 & 3 \end{pmatrix} = \begin{pmatrix} -31 & 23 \\ -11 & 9 \end{pmatrix}$$

Multiplication of rectangular matrices. Although in the preceding section matrix multiplication was only defined for square matrices of the same order, it carries over to the case of rectangular matrices A and B, provided it is possible to apply formula (3) of the preceding section, i.e., if any row of matrix A contains as many elements as there are in any column of matrix B. In other words, *one can speak of the product of rectangular matrices A and B if the number of columns of matrix A is equal to the number of rows of matrix B, the number of rows of matrix AB being equal to the number of rows of A, and the number of columns of matrix AB to the number of columns of B.*

Examples.

(1) $$\begin{pmatrix} 5 & -1 & 3 & 1 \\ 2 & 0 & -1 & 4 \end{pmatrix} \cdot \begin{pmatrix} -1 & 3 & 0 \\ -2 & 1 & 1 \\ 3 & 0 & -2 \\ 4 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 10 & 15 & -5 \\ 11 & 10 & 10 \end{pmatrix}$$

(2) $$\begin{pmatrix} 0 & -3 & 1 \\ 2 & 1 & 5 \\ -4 & 0 & -2 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ -2 \\ 2 \end{pmatrix} = \begin{pmatrix} 8 \\ 14 \\ -16 \end{pmatrix}$$

(3) $$(5\ 1\ 0\ -3) \cdot \begin{pmatrix} 2 & 0 \\ 1 & -4 \\ 3 & 1 \\ 0 & -1 \end{pmatrix} = (11\ \ -1)$$

Multiplication of rectangular matrices may be related to a successive performance of linear transformations of the unknowns, provided only that in the definition of the latter we give up the assumption that the number of unknowns is preserved under the linear transformation.

It is also easy to verify, by repeating word for word the proof given above for the case of square matrices, that *associativity holds true also for the multiplication of rectangular matrices.*

We now take advantage of the multiplication of rectangular matrices and of properties of the inverse matrix for a new derivation of Cramer's rule, which does not require the involved computations that were carried out in Sec. 7. Let there be given a system of n linear equations in n unknowns:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{aligned}\right\} \qquad (6)$$

The determinant of this system is different from zero. Denote by A the coefficient matrix of system (6); this matrix is nonsingular

since, by hypothesis, $d = |A| \neq 0$. Denote by X the column of unknowns, by B the column of constant terms of (6); thus

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \qquad B = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

The product AX is meaningful since the number of columns of matrix A is equal to the number of rows of matrix X; this product will be a column composed of the left-hand sides of the equations of system (6). Thus, (6) may be written in the form of a single matrix equation

$$AX = B \tag{7}$$

Multiplying both sides of (7) on the left by the matrix A^{-1}, the existence of which follows from the nonsingular nature of the square matrix A, we get

$$X = A^{-1}B \tag{8}$$

The product on the right is a matrix of one column; its jth element is equal to the sum of the products of the elements of the jth row of matrix A^{-1} by the corresponding elements of matrix B, that is, it is equal to the number

$$\frac{A_{1j}}{d} b_1 + \frac{A_{2j}}{d} b_2 + \ldots + \frac{A_{nj}}{d} b_n = \frac{1}{d}(A_{1j}b_1 + A_{2j}b_2 + \ldots + A_{nj}b_n)$$

The parenthesis on the right is, however, an expansion about the jth column of determinant d_j, which is obtained by replacing the jth column of d by the column B. Thus, formulas (8) are equivalent to formulas (3), Sec. 7, which express the solution obtained by Cramer's rule to system (6).

It remains to show that the values of the unknowns thus obtained are indeed the solution of system (6). To do this, put expression (8) into the matrix equation (7); it obviously yields the identity $B = B$.

The rank of a product of matrices. In the case of singular matrices, the multiplication theorem for determinants does not lead to any utterance beyond the fact that their product will also be singular, although singular square matrices can be distinguished according to rank as well. Note that there is no completely definite relationship between the ranks of the factors and the rank of the product, as is evident from a glance at the following examples:

$$\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 3 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 6 & 0 \\ 0 & 0 \end{pmatrix},$$

$$\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

In both cases, matrices of rank 1 are multiplied, but in one case the product has rank 1, in the other, rank 0. It is only the following theorem which holds true (and not only for square but for rectangular matrices as well).

The rank of a product of matrices does not exceed the rank of each of the factors.

It will suffice to prove this theorem for the case of two factors. Suppose we have matrices A and B for which the product AB is meaningful: $AB = C$. We consider formula (3), Sec. 13, which yields an expression for the elements of matrix C. Taking this formula for the given k and for all possible i ($i = 1, 2, \ldots$), we find that the kth column of matrix C is the sum of all the columns of matrix A taken with certain coefficients (namely, with the coefficients $b_{1k}, b_{2k}, \ldots$). This is proof that the system of columns of matrix C is expressed linearly in terms of the system of columns of matrix A and, therefore, as shown in Sec. 9, the rank of the first system is less than or equal to the rank of the second system; in other words, the rank of matrix C does not exceed the rank of matrix A. On the other hand, since from this same formula (3), Sec. 13, there follows, for a given i and for all k, that each ith row of matrix C is a linear combination of the rows of matrix B, we find by analogous reasoning that the rank of C is not greater than the rank of B.

A more precise result is obtained in the case when one of the factors is a nonsingular square matrix.

The rank of the product obtained by pre- or postmultiplication of an arbitrary matrix A by a nonsingular square matrix Q is equal to the rank of matrix A.

For example, suppose

$$AQ = C \tag{9}$$

From the preceding theorem it follows that the rank of matrix C is not greater than the rank of matrix A. However, multiplying (9) on the right by Q^{-1}, we arrive at the equation

$$A = CQ^{-1}$$

and for this reason, again on the basis of the preceding theorem, the rank of A does not exceed that of C. A comparison of these two results proves the coincidence of the ranks of matrices A and C.

15. Matrix Addition and Multiplication of a Matrix by a Scalar

For square matrices of order n, addition is defined as follows.

The sum $A + B$ of two square matrices $A = (a_{ij})$ and $B = (b_{ij})$ of order n is the matrix $C = (c_{ij})$, each element of which is equal

to the sum of the corresponding elements of matrices A and B:

$$c_{ij} = a_{ij} + b_{ij}*$$

The addition of matrices thus defined will obviously be commutative and associative. The inverse operation also exists; the difference between the matrices A and B is a matrix composed of the differences of the corresponding elements of the given matrices. The role of zero is played by the *zero matrix*, composed entirely of zeros; this matrix will from now on be denoted by the symbol 0. There is no real danger of confusing a zero matrix and the number zero.

The addition of square matrices and their multiplication as defined in Sec. 13 are related by the distributive laws.

Indeed, suppose we have three matrices of order n, $A = (a_{ij})$, $B = (b_{ij})$, $C = (c_{ij})$. Then for any i and j we have the obvious equality

$$\sum_{s=1}^{n} (a_{is} + b_{is})\, c_{sj} = \sum_{s=1}^{n} a_{is}c_{sj} + \sum_{s=1}^{n} b_{is}c_{sj}$$

However the left side of this equation is an element in the ith row and jth column of the matrix $(A + B)\, C$, the right side is an element in the same position in the matrix $AC + BC$. This proves the equation

$$(A + B)\, C = AC + BC$$

The equation $C\,(A + B) = CA + CB$ is proved in exactly the same way: the noncommutativity of matrix multiplication quite naturally requires proof of these two distributive laws.

Let us introduce the following definition of multiplication of matrices by a scalar.

The product kA of a square matrix $A = (a_{ij})$ by a scalar k is the matrix $A' = (a'_{ij})$ obtained by multiplying all elements of the matrix A by k:

$$a'_{ij} = ka_{ij}$$

We have already encountered (Sec. 14) one such example of multiplication of a matrix by a scalar: if matrix A is nonsingular, and $|A| = d$, then its inverse, A^{-1}, and the adjoint A^* are connected by the equation

$$A^{-1} = d^{-1}A^*$$

As we know, any square matrix of order n may be regarded as an n^2-dimensional vector: this correspondence between matrices

---

* Of course, one could define the matrix product in just as natural a way multiplying the corresponding elements. However, such multiplication, unlike that defined in Sec. 13, would not find any serious applications.

and vectors is one-to-one. The addition of matrices and the multiplication of a matrix by a scalar defined here are then converted into the addition of vectors and the multiplication of a vector by a scalar. Thus, *the collection of square matrices of order n may be regarded as an n^2-dimensional vector space.*

From this follows the truth of the following equations (here, A and B are matrices of order n; k, l are scalars and 1 is the number unity):

$$k(A + B) = kA + kB, \tag{1}$$

$$(k + l)A = kA + lA, \tag{2}$$

$$k(lA) = (kl)A, \tag{3}$$

$$1 \cdot A = A \tag{4}$$

Properties (1) and (2) connect multiplication of a matrix by a scalar with addition of matrices. At the same time, there is a very important relationship between the multiplication of a matrix by a scalar and multiplication of the matrices alone, namely,

$$(kA)B = A(kB) = k(AB) \tag{5}$$

In words, *if one of the factors in a product of matrices is multiplied by a scalar k, then the whole product is multiplied by k.*

Let there be matrices $A = (a_{ij})$ and $B = (b_{ij})$ and a scalar k. Then for any i and j,

$$\sum_{s=1}^{n} (ka_{is})\, b_{sj} = k \sum_{s=1}^{n} a_{is}b_{sj}$$

The left side of this equation, however, is an element in the ith row and the jth column of matrix $(kA)B$, the right side is an element in the same place in matrix $k(AB)$. This proves the equation

$$(kA)B = k(AB)$$

The equation $A(kB) = k(AB)$ is proved in the same way.

The operation of multiplication of a matrix by a scalar permits introducing a new mode of matrix notation. Denote by E_{ij} the matrix in which unity lies at the intersection of the ith row and the jth column, all other elements being zero. Setting $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, n$, we obtain n^2 such matrices E_{ij}, which are connected, as may easily be verified, by the following multiplication table:

$$E_{is}E_{sj} = E_{ij},\ E_{is}E_{tj} = 0 \text{ for } \quad s \neq t$$

The matrix kE_{ij} differs from the matrix E_{ij} solely in the fact that it has the scalar k at the intersection of the ith row and the jth column. Taking this into consideration and using the definition

of matrix addition, we get the following notation for an arbitrary square matrix A:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} E_{ij} \tag{6}$$

The matrix A obviously possesses only the notation (6).

The matrix kE, where E is the unit matrix, has, by the definition of multiplication of a matrix by a scalar, the following form:

$$kE = \begin{pmatrix} k & & & 0 \\ & k & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & k \end{pmatrix}$$

that is to say, one and the same scalar k on the principal diagonal and all other elements zero. Such matrices are called *scalar matrices.*

The definition of matrix addition leads to the equation

$$kE + lE = (k + l)\, E \tag{7}$$

On the other hand, using the definition of matrix multiplication or proceeding from (5), we get

$$kE \cdot lE = (kl)\, E \tag{8}$$

Multiplication of matrix A by a scalar k may be interpreted as multiplication of A by a scalar matrix kE in the meaning of multiplication of matrices. Indeed, by (5),

$$(kE)\, A = A\, (kE) = kA$$

The conclusion to be drawn here is that *every scalar matrix commutes with any matrix A.* It is very important to point out that scalar matrices are the only ones with this property.

If a matrix $C = (c_{ij})$ of order n commutes with any matrix of the same order, then C is a scalar matrix.

Indeed, set $i \neq j$ and consider the products CE_{ij} and $E_{ij}C$ (which by hypothesis are equal; see above definition of the matrix E_{ij}). It is clear that all columns of matrix CE_{ij}, except the jth, consist of zeros, and the jth column coincides with the ith column of matrix C; in particular, element c_{ii} lies at the intersection of the ith row and the jth column of matrix CE_{ij}. Similarly all the rows of matrix $E_{ij}C$, except the ith, consist of zeros, and the ith row coincides with the jth row of matrix C; at the intersection of the ith row and the jth column of matrix $E_{ij}\, C$ lies the element c_{jj}.

Using the equality $CE_{ij} = E_{ij}C$, we find that $c_{ii} = c_{jj}$ (as elements in the same positions of equal matrices), which is to say that all elements of the principal diagonal of matrix C are equal. On the other hand, element c_{ji} lies at the intersection of the jth row and the ith column of matrix CE_{ij}; but in matrix $E_{ij}C$ we have a zero at this site (because $i \neq j$), and therefore $c_{ji} = 0$, or every off-diagonal element of matrix C is zero. The theorem is proved.

16. An Axiomatic Construction of the Theory of Determinants

An nth-order determinant is a number which is uniquely defined by a given square matrix of order n. The definition of this concept given in Sec. 4 points to a rule by which a determinant can be expressed in terms of the elements of the given matrix. This constructive definition may, however, be replaced by an axiomatic definition. In other words, it is possible to point out, among the properties of a determinant that were established in Secs. 4 and 6, such properties that the determinant is the sole function of a real matrix having these properties.

The simplest definition of this kind consists in utilizing the expansion of a determinant in terms of a row. Let us consider square matrices of any order and let us assume that any such matrix M is associated with a number d_M and the following conditions hold.

(1) If the matrix M is of order one, that is, if it consists of a single element a, then $d_M = a$.

(2) If the first row of a matrix M of order n is made up of the elements $a_{11}, a_{12}, \ldots, a_{1n}$ and if M_i, $i = 1, 2, \ldots, n$, denotes a matrix of order $n - 1$ which remains after deleting from M the first row and the ith column, then

$$d_M = a_{11}d_{M_1} - a_{12}d_{M_2} + a_{13}d_{M_3} - \ldots + (-1)^{n-1}a_{1n}d_{M_n}$$

Then for any matrix M, the number d_M is equal to the determinant of that matrix. We leave it to the reader to carry out the proof of this assertion, which is done by induction with respect to n and utilizes the results of Sec. 6.

Much more interesting are some other forms of an axiomatic definition of a determinant which refer solely to the case of a given order n and have for a basis some of the simplest determinant properties that were established in Sec. 4. Let us examine one of these definitions.

Let any square matrix M of order n be associated with a number d_M, and let the following conditions hold true.

I. *If one of the rows of matrix M is a multiple of k, then the number d_M is also a multiple of k.*

II. *The number d_M is not changed if to one of the rows of M we add another row of this matrix.*

III. *If E is the unit matrix, then $d_E = 1$.*

We shall prove that *for any matrix M the number d_M is equal to the determinant of the matrix.*

Let us first derive from the conditions I to III certain properties of the number d_M that are analogous to the corresponding properties of a determinant.

(1) *If one of the rows of matrix M consists of zeros, then $d_M = 0$.*

Indeed, by multiplying a row consisting of zeros by the number 0, we do not change the matrix, but because of Condition I, the number d_M acquires the factor 0. Therefore

$$d_M = 0 \cdot d_M = 0$$

(2) *The number d_M does not change if to the ith row of matrix M we add its jth row, $j \neq i$, multiplied by a scalar k.*

If $k = 0$, then that is the proof. If $k \neq 0$, then we multiply the jth row by k and obtain a matrix M' for which, because of Condition I, $d_{M'} = kd_M$. Then to the ith row of matrix M' we add the jth row and obtain the matrix M'', and, because of Condition II, $d_{M''} = d_{M'}$. Finally, we multiply the jth row of matrix M'' by the scalar k^{-1}. We arrive at matrix M''', which is actually obtained from M by the transformation indicated in the formulation of the property being proved; note that

$$d_{M'''} = k^{-1}d_{M''} = k^{-1}d_{M'} = k^{-1} \cdot kd_M = d_M$$

(3) *If the rows of matrix M are linearly dependent, then $d_M = 0$.*

Indeed, if one of the rows, say the ith, is a linear combination of the other rows, then, applying transformation (2) several times, it is possible to replace the ith row by a row of zeros. Transformation (2) does not change the number d_M and so, by Property (1), $d_M = 0$.

(4) *If the ith row of matrix M is a sum of two vectors β and γ and if matrices M' and M'' are obtained from M by replacing its ith row by the vectors β and γ, respectively, then*

$$d_M = d_{M'} + d_{M''}$$

Let S be the system of all rows of matrix M, except the ith. If there is a linear dependence in S, then the rows of each one of the matrices M, M', M'' are linearly dependent, and therefore, by Property (3), $d_M = d_{M'} = d_{M''} = 0$, whence in that case follows the truth of the property being proved. Now if a system S consisting of $n - 1$ vectors is linearly independent, then as the results of Sec. 9 show, a vector α may be adjoined to form a maximal linearly independent system of vectors of n-dimensional vector space. It is possible to express the vectors β and γ linearly in terms

of this system. Let vector α enter these expressions with the coefficients k and l, respectively; vector α will then enter the expression for the vector $\beta + \gamma$, that is, for the ith row of matrix M, with the coefficient $k + l$. Matrices M, M' and M'' can now be transformed by subtracting from their ith rows certain linear combinations of other rows so that the vectors $(k + l)\,\alpha$, $k\alpha$ and $l\alpha$ will serve respectively as their ith rows. Therefore, denoting by M^0 the matrix obtained from matrix M by replacing its ith row by the vector α and taking into account Properties (2) and I, we arrive at the equations

$$d_M = (k + l)\,d_{M^0}, \quad d_{M'} = kd_{M^0}, \quad d_{M''} = ld_{M^0}$$

The proof of Property (4) is complete.

(5) *If matrix $\overline{M}$ is obtained from matrix M by interchanging two rows, then $d_{\overline{M}} = -d_M$.*

Suppose it is necessary in matrix M to interchange the rows with subscripts i and j. This can be achieved by a chain of transformations: first add to the ith row of M its jth row and get matrix M'; by Condition II, $d_{M'} = d_M$. Then from the jth row of M' subtract its ith row and arrive at the matrix M'', for which, by Property (2), we have $d_{M''} = d_{M'}$; the jth row of M'' will differ in sign from the ith row of M. Now add to the ith row of M'' its jth row. For matrix M''', which this manipulation yields, we have, by Condition II, $d_{M'''} = d_{M''}$, and the ith row of this matrix coincides with the jth row of matrix M. Finally, multiplying the jth row of M''' by -1, we arrive at the desired matrix $\overline{M}$. Therefore, by Condition I,

$$d_{\overline{M}} = -d_{M'''} = -d_M$$

(6) *If matrix M' is obtained from matrix M by interchanging rows, the α_i-th row of matrix M serving as the ith row of matrix M', $i = 1, 2, \ldots, n$, then*

$$d_{M'} = \pm d_M$$

The plus sign corresponds to the case when the permutation

$$\begin{pmatrix} 1 & 2 & \ldots & n \\ \alpha_1 & \alpha_2 & \ldots & \alpha_n \end{pmatrix}$$

is even; the minus sign, to the case when it is odd.

Indeed, matrix M' may be obtained from matrix M by a number of transpositions of two rows, and for this reason we can take advantage of Property (5). The parity of the number of these transpositions determines, as we know from Sec. 3, the parity of the above-given permutation.

Now let us consider the matrices $M = (a_{ij})$, $N = (b_{ij})$ and their product $Q = MN$ in the meaning of Sec. 13. We find the number d_Q. We know that any ith row of matrix Q is the sum of

all the rows of matrix N taken, respectively, with the coefficients $a_{i1}, a_{i2}, \ldots, a_{in}$ (see, for example, Sec. 14). Replace all the rows of Q by their indicated linear expressions in terms of the rows of matrix N and take advantage of Property (4) several times. We find that the number d_Q will equal the sum of the numbers d_T for all possible matrices T of the following kind: the ith row of T, $i = 1, 2, \ldots, n$, is equal to the α_i th row of matrix N multiplied by a scalar $a_{i\alpha i}$. Here, because of Property (3), we can disregard all matrices T for which there exist subscripts i and j, $i \neq j$, such that $\alpha_i = \alpha_j$; in other words, what remain are only matrices T for which the subscripts $\alpha_1, \alpha_2, \ldots, \alpha_n$ constitute an arrangement of the numbers $1, 2, \ldots, n$. Because of Properties I and (6), the number d_T for such a matrix is of the form

$$d_T = \pm a_{1\alpha_1} a_{2\alpha_2} \cdots a_{n\alpha_n} d_N$$

where the sign is determined by the parity of the permutation formed from the subscripts. Whence we arrive at the expression for the number d_Q: after factoring the common factor d_N out of all summands of the type d_T, what we obviously have left in the parentheses is the determinant $|M|$ of the matrix M in the sense of the constructive definition as given in Sec. 4, i.e.,

$$d_Q = |M| \cdot d_N \qquad (*)$$

If we now take the unit matrix E for the matrix N, then $Q = M$, and, by Property III, $d_N = d_E = 1$, that is *for any matrix* M *we have the equality*

$$d_M = |M|$$

which is what we set out to prove. At the same time, once again, and without the use of the Laplace theorem, we have proved the multiplication theorem for determinants: all that needs to be done is, in equation (*), to replace the numbers d_Q and d_N by the determinants of the respective matrices.

We conclude these axiomatic considerations with proof of the independence of Conditions I to III, that is proof that none of these conditions is a consequence of the other two.

To prove the independence of Condition III, assume that $d_M = 0$ for any matrix M of order n. Conditions I and II will obviously be fulfilled, but III breaks down.

To prove the independence of Condition II assume that for any matrix M the number d_M is equal to the product of the elements in the principal diagonal of the matrix. Conditions I and III are fulfilled, Condition II breaks down.

Finally, to prove the independence of Condition I, assume that $d_M = 1$ for any matrix M. Conditions II and III will be fulfilled but Condition I fails.

CHAPTER 4

COMPLEX NUMBERS

17. The System of Complex Numbers

During the course of elementary algebra the range of numbers is expanded several times. The beginning student of algebra brings with him from arithmetic a knowledge of positive integers and fractions. Algebra actually begins with the introduction of negative numbers, thus establishing the first of the important number systems, the system of *integers*, which consists of all the positive and all the negative integers and zero, and the broader system of *rational numbers* consisting of all integers and all fractions (both positive and negative).

A further extension of the number realm is the introduction of the irrational numbers. The system consisting of all rational and all irrational numbers is the system of *real numbers.* A university course of mathematical analysis usually contains a rigorous construction of the system of real numbers; however, for our purposes in this course the knowledge of the real numbers that the reader has when he takes up the study of higher algebra will suffice.

Finally, at the very end of the course of elementary algebra, the system of real numbers is extended to the system of *complex numbers.* Of course this system of numbers is less common than the system of real numbers, though actually it possesses many very good properties. In this chapter we recapitulate with sufficient completeness the theory of complex numbers.

Complex numbers are introduced in connection with the following problem. We know that the real numbers do not suffice for us to solve every quadratic equation with real coefficients. The simplest of the quadratics that does not have any roots in the class of real numbers is

$$x^2 + 1 = 0 \tag{1}$$

We will only be interested in this equation for the present. The problem confronting us is: *to extend the system of real numbers to a system of numbers that will supply us with a root for equation* (1).

As construction material for this new system of numbers, let us take advantage of points in a plane. It will be recalled that the depicting of real numbers by points of the straight line (this is based on the fact that we obtain a one-to-one correspondence between the set of all points of the line and the set of all real numbers if, for a given origin of coordinates and a scale unit, every point of the line is associated with an abscissa) is systematically utilized in all divisions of mathematics and is so customary that ordinarily we do not make any distinction between a real number and the point that depicts it.

Thus, *we wish to define a system of numbers correlated with all points in the plane.* Up till now we have not had to add or multiply points of a plane, and so we can define the operations involving points, taking care only that the new system of numbers should possess all the properties intended for it. These definitions, particularly for products, will at first appear to be rather artificial. In Chapter 10, it will be shown however that no other definitions of operations, which at first glance may seem more natural, would give us what we want; that is, they would not result in the construction of an extension of the system of real numbers containing the root of equation (1). It will also be demonstrated there that replacing the points of a plane by any other material would not have led to a system of numbers whose algebraic properties differ from the system of complex numbers which we will construct below.

We have a plane and we choose a rectangular system of coordinates. Let us agree to denote points of the plane by the letters α, β, γ, . . . and write a point α with abscissa a and ordinate b as (a, b), that is, departing somewhat from what is accepted in analytic geometry, and write $\alpha = (a, b)$. If we have points $\alpha = (a, b)$ and $\beta = (c, d)$, then the *sum* of these points will be a point with abscissa $a + c$ and ordinate $b + d$, or

$$(a, b) + (c, d) = (a + c, b + d) \tag{2}$$

For the *product* of the points $\alpha = (a, b)$ and $\beta = (c, d)$ we will have the point with abscissa $ac - bd$ and with ordinate $ad + bc$, or

$$(a, b)(c, d) = (ac - bd, ad + bc) \tag{3}$$

We have thus defined two algebraic operations on the set of all points in the plane. We will show that *these operations have all the basic properties possessed by operations in the system of real numbers or in the system of rational numbers; both are commutative and associative, connected by the distributive law, and have inverse operations—subtraction and division (except by zero).*

Commutativity and associativity of addition are obvious (more precisely, they follow from the corresponding properties of the addition of real numbers) since in the process of adding points of

the plane we separately add their abscissas and their ordinates. The commutativity of multiplication is based on the fact that the points α and β enter the definition of a product symmetrically. The following equations prove associativity of multiplication:

$$[(a, b)\,(c, d)]\,(e, f) = (ac - bd,\ ad + bc)\,(e, f)$$
$$= (ace - bde - adf - bcf,\ acf - bdf + ade + bce),$$

$$(a, b)\,[(c, d)\,(e, f)] = (a, b)\,(ce - df,\ cf + de)$$
$$= (ace - adf - bcf - bde,\ acf + ade + bce - bdf)$$

The distributive law follows from the equations

$$[(a, b) + (c, d)]\,(e, f) = (a + c,\ b + d)\,(e, f)$$
$$= (ae + ce - bf - df,\ af + cf + be + de),$$

$$(a, b)\,(e, f) + (c, d)\,(e, f) = (ae - bf,\ af + be) + (ce - df,\ cf + de)$$
$$= (ae - bf + ce - df,\ af + be + cf + de)$$

Let us examine the inverse operations. If we have the points $\alpha = (a, b)$ and $\beta = (c, d)$, then their difference is a point (x, y) such that

$$(c, d) + (x, y) = (a, b)$$

Whence, by (2),

$$c + x = a, \quad d + y = b$$

Thus, the *difference* of the points $\alpha = (a, b)$ and $\beta = (c, d)$ is the point

$$\alpha - \beta = (a - c,\ b - d) \tag{4}$$

and this difference is defined in unique fashion. In particular, *zero* is the coordinate origin (0, 0); the *opposite* point of $\alpha = (a, b)$ is the point

$$-\alpha = (-a,\ -b) \tag{5}$$

Now, suppose we have the points $\alpha = (a, b)$ and $\beta = (c, d)$, and suppose point β is nonzero; that is, at least one of coordinates c, d is nonzero, and therefore, $c^2 + d^2 \neq 0$. The quotient of α divided by β must be a point (x, y) such that $(c, d)\,(x, y) = (a, b)$. Whence, by (3),

$$cx - dy = a,$$
$$dx + cy = b$$

Solving this system of equations, we obtain

$$x = \frac{ac + bd}{c^2 + d^2}, \quad y = \frac{bc - ad}{c^2 + d^2}$$

Thus, for $\beta \neq 0$ the *quotient* $\frac{\alpha}{\beta}$ exists and is unambiguously defined:

$$\frac{\alpha}{\beta} = \left(\frac{ac+bd}{c^2+d^2}, \quad \frac{bc-ad}{c^2+d^2}\right) \tag{6}$$

Assuming $\beta = \alpha$, we find that in our multiplication of points *unity* is a point (1, 0) lying on the axis of abscissas at a distance 1 to the right of the origin. Also assuming in (6) that $\alpha = 1 = (1, 0)$, we find that for $\beta \neq 0$, the *inverse* of β is

$$\beta^{-1} = \left(\frac{c}{c^2+d^2}, \frac{-d}{c^2+d^2}\right) \tag{7}$$

We have thus constructed a system of numbers that can be depicted by points in the plane, and the operations on these numbers are defined by formulas (2) and (3). This system is called *the system of complex numbers.*

Let us now show that *the system of complex numbers is an extension of the system of real numbers.* To do this, we consider points lying on the axis of abscissas, or points of the form $(a, 0)$; associating a real number a with the point $(a, 0)$, we evidently get a one-to-one correspondence between the set of points under consideration and the set of all the real numbers. Applying to these points formulas (2) and (3), we get

$$(a, 0) + (b, 0) = (a + b, 0),$$
$$(a, 0) \cdot (b, 0) = (ab, 0)$$

i.e., points $(a, 0)$ may be added and multiplied in the same way as the corresponding real numbers. Thus, *the set of points on the axis of abscissas, considered as a part of the system of complex numbers, does not differ in its algebraic properties from the system of real numbers as ordinarily depicted by points on a straight line.* This will enable us, in the future, to equate the point $(a, 0)$ and the real number a, i.e., we will always assume $(a, 0) = a$. In particular, zero (0, 0) and unity (1, 0) of the system of complex numbers turn out to be the real numbers 0 and 1.

We now have to demonstrate that *the complex numbers contain the root of equation* (1), that is, a number whose square is equal to the real number -1. This is the point (0, 1), i.e., a point lying on the axis of ordinates at a distance 1 upwards from the origin. Indeed, using (3), we get

$$(0, 1) \cdot (0, 1) = (-1, 0) = -1$$

Let us agree to denote this point by the letter i, so that $i^2 = -1$.

Finally, let us show *how the customary notation of the complex numbers we have constructed can be obtained.* First find the product

of a real number b and the point i:

$$bi = (b, 0)\cdot(0, 1) = (0, b)$$

This is a point, consequently, which lies on the ordinate axis and has ordinate b; all points of the ordinate axis may be represented by such products. Now if (a, b) is an arbitrary point, then because of the equation

$$(a, b) = (a, 0) + (0, b)$$

we get

$$(a, b) = a + bi$$

In other words we have arrived at the customary notation of complex numbers; the product and sum in the expression $a + bi$ are to be understood, of course, in the sense of operations defined in the system of complex numbers we have constructed.

Now that we have constructed the complex numbers, the reader will have no difficulty in verifying that *all the preceding chapters of this book*—the theory of determinants, the theory of systems of linear equations, the theory of the linear dependence of vectors, and the theory of matrix operations—*carry over without any restrictions from real numbers to all complex numbers.*

Note, in conclusion, that the foregoing construction of the system of complex numbers raises the following question. Is it possible to define addition and multiplication of points in three-dimensional space so that the collection of these points becomes a system of numbers containing within it the system of complex numbers or at least the system of real numbers? This question goes beyond the scope of the present text, but the answer is no.

On the other hand, noting that the addition of complex numbers as defined above actually coincides with the addition of vectors (in a plane) emanating from a coordinate origin (see following section), it is natural to pose the question: is it possible, for a certain n, to define the multiplication of vectors in an n-dimensional real vector space so that, relative to this multiplication and to ordinary addition of vectors, our space proves to be a number system containing the system of real numbers? It may be demonstrated that this cannot be done if we require the fulfillment of all the properties of the operations which are valid in the systems of rational, real and complex numbers. However, if we reject commutativity of multiplication, then such a construction is possible in four-dimensional space; the resulting system of numbers is called the *system of quaternions.* A similar construction is also possible in eight-dimensional space. This yields what is called the *system of Cayley numbers.* In this case, however, we have to give up not only the commutativity of multiplication but also associativity, and replace the latter by a weaker requirement.

18. A Deeper Look at Complex Numbers

In keeping with historically evolved traditions, we call the complex number i the *imaginary unit*, and numbers of the form bi, *pure imaginaries*, although we have no doubt about the existence of such numbers and we can indicate points of the plane (points on the axis of ordinates) which depict these numbers. In the complex notation of the number α, as $\alpha = a + bi$, the a is called the *real part* of α and bi is called its *imaginary part.* A plane with points identified with complex numbers as indicated in Sec. 17 is called the *complex plane.* The axis of abscissas (x-axis) is called the *axis of reals* since its points depict the real numbers, and the axis of ordinates (y-axis) of the complex plane is termed the *axis of imaginaries.*

The addition, multiplication, subtraction and division of complex numbers written in the form $a + bi$ are performed in the following manner, as follows from formulas (2), (4), (3) and (6) of the preceding section:

$$(a + bi) + (c + di) = (a + c) + (b + d)\, i,$$

$$(a + bi) - (c + di) = (a - c) + (b - d)\, i,$$

$$(a + bi)\,(c + di) = (ac - bd) + (ad + bc)i,$$

$$\frac{a+bi}{c+di} = \frac{ac+bd}{c^2+d^2} + \frac{bc-ad}{c^2+d^2}\, i$$

In the addition of complex numbers, add separately the real parts and the imaginary parts. Similarly for subtraction. The formulas for multiplication and division would be too involved if given verbally. The last formula need not be memorized; simply bear in mind that it may be derived by multiplying the numerator and denominator of the given fraction by a number different from the denominator solely in the sign of the imaginary part. Indeed,

$$\frac{a+bi}{c+di} = \frac{(a+bi)\,(c-di)}{(c+di)\,(c-di)} = \frac{(ac+bd)+(bc-ad)\,i}{c^2+d^2} = \frac{ac+bd}{c^2+d^2} + \frac{bc-ad}{c^2+d^2}\, i$$

Examples.

(1) $(2 + 5i) + (1 - 7i) = (2 + 1) + (5 - 7)\, i = 3 - 2i.$

(2) $(3 - 9i) - (7 + i) = (3 - 7) + (-9 - 1)\, i = -4 - 10i.$

(3) $(1 + 2i)\,(3 - i) = [1 \cdot 3 - 2\,(-1)] + [1 \cdot (-1) + 2 \cdot 3]\, i = 5 + 5i.$

(4) $\dfrac{23+i}{3+i} = \dfrac{(23+i)\,(3-i)}{(3+i)\,(3-i)} = \dfrac{70-20i}{10} = 7 - 2i.$

The portrayal of complex numbers as points in a plane result in a natural desire to have a geometric interpretation of the operations involving complex numbers. For addition, this interpretation is simple. Suppose we have the numbers $\alpha = a + bi$ and $\beta = c + di$. Join the corresponding points (a, b) and (c, d) with line segments

to the origin and construct a parallelogram on these segments, as sides, as shown in Fig. 2. The fourth vertex of the parallelogram will obviously be the point $(a + c, b + d)$. Thus, *the addition of complex numbers geometrically is accomplished in accord with the parallelogram rule, which is to say by the rule of addition of vectors emanating from the coordinate origin.* Also, the *number opposite to* $\alpha = a + bi$ *is a point in the complex plane that is symmetric to* α *about the origin* (Fig. 3). This gives the geometric interpretation of subtraction.

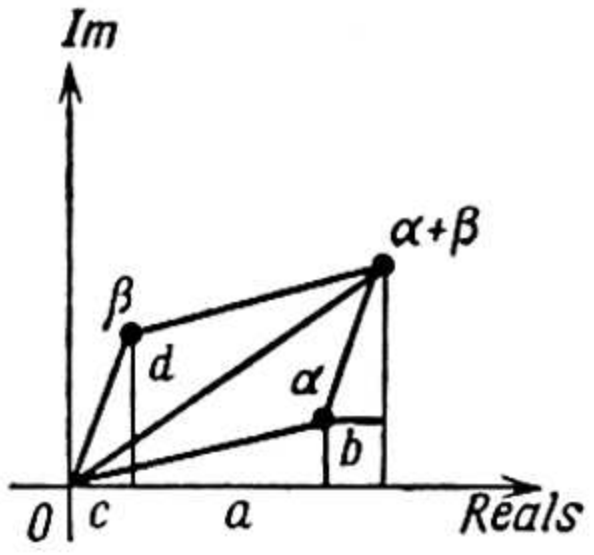

Fig. 2

The geometric meaning of multiplication and division of complex numbers will become clear only after we introduce a new notation for them that differs from that used heretofore. The notation of α as $\alpha = a + bi$ makes use of the Cartesian coordinates of a point corresponding to that number. However, the position of a point in the plane is also completely defined by specifying its polar coordinates: the distance of r from the origin to the point and the angle φ between the positive x-axis (axis of abscissas) and the direction from the origin to the point (Fig. 4).

The number r is a nonnegative real number which is zero only at the point 0. For α on the real axis (that is to say, for α a real

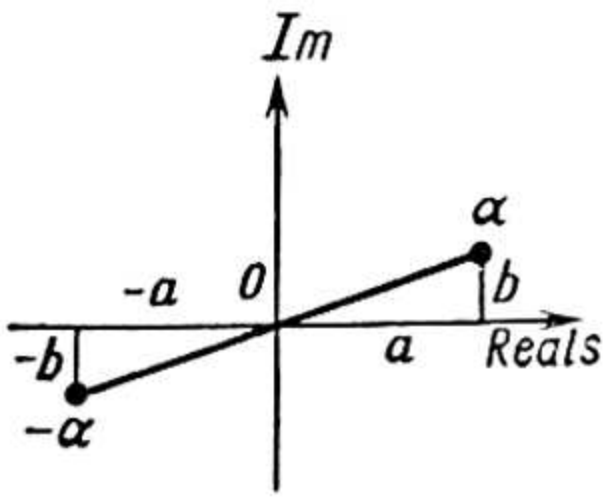

Fig. 3

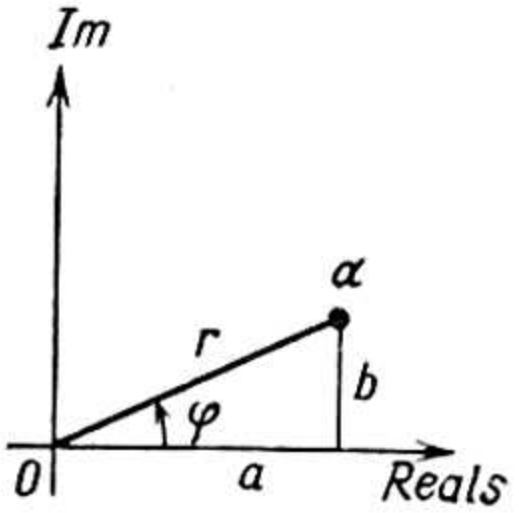

Fig. 4

number), the number r is the absolute value of α; for this reason, for any complex number α, the number r is sometimes called the *absolute value* of α; more often, however, the number r is called the *modulus* of the number α and is denoted by $|\alpha|$.

The angle φ is called the *argument* of the number α and is denoted by $\arg \alpha$ [we thus dispense with the customary names of the polar coordinates of a point: the radius vector and the polar (or vectorial) angle]. The angle φ can take on any real values (positive or negative), the positive angles being reckoned counterclockwise. But if the angles differ by 2π or a multiple of 2π, then the points they depict in the plane will be coincident.

Thus, the argument of a complex number α has an infinity of values differing by integral multiples of the number 2π; from the equality of two complex numbers specified by their moduli and arguments one can only conclude, consequently, that the arguments differ by an integral multiple of 2π, whereas the moduli are the same. It is only for the number 0 that the argument is not defined. However, this number is fully determined by the equation $|0| = 0$.

The argument of a complex number is a natural generalization of the sign of a real number. The argument of a positive real number is zero, the argument of a negative real number is π. There are only two directions out of the origin on the axis of reals and they may be distinguished by two symbols: + and —. Now in the complex plane, there are infinitely many directions issuing from the point 0, and they differ in the angle formed with the positive direction of the real axis.

The Cartesian and polar coordinates of a point are connected by the following relation which holds true for any position of points in the plane:

$$a = r \cos \varphi, \quad b = r \sin \varphi \tag{1}$$

Whence

$$r = + \sqrt{a^2 + b^2} \tag{2}$$

Let us apply formulas (1) to an arbitrary complex number $\alpha = a + bi$:

$$\alpha = a + bi = r \cos \varphi + (r \sin \varphi)\, i$$

or

$$\alpha = r (\cos \varphi + i \sin \varphi) \tag{3}$$

Conversely, let the number $\alpha = a + bi$ admit a notation of the form $\alpha = r_0 (\cos \varphi_0 + i \sin \varphi_0)$, where r_0 and φ_0 are certain real numbers and $r_0 \geqslant 0$. Then $r_0 \cos \varphi_0 = a$, $r_0 \sin \varphi_0 = b$, whence $r_0 = + \sqrt{a^2 + b^2}$, that is, by (2), $r_0 = |\alpha|$. Whence, using (1), we get $\cos \varphi_0 = \cos \varphi$, $\sin \varphi_0 = \sin \varphi$, that is $\varphi_0 = \arg \alpha$. Thus, *any complex number* α *is uniquely defined by* (3), *where* $r = |\alpha|$, $\varphi = \arg \alpha$ (the argument φ being of course defined only to within multiples of 2π). This notation of the number α is called the *trigonometric form* and will be used very often in the sequel.

The numbers

$$\alpha = 3 \left(\cos \frac{\pi}{4} + i \sin \frac{\pi}{4} \right), \quad \beta = \cos \frac{19}{3} \pi + i \sin \frac{19}{3} \pi$$

and

$$\gamma = \sqrt{3} \left[\cos \left(-\frac{\pi}{7} \right) + i \sin \left(-\frac{\pi}{7} \right) \right]$$

are given in trigonometric form; here $|\alpha|=3$, $|\beta|=1$, $|\gamma|=\sqrt{3}$; $\arg\alpha=\frac{\pi}{4}$, $\arg\beta=\frac{19}{3}\pi$, $\arg\gamma=-\frac{\pi}{7}$ $\left(\text{or } \arg\beta=\frac{\pi}{3},\ \arg\gamma=\frac{13}{7}\pi\right)$.

On the other hand, the complex numbers

$$\alpha'=(-2)\left(\cos\frac{\pi}{5}+i\sin\frac{\pi}{5}\right),\quad \beta'=3\left(\cos\frac{2}{3}\pi-i\sin\frac{2}{3}\pi\right),$$

$$\gamma'=2\left(\cos\frac{\pi}{3}+i\sin\frac{3}{4}\pi\right),\qquad \delta'=\sin\frac{3}{4}\pi+i\cos\frac{3}{4}\pi$$

are not given in trigonometric form, although their notations resemble that of (3). In trigonometric form, these numbers look like

$$\alpha'=2\left(\cos\frac{6}{5}\pi+i\sin\frac{6}{5}\pi\right),\quad \beta'=3\left(\cos\frac{4}{3}\pi+i\sin\frac{4}{3}\pi\right),$$

$$\delta'=\cos\frac{7}{4}\pi+i\sin\frac{7}{4}\pi$$

Finding the trigonometric form of a number γ' involves difficulties that are almost always encountered when passing from the customary notation of a complex number to its trigonometric notation and vice versa: with the exception of a few cases, it is impossible to find the exact angle on the basis of given numerical values of the sine and cosine, and it is impossible for a given angle to write the *exact* values of its sine and cosine.

Let the complex numbers α and β be given in trigonometric form: $\alpha = r\ (\cos\varphi + i\sin\varphi)$, $\beta = r'\ (\cos\varphi' + i\sin\varphi')$. Multiplying these numbers together, we get

$$\alpha\beta = [r\ (\cos\varphi + i\sin\varphi)]\cdot[r'\ (\cos\varphi' + i\sin\varphi')]$$

$$= rr'\ (\cos\varphi\cos\varphi' + i\cos\varphi\sin\varphi' + i\sin\varphi\cos\varphi' - \sin\varphi\sin\varphi')$$

or

$$\alpha\beta = rr'\ [\cos(\varphi+\varphi') + i\sin(\varphi+\varphi')] \tag{4}$$

We have the product $\alpha\beta$ written in trigonometric form and so $|\alpha\beta| = rr'$ or

$$|\alpha\beta| = |\alpha|\,|\beta| \tag{5}$$

In words, *the modulus of a product of complex numbers is equal to the product of the moduli of the factors.* Also, $\arg(\alpha\beta) = \varphi + \varphi'$ or

$$\arg(\alpha\beta) = \arg\alpha + \arg\beta \tag{6}$$

The argument of a product of complex numbers is equal to the sum of the arguments of the factors (note that equality here means to within a multiple of 2π). These rules obviously carry over to any finite number of factors. As applied to real numbers, formula (5) yields the familiar property of absolute values of the numbers, and (6), as can readily be verified, turns into the rule of signs in the multiplication of real numbers.

Analogous rules are valid in the case of a quotient. Indeed, let $\alpha = r(\cos\varphi + i\sin\varphi)$, $\beta = r'(\cos\varphi' + i\sin\varphi')$, $\beta \neq 0$; that is $r' \neq 0$. Then

$$\frac{\alpha}{\beta} = \frac{r(\cos\varphi + i\sin\varphi)}{r'(\cos\varphi' + i\sin\varphi')} = \frac{r(\cos\varphi + i\sin\varphi)(\cos\varphi' - i\sin\varphi')}{r'(\cos^2\varphi' + \sin^2\varphi')}$$

$$= \frac{r}{r'}(\cos\varphi\cos\varphi' + i\sin\varphi\cos\varphi' - i\cos\varphi\sin\varphi' + \sin\varphi\sin\varphi')$$

or

$$\frac{\alpha}{\beta} = \frac{r}{r'}[\cos(\varphi - \varphi') + i\sin(\varphi - \varphi')] \tag{7}$$

Whence it follows that $\left|\frac{\alpha}{\beta}\right| = \frac{r}{r'}$ or

$$\left|\frac{\alpha}{\beta}\right| = \frac{|\alpha|}{|\beta|} \tag{8}$$

The modulus of a quotient of two complex numbers is equal to the modulus of the dividend divided by the modulus of the divisor. Also, $\arg\left(\frac{\alpha}{\beta}\right) = \varphi - \varphi'$ or

$$\arg\left(\frac{\alpha}{\beta}\right) = \arg\alpha - \arg\beta \tag{9}$$

The argument of a quotient of two complex numbers is obtained by subtracting the argument of the divisor from the argument of the dividend.

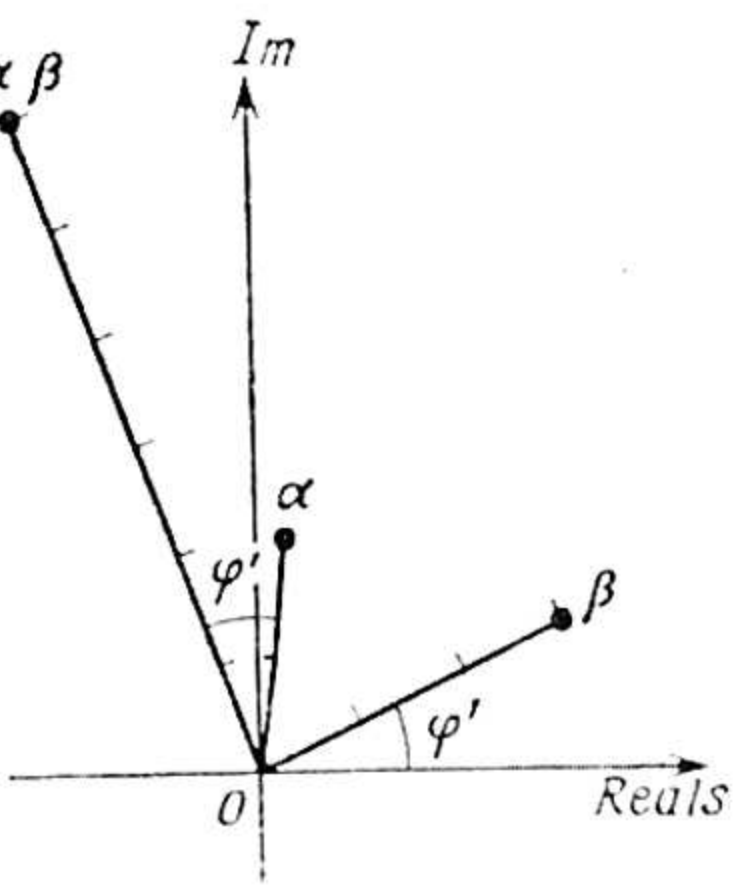

Fig. 5

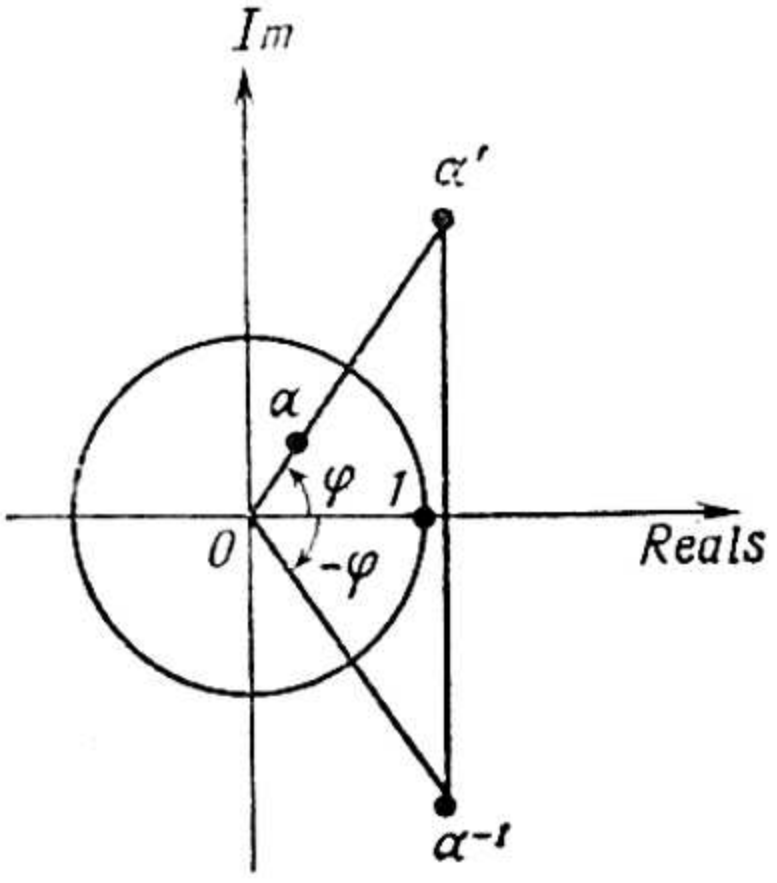

Fig. 6

It is not difficult now to grasp the geometric meaning of multiplication and division. Because of (5) and (6), we get a point depicting the product of the number α by the number $\beta = r'(\cos\varphi' + i\sin\varphi')$ if the vector from 0 to α (Fig. 5) is rotated counterclockwise through an angle $\varphi' = \arg\beta$ and then stretched by a factor $r' = |\beta|$ (for

$0 \leqslant r' < 1$ it will be a compression instead of a dilation). Also, from (7) it follows that for $\alpha = r\ (\cos \varphi + i \sin \varphi) \neq 0$ we have

$$\alpha^{-1} = r^{-1} [\cos (-\varphi) + i \sin (-\varphi)] \tag{10}$$

i.e., $|\alpha^{-1}| = |\alpha|^{-1}$, $\arg(\alpha^{-1}) = -\arg \alpha$. We thus obtain point α^{-1} if from point α we go to point α' at a distance r^{-1} from zero on the same half-line emanating from zero as is point α (Fig. 6),* and then go to a point symmetric to α' about the real axis.

A sum and difference of complex numbers given in trigonometric form cannot be expressed by formulas similar to (4) and (7). However, for the modulus of a sum we have the following important inequalities:

$$|\alpha| - |\beta| \leqslant |\alpha + \beta| \leqslant |\alpha| + |\beta| \tag{11}$$

In words, *the modulus of a sum of two complex numbers is less than or equal to the sum of the moduli of the terms but greater than or equal to the difference of these moduli.* Inequalities (11) follow from the familiar theorem of elementary geometry concerning the sides of a triangle because $|\alpha + \beta|$ is, as we know, equal to the diagonal of a parallelogram with sides $|\alpha|$ and $|\beta|$. Incidentally, the case for points α, β and 0 lying on one straight line requires a special investigation, which we leave to the reader. It is only in this case that the equalities are attained in formulas (11).

From (11), because $\alpha - \beta = \alpha + (-\beta)$ and

$$|-\beta| = |\beta| \tag{12}$$

(this equation follows at the very least from the geometric interpretation of the number $-\beta$), also follow the inequalities

$$|\alpha| - |\beta| \leqslant |\alpha - \beta| \leqslant |\alpha| + |\beta| \tag{13}$$

That is, the same inequalities hold for the modulus of a difference as for the modulus of a sum.

Inequalities (11) might be obtained in the following manner. Let $\alpha = r\ (\cos \varphi + i \sin \varphi)$, $\beta = r'\ (\cos \varphi' + i \sin \varphi')$ and let the trigonometric form of the number $\alpha + \beta$ be $\alpha + \beta = R\ (\cos \psi + i \sin \psi)$. Adding the real and imaginary parts separately, we obtain

$$r \cos \varphi + r' \cos \varphi' = R \cos \psi,$$
$$r \sin \varphi + r' \sin \varphi' = R \sin \psi$$

---

* $|\alpha'| = |\alpha|$ if and only if $|\alpha| = 1$, that is, if the point α lies on the circumference of the *unit circle*. If α lies inside the unit circle, then α' will be outside it, and vice versa. In this way we obviously obtain a one-to-one correspondence between all points of the complex plane outside the unit circle and all *nonzero* points within the unit circle.

Multiplying both sides of the first equation by $\cos\psi$ and both sides of the second by $\sin\psi$ and then adding, we get

$$r(\cos\varphi\cos\psi+\sin\varphi\sin\psi)+r'(\cos\varphi'\cos\psi+\sin\varphi'\sin\psi)$$
$$=R(\cos^2\psi+\sin^2\psi)$$

That is,

$$r\cos(\varphi-\psi)+r'\cos(\varphi'-\psi)=R$$

Whence, since the cosine is never greater than unity, follows the inequality $r+r'\geqslant R$, or $|\alpha|+|\beta|\geqslant|\alpha+\beta|$. On the other hand, $\alpha=(\alpha+\beta)-\beta=(\alpha+\beta)+(-\beta)$, whence, by what has been proved and by virtue of (12),

$$|\alpha|\leqslant|\alpha+\beta|+|-\beta|=|\alpha+\beta|+|\beta|$$

From this, $|\alpha|-|\beta|\leqslant|\alpha+\beta|$.

It is well to note that for complex numbers the concepts of "more than" and "less than" cannot be reasonably defined because these numbers, in contrast to the real numbers, are not located on a straight line, whose points are naturally ordered, but in a plane. For this reason, *complex numbers as such (not their moduli) can never be connected by an inequality sign.*

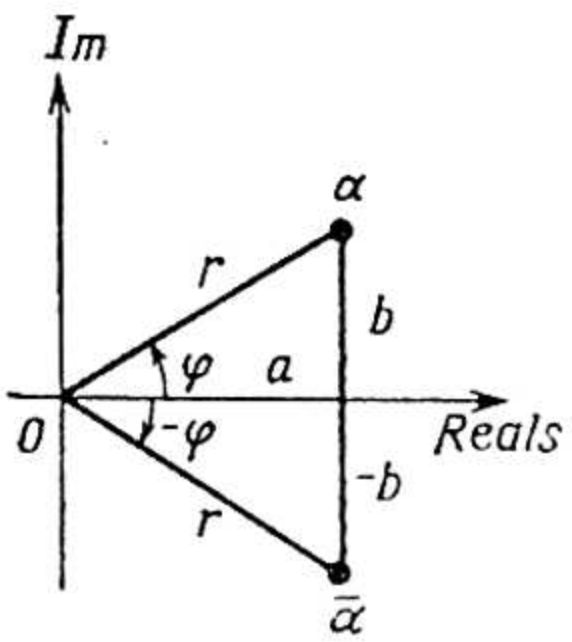

Fig. 7

Conjugate numbers. Suppose we have a complex number $\alpha=a+bi$. The number $a-bi$, which differs from α solely in the sign in front of the imaginary part, is called the *conjugate* of α and is denoted by $\bar{\alpha}$.

It will be recalled that when considering the division of complex numbers we resorted to conjugate numbers but did not introduce that term.

The conjugate number of $\bar{\alpha}$ is obviously α; in other words, we can speak of a pair of conjugate numbers. The real numbers are the only numbers which are conjugate to themselves.

Geometrically, conjugate numbers are points symmetric about the real axis (Fig. 7). Whence follow the equations

$$|\bar{\alpha}|=|\alpha|,\qquad \arg\bar{\alpha}=-\arg\alpha \tag{14}$$

The sum and product of conjugate complex numbers are real numbers. Indeed,

$$\left.\begin{aligned}\alpha+\bar{\alpha}&=2a,\\ \alpha\bar{\alpha}&=a^2+b^2=|\alpha|^2\end{aligned}\right\} \tag{15}$$

The last equation shows that the number $\alpha\bar{\alpha}$ is positive even for $\alpha \neq 0$. In Sec. 24 we will derive a theorem which shows that the property proved here is characteristic of conjugate numbers.

The equation

$$(a - bi) + (c - di) = (a + c) - (b + d)\, i$$

shows that *the conjugate of a sum of two numbers is equal to the sum of the conjugates of the numbers*:

$$\overline{\alpha + \beta} = \bar{\alpha} + \bar{\beta} \tag{16}$$

Similarly, from the equation

$$(a - bi)(c - di) = (ac - bd) - (ad + bc)\, i$$

it follows that *the conjugate of a product is equal to the product of the conjugates of the factors*:

$$\overline{\alpha\beta} = \bar{\alpha} \cdot \bar{\beta} \tag{17}$$

Direct verification also shows the following formulas to be valid:

$$\overline{\alpha - \beta} = \bar{\alpha} - \bar{\beta}, \tag{18}$$

$$\overline{\left(\frac{\alpha}{\beta}\right)} = \frac{\bar{\alpha}}{\bar{\beta}} \tag{19}$$

We will now prove the following assertion: *if a number α is in some way expressed in terms of the complex numbers $\beta_1, \beta_2, \ldots, \beta_n$ by means of addition, multiplication, subtraction and division, then by replacing all the numbers β_k in this expression by their conjugates, we obtain the conjugate of α*; in particular, if α is a real number, it does not change when all the complex numbers β_k are replaced by their conjugates.

We shall prove this assertion by means of induction with respect to n, since for $n = 2$ it follows from formulas (16)-(19).

Let the number α be expressed by the numbers $\beta_1, \beta_2, \ldots, \beta_n$ not necessarily distinct. This expression gives a definite order in which the operations of addition, multiplication, subtraction and division are applied. The last step will be to apply one of these operations to the number γ_1 expressed in terms of the numbers $\beta_1, \beta_2, \ldots, \beta_k$, where $1 \leqslant k \leqslant n - 1$, and to the number γ_2 expressed in terms of the numbers $\beta_{k+1}, \ldots, \beta_n$. By the induction hypothesis, replacement of the numbers $\beta_1, \beta_2, \ldots, \beta_k$ by their conjugates implies a replacement of the number γ_1 by the number $\bar{\gamma}_1$, and a replacement of the numbers $\beta_{k+1}, \beta_{k+2}, \ldots, \beta_n$ by their conjugates implies substitution of γ_2 by $\bar{\gamma}_2$. However, by one of the formulas (16)-(19), the transition from γ_1 and γ_2 to $\bar{\gamma}_1$ and $\bar{\gamma}_2$ converts the number α to $\bar{\alpha}$.

19. Taking Roots of Complex Numbers

Let us now examine the raising of complex numbers to a power and the taking of roots. To raise a number $\alpha = a + bi$ to a positive integral power n, it suffices to apply Newton's binomial theorem to the expression $(a + bi)^n$ (this formula holds true for complex numbers as well, since its proof is based solely on the distributive law) and then take advantage of the equations $i^2 = -1$, $i^3 = -i$, $i^4 = 1$, whence, generally,

$$i^{4k} = 1, \quad i^{4k+1} = i, \quad i^{4k+2} = -1, \quad i^{4k+3} = -i$$

If a number α is given in trigonometric form, then for a positive integral n, there follows from (4) of Sec. 18 the following formula called *De Moivre's formula*:

$$[r(\cos\varphi + i\sin\varphi)]^n = r^n(\cos n\varphi + i\sin n\varphi) \qquad (1)$$

In raising a complex number to a power, raise the modulus to that power and multiply the argument by the exponent. Formula (1) holds true for negative integral exponents as well. Indeed, since $\alpha^{-n} = (\alpha^{-1})^n$, it is sufficient to apply the De Moivre formula to the number α^{-1}, the trigonometric form of which is given by (10), Sec. 18.

Examples.

(1) $i^{37} = i, \quad i^{122} = -1.$

(2) $(2 + 5i)^3 = 2^3 + 3\cdot 2^2\cdot 5i + 3\cdot 2\cdot 5^2 i^2 + 5^3 i^3$
$= 8 + 60i - 150 - 125i = -142 - 65i.$

(3) $\left[\sqrt{2}\left(\cos\frac{\pi}{4} + i\sin\frac{\pi}{4}\right)\right]^4 = (\sqrt{2})^4(\cos\pi + i\sin\pi) = -4.$

(4) $\left[3\left(\cos\frac{\pi}{5} + i\sin\frac{\pi}{5}\right)\right]^{-3}$
$= 3^{-3}\left[\cos\left(-\frac{3}{5}\pi\right) + i\sin\left(-\frac{3}{5}\pi\right)\right] = \frac{1}{27}\left(\cos\frac{7}{5}\pi + i\sin\frac{7}{5}\pi\right).$

A special case of De Moivre's formula, namely, the equation

$$(\cos\varphi + i\sin\varphi)^n = \cos n\varphi + i\sin n\varphi$$

permits finding with ease formulas for the sine and cosine of a multiple angle. Indeed, expanding the left member of this equation by the binomial formula and equating the real and imaginary parts of both sides separately, we obtain

$$\cos n\varphi = \cos^n\varphi - \binom{n}{2}\cos^{n-2}\varphi\cdot\sin^2\varphi + \binom{n}{4}\cos^{n-4}\varphi\cdot\sin^4\varphi - \ldots,$$

$$\sin n\varphi = \binom{n}{1}\cos^{n-1}\varphi\cdot\sin\varphi - \binom{n}{3}\cos^{n-3}\varphi\cdot\sin^3\varphi + \binom{n}{5}\cos^{n-5}\varphi\sin^5\varphi - \ldots$$

Here, $\binom{n}{k}$ is the usual notation for a binomial coefficient:

$$\binom{n}{k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{1\cdot 2\cdot 3\dots k}$$

For $n = 2$ we arrive at the familiar formulas

$$\cos 2\varphi = \cos^2 \varphi - \sin^2 \varphi,$$
$$\sin 2\varphi = 2 \cos \varphi \sin \varphi$$

and for $n = 3$ we obtain the formulas

$$\cos 3\varphi = \cos^3 \varphi - 3 \cos \varphi \sin^2 \varphi,$$
$$\sin 3\varphi = 3 \cos^2 \varphi \sin \varphi - \sin^3 \varphi$$

Extracting roots of complex numbers is a far more difficult task. Let us start with the square root of the number $\alpha = a + bi$. As yet we do not know whether there exists a complex number whose square is equal to α. Let us assume that such a number $u + vi$ exists; that is, using conventional symbols, we can write

$$\sqrt{a + bi} = u + vi$$

From the equation

$$(u + vi)^2 = a + bi$$

it follows that

$$\left.\begin{aligned} u^2 - v^2 &= a, \\ 2uv &= b \end{aligned}\right\} \qquad (2)$$

Squaring both sides of each of the equations of (2) and then adding, we get

$$(u^2 - v^2)^2 + 4u^2v^2 = (u^2 + v^2)^2 = a^2 + b^2$$

whence

$$u^2 + v^2 = + \sqrt{a^2 + b^2}$$

The plus sign is taken because the numbers u and v are real and therefore the left member of the equation is positive. From this equation and from the first of the equations of (2), we get

$$u^2 = \frac{1}{2}\left(a + \sqrt{a^2 + b^2}\right),$$

$$v^2 = \frac{1}{2}\left(-a + \sqrt{a^2 + b^2}\right)$$

Thus, extracting the square roots we get two values for u which differ in sign and also two values for v. All these values will be real since the square roots are extracted from positive numbers for any a and b. The values obtained for u and v cannot be combined in arbitrary fashion, since, by the second equation of (2), the sign of the

product uv must coincide with the sign of b. This yields two possible combinations of values of u and v, that is, two numbers of the form $u + vi$ which can serve as values of the square root of the number α; these numbers differ in sign. An elementary though unwieldy check (squaring the resulting numbers separately for the case $b > 0$ and $b < 0$) shows that the numbers we found are indeed the values of the square root of the number α. Thus, *taking the square root of a complex number is always possible and yields two values which differ in sign.*

In particular, it now becomes possible to extract the square root of a negative real number; the values of this root will be pure imaginaries. Indeed, if $a < 0$ and $b = 0$, then $\sqrt{a^2 + b^2} = -a$, since this root must be positive, but then $u^2 = \frac{1}{2}(a - a) = 0$, that is, $u = 0$, whence $\sqrt{a} = \pm vi$.

Example. Let $\alpha = 21 - 20i$. Then $\sqrt{a^2 + b^2} = \sqrt{441 + 400} = 29$. Therefore, $u^2 = \frac{1}{2}(21 + 29) = 25$, $v^2 = \frac{1}{2}(-21 + 29) = 4$, whence $u = \pm 5$, $v = \pm 2$. The signs of u and v must be different since b is negative, therefore

$$\sqrt{21 - 20i} = \pm(5 - 2i)$$

Attempts to extract higher (than second) roots of complex numbers given in the form $a + bi$ encounter insuperable difficulties. Thus, if we wished to extract the cube root of a number $a + bi$, we would first have to solve some auxiliary cubic equation, which we are as yet unable to do, and which in turn would require, as we shall see in Sec. 38, the extraction of the cube root of a complex number. On the other hand, the trigonometric form is extremely well suited to extracting roots of any degree. Using the trigonometric form we will now exhaust this problem completely.

Let it be required to extract the nth root of a number $\alpha = r(\cos\varphi + i\sin\varphi)$. Let us assume that this is possible and that we get the number $\rho(\cos\theta + i\sin\theta)$, that is

$$[\rho(\cos\theta + i\sin\theta)]^n = r(\cos\varphi + i\sin\varphi) \qquad (3)$$

Then, by De Moivre's formula, $\rho^n = r$, that is $\rho = \sqrt[n]{r}$, where the right member contains a uniquely determined positive value of the nth root of the positive real number r. On the other hand, the argument of the left member of (3) is $n\theta$. We cannot assert, however, that $n\theta$ is equal to φ, since these angles may actually differ by some integral multiple of 2π. Therefore, $n\theta = \varphi + 2k\pi$, where k is an integer, whence

$$\theta = \frac{\varphi + 2k\pi}{n}$$

Conversely, if we take the number $\sqrt[n]{r}\left(\cos\frac{\varphi+2k\pi}{n} + i\sin\frac{\varphi+2k\pi}{n}\right)$, then for any integral k, positive or negative, the nth power of

this number is equal to α. Thus

$$\sqrt[n]{r(\cos\varphi+i\sin\varphi)}=\sqrt[n]{r}\left(\cos\frac{\varphi+2k\pi}{n}+i\sin\frac{\varphi+2k\pi}{n}\right) \qquad (4)$$

Assigning different values to k, we will not always get distinct values of the required root. Indeed, for

$$k=0,\ 1,\ 2,\ \ldots,\ n-1 \qquad (5)$$

we get n values of the root, all distinct, since increasing k by unity implies increasing the argument by $\frac{2\pi}{n}$. Now let k be arbitrary. If $k=nq+r,\ 0\leqslant r\leqslant n-1$, then

$$\frac{\varphi+2k\pi}{n}=\frac{\varphi+2(nq+r)\pi}{n}=\frac{\varphi+2r\pi}{n}+2q\pi$$

In other words, the value of the argument for our k differs from the value of the argument for $k=r$ by a multiple of 2π. We thus obtain the same value of the root as for the value of k equal to r, that is, such as lies in the set (5).

Thus, *extracting the nth root of a complex number α is always possible and yields n distinct values. All values of the nth root lie on a circle of radius $\sqrt[n]{|\alpha|}$ with centre at zero and divide the circle into n equal parts.*

In particular, the nth root of a real number a also has n distinct values, of which two, one, or none will be real, depending on the sign of a and the parity of n.

Examples.

(1) $\beta=\sqrt[3]{2\left(\cos\frac{3}{4}\pi+i\sin\frac{3}{4}\pi\right)}=\sqrt[3]{2}\left(\cos\frac{\frac{3}{4}\pi+2k\pi}{3}+i\sin\frac{\frac{3}{4}\pi+2k\pi}{3}\right)$;

$$k=0:\quad \beta_0=\sqrt[3]{2}\left(\cos\frac{\pi}{4}+i\sin\frac{\pi}{4}\right);$$

$$k=1:\quad \beta_1=\sqrt[3]{2}\left(\cos\frac{11}{12}\pi+i\sin\frac{11}{12}\pi\right);$$

$$k=2:\quad \beta_2=\sqrt[3]{2}\left(\cos\frac{19}{12}\pi+i\sin\frac{19}{12}\pi\right).$$

(2) $\beta=\sqrt{i}=\sqrt{\cos\frac{\pi}{2}+i\sin\frac{\pi}{2}}=\cos\frac{\frac{\pi}{2}+2k\pi}{2}+i\sin\frac{\frac{\pi}{2}+2k\pi}{2}$;

$$\beta_0=\cos\frac{\pi}{4}+i\sin\frac{\pi}{4}=\frac{\sqrt{2}}{2}+i\frac{\sqrt{2}}{2};\quad \beta_1=\cos\frac{5}{4}\pi+i\sin\frac{5}{4}\pi=-\beta_0.$$

(3) $\beta=\sqrt[3]{-8}=\sqrt[3]{8(\cos\pi+i\sin\pi)}=2\left(\cos\frac{\pi+2k\pi}{3}+i\sin\frac{\pi+2k\pi}{3}\right)$;

$$\beta_0=2\left(\cos\frac{\pi}{3}+i\sin\frac{\pi}{3}\right)=1+i\sqrt{3};$$

$$\beta_1=2(\cos\pi+i\sin\pi)=-2;$$

$$\beta_2=2\left(\cos\frac{5\pi}{3}+i\sin\frac{5\pi}{3}\right)=1-i\sqrt{3}.$$

Roots of unity. Of particular importance is the case of extracting the nth root of unity. This root has n values, and, because of the equation $1 = \cos 0 + i \sin 0$, and formula (4), all these values or, as we shall say, all the *nth roots of unity*, are given by the formula

$$\sqrt[n]{1} = \cos\frac{2k\pi}{n} + i\sin\frac{2k\pi}{n}, \qquad k = 0, 1, \ldots, n-1 \tag{6}$$

The real values of the nth root of unity are obtained from formula (6) for the values $k = 0$, and $\frac{n}{2}$, if n is even, and for $k = 0$ if n is odd. In the complex plane, the nth roots of unity are located on the circumference of the unit circle and divide it into n equal arcs: one of the division points is the number 1. From this it follows that those of the nth roots of unity which are not real are situated symmetrically about the real axis (that is, are pairwise conjugate).

The square root of unity has two values: 1 and -1; the fourth root of unity has four values: 1, -1, i and $-i$. It is advisable for what follows to memorize the values of the *cube root of unity*. By (6), the roots are $\cos\frac{2k\pi}{3} + i\sin\frac{2k\pi}{3}$, where $k = 0, 1, 2$; that is, besides unity, the conjugate numbers

$$\left.\begin{aligned} \varepsilon_1 &= \cos\frac{2\pi}{3} + i\sin\frac{2\pi}{3} = -\frac{1}{2} + i\frac{\sqrt{3}}{2}, \\ \varepsilon_2 &= \cos\frac{4\pi}{3} + i\sin\frac{4\pi}{3} = -\frac{1}{2} - i\frac{\sqrt{3}}{2} \end{aligned}\right\} \tag{7}$$

as well.

All values of the nth root of a complex number α may be obtained by multiplying one of these values by all the nth roots of unity. Indeed, let β be one of the values of the nth root of the number α, i.e., $\beta^n = \alpha$ and let ε be an arbitrary value of the nth root of unity, that is, $\varepsilon^n = 1$. Then $(\beta\varepsilon)^n = \beta^n\varepsilon^n = \alpha$. Thus $\beta\varepsilon$ is also one of the values for $\sqrt[n]{\alpha}$. Multiplying β by each of the nth roots of unity, we get n distinct values of the nth root of the number α, that is, all the values of this root.

Example 1. One of the values of the cube root of -8 is -2. The two others are, by (7), the numbers $-2\varepsilon_1 = 1 - i\sqrt{3}$ and $-2\varepsilon_2 = 1 + i\sqrt{3}$ (see Example 3 above).

Example 2. $\sqrt[4]{81}$ has four values: 3, -3, $3i$, $-3i$.

The product of two nth roots of unity is itself an nth root of unity. Indeed, if $\varepsilon^n = 1$ and $\eta^n = 1$, then $(\varepsilon\eta)^n = \varepsilon^n\eta^n = 1$. Also, *the reciprocal of an nth root of unity is itself that root.* Let $\varepsilon^n = 1$. Then from $\varepsilon \cdot \varepsilon^{-1} = 1$ it follows that $\varepsilon^n \cdot (\varepsilon^{-1})^n = 1$, that is, $(\varepsilon^{-1})^n = 1$. Generally, *any power of the nth root of unity is also an nth root of unity.*

Any kth root of unity will also be an lth root of unity for any l that is a multiple of k. Whence it follows that if we regard the entire collection of nth roots of unity, then some of these roots will already be n'-th roots of unity for some n' which are divisors of the number n. However, for any n, there exist nth roots of unity such that they are not any lesser roots of unity. These roots are termed *primitive* nth roots of unity. Their existence follows from formula (6): if the value of a root corresponding to a given value of k is denoted by ε_k (so that $\varepsilon_0 = 1$), then on the basis of De Moivre's formula (1),

$$\varepsilon_1^k = \varepsilon_k$$

Thus, no power of ε_1 less than the nth will be equal to 1, that is $\varepsilon_1 = \cos \frac{2\pi}{n} + i \sin \frac{2\pi}{n}$ is a primitive root.

An nth root ε of unity is a primitive nth root if and only if its powers ε^k, $k = 0, 1, \ldots, n-1$, are distinct, that is, if they exhaust all the nth roots of unity.

Indeed, if all the indicated powers of the number ε are distinct, then ε is obviously an nth primitive root. But if, for example, $\varepsilon^k = \varepsilon^l$ for $0 \leqslant k < l \leqslant n-1$, then $\varepsilon^{l-k} = 1$; that is, because of the inequalities $1 \leqslant l - k \leqslant n - 1$, the root ε will not be primitive.

The number ε_1 found above is not, in the general case, the only primitive nth root. The following theorem is used to find all of these roots.

If ε is a primitive nth root of unity, then the number ε^k is a primitive nth root if and only if k is relatively prime to n.

Let d be the largest common divisor of the numbers k and n. If $d > 1$ and $k = dk'$, $n = dn'$, then

$$(\varepsilon^k)^{n'} = \varepsilon^{kn'} = \varepsilon^{k'n} = (\varepsilon^n)^{k'} = 1$$

that is, the root ε^k is an n'-th root of unity.

On the other hand, let $d = 1$ and at the same time let the number ε^k be an mth root of unity, $1 \leqslant m < n$. Thus,

$$(\varepsilon^k)^m = \varepsilon^{km} = 1$$

Since the number ε is a *primitive* nth root of unity, that is, only its powers with exponents that are multiples of n can be equal to unity, it follows that the number km is a multiple of n. But since $1 \leqslant m < n$, the numbers k and n cannot be relatively prime; this contradicts the assumption.

Thus, the number of primitive nth roots of unity is equal to the number of positive integers k less than n and relatively prime to n. The expression for this number, which is ordinarily denoted by $\varphi(n)$, may be found in any course of number theory.

If p is a prime number, then all these roots except unity itself will be primitive pth roots of unity. On the other hand, i and $-i$ (not 1 and -1) will be among the primitive fourth roots of unity.

CHAPTER 5

POLYNOMIALS AND THEIR ROOTS

20. Operations on Polynomials

The content of the first two chapters of this book—the theory of determinants and the theory of systems of linear equations—grew out of the elementary school course of algebra which proceeds from one equation of the first degree in one unknown to systems of two and three equations of the first degree in two and three unknowns respectively. The second branch of elementary algebra, which in that setting appeared to be the more important one, consisted in passing from first-degree equations in one unknown to an arbitrary quadratic equation again in one unknown, and on to certain special types of equations of the third and fourth degree. This trend is further developed into a very extensive and rich branch of higher algebra devoted to the study of arbitrary equations of the nth degree in one unknown. This division of algebra, which is historically the earlier one, is treated in the present chapter and in some of the later chapters of this text.

The general form of an nth-degree equation (n a positive integer) is

$$a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n = 0 \qquad (1)$$

The coefficients $a_0, a_1, \ldots, a_{n-1}, a_n$ of this equation will be considered to be arbitrary complex numbers and the *leading coefficient* a_0 must be nonzero.

If an equation like (1) is written, it is assumed that we have to solve it. In other words, we have to find numerical values for the unknown x that satisfy the equation, that is, values, which, when substituted in place of the unknown and after all indicated operations have been carried out, reduce the left member of (1) to zero.

However, it is advisable to replace the problem of solving equation (1) by the more general one of studying the left member of this equation:

$$a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n \qquad (2)$$

which is called *a polynomial of degree n in the unknown x*. Remember that only expressions like (2) are polynomials, that is, only the sum of integral nonnegative powers of the unknown x taken with certain numerical coefficients, and not just any sum of monomials, as was the case in elementary algebra. In particular, we will not consider as polynomials expressions which contain negative or fractional powers of the unknown x, such as $2x^2 - \frac{1}{x} + 3$ or $ax^{-3} + bx^{-2} + cx^{-1} + d + ex + fx^2$ or $x^{\frac{1}{2}} + 1$. For brevity, we will denote polynomials by the symbols $f(x)$, $g(x)$, $\varphi(x)$, and so on.

Two polynomials $f(x)$ and $g(x)$ will be considered *equal* (or *identically equal*), $f(x) = g(x)$, only when the coefficients of like powers of the unknown are equal. To be specific, no polynomial can be equal to zero if at least one coefficient is nonzero and for this reason, the equality sign used in the notation (1) of an nth-degree equation has no connection with the above-defined equality of polynomials. The $=$ sign connecting polynomials will always be understood in the sense of an identical equality of these polynomials.

Thus, we look upon the nth-degree polynomial (2) as a certain formal expression, fully defined by the set of its coefficients a_0, a_1, . . ., a_n, where $a_0 \neq 0$. The exact meaning of these words will be explained in Chapter 10. Note that aside from the notation of a polynomial given in (2) (in descending powers of the unknown x), we may use other notations obtainable from (2) by a rearrangement of the terms, say, in ascending powers of the unknown.

There is of course the possibility of regarding the polynomial (2) from the viewpoint of mathematical analysis and of considering it to be a complex function of a complex variable x. However, we have to bear in mind that two functions are considered equal if their values for all values of the variable x are equal. It is clear that two polynomials which are equal in the above-mentioned formal algebraic sense will also be equal as functions of x. The converse will be proved only in Sec. 24 however. After that the algebraic and function-theoretic viewpoints on the concept of a polynomial with numerical coefficients will indeed be equivalent. For the time being, however, each time we have to indicate precisely which sense is meant. In the present section and the two following sections we will look upon the polynomial as a formal-algebraic expression.

Naturally, there are nth-degree polynomials for any natural number n. We consider all possible polynomials of this kind: first-degree (or linear), quadratic, cubic, etc. We will also encounter *polynomials of degree zero*, which are nonzero complex numbers. The number zero will also be taken to be a polynomial. This is the only polynomial whose degree is not defined.

For polynomials with complex coefficients we now define the operations of addition and multiplication. These operations will be

introduced using the pattern of operations involving polynomials with real coefficients, which are familiar from the course of elementary algebra.

If we are given polynomials $f(x)$ and $g(x)$ with complex coefficients (written, for convenience, in ascending powers of x):

$$f(x) = a_0 + a_1x + \ldots + a_{n-1}x^{n-1} + a_nx^n, \quad a_n \neq 0,$$
$$g(x) = b_0 + b_1x + \ldots + b_{s-1}x^{s-1} + b_sx^s, \quad b_s \neq 0$$

and if, for example, $n \geqslant s$, then their *sum* is the polynomial

$$f(x) + g(x) = c_0 + c_1x + \ldots + c_{n-1}x^{n-1} + c_nx^n$$

whose coefficients are obtained by adding the coefficients of the polynomials $f(x)$ and $g(x)$ of like powers of the unknown, i.e.,

$$c_i = a_i + b_i, \qquad i = 0, 1, \ldots, n \tag{3}$$

For $n > s$, the coefficients $b_{s+1}, b_{s+2}, \ldots, b_n$ are to be taken equal to zero. The degree of the sum will be equal to n if n is greater than s, but for $n = s$ it may accidentally prove less than n, namely, when $b_n = -a_n$.

The *product* of polynomials $f(x)$ and $g(x)$ is the polynomial

$$f(x) \cdot g(x) = d_0 + d_1x + \ldots + d_{n+s-1}x^{n+s-1} + d_{n+s}x^{n+s}$$

whose coefficients are determined as follows:

$$d_i = \sum_{k+l=i} a_kb_l, \qquad i = 0, 1, \ldots, n+s-1, n+s \tag{4}$$

That is, the coefficient d_i is the result of multiplying those coefficients of the polynomials $f(x)$ and $g(x)$ whose sum of indices is equal to i and of adding all such products; in particular, $d_0 = a_0b_0$, $d_1 = a_0b_1 + a_1b_0, \ldots, d_{n+s} = a_nb_s$. From the latter equality follows the inequality $d_{n+s} \neq 0$ and therefore *the degree of the product of two polynomials is equal to the sum of the degrees of these polynomials.*

From this it follows that *the product of polynomials different from zero can never be equal to zero.*

What properties do these operations that we have introduced for polynomials have? The commutative and associative laws for addition follow immediately from the validity of these properties for addition of numbers, since we add the coefficients of each power of the unknown separately. Subtraction is possible: the role of zero is played by the number zero, which we have included in the class of polynomials, and the opposite of $f(x)$ will be the polynomial

$$-f(x) = -a_0 - a_1x - \ldots - a_{n-1}x^{n-1} - a_nx^n$$

The commutative law for multiplication follows from the commutativity of multiplication of numbers and from the fact that

in the definition of a product of polynomials, the coefficients of both factors $f(x)$ and $g(x)$ are of an equal status. The associativity of multiplication is proved as follows: if besides the above-written polynomials $f(x)$ and $g(x)$, we are given the polynomial

$$h(x) = c_0 + c_1 x + \ldots + c_{t-1} x^{t-1} + c_t x^t, \qquad c_t \neq 0$$

then the coefficient of x^i, $i = 0, 1, \ldots, n + s + t$, in the product $[f(x)\, g(x)]\, h(x)$ is the number

$$\sum_{j+m=i} \Big(\sum_{k+l=j} a_k b_l \Big) c_m = \sum_{k+l+m=i} a_k b_l c_m$$

and in the product $f(x)\, [g(x)\, h(x)]$ the equivalent number

$$\sum_{k+j=i} a_k \Big(\sum_{l+m=j} b_l c_m \Big) = \sum_{k+l+m=i} a_k b_l c_m$$

Finally, the validity of the distributive law follows from the equation

$$\sum_{k+l=i} (a_k + b_k)\, c_l = \sum_{k+l=i} a_k c_l + \sum_{k+l=i} b_k c_l$$

since the left-hand member of this equation is the coefficient of x^i in the polynomial $[f(x) + g(x)]\, h(x)$ and the right-hand member is the coefficient of the same power of the unknown in the polynomial $f(x)\, h(x) + g(x)\, h(x)$.

It will be noted in the multiplication of polynomials that the role of unity is played by 1, which is regarded as a polynomial of degree zero. On the other hand, *a polynomial $f(x)$ has an inverse* $f^{-1}(x)$,

$$f(x)\, f^{-1}(x) = 1 \tag{5}$$

if and only if $f(x)$ is a polynomial of degree zero. Indeed, if $f(x)$ is a nonzero number a, then the inverse polynomial is the number a^{-1}. But if $f(x)$ has degree $n \geqslant 1$, then the degree of the left side of (5) would not be less than n if the polynomial $f^{-1}(x)$ existed, whereas the polynomial on the right is a polynomial of degree zero.

Consequently, *the multiplication of polynomials has no inverse operation (division).* In this respect, the set of all polynomials with complex coefficients resembles the set of all integers. The analogy may be continued in that polynomials, like the integers, have a division algorithm (with remainder). Elementary algebra describes this algorithm for the case of polynomials with real coefficients. However, since we are dealing with polynomials with complex coefficients, it is well to review once again all the statements and to carry out the proofs.

For any two polynomials $f(x)$ and $g(x)$ we can find polynomials $q(x)$ and $r(x)$ such that

$$f(x) = g(x)\, q(x) + r(x) \tag{6}$$

the degree of $r(x)$ *being less than the degree of* $g(x)$, *or* $r(x) = 0$. *The polynomials* $q(x)$ *and* $r(x)$ *satisfying this condition are defined uniquely.*

Let us first prove the latter half of the theorem. Let there also be polynomials $\bar{q}(x)$ and $\bar{r}(x)$ such that likewise satisfy the equation

$$f(x) = g(x)\,\bar{q}(x) + \bar{r}(x) \tag{7}$$

the degree of $\bar{r}(x)$ again being less than the degree of $g(x)$*. Equating the right sides of (6) and (7), we obtain

$$g(x)\,[q(x) - \bar{q}(x)] = \bar{r}(x) - r(x)$$

The degree of the right side of this equation is less than the degree of $g(x)$, but the degree of the left side would be greater than or equal to the degree of $g(x)$ for $q(x) - \bar{q}(x) \neq 0$. Therefore, it must be true that $q(x) - \bar{q}(x) = 0$, that is, $q(x) = \bar{q}(x)$, but then $r(x) = \bar{r}(x)$, which is what we set out to prove.

We now prove the first part of the theorem. Let the polynomials $f(x)$ and $g(x)$ have degrees n and s, respectively. If $n < s$, then we can put $q(x) = 0$, $r(x) = f(x)$. But if $n \geqslant s$, then we take advantage of the same method by which in elementary algebra we divide polynomials with real coefficients (in descending powers of the unknown). Suppose

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n, \quad a_0 \neq 0,$$
$$g(x) = b_0x^s + b_1x^{s-1} + \ldots + b_{s-1}x + b_s, \quad b_0 \neq 0$$

Setting

$$f(x) - \frac{a_0}{b_0}\,x^{n-s}\,g(x) = f_1(x) \tag{8}$$

we get a polynomial whose degree is less than n. Denote this degree by n_1 and the leading coefficient of the polynomial $f_1(x)$ by a_{10}. Now, if we still have $n_1 \geqslant s$, set

$$f_1(x) - \frac{a_{10}}{b_0}\,x^{n_1-s}\,g(x) = f_2(x) \tag{8$_1$}$$

Denoting by n_2 the degree and by a_{20} the leading coefficient of the polynomial $f_2(x)$, we set

$$f_2(x) - \frac{a_{20}}{b_0}\,x^{n_2-s}\,g(x) = f_3(x) \tag{8$_2$}$$

and so forth.

Since the degrees of the polynomials $f_1(x)$, $f_2(x)$, . . . decrease, $n > n_1 > n_2 > \ldots$, we finally arrive (after a finite number of steps) at the polynomial $f_k(x)$,

$$f_{k-1}(x) - \frac{a_{k-1,0}}{b_0}\,x^{n_{k-1}-s}\,g(x) = f_k(x) \tag{8$_{k-1}$}$$

---

* Or $\bar{r}(x) = 0$. This case will not be specifically stated in the sequel.

the degree of which, n_k, is less than s. Our procedure has come to a halt. Now adding (8), (8_1), . . ., (8_{k-1}), we get

$$f(x)-\left(\frac{a_0}{b_0}x^{n-s}+\frac{a_{10}}{b_0}x^{n_1-s}+\ldots+\frac{a_{k-1,0}}{b_0}x^{n_{k-1}-s}\right)g(x)=f_k(x)$$

Thus, the polynomials

$$q(x)=\frac{a_0}{b_0}x^{n-s}+\frac{a_{10}}{b_0}x^{n_1-s}+\ldots+\frac{a_{k-1,0}}{b_0}x^{n_{k-1}-s},$$

$$r(x)=f_k(x)$$

do indeed satisfy (6), and the degree of $r(x)$ is in fact less than the degree of $g(x)$.

Note that the polynomial $q(x)$ is called the *quotient* obtained from the division of $f(x)$ by $g(x)$, and $r(x)$ is the *remainder*.

From this consideration of the division algorithm, it is easy to establish that if $f(x)$ *and* $g(x)$ *are polynomials with real coefficients, then the coefficients of all polynomials* $f_1(x)$, $f_2(x)$, . . . *and therefore also the coefficients of the quotient* $q(x)$ *and the remainder* $r(x)$ *will be real.*

21. Divisors. Greatest Common Divisor

Suppose we have nonzero polynomials $f(x)$ and $\varphi(x)$ with complex coefficients. If the remainder after dividing $f(x)$ by $\varphi(x)$ is zero, we then say that $f(x)$ is *divisible* (*exactly divisible*) by $\varphi(x)$. Here, the polynomial $\varphi(x)$ is called the *divisor* of the polynomial $f(x)$.

The polynomial $\varphi(x)$ *is a divisor of the polynomial* $f(x)$ *if and only if there exists a polynomial* $\psi(x)$ *such that satisfies the equation*

$$f(x)=\varphi(x)\psi(x) \tag{1}$$

Indeed, if $\varphi(x)$ is a divisor of $f(x)$, then for $\psi(x)$ we should take the quotient of $f(x)$ divided by $\varphi(x)$. Conversely, let there be a polynomial $\psi(x)$ which satisfies (1). From the proof given in the preceding section on the uniqueness of the polynomials $q(x)$ and $r(x)$ which satisfy the equation

$$f(x)=\varphi(x)q(x)+r(x)$$

and the condition that the degree of $r(x)$ be less than the degree of $\varphi(x)$, it follows in our case that the quotient of $f(x)$ by $\varphi(x)$ is equal to $\psi(x)$, and the remainder is zero.

Naturally, if equation (1) holds, then $\psi(x)$ is also a divisor of $f(x)$. Furthermore, it is obvious that the degree of $\varphi(x)$ does not exceed the degree of $f(x)$.

Note that if the polynomial $f(x)$ and its divisor $\varphi(x)$ both have rational or real coefficients, then the polynomial $\psi(x)$ as well will

have rational or, respectively, real coefficients since it is sought by means of the division algorithm. Of course, a polynomial with rational or real coefficients can also have divisors, not all the coefficients of which are rational (or real). This is shown for example by the equation

$$x^2 + 1 = (x - i)(x + i)$$

We indicate a few basic properties of divisibility of polynomials that will be very useful later on.

I. *If* $f(x)$ *is divisible by* $g(x)$, *and* $g(x)$ *is divisible by* $h(x)$, *then* $f(x)$ *is divisible by* $h(x)$.

Since, by hypothesis, $f(x) = g(x)\,\varphi(x)$ and $g(x) = h(x)\,\psi(x)$, it follows that $f(x) = h(x)\,[\psi(x)\,\varphi(x)]$.

II. *If* $f(x)$ *and* $g(x)$ *are divisible by* $\varphi(x)$, *then their sum and difference are also divisible by* $\varphi(x)$.

Indeed, from the equations $f(x) = \varphi(x)\,\psi(x)$ and $g(x) = \varphi(x)\,\chi(x)$ it follows that $f(x) \pm g(x) = \varphi(x)\,[\psi(x) \pm \chi(x)]$.

III. *If* $f(x)$ *is divisible by* $\varphi(x)$, *then the product of* $f(x)$ *by any polynomial* $g(x)$ *is also divisible by* $\varphi(x)$.

True enough, if $f(x) = \varphi(x)\,\psi(x)$, then it follows that $f(x)\,g(x) = \varphi(x)\,[\psi(x)\,g(x)]$.

From II and III we have the following property.

IV. *If each of the polynomials* $f_1(x)$, $f_2(x)$, . . ., $f_k(x)$ *is divisible by* $\varphi(x)$, *then the following polynomial will also be divisible by* $\varphi(x)$:

$$f_1(x)\,g_1(x) + f_2(x)\,g_2(x) + \ldots + f_k(x)\,g_k(x)$$

where $g_1(x)$, $g_2(x)$, . . ., $g_k(x)$ *are arbitrary polynomials.*

V. *Any polynomial* $f(x)$ *is divisible by any polynomial of degree zero.*

Indeed, if $f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n$ and c is an arbitrary number not equal to zero, that is, an arbitrary polynomial of degree zero, then

$$f(x) = c\left(\frac{a_0}{c}x^n + \frac{a_1}{c}x^{n-1} + \ldots + \frac{a_n}{c}\right)$$

VI. *If* $f(x)$ *is divisible by* $\varphi(x)$, *then* $f(x)$ *is divisible by* $c\varphi(x)$ *as well, where* c *is an arbitrary number different from zero.*

From the equation $f(x) = \varphi(x)\,\psi(x)$ follows the equation $f(x) = [c\varphi(x)] \cdot [c^{-1}\psi(x)]$.

VII. *The polynomials* $cf(x)$, $c \neq 0$, *and only such polynomials are divisors of the polynomial* $f(x)$ *that have the same degree as* $f(x)$.

Indeed, $f(x) = c^{-1}\,[cf(x)]$, or $f(x)$ is divisible by $cf(x)$.

If, on the other hand, $f(x)$ is divisible by $\varphi(x)$, and the degrees of $f(x)$ and $\varphi(x)$ coincide, then the degree of the quotient of $f(x)$ by $\varphi(x)$ must be zero, i. e., $f(x) = d\varphi(x)$, $d \neq 0$, whence $\varphi(x) = d^{-1}f(x)$.

From this we get the following property.

VIII. *The polynomials* $f(x)$, $g(x)$ *are simultaneously divisible one by the other if and only if* $g(x) = cf(x)$, $c \neq 0$.

Finally, from VIII and I we get the property

IX. *Any divisor of one of two polynomials* $f(x)$, $cf(x)$, *where* $c \neq 0$, *is a divisor of the other polynomial as well.*

Greatest common divisor. Suppose we have arbitrary polynomials $f(x)$ and $g(x)$. The polynomial $\varphi(x)$ is called the *common divisor* of $f(x)$ and $g(x)$ if it is a divisor of each of them. Property V (see above) shows that the common divisors of the polynomials $f(x)$ and $g(x)$ include all polynomials of degree zero. If there are no other common divisors of these two polynomials, then the polynomials are called *relatively prime.*

But in the general case, the polynomials $f(x)$ and $g(x)$ may have divisors which depend on x; we wish to introduce the concept of the *greatest* common divisor of these polynomials.

It would be inconvenient to take a definition stating that the greatest common divisor of the polynomials $f(x)$ and $g(x)$ is their common divisor of highest degree. On the one hand, as yet we do not know whether $f(x)$ and $g(x)$ have many different common divisors of highest degree which differ not only in a zero-degree factor. In other words, isn't this definition too indeterminate? On the other hand, the reader will recall from elementary arithmetic the problem of finding the greatest common divisor of integers and also that the greatest common divisor 6 of the integers 12 and 18 is not only the greatest among the common divisors of these numbers but is even divisible by any other of their common divisors; the other common divisors of 12 and 18 are 1, 2, 3, −1, −2, −3, −6.

That is why, for polynomials, we have the following definition.

The *greatest common divisor* of the nonzero polynomials $f(x)$ and $g(x)$ is a polynomial $d(x)$, which is their common divisor and, also, is itself divisible by any other common divisor of these polynomials. The greatest common divisor of the polynomials $f(x)$ and $g(x)$ is symbolized as $(f(x), g(x))$.

This definition leaves open the question of whether there exists a greatest common divisor of any polynomials $f(x)$ and $g(x)$. We will now answer this question in the affirmative. At the same time we will give a practical method for finding the greatest common divisor of the given polynomials. Quite naturally, we cannot carry over the procedure used for finding the greatest common divisor of integers, since we do not as yet have anything analogous in polynomials to the decomposition of an integer into a product of prime factors. However, for integers there is also another method called the *algorithm of successive division*, or *Euclid's algorithm*. This procedure is quite applicable to polynomials.

Euclid's algorithm for polynomials consists in the following. Let there be given the polynomials $f(x)$ and $g(x)$. We divide $f(x)$ by $g(x)$ and obtain, generally speaking, a remainder $r_1(x)$. Then divide $g(x)$ by $r_1(x)$ and get a remainder $r_2(x)$, divide $r_1(x)$ by $r_2(x)$ and so on. Since the degrees of the remainders decrease continuously, there will come a time in this sequence of divisions when the division is exact and the procedure terminates. *The remainder* $r_k(x)$ *which divides exactly the preceding remainder* $r_{k-1}(x)$ *is the greatest common divisor of the polynomials* $f(x)$ *and* $g(x)$.

By way of proof, let us write the contents of the preceding paragraph in the form of a chain of equations:

$$\left.\begin{aligned} f(x) &= g(x)\,q_1(x) + r_1(x),\\ g(x) &= r_1(x)\,q_2(x) + r_2(x),\\ r_1(x) &= r_2(x)\,q_3(x) + r_3(x),\\ &\cdots\cdots\cdots\cdots\cdots\\ r_{k-3}(x) &= r_{k-2}(x)\,q_{k-1}(x) + r_{k-1}(x),\\ r_{k-2}(x) &= r_{k-1}(x)\,q_k(x) + r_k(x),\\ r_{k-1}(x) &= r_k(x)\,q_{k+1}(x) \end{aligned}\right\} \qquad (2)$$

The last equation shows that $r_k(x)$ is a divisor of $r_{k-1}(x)$. Whence it follows that both terms of the right member of the second last equation are divisible by $r_k(x)$ and so $r_k(x)$ is also a divisor of $r_{k-2}(x)$. Rising upwards in this fashion, we find that $r_k(x)$ is also a divisor of $r_{k-3}(x), \ldots, r_2(x), r_1(x)$. Whence, by virtue of the second equation, it will follow that $r_k(x)$ is a divisor of $g(x)$ and therefore, on the basis of the first equation, a divisor of $f(x)$ as well. Thus, $r_k(x)$ is a common divisor of $f(x)$ and $g(x)$.

Now let us take an arbitrary common divisor $\varphi(x)$ of the polynomials $f(x)$ and $g(x)$. Since the left side and the first term of the right side of the first of the equations (2) are divisible by $\varphi(x)$, it follows that $r_1(x)$ is also divisible by $\varphi(x)$. Passing to the second and successive equations, we find in the same way that the polynomials $r_2(x), r_3(x), \ldots$ are divisible by $\varphi(x)$. Finally, if it is proved that $r_{k-2}(x)$ and $r_{k-1}(x)$ are divisible by $\varphi(x)$, then from the second last equation we find that $r_k(x)$ is divisible by $\varphi(x)$. Thus, $r_k(x)$ is indeed the greatest common divisor of $f(x)$ and $g(x)$.

We have thus proved that any two polynomials have a greatest common divisor, and we have a procedure for computing it. This method shows that *if the polynomials* $f(x)$ *and* $g(x)$ *both have rational or real coefficients, then the coefficients of their greatest common divisor will also be rational or real*, though of course these polynomials may also have other divisors, not all coefficients of which are rational (real). Thus, the polynomials with rational coefficients

$$f(x) = x^3 - 3x^2 - 2x + 6, \quad g(x) = x^3 + x^2 - 2x - 2$$

have as greatest common divisor the polynomial with rational coefficients $x^2 - 2$, though they have a common divisor $x - \sqrt{2}$, not all the coefficients of which are rational.

If $d(x)$ is the greatest common divisor of the polynomials $f(x)$ and $g(x)$, then, as Properties VIII and IX (see above) show, for the greatest common divisor of these polynomials we could also choose the polynomial $cd(x)$, where c is an arbitrary number different from zero. In other words, *the greatest common divisor of two polynomials is only determined to within a factor of degree zero.* In view of this fact we can agree that the leading coefficient of the greatest common divisor of two polynomials will always be considered equal to unity. Using this condition, we can say that *two polynomials are relatively prime if and only if their greatest common divisor is unity.* Indeed, for the greatest common divisor of two relatively prime polynomials we can take any number different from zero; but multiplying it by the inverse, we get unity.

Example. Find the greatest common divisor of the polynomials

$$f(x) = x^4 + 3x^3 - x^2 - 4x - 3, \qquad g(x) = 3x^3 + 10x^2 + 2x - 3$$

Applying Euclid's algorithm to polynomials with integral coefficients, we can (to avoid fractional coefficients) multiply the dividend or reduce the divisor by any nonzero number (this may be done either at the start or at any other time in the division). Quite naturally, this will distort the quotient, but the remainders that interest us will only acquire some factor of zero degree, which as we know is quite permissible when seeking the greatest common divisor.

We divide $f(x)$ by $g(x)$ but first multiply $f(x)$ by 3:

$$\begin{array}{r|l} 3x^3+10x^2+2x-3 & x+1 \\ \hline 3x^4+9x^3-3x^2-12x-9 & \\ 3x^4+10x^3+2x^2-3x & \\ \hline -x^3-5x^2-9x-9 & \end{array}$$

(multiply by -3)

$$\begin{array}{r} 3x^3+15x^2+27x+27 \\ 3x^3+10x^2+2x-3 \\ \hline 5x^2+25x+30 \end{array}$$

Thus, the first remainder, after dividing by 5, will be $r_1(x) = x^2 + 5x + 6$. We divide the polynomial $g(x)$ by it:

$$\begin{array}{r|l} x^2+5x+6 & 3x-5 \\ \hline 3x^3+10x^2+2x-3 & \\ 3x^3+15x^2+18x & \\ \hline -5x^2-16x-3 & \\ -5x^2-25x-30 & \\ \hline 9x+27 & \end{array}$$

The second remainder, after dividing by 9, is thus $r_2(x) = x + 3$. Since

$$r_1(x) = r_2(x)(x + 2)$$

it follows that $r_2(x)$ will be the last remainder which exactly divides the preceding remainder. It will consequently be the desired greatest common divisor:

$$(f(x), g(x)) = x + 3$$

We use the Euclidean algorithm to prove the following theorem.

If $d(x)$ is the greatest common divisor of the polynomials $f(x)$ and $g(x)$, then it is possible to find polynomials $u(x)$ and $v(x)$ such that

$$f(x)\,u(x) + g(x)\,v(x) = d(x) \tag{3}$$

If the degrees of the polynomials $f(x)$ and $g(x)$ exceed zero, we can then take it that the degree of $u(x)$ is less than the degree of $g(x)$, and the degree of $v(x)$ is less than the degree of $f(x)$.

The proof rests on the equations (2). If we take into consideration that $r_k(x) = d(x)$ and if we put $u_1(x) = 1$, $v_1(x) = -q_k(x)$, then the second last of the equations (2) yields

$$d(x) = r_{k-2}(x)\,u_1(x) + r_{k-1}(x)\,v_1(x)$$

Substituting the expression $r_{k-1}(x)$ in terms of $r_{k-3}(x)$ and $r_{k-2}(x)$ from the preceding equation (2), we get

$$d(x) = r_{k-3}(x)\,u_2(x) + r_{k-2}(x)\,v_2(x)$$

where, obviously, $u_2(x) = v_1(x)$, $v_2(x) = u_1(x) - v_1(x)\,q_{k-1}(x)$. Continuing upwards through the equations of (2), we finally arrive at the equation (3) being proved.

To prove the second assertion of the theorem, assume that the polynomials $u(x)$ and $v(x)$ which satisfy (3) have already been found, but that, say, the degree of $u(x)$ is greater than or equal to the degree of $g(x)$. Divide $u(x)$ by $g(x)$:

$$u(x) = g(x)\,q(x) + r(x)$$

where the degree of $r(x)$ is less than the degree of $g(x)$, and substitute this expression into (3). We get the equation

$$f(x)\,r(x) + g(x)\,[v(x) + f(x)\,q(x)] = d(x)$$

The degree of the factor of $f(x)$ is now less than the degree of $g(x)$. The degree of the polynomial in square brackets will in turn be less than the degree of $f(x)$, since otherwise the degree of the second summand in the left-hand member would not be less than the degree of the product $g(x)\,f(x)$, and since the degree of the first summand is less than the degree of this product, the entire left side would have a degree greater than or equal to the degree of $g(x)\,f(x)$, whereas the polynomial $d(x)$ is definitely (given our assumptions) of lower degree.

This proves the theorem. At the same time we see that if the polynomials $f(x)$ and $g(x)$ have rational or real coefficients, then we can also choose the polynomials $u(x)$ and $v(x)$, which satisfy (3), so that their coefficients are rational or real.

Example. Find the polynomials $u(x)$ and $v(x)$ which satisfy (3) for

$$f(x) = x^3 - x^2 + 3x - 10, \quad g(x) = x^3 + 6x^2 - 9x - 14$$

Apply Euclid's algorithm to these polynomials. This time, when performing the divisions, we cannot allow for any distortion of the quotients since these quotients are used to find the polynomials $u(x)$ and $v(x)$. We obtain the following system of equations:

$$f(x) = g(x) + (-7x^2 + 12x + 4),$$

$$g(x) = (-7x^2 + 12x + 4)\left(-\frac{1}{7}x - \frac{54}{49}\right) + \frac{235}{49}(x - 2),$$

$$-7x^2 + 12x + 4 = (x - 2)(-7x - 2)$$

Whence it follows that $(f(x), g(x)) = x - 2$ and that

$$u(x) = \frac{7}{235}x + \frac{54}{235}, \quad v(x) = -\frac{7}{235}x - \frac{5}{235}$$

Applying the above-proved theorem to relatively prime polynomials, we get the following result.

The polynomials $f(x)$ and $g(x)$ are relatively prime if and only if it is possible to find polynomials $u(x)$ and $v(x)$ such that satisfy the equation

$$f(x)\,u(x) + g(x)\,v(x) = 1 \tag{4}$$

Proceeding from this result, we can prove a number of simple but important theorems on relatively prime polynomials:

(a) *If a polynomial $f(x)$ is relatively prime to each of the polynomials $\varphi(x)$ and $\psi(x)$, then it is also relatively prime to their product.*

Indeed, by (4), there are polynomials $u(x)$ and $v(x)$ such that

$$f(x)\,u(x) + \varphi(x)\,v(x) = 1$$

Multiplying this equation by $\psi(x)$, we get

$$f(x)\,[u(x)\,\psi(x)] + [\varphi(x)\,\psi(x)]\,v(x) = \psi(x)$$

whence it follows that any common divisor $f(x)$ and $\varphi(x)\,\psi(x)$ would also be a divisor of $\psi(x)$; however, it is given that $(f(x), \psi(x)) = 1$.

(b) *If the product of the polynomials $f(x)$ and $g(x)$ is divisible by $\varphi(x)$ but $f(x)$ and $\varphi(x)$ are relatively prime, then $g(x)$ is divisible by $\varphi(x)$.*

This is true since by multiplying the equation

$$f(x)\,u(x) + \varphi(x)\,v(x) = 1$$

by $g(x)$, we get

$$[f(x)\,g(x)]\,u(x) + \varphi(x)\,[v(x)\,g(x)] = g(x)$$

Both terms of the left-hand member of this equation are divisible by $\varphi(x)$; hence $g(x)$ is divisible by $\varphi(x)$.

(c) *If the polynomial $f(x)$ is divisible by each of the polynomials $\varphi(x)$ and $\psi(x)$, which are relatively prime, then $f(x)$ is also divisible by their product.*

Indeed, $f(x) = \varphi(x)\,\overline{\varphi}(x)$ so that the product on the right is divisible by $\psi(x)$. Therefore, by (b), $\overline{\varphi}(x)$ is divisible by $\psi(x)$, $\overline{\varphi}(x) = \psi(x)\,\overline{\psi}(x)$, whence $f(x) = [\varphi(x)\,\psi(x)]\,\overline{\psi}(x)$.

The definition of greatest common divisor may be extended to the case of any finite system of polynomials: the *greatest common divisor* of the polynomials $f_1(x), f_2(x), \ldots, f_s(x)$ is that common divisor of these polynomials which is divisible by any other common divisor of these polynomials. The existence of a greatest common divisor for any finite system of polynomials is a consequence of the following theorem, which also provides a procedure for calculating it.

The greatest common divisor of the polynomials $f_1(x), f_2(x), \ldots \ldots, f_s(x)$ is equal to the greatest common divisor of the polynomial $f_s(x)$ and the greatest common divisor of the polynomials $f_1(x), f_2(x), \ldots \ldots, f_{s-1}(x)$.

Indeed, for $s = 2$ the theorem is obvious. We thus assume that for the case $s - 1$ it holds true, that is, in particular, we have already proved the existence of the greatest common divisor $d(x)$ of the polynomials $f_1(x), f_2(x), \ldots, f_{s-1}(x)$. Denote by $\bar{d}(x)$ the greatest common divisor of the polynomials $d(x)$ and $f_s(x)$. It will obviously be a common divisor of all the given polynomials. On the other hand, any other common divisor of these polynomials will also be a divisor of $d(x)$ and, for this reason, of $\bar{d}(x)$ as well.

In particular, the system of polynomials $f_1(x), f_2(x), \ldots \ldots, f_s(x)$ is called *relatively prime* if only zero-degree polynomials are the common divisors of these polynomials; that is to say, if their greatest common divisor is unity. If $s > 2$, then these polynomials may not be pairwise relatively prime. Thus, the system of polynomials

$$f(x) = x^3 - 7x^2 + 7x + 15, \quad g(x) = x^2 - x - 20,$$

$$h(x) = x^3 + x^2 - 12x$$

is relatively prime, although

$$(f(x), g(x)) = x - 5, \quad (f(x), h(x)) = x - 3, \quad (g(x), h(x)) = x + 4$$

The reader will readily obtain a generalization of the above-proved theorems (a) to (c) on relatively prime polynomials to the case of any finite number of polynomials.

22. Roots of Polynomials

We have already (Sec. 20) dealt with the values of a polynomial when we spoke of the function-theoretic approach to the concept of a polynomial. Let us recall the definition.

If

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n \tag{1}$$

is some polynomial and c is a number, then the number

$$f(c) = a_0c^n + a_1c^{n-1} + \ldots + a_n$$

obtained by replacing in (1) the unknown x by the number c and by subsequent performance of all indicated operations, is called the *value of the polynomial* $f(x)$ *for* $x = c$. Quite naturally, if $f(x) = g(x)$ in the sense of an algebraic equality of polynomials as defined in Sec. 20, then $f(c) = g(c)$ for any c.

It is also easy to see that if

$$\varphi(x) = f(x) + g(x), \quad \psi(x) = f(x)\,g(x)$$

then

$$\varphi(c) = f(c) + g(c), \quad \psi(c) = f(c)\,g(c)$$

In other words, the addition and multiplication of polynomials defined in Sec. 20 become—from the function-theoretic approach to polynomials—the addition and multiplication of functions, to be understood in the sense of addition and multiplication of the appropriate values of these functions.

If $f(c) = 0$, that is, the polynomial $f(x)$ vanishes when the number c is substituted in place of the unknown, then c is termed a *root* of the polynomial $f(x)$ [or of the equation $f(x) = 0$]. It will now be shown that this concept applies completely to the theory of divisibility of polynomials, which was the topic of discussion in the preceding section.

If we divide the polynomial $f(x)$ by an arbitrary polynomial of degree one (or, as we shall say from now on, by a *linear polynomial*), then the remainder will either be a polynomial of degree zero, or zero, which is to say some number r. The following theorem allows us to find this remainder without performing the division itself when we divide by a polynomial of the form $x - c$.

The remainder resulting from the division of a polynomial $f(x)$ *by a linear polynomial* $x - c$ *is equal to the value* $f(c)$ *of* $f(x)$ *for* $x = c$.

Let

$$f(x) = (x - c)\,q(x) + r$$

Taking the values of both sides of this equation when $x = c$, we get

$$f(c) = (c - c)\,q(c) + r = r$$

which proves the theorem.

An exceedingly important corollary follows from this fact.

The number c is a root of the polynomial $f(x)$ if and only if $f(x)$ is divisible by $x - c$.

On the other hand, if $f(x)$ is divisible by some linear polynomial $ax + b$, then evidently it is also divisible by the polynomial $x - \left(-\frac{b}{a}\right)$, that is, by a polynomial of the form $x - c$. Thus, *finding the roots of a polynomial $f(x)$ is equivalent to finding its linear divisors.*

In view of the foregoing, it is of interest to examine the method of dividing a polynomial $f(x)$ by a linear binomial $x - c$, which is simpler than the general algorithm for dividing polynomials. This method is called the *Horner method.* Let

$$f(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \ldots + a_n \tag{2}$$

and let

$$f(x) = (x - c)\,q(x) + r \tag{3}$$

where

$$q(x) = b_0x^{n-1} + b_1x^{n-2} + b_2x^{n-3} + \ldots + b_{n-1}$$

Comparing the coefficients of like powers of x in (3), we get

$$\begin{aligned} a_0 &= b_0, \\ a_1 &= b_1 - cb_0, \\ a_2 &= b_2 - cb_1, \\ &\ldots\ldots\ldots \\ a_{n-1} &= b_{n-1} - cb_{n-2}, \\ a_n &= r - cb_{n-1} \end{aligned}$$

From this it follows that $b_0 = a_0$, $b_k = cb_{k-1} + a_k$, $k = 1, 2, \ldots, n - 1$, that is, the coefficient b_k is obtained by multiplying the preceding coefficient b_{k-1} by c and by adding the corresponding coefficient a_k; finally, $r = cb_{n-1} + a_n$, that is, the remainder r, which as we know is equal to $f(c)$, is also obtained by the same rule. Thus, the remainder and the coefficients of the quotient may be successively obtained by computations of the same type, which can be arranged in a scheme, as the following examples demonstrate.

Example 1. Divide $f(x) = 2x^5 - x^4 - 3x^3 + x - 3$ by $x - 3$.

Form an array in which the coefficients of the polynomial $f(x)$ are located above the bar, and the corresponding coefficients of the quotient and the remainder (computed successively) are located below the bar; on the left is the value of c in the given example:

3	2	−1	−3	0	1	−3
	2	3·2 − 1 = 5	3·5 − 3 = 12	3·12 + 0 = 36	3·36 + 1 = 109	3·109 − 3 = 324

Thus, the desired quotient will be

$$q(x) = 2x^4 + 5x^3 + 12x^2 + 36x + 109$$

and the remainder will be $r = f(3) = 324$.

Example 2. Divide $f(x)=x^4-8x^3+x^2+4x-9$ by $x+1$.

$$\begin{array}{c|ccccc} & 1 & -8 & 1 & 4 & -9 \\ \hline -1 & 1 & -9 & 10 & -6 & -3 \end{array}$$

The quotient will therefore be

$$q(x)=x^3-9x^2+10x-6$$

and the remainder $r=f(-1)=-3$.

These examples show that *the Horner method may also be used for quick computation of the value of a polynomial for a given value of the unknown.*

Multiple roots. If c is a root of the polynomial $f(x)$, i.e., $f(c)=0$, then $f(x)$ is, as we know, divisible by $x-c$. It may turn out that the polynomial $f(x)$ is not only divisible by the first power of the linear binomial $x-c$, but by higher powers of it as well. In any case, there will be a natural number k such that $f(x)$ is exactly divisible by $(x-c)^k$, but is not divisible by $(x-c)^{k+1}$. Therefore,

$$f(x)=(x-c)^k\varphi(x)$$

where the polynomial $\varphi(x)$ is no longer divisible by $x-c$, that is, does not have c as its root. The number k is called the *multiplicity* of the root c in the polynomial $f(x)$, and the root c is the *k-fold root* of this polynomial. If $k=1$, then we say that the root c is *simple*.

The concept of a multiple root is closely related to the concept of the derivative of a polynomial. However, we are studying polynomials with any complex coefficients and for this reason we cannot simply take advantage of the concept of a derivative as introduced in the course of mathematical analysis. What follows is to be regarded as a definition of the derivative of a polynomial which is independent of that given in the course of analysis.

Suppose we have an nth-degree polynomial

$$f(x)=a_0x^n+a_1x^{n-1}+\ldots+a_{n-1}x+a_n$$

with arbitrary complex coefficients. Its *derivative* (*first derivative*) is a polynomial of degree $n-1$:

$$f'(x)=na_0x^{n-1}+(n-1)a_1x^{n-2}+\ldots+2a_{n-2}x+a_{n-1}$$

The derivative of a polynomial of degree zero and the derivative of zero are taken to be equal to zero. The derivative of the first derivative is called the *second derivative* of the polynomial $f(x)$ and is denoted by $f''(x)$, etc. It is obvious that

$$f^{(n)}(x)=n!a_0$$

and therefore $f^{(n+1)}(x)=0$; i.e., the $(n+1)$th *derivative of a polynomial of degree n is equal to zero.*

In our case of polynomials with complex coefficients, we cannot make use of the properties of a derivative as proved in the course of analysis for polynomials with real coefficients; we have to prove these properties once again using the definition of a derivative given above. We are interested in the following properties, which are called formulas for differentiating a sum and a product:

$$(f(x) + g(x))' = f'(x) + g'(x) \tag{4}$$

$$(f(x) \cdot g(x))' = f(x)\, g'(x) + f'(x)\, g(x), \tag{5}$$

These formulas can easily be verified, incidentally, by direct computation, by taking for $f(x)$ and $g(x)$ two arbitrary polynomials and applying the above definition of a derivative; we leave this to the reader.

Formula (5) can readily be extended to the case of a product of any finite number of factors and therefore we can in the ordinary fashion derive a formula for the derivative of a power

$$(f^k(x))' = kf^{k-1}(x)\, f'(x) \tag{6}$$

Our aim will be to prove the following theorem.

If the number c is a k-fold root of the polynomial $f(x)$, then for $k > 1$ it will be the $(k-1)$-fold root of the first derivative of this polynomial; but if $k = 1$, then c will not be a root of $f'(x)$.

Let

$$f(x) = (x - c)^k \varphi(x), \qquad k \geqslant 1 \tag{7}$$

where $\varphi(x)$ is no longer divisible by $x - c$. Differentiating equation (7), we get

$$\begin{aligned} f'(x) &= (x - c)^k \varphi'(x) + k(x - c)^{k-1} \varphi(x) \\ &= (x - c)^{k-1} [(x - c) \varphi'(x) + k\varphi(x)] \end{aligned}$$

The first term of the sum in the square brackets is divisible by $x - c$, the second is not divisible by $x - c$; therefore, the whole sum is not divisible by $x - c$. Taking into account that the quotient of $f(x)$ by $(x - c)^{k-1}$ is uniquely defined, we find that $(x - c)^{k-1}$ is the highest power of the binomial $x - c$ which divides the polynomial $f'(x)$. The proof is complete.

Applying this theorem several times, we find that *the k-fold root of polynomial $f(x)$ is the $(k - s)$-fold root in the sth derivative of this polynomial ($k \geqslant s$) and for the first time will not be a root of the kth derivative of $f(x)$.*

23. Fundamental Theorem

In examining the roots of polynomials in the preceding section we did not pose the question of whether every polynomial possesses roots. We know that there are polynomials with real coefficients

that do not have real roots; $x^2 + 1$ is such a polynomial. It might be expected that there are polynomials which do not have roots even in the class of complex numbers, particularly if we consider polynomials with arbitrary complex coefficients. If this were the case, then the system of complex numbers would require a further extension. Actually, however, the following **fundamental theorem of the algebra of complex numbers** is valid.

Every polynomial of degree at least one with arbitrary numerical coefficients has at least one root, which in the general case is complex.

This theorem is one of the greatest attainments of the whole of mathematics and finds application in the most diverse spheres of science. In particular, it is the starting point of everything in the theory of polynomials with numerical coefficients and for this reason it was once called (and sometimes still is) the "fundamental theorem of higher algebra". Actually, however, the fundamental theorem is not purely algebraic. All its proofs—and since Gauss first proved the theorem at the end of the eighteenth century a very large number have been found—are forced, in one degree or another, to make use of the so-called topological properties of the real and complex numbers, that is properties associated with continuity.

In the proof which we now give, the polynomial $f(x)$ with complex coefficients will be regarded as a complex function of a complex variable x. Thus, x can assume any complex values, or, taking into account the mode of constructing complex numbers given in Sec. 17, the variable x ranges over the *complex plane*. The values of the function $f(x)$ will also be complex numbers. We may consider that these values are plotted on a second complex plane, as in the case of real functions of a real variable where the values of the independent variable are plotted on one number line (axis of abscissas) while the values of the function are plotted on the other line (axis of ordinates).

The definition of a continuous function as given in the course of mathematical analysis is carried over to functions of a complex variable (in the formulation of the definition, absolute values are replaced by moduli).

Namely, the complex function $f(x)$ of a complex variable x is *continuous at a point* x_0 if for any positive real number ε there is a positive real number δ such that no matter what (generally speaking, complex) the increment h, the modulus of which satisfies the inequality $|h| < \delta$, the inequality

$$|f(x_0 + h) - f(x_0)| < \varepsilon$$

holds true. A function $f(x)$ is called *continuous* if it is continuous at all points x_0 at which it is defined, that is, if $f(x)$ is a polynomial on the entire complex plane.

The polynomial $f(x)$ is a continuous function of the complex variable x.

The proof of this theorem could be given as it is in the course of mathematical analysis, namely, by showing that the sum and the product of continuous functions are themselves continuous and then noting that a function which is constantly equal to one and the same complex number is continuous. However, we shall take a different approach.

We first prove the particular case of the theorem when the constant term of the polynomial $f(x)$ is zero; and we will only prove the continuity of $f(x)$ at the point $x_0 = 0$. In other words, we will prove the following lemma (in place of h we write x).

Lemma 1. *If the constant term of the polynomial $f(x)$ is zero*

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x$$

that is, $f(0) = 0$, then for any $\varepsilon > 0$ there is a $\delta > 0$ such that for all x for which $|x| < \delta$ it is true that $|f(x)| < \varepsilon$.

Indeed, let

$$A = \max(|a_0|, |a_1|, \ldots, |a_{n-1}|)$$

We are already given the number ε. Let us show that if for the number δ we take

$$\delta = \frac{\varepsilon}{A+\varepsilon} \tag{1}$$

then it will satisfy the required conditions.

Indeed,

$$\begin{aligned} |f(x)| &\leqslant |a_0||x|^n + |a_1||x|^{n-1} + \ldots + |a_{n-1}||x| \\ &\leqslant A(|x|^n + |x|^{n-1} + \ldots + |x|) \end{aligned}$$

that is,

$$|f(x)| \leqslant A\,\frac{|x| - |x|^{n+1}}{1-|x|}$$

Since $|x| < \delta$ and, by (1), $\delta < 1$, it follows that

$$\frac{|x| - |x|^{n+1}}{1-|x|} < \frac{|x|}{1-|x|}$$

and therefore

$$|f(x)| < \frac{A|x|}{1-|x|} < \frac{A\delta}{1-\delta} = \frac{A\dfrac{\varepsilon}{A+\varepsilon}}{1-\dfrac{\varepsilon}{A+\varepsilon}} = \varepsilon$$

which completes the proof.

Let us now derive the following formula. Suppose we have the polynomial

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n$$

with arbitrary complex coefficients. Substitute in place of x the sum $x + h$, where h is the second unknown. Using the binomial theorem, expand each of the powers $(x + h)^k$, $k \leqslant n$, in the right-hand member and collect terms with like powers of h. This yields (as the reader can readily verify) the equation

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!} f''(x) + \ldots + \frac{h^n}{n!} f^{(n)}(x)$$

In other words, we prove *Taylor's formula*, which gives the expansion of $f(x + h)$ in powers of the "increment" h.

The continuity of an arbitrary polynomial $f(x)$ at any point x_0 is now proved as follows. By Taylor's formula,

$$f(x_0 + h) - f(x_0) = c_1 h + c_2 h^2 + \ldots + c_n h^n = \varphi(h)$$

where

$$c_1 = f'(x_0), \quad c_2 = \frac{1}{2!} f''(x_0), \quad \ldots, \quad c_n = \frac{1}{n!} f^{(n)}(x_0)$$

The polynomial $\varphi(h)$ in the unknown h is a polynomial without a constant term, and so, by Lemma 1, for any $\varepsilon > 0$ there is a $\delta > 0$ such that for $|h| < \delta$ it is true that $|\varphi(h)| < \varepsilon$, i.e.,

$$|f(x_0 + h) - f(x_0)| < \varepsilon$$

which completes the proof.

From the inequality

$$||f(x_0 + h)| - |f(x_0)|| \leqslant |f(x_0 + h) - f(x_0)|$$

based on formula (13), Sec. 18, and from the continuity, just proved, of a polynomial there follows the *continuity of the modulus* $|f(x)|$ of the polynomial $f(x)$; this modulus is obviously a real nonnegative function of the complex variable x.

We shall now prove the lemmas that are used in the proof of the fundamental theorem.

Lemma on the modulus of the highest-degree term. *If we have an nth-degree polynomial,* $n \geqslant 1$,

$$f(x) = a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \ldots + a_n$$

with arbitrary complex coefficients and if k *is any positive real number, then for sufficiently large (in modulus) values of the unknown* x *the inequality*

$$|a_0 x^n| > k\,|a_1 x^{n-1} + a_2 x^{n-2} + \ldots + a_n| \qquad (2)$$

is true, that is, the modulus of the highest-degree term is greater than the modulus of the sum of all the remaining terms; it is an arbitrary number of times greater.

Indeed, let A be the largest of the moduli of the coefficients $a_1, a_2, \ldots, a_n$:

$$A = \max (|a_1|, |a_2|, \ldots, |a_n|)$$

Then (see, in Sec. 18, the properties of the moduli of a sum and a product of complex numbers)

$$|a_1x^{n-1} + a_2x^{n-2} + \ldots + a_n| \leqslant |a_1||x|^{n-1} + |a_2||x|^{n-2}$$

$$+ \ldots + |a_n| \leqslant A(|x|^{n-1} + |x|^{n-2} + \ldots + 1) = A\frac{|x|^n - 1}{|x| - 1}$$

Assuming $|x| > 1$, we get

$$\frac{|x|^n - 1}{|x| - 1} < \frac{|x|^n}{|x| - 1}$$

whence

$$|a_1x^{n-1} + a_2x^{n-2} + \ldots + a_n| < A\frac{|x|^n}{|x| - 1}$$

Thus, inequality (2) will be fulfilled if x satisfies the condition $|x| > 1$ and also the inequality

$$kA\frac{|x|^n}{|x| - 1} \leqslant |a_0x^n| = |a_0||x|^n$$

that is, if

$$|x| \geqslant \frac{kA}{|a_0|} + 1 \tag{3}$$

Since the right side of inequality (3) is greater than 1, it may be asserted that, for values of x satisfying this inequality, inequality (2) holds true. This proves the lemma.

Lemma on the increase of the modulus of a polynomial. *For every polynomial $f(x)$ of degree not less than unity with complex coefficients, and for any arbitrary large positive real number M, it is possible to find a positive real number N such that for $|x| > N$ it will be true that $|f(x)| > M$.*

Let

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n$$

By formula (11), Sec. 18,

$$|f(x)| = |a_0x^n + (a_1x^{n-1} + \ldots + a_n)|$$

$$\geqslant |a_0x^n| - |a_1x^{n-1} + \ldots + a_n| \tag{4}$$

Apply the lemma on the modulus of the highest-degree term, putting $k = 2$; there is a number N_1 such that for $|x| > N_1$ it is true that

$$|a_0x^n| > 2|a_1x^{n-1} + \ldots + a_n|$$

whence

$$|a_1x^{n-1} + \ldots + a_n| < \frac{1}{2}|a_0x^n|$$

that is, by (4),

$$|f(x)|>|a_0x^n|-\frac{1}{2}|a_0x^n|=\frac{1}{2}|a_0x^n|$$

The right side of this inequality is greater than M for

$$|x|>N_2=\sqrt[n]{\frac{2M}{|a_0|}}$$

Thus, for $|x|>N=\max(N_1, N_2)$ we have $|f(x)|>M$.

The meaning of this lemma may be illustrated geometrically (we will frequently make use of this illustration). Suppose that at every point x_0 of the complex plane a perpendicular is erected whose length (for the given scale unit) is equal to the modulus of the value of the polynomial $f(x)$ at this point, that is, is equal to $|f(x_0)|$. The endpoints of the perpendiculars will, in view of the above-proved continuity of the modulus of a polynomial, constitute some continuous curved surface situated above the complex plane.

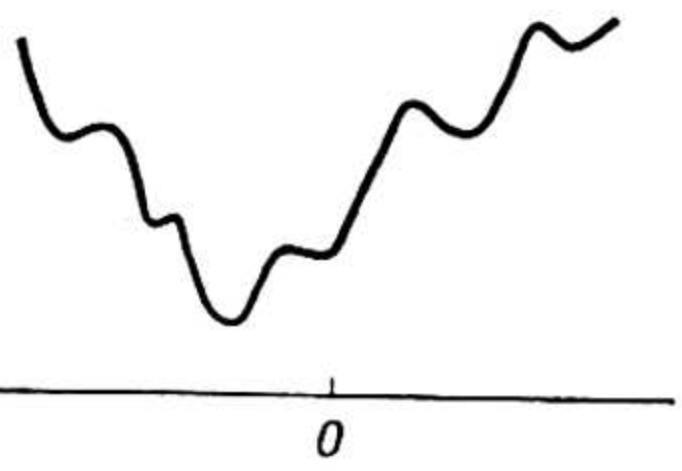

Fig. 8

The lemma on the increase of the modulus of a polynomial shows that as $|x_0|$ increases this surface recedes from the complex plane, though quite naturally the recession is not in the least monotonic. Fig. 8 is a schematic view of the line of intersection of this surface with a plane perpendicular to the complex plane and passing through the point O.

The following lemma plays a crucial role in the proof.

D'Alembert's lemma. *If for $x=x_0$ the polynomial $f(x)$ of degree n, $n\geqslant 1$, does not vanish, $f(x_0)\neq 0$ and therefore $|f(x_0)|>0$, then it is possible to find an increment h (complex in the general case) such that*

$$|f(x_0+h)|<|f(x_0)|$$

If the increment h is as yet arbitrary, then Taylor's formula yields

$$f(x_0+h)=f(x_0)+hf'(x_0)+\frac{h^2}{2!}f''(x_0)+\ldots+\frac{h^n}{n!}f^{(n)}(x_0)$$

By hypothesis, x_0 is not a root of $f(x)$. It may, however, fortuitously be a root of $f'(x)$ and perhaps also of certain other higher derivatives. Let the kth derivative ($k\geqslant 1$) be the first that does not have x_0 for a root, that is,

$$f'(x_0)=f''(x_0)=\ldots=f^{(k-1)}(x_0)=0, \quad f^{(k)}(x_0)\neq 0$$

Such a k exists since if a_0 is the leading coefficient of the polynomial $f(x)$, then

$$f^{(n)}(x_0)=n!a_0\neq 0$$

Thus,

$$f(x_0+h)=f(x_0)+\frac{h^k}{k!}f^{(k)}(x_0)+\frac{h^{k+1}}{(k+1)!}f^{(k+1)}(x_0)+\ldots+\frac{h^n}{n!}f^{(n)}(x_0)$$

Some of the numbers $f^{(k+1)}(x_0), \ldots, f^{(n-1)}(x_0)$ may also be zero, but this does not affect our reasoning in any way.

Dividing both sides of the equation by $f(x_0)$, which, by hypothesis, is different from zero, and introducing the notation

$$c_j=\frac{f^{(j)}(x_0)}{j!\,f(x_0)}, \qquad j=k,\ k+1,\ \ldots,\ n$$

we get

$$\frac{f(x_0+h)}{f(x_0)}=1+c_kh^k+c_{k+1}h^{k+1}+\ldots+c_nh^n$$

or, because $c_k \neq 0$,

$$\frac{f(x_0+h)}{f(x_0)}=(1+c_kh^k)+c_kh^k\left(\frac{c_{k+1}}{c_k}h+\ldots+\frac{c_n}{c_k}h^{n-k}\right)$$

Taking moduli, we get

$$\left|\frac{f(x_0+h)}{f(x_0)}\right|\leqslant|1+c_kh^k|+|c_kh^k|\left|\frac{c_{k+1}}{c_k}h+\ldots+\frac{c_n}{c_k}h^{n-k}\right| \qquad (5)$$

Up to this point we have not made any assumptions concerning the increment h. Now we will *choose* h: we choose the modulus and the argument separately. We choose the modulus of h in the following manner. Since

$$\frac{c_{k+1}}{c_k}h+\ldots+\frac{c_n}{c_k}h^{n-k}$$

is a polynomial in h without the constant term, it follows by Lemma 1 $\left(\text{setting } \varepsilon=\frac{1}{2}\right)$ that there is a δ_1 such that for $|h|<\delta_1$ it will be true that

$$\left|\frac{c_{k+1}}{c_k}h+\ldots+\frac{c_n}{c_k}h^{n-k}\right|<\frac{1}{2} \qquad (6)$$

On the other hand, for

$$|h|<\delta_2=\sqrt[k]{|c_k|^{-1}}$$

we have

$$|c_kh^k|<1 \qquad (7)$$

Assume that the modulus of h is chosen in accord with the inequality

$$|h|<\min(\delta_1,\ \delta_2) \qquad (8)$$

Then, because of (6), inequality (5) becomes the strict inequality

$$\left|\frac{f(x_0+h)}{f(x_0)}\right|<|1+c_kh^k|+\frac{1}{2}|c_kh^k| \qquad (9)$$

We will use Condition (7) later on.

To choose the argument of h we require that the number $c_k h^k$ be a negative real number. In other words,

$$\arg (c_k h^k) = \arg c_k + k \arg h = \pi$$

whence

$$\arg h = \frac{\pi - \arg c_k}{k} \tag{10}$$

In this choice of h, the number $c_k h^k$ will differ from its absolute value in sign:

$$c_k h^k = -| c_k h^k |$$

and therefore, using inequality (7),

$$| 1 + c_k h^k | = | 1 - | c_k h^k | | = 1 - | c_k h^k |$$

Thus, for h chosen on the basis of the Conditions (8) and (10), inequality (9) takes the form

$$\left| \frac{f(x_0+h)}{f(x_0)} \right| < 1 - | c_k h^k | + \frac{1}{2} | c_k h^k | = 1 - \frac{1}{2} | c_k h^k |$$

and all the more so

$$\left| \frac{f(x_0+h)}{f(x_0)} \right| = \frac{| f(x_0+h) |}{| f(x_0) |} < 1$$

whence it follows that

$$| f(x_0 + h) | < | f(x_0) |$$

This completes the proof of d'Alembert's lemma.

Using the geometric interpretation given earlier, we can describe d'Alembert's lemma in the following fashion. Given that $| f(x_0) | > 0$. This means that the length of the perpendicular erected to the complex plane at point x_0 is nonzero. Then, by d'Alembert's lemma, there is a point $x_1 = x_0 + h$ such that $| f(x_1) | < | f(x_0) |$; that is, the perpendicular at the point x_1 will be shorter than at the point x_0 and, consequently, the surface formed by the endpoints of the perpendiculars will at this new point be somewhat closer to the complex plane. As the proof of the lemma shows, the modulus of h may be taken as small as we wish; in other words, the point x_1 may be chosen arbitrarily close to the point x_0. However, we will not take advantage of this remark in the future.

Obviously, the roots of the polynomial $f(x)$ will be those complex numbers (or those points of the complex plane) at which the surface formed by the endpoints of the perpendiculars touches this plane. It is impossible to prove the existence of such points by relying on d'Alembert's lemma alone. Indeed, using this lemma it is possible to find an infinite sequence of points $x_0, x_1, x_2, \ldots$, such that

$$| f(x_0) | > | f(x_1) | > | f(x_2) | > \ldots \tag{11}$$

However, it does not follow from this that there exists a point $\bar{x}$ such that $f(\bar{x}) = 0$, all the more so that the decreasing sequence of positive real numbers (11) does not necessarily have to tend to zero.

The considerations that follow are based on a theorem from the theory of functions of a complex variable that generalizes the Weierstrass theorem, which is familiar to the reader from the course of mathematical analysis. It has to do with real functions of a complex variable, that is with functions of a complex variable that take on only real values. The modulus of a polynomial is an instance of such functions. For the sake of simplicity, in the statement of this theorem we will speak *about a closed circle* E to be understood as a circle in the complex plane with all boundary points included.

If a real function $g(x)$ *of a complex variable* x *is continuous at all points of a closed circle* E, *then there exists in* E *a point* x_0 *such that for all* x *in* E *the inequality* $g(x) > g(x_0)$ *holds. Consequently, the point* x_0 *is the minimum point of* $g(x)$ *in the circle* E.

The proof of this theorem is given in all courses of complex function theory and so we omit it.

We confine ourselves to the case when the function $g(x)$ is non-negative at all points of E—only this case is of interest to us—and will try to explain this theorem geometrically with the aid of the illustration used earlier. Draw a perpendicular of length $g(x_0)$ at every point x_0 of the circle E. The endpoints of these perpendiculars constitute a piece of a continuous curved surface, and due to the closed nature of the circle E the existence of minimum points of this piece of surface is geometrically clear. This illustration does not of course take the place of a proof of the theorem.

We can now take up the **proof of the fundamental theorem itself.** Let there be given a polynomial $f(x)$ of degree n, $n \geqslant 1$. If its constant term is a_n, then obviously $f(0) = a_n$. Let us apply to our polynomial the lemma on the increase of the modulus of a polynomial, assuming $M = |f(0)| = |a_n|$. Consequently, there exists an N such that for $|x| > N$ it will be true that $|f(x)| > |f(0)|$. It is then obvious that the above-indicated generalization of the Weierstrass theorem is applicable to the function $|f(x)|$ for any choice of the closed circle E. For E we take a closed circle of radius N with centre at 0. Let point x_0 be the minimum point of $|f(x)|$ in E; whence, in particular, it follows that $|f(x_0)| \leqslant |f(0)|$.

It is easy to see that x_0 *will actually serve as minimum point of* $|f(x)|$ *over the entire complex plane*: if the point x' lies outside E, then $|x'| > N$ and for this reason

$$|f(x')| > |f(0)| \geqslant |f(x_0)|$$

Whence it follows, finally, that $f(x_0) = 0$, or that x_0 *serves as a root of* $f(x)$. If we had had $f(x_0) \neq 0$, then, by d'Alembert's lemma, there

would be a point x_1 such that $|f(x_1)| < |f(x_0)|$. However, this contradicts the property of point x_0 that we have just established.

Another proof of the fundamental theorem will be given in Sec. 55.

24. Corollaries to the Fundamental Theorem

Suppose we have a polynomial of degree n, $n \geqslant 1$,

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n \qquad (1)$$

with arbitrary complex coefficients. We again regard it as a formal-algebraic expression which is fully defined by the set of its coefficients. The fundamental theorem on the existence of a root that was proved in the preceding section permits asserting the existence of a complex or real root α_1 of $f(x)$. Therefore, the polynomial $f(x)$ has the factorization

$$f(x) = (x - \alpha_1)\,\varphi(x)$$

The coefficients of the polynomial $\varphi(x)$ are again real or complex numbers, and therefore $\varphi(x)$ has a root α_2 whence

$$f(x) = (x - \alpha_1)(x - \alpha_2)\,\psi(x)$$

Continuing in similar fashion, we arrive—after a finite number of steps—at a factorization of the nth-degree polynomial $f(x)$ into a product of n linear factors,

$$f(x) = a_0(x - \alpha_1)(x - \alpha_2)\ldots(x - \alpha_n) \qquad (2)$$

The coefficient a_0 is a result of the following: if we had a coefficient b on the right of (2), then after removal of parentheses the highest-degree term of the polynomial $f(x)$ would be of the form bx^n, though in reality, by (1), it is the term a_0x^n. Therefore, $b = a_0$.

For the polynomial $f(x)$, expansion (2) is, to within the order of the factors, a unique expansion of that type.

Let there be yet another expansion

$$f(x) = a_0(x - \beta_1)(x - \beta_2)\ldots(x - \beta_n) \qquad (3)$$

From (2) and (3) follows the equation

$$(x - \alpha_1)(x - \alpha_2)\ldots(x - \alpha_n) = (x - \beta_1)(x - \beta_2)\ldots(x - \beta_n) \qquad (4)$$

If the root α_i were different from all β_j, $j = 1, 2, \ldots, n$, then, substituting α_i in place of the unknown into (4), we would have zero on the left and a nonzero number on the right. Thus, *every root α_i is equal to some root β_j and conversely.*

From this it does not yet follow that the expansions (2) and (3) are coincident. Indeed, there may be equal roots among the roots α_i, $i = 1, 2, \ldots, n$. For example, let s of these roots be

equal to α_1 and, on the other hand, let there be t roots equal to the root α_1 among the roots β_j, $j = 1, 2, \ldots, n$. We have to show that $s = t$.

Since the degree of a product of polynomials is equal to the sum of the degrees of the factors, the product of two polynomials different from zero cannot be zero. It then follows that if two products of polynomials are equal, *then a common multiple can be cancelled from both sides of the equation*: if

$$f(x)\,\varphi(x) = g(x)\,\varphi(x)$$

and $\varphi(x) \neq 0$, then from

$$[f(x) - g(x)]\,\varphi(x) = 0$$

it follows that

$$f(x) - g(x) = 0$$

that is,

$$f(x) = g(x)$$

Let us apply this to equation (4). If, for instance, $s > t$, then by cancelling the factor $(x - \alpha_1)^t$ out of both sides of (4), we arrive at an equation whose left side contains the factor $x - \alpha_1$ and whose right side does not contain it. But it has been shown that this is a contradiction, which proves the uniqueness of the expansion (2) of the polynomial $f(x)$.

Collecting like factors, we can write (2) as

$$f(x) = a_0 (x - \alpha_1)^{k_1} (x - \alpha_2)^{k_2} \ldots (x - \alpha_l)^{k_l} \qquad (5)$$

where

$$k_1 + k_2 + \ldots + k_l = n$$

It is now assumed that there are no equal roots among the roots $\alpha_1, \alpha_2, \ldots, \alpha_l$.

We will prove that *the number* k_i *of* (5), $i = 1, 2, \ldots, l$, *is the multiplicity of the root* α_i *in the polynomial* $f(x)$. Indeed, if this multiplicity is equal to s_i, then $k_i \leqslant s_i$. However, let $k_i < s_i$. By virtue of the definition of multiplicity of a root of $f(x)$, we have the expansion

$$f(x) = (x - \alpha_i)^{s_i} \varphi(x)$$

Replacing in this expansion the factor $\varphi(x)$ by its factorization into linear factors, we would get for $f(x)$ a factorization into linear factors that is definitely different from (2); in other words, it would contradict the above-proved uniqueness of the expansion.

We have thus proved the following important result.

Any polynomial $f(x)$ *of degree* n, $n \geqslant 1$, *with arbitrary numerical coefficients has* n *roots if each of the roots is counted to the degree of its multiplicity.*

Note that this theorem holds true for $n = 0$ as well, since a polynomial of zero degree quite naturally has no roots. This theorem is not applicable only to the polynomial 0, which has no degree and is equal to zero for any value of x. We use this last remark in the proof of the following theorem.

If the polynomials $f(x)$ and $g(x)$ whose degrees do not exceed n have equal values for more than n distinct values of the unknown, then $f(x) = g(x)$.

Indeed, the polynomial $f(x) - g(x)$ has, by hypothesis, more roots than n, and since its degree does not exceed n, the equation $f(x) - g(x) = 0$ must be true.

Thus, taking into account that there is an infinity of different numbers, we can assert that *for any two distinct polynomials $f(x)$ and $g(x)$ there will be values c of the unknown x such that $f(c) \neq g(c)$*. Such c may be found not only among the complex numbers but also among the real numbers, rational numbers and even the integers.

Consequently, two polynomials with numerical coefficients having different coefficients of at least one power of the unknown x will be distinct complex functions of the complex variable x. Finally, this proves *the equivalence, for polynomials with numerical coefficients, of the two definitions of equality of polynomials given in Sec. 20: the algebraic definition and the function-theoretic definition.*

The theorem proved above permits us to assert that a *polynomial whose degree does not exceed n is completely determined by its values for any distinct values of the unknown whose number is greater than n.* Can these values of the polynomial be specified arbitrarily? If we assume that the values of a polynomial are given for $n + 1$ distinct values of the unknown, then the answer is yes: *there always exists a polynomial of degree not higher than n which takes on preassigned values for $n + 1$ specified distinct values of the unknown.*

Indeed, let it be necessary to construct a polynomial of degree not higher than n, which, for values of the unknown $a_1, a_2, \ldots \ldots, a_{n+1}$ (assumed distinct), takes on, respectively, the values $c_1, c_2, \ldots, c_{n+1}$. The polynomial will be

$$f(x) = \sum_{i=1}^{n+1} \frac{c_i (x-a_1) \ldots (x-a_{i-1}) (x-a_{i+1}) \ldots (x-a_{n+1})}{(a_i-a_1) \ldots (a_i-a_{i-1}) (a_i-a_{i+1}) \ldots (a_i-a_{n+1})} \qquad (6)$$

Indeed, its degree does not exceed n and the value of $f(a_i)$ is equal to c_i.

Formula (6) is called the *Lagrange interpolation formula*. The term "interpolation" is due to the fact that, using this formula and knowing the values of the polynomial at $n + 1$ points, it is possible to compute its values at all other points.

Vieta's formulas. Let there be given a polynomial $f(x)$ of degree n with leading coefficient 1,

$$f(x) = x^n + a_1x^{n-1} + a_2x^{n-2} + \ldots + a_{n-1}x + a_n \qquad (7)$$

and let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be its roots (counting multiplicities). Then $f(x)$ has the following expansion:

$$f(x) = (x - \alpha_1)(x - \alpha_2) \ldots (x - \alpha_n)$$

Multiplying out the parentheses on the right, and then collecting like terms and comparing the resulting coefficients with the coefficients of (7), we get the following equations, called *Vieta's formulas*, which express the coefficients of a polynomial in terms of its roots:

$$\begin{aligned}
a_1 &= -(\alpha_1 + \alpha_2 + \ldots + \alpha_n),\\
a_2 &= \alpha_1\alpha_2 + \alpha_1\alpha_3 + \ldots + \alpha_1\alpha_n + \alpha_2\alpha_3 + \ldots + \alpha_{n-1}\alpha_n,\\
a_3 &= -(\alpha_1\alpha_2\alpha_3 + \alpha_1\alpha_2\alpha_4 + \ldots + \alpha_{n-2}\alpha_{n-1}\alpha_n),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
a_{n-1} &= (-1)^{n-1}(\alpha_1\alpha_2 \ldots \alpha_{n-1}\\
&\qquad + \alpha_1\alpha_2 \ldots \alpha_{n-2}\alpha_n + \ldots + \alpha_2\alpha_3 \ldots \alpha_n,\\
a_n &= (-1)^n \alpha_1\alpha_2 \ldots \alpha_n
\end{aligned}$$

Thus, the right side of the kth equation, $k = 1, 2, \ldots, n$, contains a sum of all possible products of k roots taken with the plus sign or minus sign, according as k is even or odd.

For $n = 2$, these formulas become the familiar (from elementary algebra) relationship between the roots and the coefficients of a quadratic polynomial. For $n = 3$, that is, for a cubic polynomial, these formulas take the form

$$a_1 = -(\alpha_1 + \alpha_2 + \alpha_3),\ a_2 = \alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3,\ a_3 = -\alpha_1\alpha_2\alpha_3$$

The Vieta formulas simplify writing a polynomial, given its roots. For instance, find the fourth-degree polynomial $f(x)$ which has the simple roots 5 and -2 and the double root 3. We get

$$\begin{aligned}
a_1 &= -(5 - 2 + 3 + 3) = -9,\\
a_2 &= 5\cdot(-2) + 5\cdot3 + 5\cdot3 + (-2)\cdot3 + (-2)\cdot3 + 3\cdot3 = 17,\\
a_3 &= -[5\cdot(-2)\cdot3 + 5\cdot(-2)\cdot3 + 5\cdot3\cdot3 + (-2)\cdot3\cdot3] = 33,\\
a_4 &= 5\cdot(-2)\cdot3\cdot3 = -90
\end{aligned}$$

and therefore

$$f(x) = x^4 - 9x^3 + 17x^2 + 33x - 90$$

If the leading coefficient a_0 of the polynomial $f(x)$ is different from unity, then in order to make use of Vieta's formulas, it is first necessary to divide all the coefficients by a_0; this has no effect on

the roots of the polynomial. Thus, in this case the Vieta formulas yield an expression for the relation of all coefficients to the leading coefficient.

Polynomials with real coefficients. We now derive some corollaries to the fundamental theorem of algebra which refer to polynomials with real coefficients. Actually, it is precisely from these corollaries that the great significance of the fundamental theorem of the algebra of complex numbers stems.

Let the following polynomial with real coefficients

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n$$

have a complex root α, that is,

$$a_0\alpha^n + a_1\alpha^{n-1} + \ldots + a_{n-1}\alpha + a_n = 0$$

We know that this equation is unaffected by changing all the numbers to their conjugates; but all the coefficients $a_0, a_1, \ldots, a_{n-1}, a_n$ and also the number 0 on the right, being real, will remain unchanged in such a substitution, and we arrive at the equation

$$a_0\bar{\alpha}^n + a_1\bar{\alpha}^{n-1} + \ldots + a_{n-1}\bar{\alpha} + a_n = 0$$

that is,

$$f(\bar{\alpha}) = 0$$

Thus, *if a complex (but not real) number α serves as a root of a polynomial $f(x)$ with real coefficients, then the conjugate number $\bar{\alpha}$ will also be a root of $f(x)$.*

Consequently, the polynomial $f(x)$ will be divisible by the quadratic trinomial

$$\varphi(x) = (x - \alpha)(x - \bar{\alpha}) = x^2 - (\alpha + \bar{\alpha})x + \alpha\bar{\alpha} \qquad (8)$$

whose coefficients, as we know from Sec. 18, are real. Taking advantage of this fact, we will prove that *the roots α and $\bar{\alpha}$ have one and the same multiplicity in the polynomial $f(x)$.*

Indeed, let these roots have, respectively, the multiplicities k and l and, say, let $k > l$. Then $f(x)$ is divisible by the lth power of the polynomial $\varphi(x)$,

$$f(x) = \varphi^l(x)\, q(x)$$

The polynomial $q(x)$, as a quotient of two polynomials with real coefficients, also has real coefficients, but, in conflict with what was proved above, it has the number α for its $(k - l)$-fold root, whereas the number $\bar{\alpha}$ is not one of its roots. This means that $k = l$.

Now we can say that *the complex roots of any polynomial with real coefficients are pairwise conjugate.* From this fact and from the earlier proved uniqueness of expansions of type (2) follows the final result.

Any polynomial $f(x)$ with real coefficients can be expressed uniquely (to within the order of the factors) in the form of a product of its leading coefficient a_0 and several linear polynomials with real coefficients—of the form $x - \alpha$ that correspond to its real roots—and quadratic polynomials of the form (8) *that correspond to pairs of conjugate complex roots.*

For what follows it will be useful to stress that among polynomials with real coefficients and leading coefficient 1, only linear polynomials of the form $x - \alpha$ and quadratic polynomials of the form (8) are *irreducible* (that is, cannot be decomposed into factors of lower degree).

25. Rational Fractions

The course of mathematical analysis deals with integral rational functions (which we have called polynomials) and also *fractional rational functions.* The latter are quotients $\frac{f(x)}{g(x)}$ of two integral rational functions, where $g(x) \not\equiv 0$. Algebraic operations are performed on these functions in accord with the same laws as are used to manipulate rational numbers, that is to say, fractions with integral numerators and denominators. The equality of two fractional rational functions, or, as we will now term them, *rational fractions,* is to be understood in the same sense as the equality of fractions in elementary arithmetic. For the sake of definiteness, we consider rational fractions with real coefficients. The reader will easily note that this whole section can almost literally be extended to the case of rational fractions with complex coefficients.

A rational fraction is *in lowest terms* (simplified) if the numerator is relatively prime to the denominator.

Any rational fraction is equal to some fraction in lowest terms which is uniquely defined to within a zero-power factor common to both numerator and denominator.

Indeed, any rational fraction may be reduced by dividing numerator and denominator by the greatest common divisor; this yields an equivalent fraction in lowest terms. If, moreover, we have two simplified fractions $\frac{f(x)}{g(x)}$ and $\frac{\varphi(x)}{\psi(x)}$ that are equal, that is

$$f(x)\,\psi(x) = g(x)\,\varphi(x) \tag{1}$$

then it follows from the relative primality of $f(x)$ and $g(x)$ [by Property (b) of Sec. 21] that $f(x)$ divides $\varphi(x)$, and from the relative primality of $\varphi(x)$ and $\psi(x)$ that $\varphi(x)$ divides $f(x)$. Thus, $f(x) = c\varphi(x)$, and then from (1) it follows that $g(x) = c\psi(x)$.

A rational fraction is *proper* if the degree of the numerator is less than the degree of the denominator. If we include the polyno-

mial zero in the set of proper fractions, then the following theorem holds.

Any rational fraction may be represented uniquely in the form of a sum of a polynomial and a proper fraction.

If there is a rational fraction $\frac{f(x)}{g(x)}$ and if, dividing the numerator by the denominator, we get the equation

$$f(x) = g(x)\, q(x) + r(x)$$

where the degree of $r(x)$ is less than the degree of $g(x)$, then it is easy to check that

$$\frac{f(x)}{g(x)} = q(x) + \frac{r(x)}{g(x)}$$

If we also have the equation

$$\frac{f(x)}{g(x)} = \bar{q}(x) + \frac{\varphi(x)}{\psi(x)}$$

where the degree of $\varphi(x)$ is less than the degree of $\psi(x)$, then we obtain the equation

$$q(x) - \bar{q}(x) = \frac{\varphi(x)}{\psi(x)} - \frac{r(x)}{g(x)} = \frac{\varphi(x)\, g(x) - \psi(x)\, r(x)}{\psi(x)\, g(x)}$$

Since the left-hand side is a polynomial, and the right, as is easily seen, is a proper fraction, we get $q(x) - \bar{q}(x) = 0$ and

$$\frac{\varphi(x)}{\psi(x)} - \frac{r(x)}{g(x)} = 0$$

Proper rational fractions can be studied further. As was pointed out at the end of the last section, irreducible real polynomials are polynomials of the form $x - \alpha$, where the number α is real, and polynomials of the form $x^2 - (\beta + \bar{\beta})\, x + \beta\bar{\beta}$, where β and $\bar{\beta}$ are a pair of conjugate complex numbers. It is easy to verify that in the complex case a similar role is played by polynomials of the form $x - \alpha$, where α is any complex number.

A proper rational fraction $\frac{f(x)}{g(x)}$ is called a *partial fraction* if its denominator $g(x)$ is a power of the irreducible polynomial $p(x)$,

$$g(x) = p^k(x), \qquad k \geqslant 1$$

and the degree of the numerator $f(x)$ is less than that of $p(x)$.

The following fundamental theorem holds.

Any proper rational fraction can be decomposed into a sum of partial fractions.

Proof. We first consider the proper rational fraction $\frac{f(x)}{g(x)\, h(x)}$, where the polynomials $g(x)$ and $h(x)$ are relatively prime,

$$(g(x),\ h(x)) = 1$$

Thus, by Sec. 21, there are polynomials $\bar{u}(x)$ and $\bar{v}(x)$ such that

$$g(x)\,\bar{u}(x) + h(x)\,\bar{v}(x) = 1$$

Whence

$$g(x)\,[\bar{u}(x)\,f(x)] + h(x)\,[\bar{v}(x)\,f(x)] = f(x) \qquad (2)$$

Suppose, in dividing the product $\bar{u}(x)\,f(x)$ by $h(x)$, we get a remainder $u(x)$ whose degree is less than the degree of $h(x)$. Then (2) may be rewritten in the form

$$g(x)\,u(x) + h(x)\,v(x) = f(x) \qquad (3)$$

where $v(x)$ is a polynomial whose expression could readily be written. Since the degree of the product $g(x)\,u(x)$ is less than the degree of the product $g(x)\,h(x)$ and this, by hypothesis, is true for the polynomial $f(x)$, it follows that the product $h(x)\,v(x)$ also has degree less than that of $g(x)\,h(x)$, and therefore the degree of $v(x)$ is less than that of $g(x)$. From (3) there now follows the equation

$$\frac{f(x)}{g(x)\,h(x)} = \frac{v(x)}{g(x)} + \frac{u(x)}{h(x)}$$

the right member of which is a sum of proper fractions.

If even one of the denominators $g(x)$, $h(x)$ can be factored into a product of prime factors, then a further decomposition is possible. Continuing in the same manner, we find that *any proper fraction can be decomposed into a sum of several proper fractions, each of which has for the denominator a power of some irreducible polynomial.* More precisely, if we are given a proper fraction $\frac{f(x)}{g(x)}$, whose denominator can be factored into the irreducible factors

$$g(x) = p_1^{k_1}(x)\,p_2^{k_2}(x)\,\ldots\,p_l^{k_l}(x)$$

(of course, one can always say that the leading coefficient of the denominator of a rational fraction is unity), and $p_i(x) \neq p_j(x)$ for $i \neq j$, then it follows that

$$\frac{f(x)}{g(x)} = \frac{u_1(x)}{p_1^{k_1}(x)} + \frac{u_2(x)}{p_2^{k_2}(x)} + \ldots + \frac{u_l(x)}{p_l^{k_l}(x)}$$

All the terms on the right of this equation are proper fractions.

It remains to consider a proper fraction of the form $\frac{u(x)}{p^k(x)}$, where $p(x)$ is an irreducible polynomial. Applying the division algorithm, divide $u(x)$ by $p^{k-1}(x)$, divide the remainder by $p^{k-2}(x)$, and so on.

We arrive at the following equalities:

$$u(x) = p^{k-1}(x)\, s_1(x) + u_1(x),$$
$$u_1(x) = p^{k-2}(x)\, s_2(x) + u_2(x),$$
$$\dots\dots\dots\dots\dots\dots$$
$$u_{k-2}(x) = p(x)\, s_{k-1}(x) + u_{k-1}(x)$$

Since the degree of $u(x)$ is, by hypothesis, less than the degree of $p^k(x)$, and the degree of each of the remainders $u_i(x)$, $i = 1, 2, \dots, k-1$, is less than the degree of the corresponding divisor $p^{k-i}(x)$, it follows that the degrees of all quotients $s_1(x)$, $s_2(x)$, $\dots, s_{k-1}(x)$ will be strictly less than the degree of the polynomial $p(x)$. The degree of the last remainder $u_{k-1}(x)$ is also less than the degree of $p(x)$. It follows from the equations obtained that

$$u(x) = p^{k-1}(x)\, s_1(x) + p^{k-2}(x)\, s_2(x) + \dots$$
$$\dots + p(x)\, s_{k-1}(x) + u_{k-1}(x)$$

whence we arrive at the desired representation of the rational fraction $\frac{u(x)}{p^k(x)}$ as a sum of partial fractions:

$$\frac{u(x)}{p^k(x)} = \frac{u_{k-1}(x)}{p^k(x)} + \frac{s_{k-1}(x)}{p^{k-1}(x)} + \dots + \frac{s_2(x)}{p^2(x)} + \frac{s_1(x)}{p(x)}$$

The proof of the fundamental theorem is complete. It may be supplemented by the following **uniqueness theorem.**

Every proper rational fraction has a unique decomposition into a sum of partial fractions.

Let some proper fraction be decomposable into sums of partial fractions in two ways. Subtracting one of these representations from the other and collecting like terms, we get a sum of partial fractions identically equal to zero. Let the denominators of the partial fractions which constitute this sum be certain powers of distinct irreducible polynomials $p_1(x)$, $p_2(x)$, $\dots$, $p_s(x)$ and let the highest power of the polynomial $p_i(x)$, $i = 1, 2, \dots, s$, which is one of these denominators, be $p_i^{h_i}(x)$. Multiply both sides of the equality at hand by the product $p_1^{h_1-1}(x)\, p_2^{h_2}(x) \dots p_s^{h_s}(x)$. Then all the terms of our sum, except one, become polynomials. The term $\frac{u(x)}{p_1^{h_1}(x)}$ is converted into a fraction whose denominator is $p_1(x)$ and whose numerator is the product $u(x)\, p_2^{h_2}(x) \dots p_s^{h_s}(x)$. The numerator is not exactly divisible by the denominator since the polynomial $p_1(x)$ is irreducible, and all the factors of the numerator are relatively prime to it. Performing division with a remainder, we find that the sum of a polynomial and a nonzero proper fraction is equal to zero, which is impossible.

Example. Decompose into a sum of partial fractions the real proper fraction $\frac{f(x)}{g(x)}$ where

$$f(x) = 2x^4 - 10x^3 + 7x^2 + 4x + 3,$$
$$g(x) = x^5 - 2x^3 + 2x^2 - 3x + 2$$

It is easy to check that

$$g(x) = (x+2)(x-1)^2(x^2+1)$$

Each of the polynomials $x+2$, $x-1$, x^2+1 is irreducible. From the foregoing theory it follows that the desired decomposition should be of the form

$$\frac{f(x)}{g(x)} = \frac{A}{x+2} + \frac{B}{(x-1)^2} + \frac{C}{x-1} + \frac{Dx+E}{x^2+1} \tag{4}$$

where the numbers A, B, C, D and E have still to be found.

From (4) follows the equation

$$f(x) = A(x-1)^2(x^2+1) + B(x+2)(x^2+1) + C(x+2)(x-1)(x^2+1) + Dx(x+2)(x-1)^2 + E(x+2)(x-1)^2 \tag{5}$$

Equating coefficients of like powers of the unknown x in both members of (5), we would get a system of five linear equations in five unknowns A, B, C, D, E; and, as follows from what has been said, this system has a unique solution. However, we will take a different approach.

Assuming $x = -2$ in (5), we get the equation $45A = 135$, whence

$$A = 3 \tag{6}$$

Putting $x = 1$ in (5), we get $6B = 6$, or

$$B = 1 \tag{7}$$

Now, in succession, set $x = 0$ and $x = -1$ in (5). Using (6) and (7), we get the equations

$$\left.\begin{aligned} -2C + 2E &= -2, \\ -4C - 4D + 4E &= -8 \end{aligned}\right\} \tag{8}$$

whence

$$D = 1 \tag{9}$$

Now, finally, set $x = 2$ in (5). Using (6), (7), and (9), we arrive at the equation

$$20C + 4E = -52$$

which, together with the first equation of (8), yields

$$C = -2, \qquad E = -3$$

Thus,

$$\frac{f(x)}{g(x)} = \frac{3}{x+2} + \frac{1}{(x-1)^2} - \frac{2}{x-1} + \frac{x-3}{x^2+1}$$

CHAPTER 6

QUADRATIC FORMS

26. Reducing a Quadratic Form to Canonical Form

The genesis of the theory of quadratic forms lies in analytic geometry, namely, in the theory of quadric curves and surfaces. It will be recalled that the equation of a central quadric curve in a plane, after translating the origin of the rectangular coordinate system to the centre of the curve, is of the form

$$Ax^2 + 2Bxy + Cy^2 = D \tag{1}$$

It is also possible to perform a rotation of the coordinate axes through an angle α, such that we have the following transformation from the coordinates x, y to the coordinates x', y':

$$\left.\begin{aligned} x &= x' \cos \alpha - y' \sin \alpha, \\ y &= x' \sin \alpha + y' \cos \alpha \end{aligned}\right\} \tag{2}$$

Then the equation of our curve in the new coordinates will be of "canonical" form:

$$A'x'^2 + C'y'^2 = D \tag{3}$$

In this equation, the coefficient of the product of unknowns $x'y'$ is, thus, zero. The transformation of coordinates (2) may obviously be interpreted as a linear transformation of the unknowns (see Sec. 13); the transformation is nonsingular since the determinant of its coefficients is equal to unity. This transformation is applied to the left side of (1) and for this reason we can say that the left member of (1) is converted into the left side of (3) by the nonsingular linear transformation (2).

Numerous applications required the construction of a similar theory for the case when the number of unknowns is equal to an arbitrary n instead of two, and the coefficients are either real or any complex numbers.

Generalizing the expression on the left of (1), we arrive at the following concept.

A *quadratic form* f in n unknowns $x_1, x_2, \ldots, x_n$ is a sum, each term of which is either a square of one of the unknowns or a product of two different unknowns. A quadratic form is called *real* or *complex* according as its coefficients are real or complex numbers.

If we take it that like terms in the quadratic form f have already been collected, we can introduce the following notations for the coefficients of this form: we denote by a_{ii} the coefficient of x_i^2, and by $2a_{ij}$ [compare with (1)!] the coefficient of the product x_ix_j for $i \neq j$. However, since $x_ix_j = x_jx_i$, the coefficient of this product could be written as $2a_{ji}$, that is, the designations we have proposed presume the validity of the equality

$$a_{ji} = a_{ij} \tag{4}$$

The term $2a_{ij}x_ix_j$ may now be written as

$$2a_{ij}x_ix_j = a_{ij}x_ix_j + a_{ji}x_jx_i$$

and the entire quadratic form f may be written in the form of a sum of all possible terms $a_{ij}x_ix_j$, where i and j independently take on the values from 1 to n:

$$f = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}x_ix_j \tag{5}$$

In particular, for $i = j$ we have the term $a_{ii}x_i^2$.

Obviously, we can construct a square matrix $A = (a_{ij})$ of order n out of the coefficients a_{ij}; it is called the *matrix of the quadratic form* f, and its rank r is called the *rank* of the quadratic form. If, say, $r = n$, that is, the matrix is nonsingular, then the quadratic form f is termed *nonsingular* too. Due to (4), the elements of matrix A which are symmetric about the principal diagonal are equal; that is, matrix A is a *symmetric* matrix. Conversely, for any symmetric matrix A of order n there is a definite quadratic form (5) in n unknowns having for coefficients the elements of the matrix A.

The quadratic form (5) may be written differently by using the multiplication of rectangular matrices introduced in Sec. 14. Let us make the following convention: if we have a square or, generally, rectangular matrix A, then A' will denote the transpose of A. If matrices A and B are such that their product is defined, then we have the equality

$$(AB)' = B'A' \tag{6}$$

Thus, *the transpose of a product of matrices is equal to the product of the transposes of the matrices in reverse order.*

Indeed, if the product AB is defined, then, as may easily be verified, the product $B'A'$ will also be defined: the number of columns of matrix B' is equal to the number of rows of matrix A'. The element of matrix $(AB)'$ in the ith row and jth column lies in the jth

row and ith column of the matrix AB. It is therefore equal to the sum of the products of the corresponding elements of the jth row of matrix A and the ith column of matrix B, which is to say it is equal to the sum of the products of the corresponding elements of the jth column of matrix A' and the ith row of matrix B'. This proves (6).

Note that *the matrix A is symmetric if and only if it coincides with its transpose*, i.e., if

$$A' = A$$

Now denote by X the column made up of the unknowns:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

X is a matrix with n rows and one column. Its transpose is the matrix

$$X' = (x_1, x_2, \ldots, x_n)$$

comprising a single row.

The quadratic form (5) with matrix $A = (a_{ij})$ may now be written as a product:

$$f = X'AX \tag{7}$$

Indeed, the product AX will be a matrix consisting of one column:

$$AX = \begin{Bmatrix} \sum_{j=1}^{n} a_{1j}x_j \\ \sum_{j=1}^{n} a_{2j}x_j \\ \vdots \\ \sum_{j=1}^{n} a_{nj}x_j \end{Bmatrix}$$

Multiplying this matrix on the left by the matrix X', we get a "matrix" consisting of one row and one column, namely, the right side of (5).

What will happen to the quadratic form f if the unknowns $x_1, x_2, \ldots, x_n$ in it are subjected to the linear transformation

$$x_i = \sum_{k=1}^{n} q_{ik}y_k, \quad i = 1, 2, \ldots, n \tag{8}$$

with the matrix $Q = (q_{ik})$? We will assume here that if the form f is real, then the elements of the matrix Q must be real. Denoting

by Y the column of unknowns $y_1, y_2, \ldots, y_n$, let us write the linear transformation (8) in the form of a matrix equation:

$$X = QY \tag{9}$$

whence, from (6),

$$X' = Y'Q' \tag{10}$$

Substituting (9) and (10) into (7), we get

$$f = Y' (Q'AQ) Y$$

or

$$f = Y'BY$$

where

$$B = Q'AQ$$

The matrix B is symmetric since, because of (6), which is obviously true for any number of factors, and due to the equality $A' = A$, which is equivalent to the symmetry of matrix A, we have

$$B' = Q'A'Q = Q'AQ = B$$

This is proof of the following theorem.

A quadratic form in n unknowns having a matrix A is converted (after performing a linear transformation of the unknowns with matrix Q) into a quadratic form in new unknowns, the product Q'AQ serving as the matrix of this form.

Now assume that we perform a nonsingular linear transformation; that is, Q and, therefore, Q' too are nonsingular matrices. In this case, the product $Q'AQ$ is obtained by multiplying matrix A by the nonsingular matrices; for this reason, as follows from the results of Sec. 14, the rank of this product is equal to the rank of matrix A. Thus, *the rank of a quadratic form does not change under a nonsingular linear transformation.*

By analogy with the geometric problem, indicated at the beginning of this section, of reducing the equation of a central quadric curve to canonical form (3), let us now consider the question of reducing an arbitrary quadratic form (by some nonsingular linear transformation) to a sum of squares of the unknowns, that is to say, to a form where all coefficients of products of distinct unknowns are zero. This special form of the quadratic form is called *canonical.* First, let us suppose that a quadratic form f in n unknowns $x_1, x_2, \ldots, x_n$ has already been reduced (via a nonsingular linear transformation) to the canonical form

$$f = b_1y_1^2 + b_2y_2^2 + \ldots + b_ny_n^2 \tag{11}$$

where $y_1, y_2, \ldots, y_n$ are the new unknowns. Some of the coefficients $b_1, b_2, \ldots, b_n$ may of course be zeros. We will prove that

the number of nonzero coefficients in (11) *is invariably equal to the rank r of the form f.*

Indeed, since we reached (11) by means of a nonsingular transformation, the quadratic form on the right of (11) must also be of rank r. But the matrix of this quadratic form is diagonal:

$$\begin{pmatrix} b_1 & & & 0 \\ & b_2 & & \\ & & \ddots & \\ 0 & & & b_n \end{pmatrix}$$

and a requirement that this matrix have rank r is equivalent to supposing that its principal diagonal contains exactly r nonzero elements.

We now take up the proof of the following **fundamental theorem on quadratic forms.**

Any quadratic form may be reduced to canonical form by means of a nonsingular linear transformation. If a real quadratic form is under consideration, then all the coefficients of this linear transformation may be taken to be real.

This theorem is true for the case of quadratic forms in one unknown since every such form has the form ax^2, which is canonical. We can therefore carry out the proof by induction with respect to the number of unknowns; that is, we can prove the theorem for quadratic forms in n unknowns, assuming it proved for forms with a smaller number of unknowns.

Suppose we have the quadratic form

$$f = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \tag{12}$$

in the n unknowns $x_1, x_2, \ldots, x_n$. We try to find a nonsingular linear transformation that isolates from f a square of one of the unknowns, that is, such that reduces f to the form of a sum of this square and some quadratic form in the remaining unknowns. This is readily achieved if among the coefficients $a_{11}, a_{22}, \ldots, a_{nn}$ in the principal diagonal of the matrix of the form f there are some nonzero coefficients, that is to say, if the square of at least one of the unknowns x_i enters into (12) with a nonzero coefficient.

For example, let $a_{11} \neq 0$. Then it will be easy to see that the expression $a_{11}^{-1}(a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n)^2$, which is a quadratic form, contains the same terms with the unknown x_1 as our form f, and so the difference

$$f - a_{11}^{-1}(a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n)^2 = g$$

is a quadratic form containing only the unknowns $x_2, \ldots, x_n$, but not x_1. Whence

$$f = a_{11}^{-1}(a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n)^2 + g$$

If we introduce the designations

$$y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n, \quad y_i = x_i \quad \text{for} \quad i = 2, 3, \ldots, n \tag{13}$$

we obtain

$$f = a_{11}^{-1}y_1^2 + g \tag{14}$$

where g is now a quadratic form in the unknowns $y_2, y_3, \ldots, y_n$. Expression (14) is the desired expression for the form f, since it was obtained from (12) by a nonsingular linear transformation, namely, by a transformation inverse to the linear transformation (13), which has a_{11} for its determinant and is therefore not singular.

However, if we have the equalities $a_{11} = a_{22} = \ldots = a_{nn} = 0$, then we first have to perform an auxiliary linear transformation that leads to the appearance, in our form f, of squares of the unknowns. Since there must be nonzero coefficients among those in (12) of this form—otherwise there would be nothing to prove—suppose, say, that $a_{12} \neq 0$, i.e., f is the sum of the term $2a_{12}x_1x_2$ and of terms such that each contains at least one of the unknowns $x_3, \ldots, x_n$.

Let us now perform the linear transformation

$$x_1 = z_1 - z_2, \quad x_2 = z_1 + z_2, \quad x_i = z_i \quad \text{for} \quad i = 3, \ldots, n \tag{15}$$

It will be nonsingular since it has the determinant

$$\begin{vmatrix} 1 & -1 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \end{vmatrix} = 2 \neq 0$$

As a result of this transformation, the term $2a_{12}x_1x_2$ of our form becomes

$$2a_{12}x_1x_2 = 2a_{12}(z_1 - z_2)(z_1 + z_2) = 2a_{12}z_1^2 - 2a_{12}z_2^2$$

In other words in form f there will appear the squares of two unknowns at once with nonzero coefficients; what is more, they do not cancel with any one of the remaining terms, since each one of the latter contains at least one of the unknowns $z_3, \ldots, z_n$. We are now in the conditions of the case that has already been considered; one more nonsingular linear transformation will reduce the form f to the form (14).

To conclude the proof, note that the quadratic form g depends on a smaller (than n) number of unknowns and for this reason, by the induction hypothesis, it is reducible to the canonical form by means of a nonsingular transformation of the unknowns $y_2, y_3, \ldots$ $\ldots, y_n$. This transformation, which we regard as a (nonsingular, quite obviously) transformation of all n unknowns under which y_1 remains unchanged, consequently reduces (14) to canonical form. Thus, by means of two or three nonsingular linear transformations, which may be replaced by a single nonsingular transformation (their product), a quadratic form f may be reduced to a sum of squares of the unknowns with certain coefficients. And, as we know, the number of such squares is equal to the rank r of the form. If, besides, the quadratic form f is real, then the coefficients both in the canonical form of f and in the linear transformation which reduces f to this canonical form will be real; indeed, both the linear transformation which is inverse to (13) and the linear transformation (15) have real coefficients.

The proof of the fundamental theorem is complete. The method employed in this proof can be used in specific examples for an actual reduction of a quadratic form to canonical form. It is only necessary, in place of the induction we used in the proof, to isolate the squares of the unknowns successively by the method given above.

Example. Reduce to canonical form the quadratic form

$$f = 2x_1x_2 - 6x_2x_3 + 2x_3x_1 \tag{16}$$

Since there are no squares of the unknowns in this form, we first perform a nonsingular linear transformation

$$x_1 = y_1 - y_2, \quad x_2 = y_1 + y_2, \quad x_3 = y_3$$

with the matrix

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

This yields

$$f = 2y_1^2 - 2y_2^2 - 4y_1y_3 - 8y_2y_3$$

Now the coefficient of y_1^2 is nonzero, and so we can isolate the square of one unknown. Setting

$$z_1 = 2y_1 - 2y_3, \quad z_2 = y_2, \quad z_3 = y_3$$

that is, performing a linear transformation, the inverse of which has the matrix

$$B = \begin{pmatrix} \frac{1}{2} & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

we reduce f to the form

$$f = \frac{1}{2}z_1^2 - 2z_2^2 - 2z_3^2 - 8z_2z_3$$

So far only the square of the unknown z_1 has been isolated, since the form still contains the product of two other unknowns. Using the fact that the coefficient of z_2^2 is nonzero, we again apply the method described above. Performing the linear transformation

$$t_1 = z_1, \quad t_2 = -2z_2 - 4z_3, \quad t_3 = z_3$$

the inverse of which has the matrix

$$C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{2} & -2 \\ 0 & 0 & 1 \end{pmatrix}$$

we finally reduce the form f to canonical form:

$$f = \frac{1}{2} t_1^2 - \frac{1}{2} t_2^2 + 6t_3^2 \tag{17}$$

The linear transformation that immediately reduces (16) to (17) will have for its matrix the product

$$ABC = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 3 \\ \frac{1}{2} & -\frac{1}{2} & -1 \\ 0 & 0 & 1 \end{pmatrix}$$

It is also possible, by direct substitution, to verify that the nonsingular $\left(\text{since the determinant is equal to } -\frac{1}{2}\right)$ linear transformation

$$x_1 = \frac{1}{2} t_1 + \frac{1}{2} t_2 + 3t_3,$$

$$x_2 = \frac{1}{2} t_1 - \frac{1}{2} t_2 - t_3,$$

$$x_3 = \qquad\qquad t_3$$

converts (16) into (17).

The theory of reducing a quadratic form to canonical form is based on an analogy with the geometric theory of central quadric curves but it cannot be considered a generalization of this latter theory. Actually, in our theory we are allowed to use any nonsingular linear transformations, whereas reducing a quadric to canonical form is achieved by applying linear transformations of a very special kind (2); these transformations are rotations of the plane. However, this geometric theory can be generalized to the case of quadratic forms in n unknowns with real coefficients. The generalization, which goes by the name of reduction of quadratic forms to principal axes, will be given in Chapter 8.

27. Law of Inertia

The canonical form to which a given quadratic form is reduced is by no means uniquely determined: any quadratic form may be reduced to canonical form in many different ways. Thus, the quadratic form $f = 2x_1x_2 - 6x_2x_3 + 2x_3x_1$ that was considered in the preceding section can, by the following nonsingular linear transformation,

$$\begin{aligned} x_1 &= t_1 + 3t_2 + 2t_3, \\ x_2 &= t_1 - t_2 - 2t_3, \\ x_3 &= t_2 \end{aligned}$$

be reduced to the canonical form

$$f = 2t_1^2 + 6t_2^2 - 8t_3^2$$

which is different from the earlier obtained form.

The question arises as to what these different canonical quadratic forms to which the given form f is reduced have in common. As we shall see, this question is closely connected with the following one: under what condition can one of the two given quadratic forms be carried into the other by a nonsingular linear transformation? The answer depends on whether we are considering complex or real quadratic forms.

First suppose we are considering arbitrary complex quadratic forms; at the same time, let us assume we admit the use of nonsingular linear transformations also with arbitrary complex coefficients. We know that any quadratic form f in n unknowns having rank r can be reduced to the canonical form

$$f = c_1y_1^2 + c_2y_2^2 + \ldots + c_ry_r^2$$

where all the coefficients $c_1, c_2, \ldots, c_r$ are nonzero. Using the fact that we can take the square root of any complex number, let us perform the following nonsingular linear transformation:

$$z_i = \sqrt{c_i}y_i \text{ for } i = 1, 2, \ldots, r; \; z_j = y_j \text{ for } j = r + 1, \ldots, n$$

It reduces f to the form

$$f = z_1^2 + z_2^2 + \ldots + z_r^2 \tag{1}$$

which is called *normal*. This is simply the sum of the squares of r unknowns with coefficients equal to unity.

The normal form depends solely on the rank r of the form f, that is, all quadratic forms of rank r can be reduced to one and the same normal form (1). Consequently, if forms f and g in n unknowns have the same rank r, then we can transform f to (1) and then (1) to g; in other words, there exists a nonsingular linear transformation

that takes f into g. Since, on the other hand, no nonsingular linear transformation alters the rank of the form, we arrive at the following result.

Two complex quadratic forms in n unknowns can be carried one into the other by means of nonsingular linear transformations with complex coefficients if and only if these forms have one and the same rank.

It very easily follows from this theorem that *any sum of squares of r unknowns with any nonzero complex coefficients can serve as the canonical form of a complex quadratic form of rank r.*

The situation is somewhat more complicated if we consider real quadratic forms and—this is particularly important—if we allow only for linear transformations with real coefficients. Now not every form can be reduced to (1), since this might require taking the square root of a negative number. However, if we now use the term *normal form* of a quadratic form for the sum of squares of several unknowns with coefficients $+1$ or -1, then it is easy to show that *any real quadratic form f may be reduced to the normal form via a nonsingular linear transformation with real coefficients.*

Indeed, the form f of rank r in n unknowns can be reduced to a canonical form that can be written as follows (the numbering of the unknowns may be changed if necessary):

$$f = c_1 y_1^2 + \ldots + c_k y_k^2 - c_{k+1} y_{k+1}^2 - \ldots - c_r y_r^2, \quad 0 \leqslant k \leqslant r$$

where all the numbers $c_1, \ldots, c_k, c_{k+1}, \ldots, c_r$ are nonzero and positive. Then the nonsingular linear transformation with real coefficients

$$z_i = \sqrt{c_i}\, y_i \text{ for } i = 1, 2, \ldots, r, \quad z_j = y_j \text{ for } j = r+1, \ldots, n$$

reduces f to normal form:

$$f = z_1^2 + \ldots + z_k^2 - z_{k+1}^2 - \ldots - z_r^2$$

The total number of squares here is equal to the rank of the form.

A real quadratic form may be reduced to normal form by many different transformations; however, to within the numbering of the unknowns, it can be reduced only to one normal form. This is demonstrated by the following important theorem, which is called *the law of inertia of real quadratic forms.*

The number of positive and the number of negative squares in the normal form to which a given quadratic form with real coefficients can be reduced by a real nonsingular linear transformation is independent of the choice of the transformation.

Indeed, let a quadratic form f of rank r in n unknowns $x_1, x_2, \ldots, x_n$ be reduced to the following normal form in two ways:

$$\begin{aligned} f &= y_1^2 + \ldots + y_k^2 - y_{k+1}^2 - \ldots - y_r^2 \\ &= z_1^2 + \ldots + z_l^2 - z_{l+1}^2 - \ldots - z_r^2 \end{aligned} \tag{2}$$

Since the transition from the unknowns $x_1, x_2, \ldots, x_n$ to the unknowns $y_1, y_2, \ldots, y_n$ was a nonsingular linear transformation, it follows, conversely, that the second set of unknowns will also be expressed linearly in terms of the first set with a nonzero determinant:

$$y_i = \sum_{s=1}^{n} a_{is} x_s, \qquad i = 1, 2, \ldots, n \tag{3}$$

Similarly,

$$z_j = \sum_{t=1}^{n} b_{jt} x_t, \qquad j = 1, 2, \ldots, n \tag{4}$$

the determinant of the coefficients again being different from zero. The coefficients are real numbers both in (3) and in (4).

Now suppose that $k < l$. Write the system of equalities

$$y_1 = 0, \ldots, y_k = 0, z_{l+1} = 0, \ldots, z_r = 0, \ldots, z_n = 0 \tag{5}$$

If the left members of these equalities are replaced by their expressions taken from (3) and (4), we get a system of $n - l + k$ homogeneous linear equations in n unknowns $x_1, x_2, \ldots, x_n$. The number of equations in this system is less than the number of unknowns. For this reason, as we know from Sec. 1, our system has a nonzero real solution $\alpha_1, \alpha_2, \ldots, \alpha_n$.

Now in (2) let us replace all y's and all z's by their expressions (3) and (4), and then let us substitute for the unknowns the numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$. If for brevity the values of the unknowns y_i and z_j obtained in this substitution are denoted by $y_i(\alpha)$ and $z_j(\alpha)$, then, by (5), (2) becomes

$$-y_{k+1}^2(\alpha) - \ldots - y_r^2(\alpha) = z_1^2(\alpha) + \ldots + z_l^2(\alpha) \tag{6}$$

Since all the coefficients in (3) and (4) are real, all the squares in (6) are positive and for this reason (6) implies that all these squares are zero, whence follow the equalities

$$z_1(\alpha) = 0, \ldots, z_l(\alpha) = 0 \tag{7}$$

On the other hand, by the very choice of the numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$,

$$z_{l+1}(\alpha) = 0, \ldots, z_r(\alpha) = 0, \ldots, z_n(\alpha) = 0 \tag{8}$$

Thus, the system of n homogeneous linear equations

$$z_i = 0, \quad i = 1, 2, \ldots, n$$

in n unknowns $x_1, x_2, \ldots, x_n$ has, by (7) and (8), the nontrivial solution $\alpha_1, \alpha_2, \ldots, \alpha_n$; that is, the determinant of this system must be zero. This however contradicts the fact that the transformation (4) was presumed to be nonsingular. We have the same contradiction for $l < k$, whence follows $k = l$ which proves the theorem.

The number of positive squares in the normal form to which a given real quadratic form f is reduced is called the *positive index of inertia* of this form; the number of negative squares is termed the *negative index of inertia*, and the number of positive indices diminished by the numbers of negative indices of inertia is the *signature* of the form f. Clearly, if we are given the rank of a form, any one of the three numbers just defined will fully determine the other two, and for this reason, we can speak of any one of the three numbers in subsequent formulations.

We now prove the following theorem.

Two quadratic forms in n unknowns with real coefficients are carried one into the other by real nonsingular linear transformations if and only if the forms have the same ranks and the same signatures.

Indeed, let a form f be carried into a form g by a real nonsingular transformation. We know that this transformation does not alter the rank of the form. Neither can it change the signature, for then f and g would reduce to different normal forms, but then f would reduce—in conflict with the law of inertia—to both these normal forms. Conversely, if the forms f and g have the same ranks and the same signatures, then they reduce to one and the same normal form and therefore can be carried into one another.

If we have a quadratic form g in canonical form with nonzero real coefficients

$$g = b_1y_1^2 + b_2y_2^2 + \ldots + b_ry_r^2 \tag{9}$$

then the rank of this form is obviously equal to r. Taking advantage of the procedure used earlier of reducing such a form to the normal form, it is easy to see that the positive index of inertia of form g is equal to the number of positive coefficients in the right member of (9). From this and from the preceding theorem we obtain the following result.

A quadratic form f has form (9) *as its canonical form if and only if the rank of f is equal to r and the positive index of inertia of this form coincides with the number of positive coefficients in* (9).

Decomposable quadratic forms. By multiplying any two linear forms in n unknowns,

$$\varphi = a_1x_1 + a_2x_2 + \ldots + a_nx_n, \qquad \psi = b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

we obviously get another quadratic form. Not every quadratic form can be represented as a product of two linear forms and we wish to derive the conditions under which this occurs, that is, the conditions under which a quadratic form is decomposable.

A complex quadratic form $f(x_1, x_2, \ldots, x_n)$ *is decomposable if and only if its rank is less than or equal to two. A real quadratic form* $f(x_1, x_2, \ldots, x_n)$ *is decomposable if and only if either its rank does not exceed unity or the rank is equal to two and the signature is zero.*

Let us first consider the product of the linear forms φ and ψ. If at least one of them is a zero form, then their product will be a quadratic form with zero coefficients, which means it has rank 0. If the linear forms φ and ψ are proportional,

$$\psi = c\varphi$$

and $c \neq 0$ and the form φ is nonzero, then, for example, let the coefficient a_1 be different from zero. Then the nonsingular linear transformation

$$y_1 = a_1x_1 + \ldots + a_nx_n, \quad y_i = x_i \quad \text{for} \quad i = 2, 3, \ldots, n$$

reduces the quadratic form $\varphi\psi$ to

$$\varphi\psi = cy_1^2$$

On the right is a quadratic form of rank 1, and so the quadratic form $\varphi\psi$ has rank 1. Finally, if the linear forms φ and ψ are not proportional then, say, let

$$\begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix} \neq 0$$

Then the linear transformation

$$\begin{aligned} y_1 &= a_1x_1 + a_2x_2 + \ldots + a_nx_n, \\ y_2 &= b_1x_1 + b_2x_2 + \ldots + b_nx_n, \\ y_i &= x_i \text{ for } i = 3, 4, \ldots, n \end{aligned}$$

will be nonsingular; it reduces the quadratic form $\varphi\psi$ to

$$\varphi\psi = y_1y_2$$

On the right is a quadratic form of rank 2, which in the case of real coefficients has a signature of 0.

Let us now prove the converse. A quadratic form of rank 0 can of course be regarded as a product of two linear forms, one of which is a zero form. Next, a quadratic form $f(x_1, x_2, \ldots, x_n)$ of rank 1 is reduced by a nonsingular linear transformation to

$$f = cy_1^2, \qquad c \neq 0$$

that is, to the form

$$f = (cy_1)\, y_1$$

Expressing y_1 linearly in terms of $x_1, x_2, \ldots, x_n$, we get a representation of the form f as a product of two linear forms. Finally, the real quadratic form $f(x_1, x_2, \ldots, x_n)$ of rank 2 and signature 0 is reduced by a nonsingular linear transformation to

$$f = y_1^2 - y_2^2$$

Any complex quadratic form of rank 2 can be reduced to this same form. However,

$$y_1^2 - y_2^2 = (y_1 - y_2)(y_1 + y_2)$$

but after replacing y_1 and y_2 by their linear expressions in terms of $x_1, x_2, \ldots, x_n$, we will have on the right a product of two linear forms. This proves the theorem.

28. Positive Definite Forms

A quadratic form f in n unknowns with real coefficients is called *positive definite* if it can be reduced to a normal form consisting of n positive squares, that is, if both the rank and the positive index of inertia of this form are equal to the number of unknowns.

The following theorem enables us to characterize positive definite forms without reducing them to normal form or canonical form.

A quadratic form f in n unknowns $x_1, x_2, \ldots, x_n$ with real coefficients is positive definite if and only if for all real values of the unknowns, at least one of which is nonzero, the form receives positive values.

Proof. Let the form f be positive definite, i. e., reducible to the normal form

$$f = y_1^2 + y_2^2 + \ldots + y_n^2 \tag{1}$$

and let

$$y_i = \sum_{j=1}^{n} a_{ij}x_j, \qquad i = 1, 2, \ldots, n \tag{2}$$

with a nonzero determinant of the real coefficients a_{ij}. If we want to substitute, into f, arbitrary real values of the unknowns $x_1, x_2, \ldots, x_n$, at least one of which is nonzero, then we can first substitute them into (2) and then substitute the values obtained for all y_i into (1). It will be noted that the values obtained from (2) for $y_1, y_2, \ldots, y_n$ cannot all be zero at once, for then we would have that the system of homogeneous linear equations

$$\sum_{j=1}^{n} a_{ij}x_j = 0, \qquad i = 1, 2, \ldots, n$$

has a nontrivial solution, though its determinant is different from zero. Substituting the values found for $y_1, y_2, \ldots, y_n$ into (1), we get the value of the form f equal to the sum of the squares of n real numbers, not all zero. This value will consequently be strictly positive.

Conversely, suppose the form f is not positive definite, that is either its rank or the positive index of inertia is less than n. This means that in the normal form of f, to which it is reduced, say, by

the nonsingular linear transformation (2), the square of at least one of the new unknowns, say y_n, is either absent altogether or is present with a minus sign. We will show that in this case it is possible to choose real values for the unknowns $x_1, x_2, \ldots, x_n$, not all zero, such that the value of the form f for these values of the unknowns is equal to zero or is even negative. Such, for instance, are the values for $x_1, x_2, \ldots, x_n$ which we obtain when solving, by Cramer's rule, the system of linear equations obtained from (2) for $y_1 = y_2 = \ldots = y_{n-1} = 0$, $y_n = 1$. Indeed, for these values of the unknowns $x_1, x_2, \ldots, x_n$, the form f is zero if y_n^2 does not enter into the normal form of f, and is equal to -1 if y_n^2 enters into the normal form with a minus sign.

The theorem that has just been proved is used wherever positive definite quadratic forms are employed. However, it cannot be used to establish from the coefficients whether a form is positive definite or not. This is handled by a different theorem which we will state and prove after introducing an auxiliary notion.

Suppose we have a quadratic form f in n unknowns with the matrix $A = (a_{ij})$. The minors of order $1, 2, \ldots, n$ of this matrix situated in the upper left corner, that is, the minors

$$a_{11}, \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \ldots, \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1k} \\ a_{21} & a_{22} & \ldots & a_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ a_{k1} & a_{k2} & \ldots & a_{kk} \end{vmatrix}, \ldots, \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix}$$

of which the last obviously coincides with the determinant of matrix A are called the *principal minors* of the form f.

The following theorem holds true.

A quadratic form f in n unknowns with real coefficients is positive definite if and only if all its principal minors are strictly positive.

Proof. For $n = 1$, the theorem is true since the form then is ax^2 and therefore is positive definite if and only if $a > 0$. For this reason, we prove the theorem for the case of n unknowns on the assumption that it has already been proved for quadratic forms in $n - 1$ unknowns.

Note the following.

If a quadratic form f with real coefficients constituting a matrix A is subjected to a nonsingular linear transformation with a real matrix Q, then *the sign of the determinant of the form (that is, the determinant of its matrix) remains unchanged.*

Indeed, after the transformation we obtain a quadratic form with the matrix $Q'AQ$; however, due to $|Q'| = |Q|$,

$$|Q'AQ| = |Q'| \cdot |A| \cdot |Q| = |A| \cdot |Q|^2$$

that is, the determinant $|A|$ is multiplied by a positive number.

Now suppose we have the quadratic form

$$f = \sum_{i,\, j=1}^{n} a_{ij} x_i x_j$$

It can be written as

$$f = \varphi(x_1, x_2, \ldots, x_{n-1}) + 2 \sum_{i=1}^{n-1} a_{in} x_i x_n + a_{nn} x_n^2 \qquad (3)$$

where φ is a quadratic form in $n - 1$ unknowns composed of those terms of form f which do not contain the unknown x_n. The principal minors of the form φ evidently coincide with all principal minors of the form f except the last.

Let the form f be positive definite. Then the form φ will also be positive definite: if there existed values of the unknowns x_1, x_2, . . ., x_{n-1}, not all zero, for which the form φ receives a nonstrictly positive value, then, additionally assuming $x_n = 0$, we would also obtain, by (3), a nonstrictly positive value of the form f, although not all the values of the unknowns x_1, x_2, . . ., x_{n-1}, x_n are equal to zero. For this reason, by the induction hypothesis, all the principal minors of the form φ that is, all the principal minors of the form f, except the last, are strictly positive. As for the last principal minor of f (that is the determinant of the matrix A itself), its positivity is a consequence of the following reasoning: because of its positive definiteness, form f is reduced by a nonsingular linear transformation to a normal form consisting of n positive squares. The determinant of this normal form is strictly positive, and so, by the remark made above, the determinant of the form f itself is positive.

Now let all the principal minors of the form f be strictly positive. From this follows the positivity of all the principal minors of the form φ, that is, by the induction hypothesis, the positive definiteness of this form. Therefore, there is a nonsingular linear transformation of the unknowns x_1, x_2, . . ., x_{n-1} such that reduces the form φ to a sum of $n - 1$ positive squares in the new unknowns y_1, y_2, ..., y_{n-1}. By setting $x_n = y_n$, this linear transformation may be completed to form a (nonsingular) linear transformation of all the unknowns x_1, x_2, . . ., x_n. By (3), form f is reduced by the indicated transformation to

$$f = \sum_{i=1}^{n-1} y_i^2 + 2 \sum_{i=1}^{n-1} b_{in} y_i y_n + b_{nn} y_n^2 \qquad (4)$$

The exact expressions of the coefficients b_{in} are not essential to us Since

$$y_i^2 + 2 b_{in} y_i y_n = (y_i + b_{in} y_n)^2 - b_{in}^2 y_n^2$$

it follows that the nonsingular linear transformation

$$z_i = y_i + b_{in}y_n, \quad i = 1, 2, \ldots, n-1,$$
$$z_n = \qquad y_n$$

reduces the form f by (4) to the canonical form

$$f = \sum_{i=1}^{n-1} z_i^2 + cz_n^2 \tag{5}$$

To prove the positive definiteness of the form f, it remains to prove that the number c is positive. The determinant of the form in the right member of (5) is equal to c. However, this determinant should be positive since the right side of (5) is obtained from f by two nonsingular linear transformations, and the determinant of the form f was positive (being the last of the principal minors of this form).

This completes the proof of the theorem.

Example 1. The quadratic form

$$f = 5x_1^2 + x_2^2 + 5x_3^2 + 4x_1x_2 - 8x_1x_3 - 4x_2x_3$$

is positive definite since its principal minors

$$5, \quad \begin{vmatrix} 5 & 2 \\ 2 & 1 \end{vmatrix} = 1, \quad \begin{vmatrix} 5 & 2 & -4 \\ 2 & 1 & -2 \\ -4 & -2 & 5 \end{vmatrix} = 1$$

are positive.

Example 2. The quadratic form

$$f = 3x_1^2 + x_2^2 + 5x_3^2 + 4x_1x_2 - 8x_1x_3 - 4x_2x_3$$

is not positive definite since its second principal minor is negative:

$$\begin{vmatrix} 3 & 2 \\ 2 & 1 \end{vmatrix} = -1$$

Note that by analogy with positive definite quadratic forms we can introduce *negative definite forms*, that is, nonsingular quadratic forms with real coefficients whose normal form contains only negative squares of the unknowns. Singular quadratic forms whose normal form consists of the squares of one sign are sometimes termed *semidefinite*. Finally, *indefinite* quadratic forms are those whose normal form contains both positive and negative squares of the unknowns.

CHAPTER 7

LINEAR SPACES

29. Definition of a Linear Space. An Isomorphism

The definition of an n-dimensional vector space given in Sec. 8 began with a definition of an n-dimensional vector as an ordered set of n numbers (n-tuple). For n-dimensional vectors we then introduced addition and multiplication by scalars, which is what led to the concept of an n-dimensional vector space. The first instances of vector spaces are collections of vector segments emanating from a coordinate origin in the plane or in three-dimensional space. However, when we encounter such cases in geometry, we do not always find it necessary to specify the vectors via their components in some fixed system of coordinates, since both addition of vectors and their multiplication by a scalar are determined geometrically, irrespective of the choice of any coordinate system. Namely, the addition of vectors in the plane or in space is accomplished by the parallelogram rule, while the multiplication of a vector by a scalar α signifies a stretching of the vector by the factor α (the direction is reversed if α is negative). It is advisable to give a "coordinateless" definition of a vector space in the general case as well. By this is meant a definition which does not require specifying vectors by ordered sets of numbers. We now give such a definition. This definition is axiomatic; nothing will be said about the properties of a separate vector, but we will enumerate the properties of operations involving vectors.

Suppose we have a set V. We denote its elements by lower-case Latin letters: a, b, c, ...* Now, in set V we define the *operation of addition*, which associates every pair of elements a, b in V with a uniquely defined element $a + b$ in V, called the *sum*, and the *operation of multiplication by a real number* (scalar); the *product* αa *of element* a by a scalar α is uniquely defined and belongs to V.

The elements of V will be termed *vectors*, and V itself will be called a *real linear* (or *vector*, or *affine*) *space* if the indicated operations have the following properties (I to VIII).

---

* In contrast to Chapter 2, here and in the sequel, vectors will be designated by lower-case Latin letters, scalars by lower-case Greek letters.

I. Addition is commutative: $a + b = b + a$.

II. Addition is associative: $(a + b) + c = a + (b + c)$.

III. There is a *zero element* 0 in V which satisfies the condition: $a + 0 = a$ for all a in V.

Using I it is easy to prove the *uniqueness of the zero element*: if 0_1 and 0_2 are two zero elements, then

$$0_1 + 0_2 = 0_1,$$

$$0_1 + 0_2 = 0_2 + 0_1 = 0_2$$

whence $0_1 = 0_2$.

IV. For any element a in V there exists an *opposite* (*inverse*) *element* $-a$, which satisfies the condition: $a + (-a) = 0$.

Using II and I, it is easy to prove the *uniqueness of the inverse element*: if $(-a)_1$ and $(-a)_2$ are two inverse elements of a, then

$$(-a)_1 + [a + (-a)_2] = (-a)_1 + 0 = (-a)_1,$$

$$[(-a)_1 + a] + (-a)_2 = 0 + (-a)_2 = (-a)_2$$

whence $(-a)_1 = (-a)_2$.

From axioms I to IV we deduce the *existence and uniqueness of the difference* $a - b$, that is, an element which satisfies the equation

$$b + x = a \tag{1}$$

We can set

$$a - b = a + (-b)$$

since

$$b + [a + (-b)] = [b + (-b)] + a = 0 + a = a.$$

Now if there is an element c such that satisfies (1),

$$b + c = a$$

then, by adding to both sides an element $-b$, we get

$$c = a + (-b)$$

Axioms V to VIII (cf. Sec. 8) relate multiplication by a scalar to addition and to operations involving scalars. Namely, for any elements a, b in V, for any real numbers α, β, and for the real number 1, the following equalities must hold:

V. $\alpha (a + b) = \alpha a + \alpha b$,

VI. $(\alpha + \beta) a = \alpha a + \beta a$,

VII. $(\alpha\beta) a = \alpha (\beta a)$,

VIII. $1 \cdot a = a$.

Elementary corollaries to these axioms are:

[1] $$\alpha \cdot 0 = 0$$

For some a in V,

$$\alpha a = \alpha (a + 0) = \alpha a + \alpha \cdot 0$$

that is

$$\alpha \cdot 0 = \alpha a - \alpha a = \alpha a + [-(\alpha a)] = 0$$

[2] $$0 \cdot a = 0$$

where the zero on the left is the number zero and the zero on the right is the zero element of V.

To prove this, take any scalar α. Then

$$\alpha a = (\alpha + 0)\, a = \alpha a + 0 \cdot a$$

whence

$$0 \cdot a = \alpha a - \alpha a = 0$$

[3] If $\alpha a = 0$, then either $\alpha = 0$ or $a = 0$.
If $\alpha \neq 0$, that is the scalar α^{-1} exists, then

$$a = 1 \cdot a = (\alpha^{-1}\alpha)\, a = \alpha^{-1}\, (\alpha a) = \alpha^{-1} \cdot 0 = 0$$

[4] $$\alpha(-a) = -\alpha a$$

Indeed,

$$\alpha a + \alpha\, (-a) = \alpha\, [a + (-a)] = \alpha \cdot 0 = 0$$

that is, the element $\alpha\, (-a)$ is the inverse of αa.

[5] $$(-\alpha)\, a = -\alpha a$$

Indeed,

$$\alpha a + (-\alpha)\, a = [\alpha + (-\alpha)]\, a = 0 \cdot a = 0$$

that is, the element $(-\alpha)\, a$ is the inverse of αa.

[6] $$\alpha\, (a - b) = \alpha a - \alpha b$$

By [4],

$$\alpha\, (a - b) = \alpha\, [a + (-b)] = \alpha a + \alpha\, (-b)$$
$$= \alpha a + (-\alpha b) = \alpha a - \alpha b$$

[7] $$(\alpha - \beta)\, a = \alpha a - \beta a$$

Indeed,

$$(\alpha - \beta)\, a = [\alpha + (-\beta)]\, a = \alpha a + (-\beta)\, a$$
$$= \alpha a + (-\beta a) = \alpha a - \beta a$$

These axioms and their corollaries will be used from now on without any special reservations.

The definition given above is for a real linear space. If we assumed, in V, multiplication not only by real numbers but also by arbitrary complex numbers, then, retaining Axioms I to VIII, we would have the definition of a *complex linear space.* For the sake of definiteness, we will consider real linear spaces; however, everything in this chapter can be extended word for word to the case of complex linear spaces.

Examples of real linear spaces come to mind immediately. They include the n-dimensional real vector spaces composed of row vectors that were studied in Chapter 2, also sets of vector segments emanating from a coordinate origin in the plane or in three-dimensional space if the operations of addition and multiplication by a scalar are understood in the geometric sense stated at the beginning of this section.

We also have linear spaces that are infinite-dimensional. Let us consider all possible sequences of real numbers; they have the form

$$a = (\alpha_1, \alpha_2, \ldots, \alpha_n, \ldots)$$

We perform operations on sequences componentwise: if

$$b = (\beta_1, \beta_2, \ldots, \beta_n, \ldots)$$

then

$$a + b = (\alpha_1 + \beta_1, \alpha_2 + \beta_2, \ldots, \alpha_n + \beta_n, \ldots)$$

On the other hand, for any real number γ,

$$\gamma a = (\gamma\alpha_1, \gamma\alpha_2, \ldots, \gamma\alpha_n, \ldots)$$

All the axioms from I to VIII are fulfilled, which means we have a real linear space.

Another instance of an infinite-dimensional space is the set of all possible real functions of a real variable if the addition of functions and their multiplication by a real number are to be understood as is conventional in the theory of functions, that is, as the addition or multiplication by the number of values of the functions for each value of the independent variable.

Isomorphisms. Our immediate aim is to select from all linear spaces those which it will be natural to call finite-dimensional. First let us introduce a general concept.

In the definition of a linear space we spoke about the properties of operations involving vectors, but we said nothing about the properties of the vectors themselves. Thus, it may happen that although the vectors of two given linear spaces are quite different as to their nature, the two spaces are indistinguishable from the standpoint of the properties of the operations. The exact definition is as follows.

Two real linear spaces V and V' are called *isomorphic* if a one-to-one correspondence can be set up between their vectors: every

vector a of V is associated with a vector a' of V', the *image* of the vector a; different vectors from V possess different images and every vector in V' serves as an image of some vector in V; and if in this correspondence the image of a sum of two vectors is the sum of the images of the two vectors,

$$(a + b)' = a' + b' \tag{2}$$

and the image of a product of a vector by a scalar is the product of the image of the vector by that scalar,

$$(\alpha a)' = \alpha a' \tag{3}$$

The one-to-one correspondence between spaces V and V' which satisfies the conditions (2) and (3) is called an *isomorphic correspondence.*

Thus, the space of vector segments (in a plane) emanating from a coordinate origin is isomorphic to a two-dimensional vector space made up of ordered pairs of real numbers: we obtain an isomorphic correspondence between these spaces if in the plane we fix some system of coordinates and associate with every vector segment an ordered pair of its coordinates.

Let us prove the following property of an isomorphism of linear spaces: *the image of zero of the space V is the zero of the space V' in an isomorphic correspondence between V and V'.*

Let a be some vector in V and a' its image in V'. Then, by (2),

$$a' = (a + 0)' = a' + 0'$$

That is to say, $0'$ is a zero of the space V'.

30. Finite-Dimensional Spaces. Bases

As the reader can verify without difficulty, the two definitions of linear dependence of row vectors given in Sec. 9, and also the proof of the equivalence of these definitions, employ only operations on vectors and for this reason can be carried over to the case of any linear spaces. Consequently, in axiomatically defined linear spaces we can speak of linearly independent systems of vectors, of maximal linearly independent systems, if such exist, and so on.

If the linear spaces V and V' are isomorphic, then the system of vectors $a_1, a_2, \ldots, a_k$ in V is linearly dependent if and only if the system of their images $a'_1, a'_2, \ldots, a'_k$ in V' is linearly dependent.

Note that if the correspondence $a \to a'$ (for all a in V) is an isomorphic correspondence between V and V', then the reverse correspondence $a' \to a$ will also be isomorphic. It is therefore sufficient to consider the case when the system $a_1, a_2, \ldots, a_k$ is linearly dependent. Let there be scalars $\alpha_1, \alpha_2, \ldots, \alpha_k$, not all zero, such

that

$$\alpha_1 a_1 + \alpha_2 a_2 + \ldots + \alpha_k a_k = 0$$

In the isomorphism under consideration, the image of the right member of this equation is, as we know, the zero $0'$ of space V'. Taking the image of the left member and applying (2) and (3) several times, we get

$$\alpha_1 a_1' + \alpha_2 a_2' + \ldots + \alpha_k a_k' = 0'$$

Thus, the system $a_1', a_2', \ldots, a_k'$ is also linearly dependent.

Finite-dimensional spaces. A linear space V is called *finite-dimensional* if in it we can find a finite maximal linearly independent system of vectors; any such system of vectors will be termed the *basis* of the space V.

A finite-dimensional linear space can have many different bases. Thus, in the space of vector segments in the plane, any pair of vectors different from zero and not lying on one straight line can serve as a basis. Note that so far our definition of a finite-dimensional space does not specify whether there can exist, in this space, bases consisting of a different number of vectors. What is more, it might even be assumed that in some finite-dimensional spaces there exist bases with an arbitrarily large number of vectors. Let us investigate this situation.

Suppose a linear space V has a basis

$$e_1, e_2, \ldots, e_n \tag{1}$$

consisting of n vectors. If a is an arbitrary vector in V, then from the maximality of the linearly independent system (1) it follows that a is expressed linearly in terms of the system:

$$a = \alpha_1 e_1 + \alpha_2 e_2 + \ldots + \alpha_n e_n \tag{2}$$

On the other hand, due to the linear independence of (1), expression (2) will be unique for the vector a: if

$$a = \alpha_1' e_1 + \alpha_2' e_2 + \ldots + \alpha_n' e_n$$

then

$$(\alpha_1 - \alpha_1')\, e_1 + (\alpha_2 - \alpha_2')\, e_2 + \ldots + (\alpha_n - \alpha_n')\, e_n = 0$$

whence

$$\alpha_i = \alpha_i', \qquad i = 1, 2, \ldots, n$$

Thus, the vector a is associated one-to-one with the row

$$(\alpha_1, \alpha_2, \ldots, \alpha_n) \tag{3}$$

of coefficients of its expression (2) in terms of the basis (1) or, as we shall say, *the row of its coordinates in the basis* (1). Conversely, every row of type (3), that is, any n-dimensional vector in the sense

of Chapter 2 serves as a row of coordinates in basis (1) for some vector of space V, namely, for the vector written in the form (2) in terms of the basis (1).

We have thus obtained a one-to-one correspondence between all vectors of the space V and all vectors of an n-dimensional vector row-space. We will show that this correspondence, which quite naturally is dependent on the choice of the basis (1), is isomorphic.

In space V let us, in addition to vector a, which is expressed in terms of the basis (1) in the form (2), also take a vector b whose expression in terms of the basis (1) is

$$b = \beta_1 e_1 + \beta_2 e_2 + \ldots + \beta_n e_n$$

Then

$$a + b = (\alpha_1 + \beta_1)\, e_1 + (\alpha_2 + \beta_2)\, e_2 + \ldots + (\alpha_n + \beta_n)\, e_n$$

that is, *the sum of the vectors a and b corresponds to the sum of the rows of their coordinates in the basis* (1). On the other hand,

$$\gamma a = (\gamma\alpha_1)\, e_1 + (\gamma\alpha_2)\, e_2 + \ldots + (\gamma\alpha_n)\, e_n$$

that is, *to the product of a vector a by a scalar* γ *corresponds the product of the row of its coordinates in the basis* (1) *by the same scalar* γ.

The foregoing proves the following theorem.

Any linear space with a basis consisting of n vectors is isomorphic to an n-dimensional vector row-space.

As we know, in an isomorphic correspondence between linear spaces, a linearly dependent system of vectors goes into a linearly dependent system and conversely; for this reason, a linearly independent system goes into a linearly independent system. From this it follows that *in an isomorphic correspondence, a basis goes into a basis.*

Indeed, let a basis $e_1, e_2, \ldots, e_n$ of a space V go (under an isomorphic correspondence between the spaces V and V') into a system of vectors $e'_1, e'_2, \ldots, e'_n$ of space V', which, though it is linearly independent, is not maximal. Consequently, in V' we can find a vector f' such that the system $e'_1, e'_2, \ldots, e'_n, f'$ remains linearly independent. However, the vector f' in this isomorphism serves as an image of some vector f in V. We find that the system of vectors $e_1, e_2, \ldots, e_n, f$ must be linearly independent, which is in contradiction to the definition of a basis.

Further, we know (see Sec. 9) that in an n-dimensional vector row-space, all maximal linearly independent systems consist of n vectors, that any system of $n + 1$ vectors is linearly dependent, and that any linearly independent system of vectors is contained in some maximal linearly independent system. Using the above-established properties of isomorphic correspondences, we arrive at the following results.

All bases of a finite-dimensional linear space V consist of one and the same number of vectors. If this number is equal to n, then V is

called an *n-dimensional linear space*, and the number n is the *dimension* of this space.

Any system of $n+1$ vectors of an n-dimensional linear space is linearly dependent.

Any linearly independent system of vectors of an n-dimensional linear space is contained in some basis of that space.

It is now easy to verify that the above-indicated examples of real linear spaces—the space of sequences and the space of functions—are not finite-dimensional spaces: in each of these spaces the reader will easily find linearly independent systems consisting of an arbitrarily large number of vectors.

Relationships between bases. We are interested in finite-dimensional linear spaces. Clearly, when studying n-dimensional linear spaces we are actually studying the n-dimensional vector row-space that was introduced back in Chapter 2. Earlier, however, we extracted one basis from this space, namely, the basis composed of unit vectors (these are vectors, one coordinate of which is equal to unity and all others are zero), all the vectors of the space were specified by the rows of their coordinates in that basis. Now, however, all bases of a space have equal status.

Let us see how many bases can be found in an n-dimensional linear space and how these bases are interrelated.

Suppose in an n-dimensional linear space V we have the bases

$$e_1, e_2, \ldots, e_n \tag{4}$$

and

$$e'_1, e'_2, \ldots, e'_n \tag{5}$$

Each vector of basis (5), like any vector of the space V, is unambiguously written in terms of basis (4) as

$$e'_i = \sum_{j=1}^{n} \tau_{ij} e_j, \quad i = 1, 2, \ldots, n \tag{6}$$

The matrix

$$T = \begin{pmatrix} \tau_{11} & \ldots & \tau_{1n} \\ \ldots & \ldots & \ldots \\ \tau_{n1} & \ldots & \tau_{nn} \end{pmatrix}$$

whose rows are the rows of the coordinates of the vectors (5) in basis (4), is called the *change-of-basis matrix* from basis (4) to basis (5).

Because of (6), we can write the relationship between bases (4) and (5) and the change-of-basis matrix T in the form of a matrix equation:

$$\begin{pmatrix} e'_1 \\ e'_2 \\ \vdots \\ e'_n \end{pmatrix} = \begin{pmatrix} \tau_{11} & \tau_{12} & \ldots & \tau_{1n} \\ \tau_{21} & \tau_{22} & \ldots & \tau_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \tau_{n1} & \tau_{n2} & \ldots & \tau_{nn} \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \tag{7}$$

or, denoting by e and e', respectively, the bases (4) and (5) as columns:

$$e' = Te$$

On the other hand, if T' is the change-of-basis matrix from (5) to (4), then

$$e = T'e'$$

whence

$$e = (T'T)\, e,$$
$$e' = (TT')\, e'$$

or, because of the linear independence of the bases e and e',

$$T'T = TT' = E$$

whence

$$T' = T^{-1}$$

This proves that *the change-of-basis matrix is always a nonsingular matrix.*

Any nonsingular square matrix of order n with real elements can serve as a matrix for changing from a given basis of an n-dimensional real linear space to some other basis.

Suppose we have a given basis (4) and a nonsingular matrix T of order n. For (5) take a system of vectors for which the rows of matrix T serve as the rows of coordinates in basis (4); thus, we have equation (7). The vectors (5) are linearly independent (linear dependence would have implied a linear dependence of the rows of matrix T, in conflict with its nonsingularity). Therefore, system (5), as a linearly independent system consisting of n vectors, is a basis of our space, and the matrix T serves as a change-of-basis matrix from basis (4) to basis (5).

We see that in an n-dimensional linear space we can find as many distinct bases as there are distinct nonsingular square matrices of order n. True, here, two bases consisting of the same vectors but written in a different order are considered distinct.

Transformation of vector coordinates. Suppose in an n-dimensional linear space we have the bases (4) and (5) given with the change-of-basis matrix $T = (\tau_{ij})$,

$$e' = Te$$

Let us find the connection between the coordinate rows of an arbitrary vector a in these bases.

Let

$$a = \sum_{j=1}^{n} \alpha_j e_j, \tag{8}$$

$$a = \sum_{i=1}^{n} \alpha'_i e'_i$$

Using (6) we find

$$a = \sum_{i=1}^{n} \alpha'_i \left(\sum_{j=1}^{n} \tau_{ij} e_j \right) = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} \alpha'_i \tau_{ij} \right) e_j$$

Comparing with (8) and using the uniqueness of vector notation in terms of a basis, we obtain

$$\alpha_j = \sum_{i=1}^{n} \alpha'_i \tau_{ij}, \quad j = 1, 2, \ldots, n$$

Thus we have the matrix equation

$$(\alpha_1, \alpha_2, \ldots, \alpha_n) = (\alpha'_1, \alpha'_2, \ldots, \alpha'_n)\, T$$

Thus, *the row of coordinates of the vector a in the basis e is equal to the row of coordinates of this vector in the basis e′ multiplied on the right by the change-of-basis matrix from the basis e to the basis e′.*

Whence clearly follows the equation

$$(\alpha'_1, \alpha'_2, \ldots, \alpha'_n) = (\alpha_1, \alpha_2, \ldots, \alpha_n)\, T^{-1}$$

Example. Consider a three-dimensional real linear space with the basis

$$e_1, e_2, e_3 \tag{9}$$

The vectors

$$\left.\begin{aligned} e'_1 &= 5e_1 - e_2 - 2e_3, \\ e'_2 &= 2e_1 + 3e_2, \\ e'_3 &= -2e_1 + e_2 + e_3 \end{aligned}\right\} \tag{10}$$

also form a basis in this space, the matrix

$$T = \begin{pmatrix} 5 & -1 & -2 \\ 2 & 3 & 0 \\ -2 & 1 & 1 \end{pmatrix}$$

serving as the change-of-basis matrix from (9) to (10). We then have

$$T^{-1} = \begin{pmatrix} 3 & -1 & 6 \\ -2 & 1 & -4 \\ 8 & -3 & 17 \end{pmatrix}$$

The vector

$$a = e_1 + 4e_2 - e_3$$

therefore has, in basis (10), the row of coordinates

$$(\alpha'_1, \alpha'_2, \alpha'_3) = (1, 4, -1) \begin{pmatrix} 3 & -1 & 6 \\ -2 & 1 & -4 \\ 8 & -3 & 17 \end{pmatrix} = (-13, 6, -27)$$

or

$$a = -13e'_1 + 6e'_2 - 27e'_3$$

31. Linear Transformations

In Chapter 3 we dealt with the concept of a linear transformation of unknowns. The concept we now introduce bears the same name but is different in character. True, certain relationships could be established between these two notions.

Let there be given an n-dimensional real linear space, which we denote by V_n. We consider a *transformation* of this space, that is a mapping which takes *every* vector a of V_n into some vector a' of the same space. The vector a' is called the *image* of a under the given transformation.

If we use φ to denote the transformation, then the image of vector a will be written as $a\varphi$ instead of the more customary $\varphi(a)$ or φa. Thus,

$$a' = a\varphi$$

A transformation φ of a linear space V_n is called a *linear transformation* of this space if it takes the sum of any two vectors a, b into the sum of the images of these vectors:

$$(a + b)\,\varphi = a\varphi + b\varphi \tag{1}$$

and the product of any vector a by any scalar α into the product of the image of the vector a by that same scalar α:

$$(\alpha a)\,\varphi = \alpha\,(a\varphi) \tag{2}$$

From this definition, it immediately follows that *a linear transformation of a linear space carries any linear combination of given vectors $a_1, a_2, \ldots, a_k$ into a linear combination (with the same coefficients) of the images of the vectors*:

$$(\alpha_1 a_1 + \alpha_2 a_2 + \ldots + \alpha_k a_k)\,\varphi = \alpha_1\,(a_1\varphi) + \alpha_2\,(a_2\varphi) + \ldots + \alpha_k\,(a_k\varphi) \tag{3}$$

Let us prove the following assertion.

Under any linear transformation φ of a linear space V_n, the zero vector 0 remains fixed,

$$0\varphi = 0$$

and the image of the inverse of the given vector a is a vector that is inverse to the image of a:

$$(-a)\,\varphi = -a\varphi$$

Indeed, if b is an arbitrary vector, then, by (2),

$$0\varphi = (0 \cdot b)\,\varphi = 0 \cdot (b\varphi) = 0$$

On the other hand,

$$(-a)\,\varphi = [(-1)\,a]\,\varphi = (-1)\,(a\varphi) = -a\varphi$$

The concept of a linear transformation of a linear space arose as a generalization of the familiar analytic geometry concept of the affine transformation of a plane or of three-dimensional space. Indeed, conditions (1) and (2) are fulfilled under affine transformations. These conditions are also fulfilled for projections of vectors on a plane or, in three-dimensional space, on a straight line (or a plane). Thus, for example, in a two-dimensional linear space of vector segments emanating from the origin of the plane, the transformation carrying a vector into its projection on some axis passing through the origin is a linear transformation.

Examples of linear transformations in an arbitrary space V_n are the *identity transformation* ε, which leaves every vector a fixed,

$$a\varepsilon = a$$

and the *zero transformation* ω, which maps every vector a into zero,

$$a\omega = 0$$

We will now obtain a survey of all linear transformations of a linear space V_n. Let

$$e_1, e_2, \ldots, e_n \tag{4}$$

be a basis of this space. As we have already done, denote by e the basis (4) arranged in a column. Since any vector a of the space V_n is uniquely represented as a linear combination of vectors of the basis (4), it follows, by (3), that the image of vector a with the same coefficients can be expressed in terms of the images of the vectors (4). In other words, *any linear transformation* φ *of* V_n *is uniquely determined by specifying the images* $e_1\varphi, e_2\varphi, \ldots, e_n\varphi$ *of all vectors of the fixed basis* (4).

No matter what the ordered system of n *vectors of* V_n,

$$c_1, c_2, \ldots, c_n \tag{5}$$

there is a unique linear transformation φ *of this space such that* (5) *serves as the system of images of the vectors of basis* (4) *under this transformation*,

$$e_i\varphi = c_i, \qquad i = 1, 2, \ldots, n \tag{6}$$

The uniqueness of the transformation φ has already been proved; it remains to prove its existence. Let us define the transformation φ as follows: if a is an arbitrary vector of the space and

$$a = \sum_{i=1}^{n} \alpha_i e_i$$

is its notation in the basis (4), then put

$$a\varphi = \sum_{i=1}^{n} \alpha_i c_i \tag{7}$$

Let us prove the linearity of this transformation. If

$$b = \sum_{i=1}^{n} \beta_i e_i$$

is any other vector of the space, then

$$(a + b)\varphi = [\sum_{i=1}^{n} (\alpha_i + \beta_i) e_i]\varphi = \sum_{i=1}^{n} (\alpha_i + \beta_i) c_i$$

$$= \sum_{i=1}^{n} \alpha_i c_i + \sum_{i=1}^{n} \beta_i c_i = a\varphi + b\varphi$$

But if γ is any scalar, then

$$(\gamma a)\varphi = [\sum_{i=1}^{n} (\gamma\alpha_i) e_i]\varphi = \sum_{i=1}^{n} (\gamma\alpha_i) c_i = \gamma \sum_{i=1}^{n} \alpha_i c_i = \gamma (a\varphi)$$

The correctness of (6) follows from the definition (7) of the transformation φ, since all coordinates of the vector e_i in the basis (4) are zero (except the ith coordinate, which is equal to unity).

We have thus established a one-to-one correspondence between all linear transformations of the linear space V_n and all ordered systems (5) *made up of n vectors of this space.*

However, every vector c_i has a definite notation in the basis (4):

$$c_i = \sum_{j=1}^{n} \alpha_{ij} e_j, \qquad i = 1, 2, \ldots, n \tag{8}$$

We can form a square matrix of the coordinates of the vector c_i in the basis (4)

$$A = (\alpha_{ij}) \tag{9}$$

taking for its ith row the row of coordinates of the vector c_i, $i = 1, 2, \ldots, n$. Since system (5) was arbitrary, the matrix A will be an arbitrary square matrix of order n with real elements.

We thus have a one-to-one correspondence between all linear transformations of the space V_n and all square matrices of order n; this correspondence is of course dependent on the choice of basis (4).

We shall say that the matrix A *specifies* a linear transformation φ in the basis (4) or, more succinctly, that A is the *matrix of the linear transformation* φ in the basis (4). If by $e\varphi$ we denote a column composed of the images of the vectors of (4), then from (6), (8) and (9) there follows a matrix equation which completely describes the relationships existing between the linear transformation φ, the basis e and the matrix A specifying the linear transformation in that basis:

$$e\varphi = Ae \tag{10}$$

Let us show how, knowing the matrix A of a linear transformation φ in basis (4), it is possible, via the coordinates of the vector a in this basis, to find the coordinates of its image $a\varphi$. If

$$a = \sum_{i=1}^{n} \alpha_i e_i$$

then

$$a\varphi = \sum_{i=1}^{n} \alpha_i (e_i\varphi)$$

which is equivalent to the matrix equation

$$a\varphi = (\alpha_1, \alpha_2, \ldots, \alpha_n)\, e\varphi$$

Utilizing (10) and taking into account that the associativity of matrix multiplication is easy to verify when one of the matrices is a column made up of vectors, we obtain

$$a\varphi = [(\alpha_1, \alpha_2, \ldots, \alpha_n)\, A]\, e$$

Whence it follows that *the row of coordinates of a vector $a\varphi$ is equal to the row of the coordinates of the vector a multiplied on the right by the matrix A of the linear transformation φ, all in the basis* (4).

Example. Let there be a linear transformation φ given by the following matrix in a basis e_1, e_2, e_3 of three-dimensional linear space:

$$A = \begin{pmatrix} -2 & 1 & 0 \\ 1 & 3 & 2 \\ 0 & -4 & 1 \end{pmatrix}$$

If

$$a = 5e_1 + e_2 - 2e_3$$

then

$$(5, 1, -2) \begin{pmatrix} -2 & 1 & 0 \\ 1 & 3 & 2 \\ 0 & -4 & 1 \end{pmatrix} = (-9, 16, 0)$$

that is,

$$a\varphi = -9e_1 + 16e_2$$

Relationships between matrices of a linear transformation in different bases. Quite naturally, a matrix specifying a linear transformation is dependent on the choice of the basis. We will show what the relationship is between matrices that specify one and the same linear transformation in different bases.

Let there be given the bases e and e' with change-of-basis matrix T,

$$e' = Te \tag{11}$$

and let the linear transformation φ be given in these bases by matrices A and A', respectively,

$$e\varphi = Ae, \quad e'\varphi = A'e' \tag{12}$$

By (11), the second equation of (12) reduces to

$$(Te)\,\varphi = A'\,(Te)$$

However,

$$(Te)\,\varphi = T\,(e\varphi)$$

Indeed, if $(\tau_{i1}, \tau_{i2}, \ldots, \tau_{in})$ is the ith row of matrix T, then

$$(\tau_{i1}e_1 + \tau_{i2}e_2 + \ldots + \tau_{in}e_n)\,\varphi$$
$$= \tau_{i1}\,(e_1\varphi) + \tau_{i2}\,(e_2\varphi) + \ldots + \tau_{in}\,(e_n\varphi)$$

Hence, by (12),

$$(Te)\,\varphi = T\,(e\varphi) = T\,(Ae) = (TA)\,e,$$
$$A'\,(Te) = (A'T)\,e$$

that is,

$$(TA)\,e = (A'T)\,e$$

If for at least one i, $1 \leqslant i \leqslant n$, the ith row of the matrix TA is different from the ith row of the matrix $A'T$, then two distinct linear combinations of vectors $e_1, e_2, \ldots, e_n$ will be equal to each other, which contradicts the linear independence of the basis e. Thus,

$$TA = A'T$$

whence, due to the nonsingularity of the change-of-basis matrix T,

$$A' = TAT^{-1}, \qquad A = T^{-1}A'T \tag{13}$$

Note that the square matrices B and C are called *similar* if they are connected by the equation

$$C = Q^{-1}BQ$$

where Q is some nonsingular matrix. We say that the matrix C is obtained from B by a *transformation* by the matrix Q.

The equations (13) proved above may be formulated as an important theorem.

Matrices which represent one and the same linear transformation in different bases are similar. And the matrix of the linear transformation φ in the basis e' is obtained by transforming the matrix of this linear transformation in the basis e via the change-of-basis matrix from basis e' to basis e.

Let us piont out that if a matrix A represents a linear transformation φ in the basis e, then any matrix B, similar to A,

$$B = Q^{-1}AQ$$

also represents the transformation φ in some basis, namely, in the basis obtained from e by means of the change-of-basis matrix Q^{-1}.

Operations on linear transformations. Associating to every linear transformation of the space V_n its matrix in a fixed basis, we obtain (as was proved above) a one-to-one correspondence between all linear transformations and all square matrices of order n. It is natural to expect that the operations of addition and multiplication of matrices and also matrix multiplication by a scalar will be associated with analogous operations involving linear transformations.

Suppose we have the linear transformations φ and ψ in a space V_n. The *sum* of these transformations is the transformation $\varphi + \psi$ defined by the equation

$$a(\varphi + \psi) = a\varphi + a\psi \tag{14}$$

It thus carries any vector a into the sum of its images under the transformations φ and ψ.

The transformation $\varphi + \psi$ *is linear.* Indeed, for all vectors a and b and any scalar α,

$$\begin{aligned}(a+b)(\varphi+\psi) &= (a+b)\varphi + (a+b)\psi \\ &= a\varphi + b\varphi + a\psi + b\psi = a(\varphi+\psi) + b(\varphi+\psi),\\ (\alpha a)(\varphi+\psi) &= (\alpha a)\varphi + (\alpha a)\psi = \alpha(a\varphi) + \alpha(a\psi)\\ &= \alpha(a\varphi + a\psi) = \alpha[a(\varphi+\psi)]\end{aligned}$$

On the other hand, we use the term "*product*" of linear transformations φ and ψ for the transformation $\varphi\psi$ defined by the equation

$$a(\varphi\psi) = (a\varphi)\psi \tag{15}$$

that is, the transformation obtained by successive application of the transformations φ and ψ.

The transformation $\varphi\psi$ *is linear*:

$$\begin{aligned}(a+b)(\varphi\psi) &= [(a+b)\varphi]\psi = (a\varphi + b\varphi)\psi\\ &= (a\varphi)\psi + (b\varphi)\psi = a(\varphi\psi) + b(\varphi\psi),\\ (\alpha a)(\varphi\psi) &= [(\alpha a)\varphi]\psi = [\alpha(a\varphi)]\psi = \alpha[(a\varphi)\psi] = \alpha[a(\varphi\psi)]\end{aligned}$$

Finally, we use the term "*product*" *of a linear transformation* φ *by a scalar* $\varkappa$ for the transformation $\varkappa\varphi$ defined by

$$a(\varkappa\varphi) = \varkappa(a\varphi) \tag{16}$$

Thus, in the φ-transformation of all vectors, the images are multiplied by the scalar $\varkappa$.

The transformation $\varkappa\varphi$ *is linear:*

$$(a+b)(\varkappa\varphi) = \varkappa[(a+b)\varphi] = \varkappa(a\varphi + b\varphi)$$
$$= \varkappa(a\varphi) + \varkappa(b\varphi) = a(\varkappa\varphi) + b(\varkappa\varphi)$$
$$(\alpha a)(\varkappa\varphi) = \varkappa[(\alpha a)\varphi] = \varkappa[\alpha(a\varphi)] =$$
$$= \alpha[\varkappa(a\varphi)] = \alpha[a(\varkappa\varphi)]$$

Let the transformations φ and ψ be given in the basis $e_1, e_2, \ldots, e_n$, by the matrices $A = (\alpha_{ij})$ and $B = (\beta_{ij})$, respectively,

$$e\varphi = Ae, \quad e\psi = Be$$

Then, by (14),

$$e_i(\varphi+\psi) = e_i\varphi + e_i\psi = \sum_{j=1}^{n} \alpha_{ij}e_j + \sum_{j=1}^{n} \beta_{ij}e_j = \sum_{j=1}^{n} (\alpha_{ij}+\beta_{ij})e_j$$

that is,

$$e(\varphi+\psi) = (A+B)e$$

Thus, *the matrix of a sum of linear transformations in any basis is equal to the sum of the matrices of these transformations in the same basis.*

On the other hand, by (15),

$$e_i(\varphi\psi) = (e_i\varphi)\psi = \Big(\sum_{j=1}^{n} \alpha_{ij}e_j\Big)\psi = \sum_{j=1}^{n} \alpha_{ij}(e_j\psi)$$
$$= \sum_{j=1}^{n} \alpha_{ij}\Big(\sum_{k=1}^{n} \beta_{jk}e_k\Big) = \sum_{k=1}^{n}\Big(\sum_{j=1}^{n} \alpha_{ij}\beta_{jk}\Big)e_k$$

that is,

$$e(\varphi\psi) = (AB)e$$

In other words, *the matrix of a product of linear transformations in any basis is equal to the product of the matrices of these transformations in the same basis.*

Finally, due to (16),

$$e_i(\varkappa\varphi) = \varkappa(e_i\varphi) = \varkappa\sum_{j=1}^{n} \alpha_{ij}e_j = \sum_{j=1}^{n} (\varkappa\alpha_{ij})e_j$$

that is,

$$e(\varkappa\varphi) = (\varkappa A)e$$

Consequently, *a matrix which in some basis specifies the product of a linear transformation* φ *by a scalar* $\varkappa$ *is equal to the product of the matrix of the transformation* φ *in this basis by the scalar* $\varkappa$.

From the results obtained it follows that operations on linear transformations possess the same properties as operations on matrices. Thus, the addition of linear transformations is commutative and associative, while multiplication is associative but is not commutative for $n > 1$. For linear transformations there exists unique

subtraction. Also note that *in linear transformations, the identity transformation* ε *plays the role of unity, and the zero transformation* ω, *the role of zero.* In any basis, the transformation ε is given by the unit matrix, and the transformation ω is given by the zero matrix.

32. Linear Subspaces

A subset L of a linear space V is called a *linear subspace* of this space if it is a linear space with respect to the operations defined in V of addition of vectors and the multiplication of a vector by a scalar. Thus, in three-dimensional Euclidean space, the collection of vectors emanating from the coordinate origin and lying in some plane (or on some straight line) passing through the origin is a linear subspace.

For a nonempty subset L *of space* V *to be a linear subspace of* V, *the following requirements must be met.*

1. *If the vectors* a *and* b *lie in* L, *then the vector* $a + b$ *also belongs to* L.

2. *If the vector* a *belongs to* L, *then the vector* αa, *for any value of the scalar* α, *belongs to* L *too.*

Indeed, by Condition 2, the set L contains the zero vector: if vector a belongs to L, then L also contains $0 \cdot a = 0$. Furthermore, again by Property 2, L contains a vector a and the inverse vector $-a = (-1) \cdot a$, and therefore, due to Property 1, L also contains the difference of any two vectors in L. As to all the other requirements that enter into the definition of a linear space, we can say that if they are fulfilled in V, then they will likewise be fulfilled in L.

Instances of linear subspaces of the space V are: the space V itself and also the set consisting of a single zero vector, the so-called *zero subspace.* A more interesting example is the following: in the space V take any finite system of vectors

$$a_1, a_2, \ldots, a_r \tag{1}$$

and denote by L the set of all those vectors which are linear combinations of the vectors of (1). We will prove that L is a linear subspace. Indeed, if

$$b = \alpha_1 a_1 + \alpha_2 a_2 + \ldots + \alpha_r a_r, \quad c = \beta_1 a_1 + \beta_2 a_2 + \ldots + \beta_r a_r$$

then

$$b + c = (\alpha_1 + \beta_1) a_1 + (\alpha_2 + \beta_2) a_2 + \ldots + (\alpha_r + \beta_r) a_r$$

that is, the vector $b + c$ belongs to L; also in L is the vector

$$\gamma b = (\gamma \alpha_1) a_1 + (\gamma \alpha_2) a_2 + \ldots + (\gamma \alpha_r) a_r$$

for any scalar γ.

We say that this linear subspace L *is generated* by the system of vectors (1); in particular, the vectors (1) themselves belong to L.

Incidentally, *any linear subspace of a finite-dimensional linear space is generated by a finite system of vectors*, for if it is not a zero subspace, then it possesses a finite basis. The dimension of the linear subspace L is not greater than the dimension n of the space V_n itself and is equal to n only when $L = V_n$. The dimension of the zero subspace is of course the number 0.

For any k, $0 < k < n$, *in the space* V_n *there are linear subspaces of dimension* k. It is sufficient to take a subspace generated by any system of k linearly independent vectors.

Let there be given linear subspaces L_1 and L_2 in the space V. The collection L_0 of vectors belonging both to L_1 and to L_2 will be a linear subspace, as can readily be verified. It is the *intersection* of the subspaces L_1 and L_2. On the other hand, another linear subspace is the sum $\bar{L}$ of the subspaces L_1 and L_2, or the collection of all those vectors in V which can be represented as a sum of two terms, one from L_1 and the other from L_2. If the dimensions of the subspaces L_1, L_2, L_0 and $\bar{L}$ are, respectively, d_1, d_2, d_0 and $\bar{d}$, then the following formula holds:

$$\bar{d} = d_1 + d_2 - d_0 \tag{2}$$

which is to say that *the dimension of the sum of two subspaces is equal to the sum of the dimensions of these subspaces diminished by the dimension of their intersection.*

To prove this, let us take an arbitrary basis

$$a_1, a_2, \ldots, a_{d_0} \tag{3}$$

of subspace L_0 and augment it to obtain the basis

$$a_1, a_2, \ldots, a_{d_0}, \quad b_{d_0+1}, \ldots, b_{d_1} \tag{4}$$

of the subspace L_1 and also augment it to obtain the basis

$$a_1, a_2, \ldots, a_{d_0}, \quad c_{d_0+1}, \ldots, c_{d_2} \tag{5}$$

of the subspace L_2. Utilizing the definition of the subspace $\bar{L}$, it is easy to see that this subspace is generated by the system of vectors

$$a_1, a_2, \ldots, a_{d_0}, \quad b_{d_0+1}, \ldots, b_{d_1}, \quad c_{d_0+1}, \ldots, c_{d_2} \tag{6}$$

Formula (2) will thus be proved if we demonstrate the linear independence of system (6).

Suppose the equation

$$\alpha_1 a_1 + \alpha_2 a_2 + \ldots + \alpha_{d_0} a_{d_0} + \beta_{d_0+1} b_{d_0+1} + \ldots + \beta_{d_1} b_{d_1} + \gamma_{d_0+1} c_{d_0+1} + \ldots + \gamma_{d_2} c_{d_2} = 0$$

with certain numerical coefficients is true. Then

$$d = \alpha_1 a_1 + \alpha_2 a_2 + \ldots + \alpha_{d_0} a_{d_0} + \beta_{d_0+1} b_{d_0+1} + \ldots + \beta_{d_1} b_{d_1} = -\gamma_{d_0+1} c_{d_0+1} - \ldots - \gamma_{d_2} c_{d_2} \tag{7}$$

The left member of this equation lies in L_1, the right member in L_2, therefore vector d (which is equal both to the left and to the right

member of this equation) belongs to L_0 and, consequently, can be expressed linearly in terms of the basis (3). However, the right member of (7) shows that the vector d can also be expressed linearly in terms of the vectors $c_{d_0+1}, \ldots, c_{d_2}$. Whence, by the linear independence of system (5), it follows that all the coefficients $\gamma_{d_0+1}, \ldots, \gamma_{d_2}$ are zero, that is, that $d = 0$; but then, because of the linear independence of system (4), all the coefficients $\alpha_1, \ldots, \alpha_{d_0}$, $\beta_{d_0+1}, \ldots, \beta_{d_1}$ are also zero. This proves the linear independence of system (6).

The reader can verify that our proof holds true for the case when the subspace L_0 is a zero subspace, i.e., $d_0 = 0$.

The range of values and the kernel (null space) of a linear transformation. Suppose we have a linear transformation φ in a linear space V_n. If L is any linear subspace of the space V_n, then the collection $L\varphi$ of images of all vectors of L under the transformation φ will also be a linear subspace, as follows directly from the definitions of a linear subspace and a linear transformation. In particular, the collection $V_n\varphi$ of images of all vectors of the space V_n is a linear subspace. It is called the *range of values* of the transformation φ. Let us find the dimension of the range. To do this, note that since all matrices representing the transformation φ in different bases are similar, it follows, due to the last theorem of Sec. 14, that they all have one and the same rank. This number can therefore be termed the *rank* of the linear transformation φ.

The dimension of the range of values of a linear transformation φ *is equal to the rank of the transformation.*

Indeed, let φ be represented in the basis $e_1, e_2, \ldots, e_n$ by the matrix A. The subspace $V_n\varphi$ is generated by the vectors

$$e_1\varphi, e_2\varphi, \ldots, e_n\varphi \tag{8}$$

and therefore, as a particular case, any maximal linearly independent subsystem of system (8) will serve as a basis of the subspace $V_n\varphi$. However, the maximum number of linearly independent vectors in system (8) is equal to the maximum number of linearly independent rows of the matrix A, i.e., it is equal to the rank of the matrix. The theorem is proved.

We know that under the linear transformation φ the zero vector goes into itself. The collection $N(\varphi)$ of all vectors of the space V_n which under φ are mapped into the zero vector is consequently non-void and is evidently a linear subspace. This subspace is termed the *null space* of the transformation φ, and its dimension is called the *nullity* of this transformation.

For any linear transformation φ *of space* V_n, *the sum of the rank and of the nullity of the transformation is equal to the dimension* n *of the whole space.*

Indeed, if r is the rank of the transformation φ, then the subspace $V_n\varphi$ has the following basis of r vectors:

$$a_1,\ a_2,\ \ldots,\ a_r \tag{9}$$

In V_n we can select the vectors

$$b_1,\ b_2,\ \ldots,\ b_r \tag{10}$$

such that

$$b_i\varphi = a_i, \qquad i = 1,\ 2,\ \ldots,\ r$$

The choice of vectors (10) is not unambiguous, naturally. If some nontrivial linear combination of vectors (10) were mapped into zero by the transformation φ, in particular, if the vectors (10) were linearly dependent, then the vectors (9) would themselves be linearly dependent, but this runs counter to our assumption. And so the linear subspace L generated by the vectors (10) has dimension r and its intersection with the subspace $N(\varphi)$ is zero.

On the other hand, the sum of the subspaces L and $N(\varphi)$ coincides with the entire space V_n. Indeed, if c is any vector of the space, it follows that the vector $d = c\varphi$ of course belongs to the subspace $V_n\varphi$. Then in the subspace L there will be a vector b such that

$$b\varphi = d$$

The vector b is written in terms of system (10) with the same coefficients as is the vector d in terms of the basis (9). From this we have

$$c = b + (c - b)$$

and the vector $c - b$ is contained in the subspace $N(\varphi)$, since

$$(c - b)\,\varphi = c\varphi - b\varphi = d - d = 0$$

The assertion of the theorem follows from the results obtained and from the formula (2) that was proved earlier.

Nonsingular linear transformations. A linear transformation φ of a linear space V_n is called *nonsingular* if it satisfies any one of the following conditions, the equivalence of which follows directly from the theorems proved above.

1. The rank of the transformation φ is equal to n.
2. The entire space V_n serves as the range of values of the transformation φ.
3. The nullity of the transformation φ is zero.

There are many other definitions of nonsingular linear transformations that are equivalent to those given above, for instance, definitions 4 to 6.

4. Distinct vectors of the space V_n have distinct images under the transformation φ.

Indeed, if a transformation φ has Property 4, then the null space of this transformation consists of the zero vector alone, i.e., Property 3 holds. But if the vectors a and b are such that $a \neq b$, but $a\varphi = b\varphi$, then $a - b \neq 0$, but $(a - b)\,\varphi = 0$, or Property 3 does not hold.

From 2 and 4 there follows

5. The transformation φ is a one-to-one mapping of the space V_n onto this whole space.

From 5 it follows that a nonsingular linear transformation φ has an *inverse transformation* φ^{-1} which carries any vector $a\varphi$ into the vector a,

$$(a\varphi)\,\varphi^{-1} = a$$

The transformation φ^{-1} *is linear* since

$$(a\varphi + b\varphi)\,\varphi^{-1} = [(a + b)\,\varphi]\,\varphi^{-1} = a + b,$$
$$[\alpha\,(a\varphi)]\,\varphi^{-1} = [(\alpha a)\,\varphi]\,\varphi^{-1} = \alpha a$$

From the definition of the transformation φ^{-1} it follows that

$$\varphi\varphi^{-1} = \varphi^{-1}\varphi = \varepsilon \tag{11}$$

The equalities (11) can themselves be viewed as a definition of an inverse transformation. Then from this and from the last results of the preceding section it follows that *if a nonsingular linear transformation* φ *is represented in some basis by the matrix* A *(which is nonsingular due to Property 1), then the transformation* φ^{-1} *is represented in that basis by the matrix* A^{-1}.

We thus arrive at the following definition of a nonsingular linear transformation.

6. A transformation φ has an inverse linear transformation φ^{-1}.

33. Characteristic Roots and Eigenvalues

Let $A = (\alpha_{ij})$ be a square matrix of order n with real elements. On the other hand, let λ be some unknown. Then the matrix $A - \lambda E$, where E is a unit matrix of order n, is called the *characteristic matrix* of the matrix A. Since in the matrix λE the principal diagonal is occupied by λ and all other elements are zero, we have

$$A - \lambda E = \begin{pmatrix} \alpha_{11} - \lambda & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} - \lambda & \dots & \alpha_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} - \lambda \end{pmatrix}$$

The determinant of the matrix $A - \lambda E$ is a polynomial in λ of degree n. Indeed, the product of elements on the principal diagonal is a polynomial in λ with highest-degree term $(-1)^n \lambda^n$; all

the other terms of the determinant do not contain at least two of the number of elements on the principal diagonal; therefore, their degree in λ does not exceed $n-2$. It is easy to find the coefficients of this polynomial. For instance, the coefficient of λ^{n-1} is equal to $(-1)^{n-1}(\alpha_{11}+\alpha_{22}+\ldots+\alpha_{nn})$ and the constant term coincides with the determinant of matrix A.

The polynomial $|A-\lambda E|$ of degree n is called the *characteristic polynomial* of matrix A, and its roots (which may be real or complex) are termed the *characteristic roots* of the matrix.

Similar matrices have the same characteristic polynomials, and, consequently, the same characteristic roots.

To see this, let

$$B = Q^{-1}AQ$$

Then, taking into account that the matrix λE commutes with the matrix Q, and $|Q^{-1}| = |Q|^{-1}$, we have

$$|B-\lambda E| = |Q^{-1}AQ-\lambda E| = |Q^{-1}(A-\lambda E)Q|$$
$$= |Q|^{-1}\cdot|A-\lambda E|\cdot|Q| = |A-\lambda E|$$

The proof is complete.

From this result it follows (by the theorem proved in Sec. 31 on the relationship between matrices representing a linear transformation in different bases) that *although the linear transformation* φ *may be represented in different bases by different matrices, all the matrices have one and the same set of characteristic roots.* These roots can therefore be called the *characteristic roots of the transformation* φ. The set of these characteristic roots, each root being taken with the multiplicity that it has in the characteristic polynomial, is called the *spectrum* of the linear transformation φ.

Characteristic roots play a very important role in the study of linear transformations, as the reader will have ample opportunity to see. We now investigate one of the applications of characteristic roots.

Let there be given a linear transformation φ in a real linear space V_n. If a vector b (nonzero) is carried by the transformation φ into a vector proportional to b,

$$b\varphi = \lambda_0 b \tag{1}$$

where λ_0 is some real number, then the vector b is called the *eigenvector* of the transformation φ, and the number λ_0 is the *eigenvalue* of this transformation. We say that the eigenvector b *corresponds* to the eigenvalue λ_0.

Note that since $b \neq 0$, the number λ_0 which satisfies Condition (1) is uniquely defined for the vector b. Also bear in mind that the zero vector is not considered to be an eigenvector of the transformation φ, although it satisfies Condition (1) for any λ_0.

Rotation of the Euclidean plane about the origin through an angle that is not a multiple of π is an example of a linear transformation which has no eigenvectors. An instance of another extreme case is the stretching of a plane in which all vectors issuing from the origin are stretched, say, five times. This is a linear transformation and all nonzero vectors of the plane are its eigenvectors; all of them correspond to the eigenvalue 5.

Only the real characteristic roots (if they exist) of a linear transformation φ serve as eigenvalues of the transformation.

Let a transformation φ have a matrix $A = (\alpha_{ij})$ in the basis e_1, e_2, . . ., e_n and let the vector

$$b = \sum_{i=1}^{n} \beta_i e_i$$

be an eigenvector of the transformation φ

$$b\varphi = \lambda_0 b \tag{2}$$

As was proved in Sec. 31,

$$b\varphi = [(\beta_1, \beta_2, \ldots, \beta_n) A] e \tag{3}$$

Equations (2) and (3) lead to the system of equations

$$\begin{aligned} \beta_1\alpha_{11} + \beta_2\alpha_{21} + \ldots + \beta_n\alpha_{n1} &= \lambda_0\beta_1, \\ \beta_1\alpha_{12} + \beta_2\alpha_{22} + \ldots + \beta_n\alpha_{n2} &= \lambda_0\beta_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ \beta_1\alpha_{1n} + \beta_2\alpha_{2n} + \ldots + \beta_n\alpha_{nn} &= \lambda_0\beta_n \end{aligned} \tag{4}$$

Since $b \neq 0$, not all the numbers $\beta_1, \beta_2, \ldots, \beta_n$ are zero. Thus, by (4), the system of homogeneous linear equations

$$\begin{aligned} (\alpha_{11} - \lambda_0) x_1 + \alpha_{21}x_2 + \ldots + \alpha_{n1}x_n &= 0, \\ \alpha_{12}x_1 + (\alpha_{22} - \lambda_0) x_2 + \ldots + \alpha_{n2}x_n &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ \alpha_{1n}x_1 + \alpha_{2n}x_2 + \ldots + (\alpha_{nn} - \lambda_0) x_n &= 0 \end{aligned} \tag{5}$$

has a nontrivial solution and for this reason its determinant is equal to zero:

$$\begin{vmatrix} \alpha_{11} - \lambda_0, & \alpha_{21}, & \ldots, & \alpha_{n1} \\ \alpha_{12}, & \alpha_{22} - \lambda_0, & \ldots, & \alpha_{n2} \\ \ldots\ldots\ldots\ldots\ldots\ldots & & & \\ \alpha_{1n}, & \alpha_{2n}, & \ldots, & \alpha_{nn} - \lambda_0 \end{vmatrix} = 0 \tag{6}$$

Taking the transpose, we get

$$| A - \lambda_0 E | = 0 \tag{7}$$

that is to say, the eigenvalue λ_0 actually does prove to be a characteristic root (and, quite naturally, a real root) of the matrix A and, hence, of the linear transformation φ.

Conversely, let λ_0 be any real characteristic root of the transformation φ and, consequently, of the matrix A. Then we have equation (7) and therefore equation (6), which was obtained from (7) by taking the transpose. From this it follows that the system of homogeneous linear equations (5) has a nontrivial solution, and even a real one, since all the coefficients of the system are real. If we denote this solution by

$$(\beta_1, \beta_2, \ldots, \beta_n) \tag{8}$$

we have equations (4). Use b to denote the vector of space V_n having in the basis $e_1, e_2, \ldots, e_n$ the coordinate row (8). It is clear that $b \neq 0$. Then equation (3) holds and from (4) and (3) follows (2). Thus, vector b has proved to be an eigenvector (of the transformation φ) corresponding to the eigenvalue λ_0. This proves the theorem.

Note that if we considered a complex linear space, then the demand that the characteristic root be real would be superfluous. In other words, we would have proved the following theorem: *The characteristic roots of a linear transformation of a complex linear space, and only these roots, serve as eigenvalues of the transformation.* Whence it follows that *in a complex linear space, any linear transformation has eigenvectors.*

Returning to our real case, note that the collection of eigenvectors of the linear transformation φ which correspond to the eigenvalue λ_0 coincides with the collection of nontrivial real solutions of the system of homogeneous linear equations (5). Whence it follows that *the collection of eigenvectors of the linear transformation φ which correspond to the eigenvalue λ_0 will, after the zero vector has been adjoined to it, be a linear subspace of the space V_n.* Indeed, from what was proved in Sec. 12, it follows that *the collection of (real) solutions of any system of homogeneous linear equations in n unknowns is a linear subspace of the space V_n.*

Linear transformations with a simple spectrum. In many cases it is necessary to know whether a given linear transformation φ can have a diagonal matrix in some basis. As a matter of fact, by far not every linear transformation can be represented by a diagonal matrix. The necessary and sufficient conditions for this will be indicated in Sec. 61. In the meantime we wish to indicate one sufficient condition.

We will first prove the following auxiliary results.

A linear transformation φ is represented by a diagonal matrix in a basis $e_1, e_2, \ldots, e_n$ if and only if all the vectors of the basis are eigenvectors of the transformation φ.

Indeed, the equation

$$e_i\varphi = \lambda_i e_i$$

is equivalent to the fact that in the ith row of the matrix representing the transformation φ in the indicated basis all off-diagonal elements are zero and the principal diagonal has the number λ_i (in the ith position).

The eigenvectors $b_1, b_2, \ldots, b_k$ of the linear transformation φ which correspond to different eigenvalues constitute a linearly independent system.

We shall prove this assertion by induction with respect to k, since for $k = 1$ it holds true: a single eigenvector, being nonzero, constitutes a linearly independent system. Let

$$b_i\varphi = \lambda_i b_i, \qquad i = 1, 2, \ldots, k$$

and

$$\lambda_i \neq \lambda_j \quad \text{for} \quad i \neq j$$

If there exists a linear dependence

$$\alpha_1 b_1 + \alpha_2 b_2 + \ldots + \alpha_k b_k = 0 \tag{9}$$

where, for example, $\alpha_1 \neq 0$, then, applying the transformation φ to both sides of (9), we get

$$\alpha_1\lambda_1 b_1 + \alpha_2\lambda_2 b_2 + \ldots + \alpha_k\lambda_k b_k = 0$$

Subtracting equation (9) multiplied by λ_k we get

$$\alpha_1 (\lambda_1 - \lambda_k) b_1 + \alpha_2 (\lambda_2 - \lambda_k) b_2 + \ldots + \alpha_{k-1} (\lambda_{k-1} - \lambda_k) b_{k-1} = 0$$

which yields a nontrivial linear dependence between the vectors b_1, $b_2, \ldots, b_{k-1}$ since $\alpha_1 (\lambda_1 - \lambda_k) \neq 0$.

We say that a linear transformation φ of a real linear space V_n has a *simple spectrum* if all its characteristic roots are real and distinct. Consequently, the transformation φ has n distinct eigenvalues and therefore, by the theorem just proved, the space V_n has a basis composed of the eigenvectors of this transformation. Thus, *any linear transformation with a simple spectrum may be represented by a diagonal matrix.*

Passing from the linear transformation to the matrix representing it, we obtain the following result.

Any matrix whose characteristic roots are all real and distinct is similar to a diagonal matrix, or we say that *such a matrix can be reduced to diagonal form* (*diagonalized*).

CHAPTER 8

EUCLIDEAN SPACES

34. Definition of a Euclidean Space. Orthonormal Bases

The concept of an n-dimensional linear space does not by any means fully generalize the concept of a plane or three-dimensional Euclidean space: in the n-dimensional case, for $n > 3$, neither the length of a vector nor the angle between vectors is defined and it is therefore impossible to develop the rich geometrical theory so familiar to the reader for $n = 2$ and $n = 3$. It turns out, however, that we can rectify the situation in the following manner.

From analytic geometry we know that for two-dimensional (a plane) and three-dimensional space we can introduce the concept of scalar multiplication of vectors. It is defined by means of the lengths of the vectors and the angle between them; it appears, however, that both the length of a vector and the angle between vectors can, in turn, be expressed in terms of scalar products. We will therefore define the concept of scalar multiplication (we will define it axiomatically) for any n-dimensional linear space. This will be done with the aid of certain properties which we know the scalar multiplication of vectors in the plane or in three-dimensional space actually possesses. Considering the immediate reasons for this material being included in the course of higher algebra, we dispense with the definitions of the length of a vector and the angle between vectors. The reader interested in the construction of geometry in n-dimensional spaces is referred to the special literature, in particular, to more exhaustive texts on linear algebra.

The reader should bear in mind that, with the exception of the end of this section, the whole chapter deals solely with *real* linear spaces.

We shall say that *scalar multiplication* is defined in an n-dimensional real linear space V_n if to every pair of vectors a, b there is associated a real number denoted by the symbol (a, b) and called the *scalar product* of the vectors a and b. The following conditions are satisfied (here, a, b, c, are any vectors of the space V_n, and α

is any real number, or scalar):

I. $$(a, b) = (b, a).$$

II. $$(a + b, c) = (a, c) + (b, c).$$

III. $$(\alpha a, b) = \alpha (a, b).$$

IV. If $a \neq 0$, then the scalar square of the vector a is strictly positive

$$(a, a) > 0$$

Note that from III we have, for $\alpha = 0$, the equation

$$(0, b) = 0 \tag{1}$$

which states that *the scalar product of the zero vector by any vector b is zero*: in particular, the scalar square of the zero vector is also zero.

From II and III there immediately follows a formula for the scalar product of linear combinations of two systems of vectors:

$$\left(\sum_{i=1}^{k} \alpha_i a_i, \sum_{j=1}^{l} \beta_j b_j\right) = \sum_{i=1}^{k} \sum_{j=1}^{l} \alpha_i \beta_j (a_i, b_j) \tag{2}$$

If scalar multiplication is defined in an n-dimensional linear space, then the space is termed n-dimensional *Euclidean space.*

It is possible to define scalar multiplication in an n-dimensional linear space V_n for any n, which is to say that we can convert this space into a Euclidean space.

Indeed, in V_n take any basis $e_1, e_2, \ldots, e_n$. If

$$a = \sum_{i=1}^{n} \alpha_i e_i, \quad b = \sum_{i=1}^{n} \beta_i e_i$$

then put

$$(a, b) = \sum_{i=1}^{n} \alpha_i \beta_i \tag{3}$$

It is easy to see that Conditions I-IV will be fulfilled, that is, equation (1) defines scalar multiplication in the space V_n.

Generally speaking, we see that in n-dimensional linear space it is possible to specify scalar multiplication in many different ways. Naturally, definition (3) depends on the choice of the basis, but as yet we do not know whether it is possible to introduce scalar multiplication in any other fundamentally different manner or not. Our immediate purpose is to survey all possible modes of converting n-dimensional linear space into Euclidean space and of establishing the fact that in a certain sense there is only one n-dimensional Euclidean space for any n.

Suppose we have an arbitrary n-dimensional Euclidean space E_n, which means that scalar multiplication has been introduced in some

fashion into an n-dimensional linear space. The vectors a and b are *orthogonal* if their scalar product is zero,

$$(a,\ b) = 0$$

From (1) it follows that the zero vector is orthogonal to any vector; however, there can be nonzero orthogonal vectors too.

A set of vectors is called an *orthogonal system* if all the vectors are pairwise orthogonal.

Every orthogonal system of nonzero vectors is linearly independent.

Indeed, let there be a system of vectors $a_1, a_2, \ldots, a_k$ in E_n and let $a_i \neq 0$, $i = 1, 2, \ldots, k$ and

$$(a_i,\ a_j) = 0, \qquad i \neq j \tag{4}$$

If

$$\alpha_1 a_1 + \alpha_2 a_2 + \ldots + \alpha_k a_k = 0$$

then by forming the scalar product of both sides of this equation by the vector a_i, $1 \leqslant i \leqslant k$, we get [by (1), (2) and (4)]

$$\begin{aligned} 0 = (0,\ a_i) &= (\alpha_1 a_1 + \alpha_2 a_2 + \ldots + \alpha_k a_k,\ a_i) \\ &= \alpha_1 (a_1,\ a_i) + \alpha_2 (a_2,\ a_i) + \ldots + \alpha_k (a_k,\ a_i) \\ &= \alpha_i (a_i,\ a_i) \end{aligned}$$

Whence, since $(a_i,\ a_i) > 0$ by IV, it follows that $\alpha_i = 0$, $i = 1, 2, \ldots, k$, which is what we set out to prove.

We now describe the *orthogonalization process*, which is a means of passing from any linearly independent system of k vectors

$$a_1,\ a_2,\ \ldots,\ a_k \tag{5}$$

of Euclidean space E_n to an orthogonal system, also consisting of k nonzero vectors. We denote these vectors by $b_1, b_2, \ldots, b_k$.

Let us put $b_1 = a_1$, which is to say that the first vector of system (5) will enter into the orthogonal system we are building. After that, put

$$b_2 = \alpha_1 b_1 + a_2$$

Since $b_1 = a_1$ and the vectors a_1 and a_2 are linearly independent, it follows that the vector b_2 is different from zero for any scalar α_1. We choose this scalar remembering that the vector b_2 must be orthogonal to the vector b_1:

$$0 = (b_1,\ b_2) = (b_1,\ \alpha_1 b_1 + a_2) = \alpha_1 (b_1,\ b_1) + (b_1,\ a_2)$$

whence, by IV

$$\alpha_1 = -\frac{(b_1,\ a_2)}{(b_1,\ b_1)}$$

Suppose an orthogonal system of nonzero vectors $b_1, b_2, \ldots, bl$ has already been constructed; we also assume that for any i, $1 \leqslant$

$\leqslant i \leqslant l$, the vector b_i is a linear combination of the vectors a_1, a_2, . . ., a_i. Then this assumption will also hold for some vector b_{l+1} if it is chosen in the form

$$b_{l+1} = \alpha_1 b_1 + \alpha_2 b_2 + \ldots + \alpha_l b_l + a_{l+1}$$

The vector b_{l+1} will then be different from zero, since system (5) is linearly independent and the vector a_{l+1} does not enter into the notation of vectors b_1, b_2, . . ., b_l. We choose the coefficients α_i, $i = 1, 2, \ldots, l$, from the fact that the vector b_{l+1} must be orthogonal to all the vectors b_i, $i = 1, 2, \ldots, l$:

$$0 = (b_i,\ b_{l+1}) = (b_i,\ \alpha_1 b_1 + \alpha_2 b_2 + \ldots + \alpha_l b_l + a_{l+1})$$
$$= \alpha_1 (b_i,\ b_1) + \alpha_2 (b_i,\ b_2) + \ldots + \alpha_l (b_i,\ b_l) + (b_i,\ a_{l+1})$$

whence, since the vectors b_1, b_2, . . ., b_l are mutually orthogonal,

$$\alpha_i (b_i,\ b_i) + (b_i,\ a_{l+1}) = 0$$

or

$$\alpha_i = -\frac{(b_i,\ a_{l+1})}{(b_i,\ b_i)}, \quad i = 1, 2, \ldots, l$$

Continuing this process, we can construct the desired orthogonal system b_1, b_2, . . ., b_k.

Applying the orthogonalization process to an arbitrary basis of the space E_n, we obtain an orthogonal system of n nonzero vectors, that is to say, an *orthogonal basis*, since (as has been proved) this system is linearly independent. Now, using the remark made in connection with the first step of the process of orthogonalization, and also taking into account the fact that any nonzero vector may be included in some basis of the space, we can even make the following assertion.

Every Euclidean space possesses orthogonal bases, and any nonzero vector of this space enters into some orthogonal basis.

In what follows, an important role will be played by a special type of orthogonal basis. Basis of this kind correspond to the rectangular Cartesian systems of coordinates used in analytic geometry.

We shall call a vector b *normalized* if its scalar square is equal to unity

$$(b,\ b) = 1$$

If $a \neq 0$, whence $(a, a) > 0$, then the transition to the vector

$$b = \frac{1}{\sqrt{(a,\ a)}}\, a$$

is termed *normalization* of the vector a. The vector b is normalized since

$$(b, b) = \left(\frac{1}{\sqrt{(a, a)}}\, a,\ \frac{1}{\sqrt{(a, a)}}\, a\right) = \left(\frac{1}{\sqrt{(a, a)}}\right)^2 (a, a) = 1$$

A basis $e_1, e_2, \ldots, e_n$ for the Euclidean space E_n is called *orthonormal* if it is orthogonal and all its vectors are normalized, that is,

$$\begin{aligned} (e_i, e_j) &= 0, \qquad i \neq j \\ (e_i, e_i) &= 1, \qquad i = 1, 2, \ldots, n \end{aligned} \tag{6}$$

Every Euclidean space has orthonormal bases.

To prove this, it will suffice to take any orthogonal basis and to normalize all its vectors. The basis will remain orthogonal, since for any α and β it follows from $(a, b) = 0$ that

$$(\alpha a, \beta b) = \alpha\beta\,(a, b) = 0$$

A basis $e_1, e_2, \ldots, e_n$ of a Euclidean space E_n is orthonormal if and only if the scalar product of any two vectors of the space is equal to the sum of the products of the corresponding coordinates of the vectors in the indicated basis; that is, from

$$a = \sum_{i=1}^{n} \alpha_i e_i, \qquad b = \sum_{j=1}^{n} \beta_j e_j \tag{7}$$

follows

$$(a, b) = \sum_{i=1}^{n} \alpha_i \beta_i \tag{8}$$

Indeed, if equations (6) hold for our basis, then

$$(a, b) = \left(\sum_{i=1}^{n} \alpha_i e_i,\ \sum_{j=1}^{n} \beta_j e_j\right) = \sum_{i,\, j=1}^{n} \alpha_i \beta_j\,(e_i, e_j) = \sum_{i=1}^{n} \alpha_i \beta_i$$

Conversely, if our basis is such that for any vectors a and b written in this basis in the form (7), equation (8) holds true, then, taking for a and b any two vectors e_i and e_j in the basis, which are distinct or the same, we can derive (6) from (8).

Comparing the result just obtained with the earlier given proof of the existence of n-dimensional Euclidean spaces for any n, we can make the following assertion: *if an arbitrary basis is chosen in an n-dimensional linear space V_n, then in V_n we can specify scalar multiplication so that in the resulting Euclidean space the chosen basis will be one of the orthonormal bases.*

Isomorphism of Euclidean spaces. Euclidean spaces E and E' are termed *isomorphic* if we can establish between the vectors of these spaces a one-to-one correspondence such that the following requirements are met.

(1) The correspondence is an isomorphic correspondence between E and E', which are regarded as linear spaces (see Sec. 29).

(2) In this correspondence the scalar product is preserved; in other words, if for the images of the vectors a and b in E we have the corresponding vectors a' and b' in E', then

$$(a, b) = (a', b') \tag{9}$$

From Condition (1) it follows immediately that *isomorphic Euclidean spaces have one and the same dimension.* We will prove the converse.

Any Euclidean spaces E and E' having the same dimension n are isomorphic to each other.

In the spaces E and E', choose the orthonormal bases

$$e_1, e_2, \ldots, e_n \tag{10}$$

and, respectively,

$$e'_1, e'_2, \ldots, e'_n \tag{11}$$

If we associate every vector

$$a = \sum_{i=1}^{n} \alpha_i e_i$$

in E with a vector

$$a' = \sum_{i=1}^{n} \alpha_i e'_i$$

in E', having in the basis (11) the same coordinates as the vector a in the basis (10), we will obviously get an isomorphic correspondence between the linear spaces E and E'. We will show that (9) holds as well: if

$$b = \sum_{i=1}^{n} \beta_i e_i, \quad b' = \sum_{i=1}^{n} \beta_i e'_i$$

then, by (8) [use the fact that the bases (10) and (11) are orthonormal!],

$$(a, b) = \sum_{i=1}^{n} \alpha_i \beta_i = (a', b')$$

It is natural not to consider isomorphic Euclidean spaces as distinct, and so for every n there exists a unique n-dimensional Euclidean space in the same sense that for every n there exists a unique n-dimensional real linear space.

The concepts and results of this section may be extended to the case of complex linear spaces in the following manner. A complex linear space is called a *unitary space* if scalar multiplication is given and (a, b) is, in general, a complex number. Axioms II-IV

must hold true (note, in the statement of Axiom IV, that the scalar square of a nonzero vector is real and is strictly positive), and Axiom I is replaced by the axiom

$$\text{I}' \qquad (a,\ b) = \overline{(b,\ a)}$$

where, as usual, the bar denotes the complex conjugate.

Consequently, scalar multiplication will no longer be commutative. Still, an equation that is symmetric to Axiom II holds true,

$$\text{II}' \qquad (a,\ b + c) = (a,\ b) + (a,\ c)$$

since

$$(a,\ b + c) = \overline{(b + c,\ a)} = \overline{(b,\ a) + (c,\ a)}$$
$$= \overline{(b,\ a)} + \overline{(c,\ a)} = (a,\ b) + (a,\ c)$$

On the other hand

$$\text{III}' \qquad (a,\ \alpha b) = \bar{\alpha}\,(a,\ b)$$

since

$$(a,\ \alpha b) = \overline{(\alpha b,\ a)} = \bar{\alpha}\,\overline{(b,\ a)} = \bar{\alpha}\,(\overline{b,\ a}) = \bar{\alpha}\,(a,\ b)$$

The concepts of orthogonality and of an orthonormal system of vectors are carried over to the case of unitary spaces without any alterations. As before, proof is given of the existence of orthonormal bases in any finite-dimensional unitary space. Here, however, if $e_1, e_2, \ldots, e_n$ is an orthonormal basis and the vectors a, b have the notations (7) in this basis, then

$$(a,\ b) = \sum_{i=1}^{n} \alpha_i \bar{\beta}_i$$

The results of the other sections of this chapter can also be extended from Euclidean to unitary spaces, but we will not do this and will refer the interested reader to special books on linear algebra.

35. Orthogonal Matrices, Orthogonal Transformations

Let there be given a real linear transformation of n unknowns:

$$x_i = \sum_{k=1}^{n} q_{ik} y_k, \qquad i = 1,\ 2,\ \ldots,\ n \tag{1}$$

Denote the matrix of the transformation by Q. This transformation carries the sum of the squares of the unknowns $x_1, x_2, \ldots, x_n$, that is the quadratic form $x_1^2 + x_2^2 + \ldots + x_n^2$, which is the normal form of positive definite quadratic forms (see Sec. 28), into a certain quadratic form in the unknowns $y_1, y_2, \ldots, y_n$. Quite accidentally, this new quadratic form may itself turn out to be a sum of the

squares of the unknowns $y_1, y_2, \ldots, y_n$; that is, we can have the equation

$$x_1^2 + x_2^2 + \ldots + x_n^2 = y_1^2 + y_2^2 + \ldots + y_n^2 \qquad (2)$$

which, after replacing the unknowns $x_1, x_2, \ldots, x_n$ by their expressions (1), becomes an identity. The linear transformation of unknowns (1) which has this property (or, as we say, such as leaves the sum of the squares of the unknowns invariant) is called an *orthogonal transformation of the unknowns.* Its matrix Q is an *orthogonal matrix.*

There are many other definitions of an orthogonal transformation and an orthogonal matrix which are equivalent to those given above. We now give some of them that will be needed in the sequel.

In Sec. 26 we gave a rule for the transformation of the matrix of a quadratic form under a linear transformation of the unknowns. Applying it to our case and taking into account that the unit matrix E is the matrix of a quadratic form (being the sum of the squares of all the unknowns), we find that equation (2) is equivalent to the matrix equation

$$Q'EQ = E$$

that is,

$$Q'Q = E \qquad (3)$$

Whence

$$Q' = Q^{-1} \qquad (4)$$

and so the following equation holds true too:

$$QQ' = E \qquad (5)$$

Thus, by (4), *an orthogonal matrix Q may be defined as a matrix for which the transpose Q' is equal to the inverse matrix Q^{-1}.* Each one of the equations (3) and (5) can also be taken as a definition of an orthogonal matrix.

Since the columns of Q' are the rows of Q, it follows from (5) that *the square matrix Q is orthogonal if and only if the sum of the squares of all elements of any one of its rows is equal to unity, and the sum of the products of the corresponding elements of any two distinct rows is zero.* From (3) follows an analogous assertion for the columns of a matrix Q.

Taking determinants in (3), we get (since $|Q'| = |Q|$)

$$|Q|^2 = 1$$

Whence it follows that *the determinant of an orthogonal matrix is equal to* ± 1. Thus *any orthogonal transformation of unknowns is a nonsingular transformation.* We cannot, quite naturally, assert the converse: also note that by far not every matrix with determinant ± 1 is orthogonal.

A matrix that is inverse to an orthogonal matrix will itself be orthogonal. Indeed, taking transposes in (4), we obtain

$$(Q^{-1})' = (Q')' = Q = (Q^{-1})^{-1}$$

On the other hand, *a product of orthogonal matrices is orthogonal.* Indeed, if matrices Q and R are orthogonal, then, using (4), and also (6) of Sec. 26 and an analogous equation which is true for inverses, we get

$$(QR)' = R'Q' = R^{-1}Q^{-1} = (QR)^{-1}$$

In Sec. 37, use will be made of the following assertion.

The change-of-basis matrix from an orthonormal basis of a Euclidean space to any other of its orthonormal bases is orthogonal.

In a space E_n, let there be given two orthonormal bases $e_1, e_2, \ldots, e_n$ and $e'_1, e'_2, \ldots, e'_n$ with the change-of-basis matrix $Q = (q_{ij})$,

$$e' = Qe$$

Since the basis e is orthonormal, the scalar product of any two vectors (of any two vectors from the basis e', for instance), is equal to the sum of the products of the corresponding coordinates of these vectors in the basis e. However, since basis e' is also orthonormal, the scalar square of each vector of e' is equal to unity, and the scalar product of any two distinct vectors of e' is equal to zero. Whence, for the rows of coordinates of the vectors of basis e' in basis e (i.e., for the rows of matrix Q), follow the assertions which, as derived above from (5), are characteristic of an orthogonal matrix.

Orthogonal transformations of Euclidean space. It will be well at this point to make a study of an interesting special type of linear transformations of Euclidean spaces, though such transformations will not be used in the sequel.

A linear transformation φ of a Euclidean space E_n is called an *orthogonal transformation of that Euclidean space* if it preserves the scalar square of every vector, that is, for any vector a,

$$(a\varphi,\ a\varphi) = (a,\ a) \tag{6}$$

From this we derive the following more general assertion, which quite naturally can also be taken as a definition of an orthogonal transformation.

An orthogonal transformation φ of a Euclidean space preserves the scalar product of any two vectors a, b:

$$(a\varphi,\ b\varphi) = (a,\ b) \tag{7}$$

Indeed, by (6),

$$((a + b)\varphi,\ (a + b)\varphi) = (a + b,\ a + b)$$

However,

$$((a + b)\varphi,\ (a + b)\varphi) = (a\varphi + b\varphi,\ a\varphi + b\varphi)$$
$$= (a\varphi,\ a\varphi) + (a\varphi,\ b\varphi) + (b\varphi,\ a\varphi) + (b\varphi,\ b\varphi),$$
$$(a + b,\ a + b) = (a,\ a) + (a,\ b) + (b,\ a) + (b,\ b)$$

Whence, using (6) both for a and for b, and taking into account the commutativity of scalar multiplication, we obtain

$$2\,(a\varphi,\ b\varphi) = 2\,(a,\ b)$$

and so (7) holds true.

In an orthogonal transformation of a Euclidean space, the images of all vectors of any orthonormal basis themselves form an orthonormal basis. Conversely, if a linear transformation of a Euclidean space carries at least one orthonormal basis again into an orthonormal basis, then the transformation is orthogonal.

Indeed, let φ be an orthogonal transformation of the space E_n, and let $e_1, e_2, \ldots, e_n$ be an arbitrary orthonormal basis of this space. Due to (7), there follow from the equations

$$(e_i,\ e_i) = 1, \qquad i = 1, 2, \ldots, n,$$
$$(e_i,\ e_j) = 0 \quad \text{for} \quad i \neq j$$

the equations

$$(e_i\varphi,\ e_i\varphi) = 1, \qquad i = 1, 2, \ldots, n$$
$$(e_i\varphi,\ e_j\varphi) = 0, \qquad i \neq j$$

That is, the system of vectors $e_1\varphi, e_2\varphi, \ldots, e_n\varphi$ proves to be orthogonal and normal; for this reason it is an orthonormal basis of the space E_n.

Conversely, let a linear transformation φ of the space E_n carry the orthonormal basis $e_1, e_2, \ldots, e_n$ again into an orthonormal basis; that is, the system of vectors $e_1\varphi, e_2\varphi, \ldots, e_n\varphi$ is an orthonormal basis of the space E_n. If

$$a = \sum_{i=1}^{n} \alpha_i e_i$$

is an arbitrary vector of the space E_n, then

$$a\varphi = \sum_{i=1}^{n} \alpha_i (e_i\varphi)$$

The vector $a\varphi$ has the same coordinates in the basis $e\varphi$ as the vector a has in the basis e. However, both these bases are orthonormal, and for this reason the scalar square of any vector is equal to the sum

of the squares of its coordinates in any one of these bases. Thus

$$(a, a) = (a\varphi, a\varphi) = \sum_{i=1}^{n} \alpha_i^2$$

Equation (6) indeed holds true.

An orthogonal transformation of a Euclidean space in any orthonormal basis is represented by an orthogonal matrix. Conversely, if a linear transformation of a Euclidean space in at least one orthonormal basis is represented by an orthogonal matrix, then the transformation is orthogonal.

Indeed, if the transformation φ is orthogonal, and the basis e_1, e_2, . . ., e_n is orthonormal, then the system of vectors $e_1\varphi$, $e_2\varphi$, ..., $e_n\varphi$ will also be an orthonormal basis. The matrix A of the transformation φ in the basis e,

$$e\varphi = Ae \tag{8}$$

will thus be the transition matrix from the orthonormal basis e to the orthonormal basis $e\varphi$, i.e. (as proved above), it will be orthogonal.

Conversely, let a linear transformation φ be represented in an orthonormal basis e_1, e_2, . . ., e_n by the orthogonal matrix A; then (8) holds. Since the basis e is orthonormal, the scalar product of any vectors (in particular, any vectors of the system $e_1\varphi$, $e_2\varphi$, . . ., $e_n\varphi$) is equal to the sum of the products of the corresponding coordinates of these vectors in the basis e. Therefore, since matrix A is orthogonal,

$$(e_i\varphi, e_i\varphi) = 1, \qquad i = 1, 2, \ldots, n,$$
$$(e_i\varphi, e_j\varphi) = 0 \qquad \text{for } i \neq j$$

That is to say, the system $e\varphi$ is itself an orthonormal basis for the space E_n. Whence follows the orthogonality of the transformation φ.

As the reader will recall from analytic geometry, of all the affine transformations of a plane that leave the coordinate origin fixed, rotations (combined perhaps with mirror reflections) are the only ones that preserve the scalar product of the vectors. Thus, orthogonal transformations of n-dimensional Euclidean space may be regarded as "rotations" of this space.

Obviously, one of the orthogonal transformations of Euclidean space is the identity transformation. On the other hand, the relationship we have established between orthogonal transformations and orthogonal matrices, and also the relationship (presented in Sec. 31) between operations on linear transformations and on matrices, permit deriving, from familiar properties of orthogonal matrices, the following properties of orthogonal transformations of Euclidean space, which can be verified directly.

Every orthogonal transformation is nonsingular and its inverse is also orthogonal.

The product of any orthogonal transformations is orthogonal.

36. Symmetric Transformations

A linear transformation φ of n-dimensional Euclidean space is called *symmetric* (or *self-adjoint*) if for any vectors a, b of this space we have the equality

$$(a\varphi,\ b) = (a,\ b\varphi) \tag{1}$$

That is, in scalar multiplication the symbol of symmetric transformation may be carried from one factor to the other.

Obvious instances of symmetric transformations are the identity transformation ε and the zero transformation ω. A more general example is the linear transformation in which each vector is multiplied by a fixed scalar α,

$$a\varphi = \alpha a$$

Indeed, in this case

$$(a\varphi,\ b) = (\alpha a,\ b) = \alpha\,(a,\ b) = (a,\ \alpha b) = (a,\ b\varphi)$$

The role of symmetric transformations is extremely great and calls for a detailed study.

A symmetric transformation of a Euclidean space in any orthonormal basis is represented by a symmetric matrix. Conversely, if a linear transformation of a Euclidean space is represented in at least one orthonormal basis by a symmetric matrix, then the transformation is symmetric.

Indeed, let the symmetric transformation φ be represented in an orthonormal basis $e_1, e_2, \ldots, e_n$ by the matrix $A = (\alpha_{ij})$. Taking into account that in an orthonormal basis the scalar product of two vectors is equal to the sum of the products of the corresponding coordinates of these vectors, we obtain

$$(e_i\varphi,\ e_j) = \Big(\sum_{k=1}^{n} \alpha_{ik}e_k,\ e_j\Big) = \alpha_{ij}$$

$$(e_i,\ e_j\varphi) = \Big(e_i,\ \sum_{k=1}^{n} \alpha_{jk}e_k\Big) = \alpha_{ji}$$

That is, due to (1),

$$\alpha_{ij} = \alpha_{ji}$$

for all i and j. The matrix A is thus symmetric.

Conversely, let a linear transformation φ be represented in the orthonormal basis $e_1, e_2, \ldots, e_n$ by the symmetric matrix $A = (\alpha_{ij})$,

$$\alpha_{ij} = \alpha_{ji} \quad \text{for all } i \text{ and } j \tag{2}$$

If

$$b = \sum_{i=1}^{n} \beta_i e_i, \qquad c = \sum_{j=1}^{n} \gamma_j e_j$$

are any vectors of the space, then

$$b\varphi = \sum_{i=1}^{n} \beta_i (e_i\varphi) = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} \beta_i \alpha_{ij}\right) e_j$$

$$c\varphi = \sum_{j=1}^{n} \gamma_j (e_j\varphi) = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \gamma_j \alpha_{ji}\right) e_i$$

Using the fact that the e-basis is orthonormal, we get

$$(b\varphi, c) = \sum_{j,\, i=1}^{n} \beta_i \alpha_{ij} \gamma_j,$$

$$(b, c\varphi) = \sum_{i,\, j=1}^{n} \beta_i \gamma_j \alpha_{ji}$$

By (2), the right sides of the latter equalities coincide, and therefore

$$(b\varphi, c) = (b, c\varphi)$$

which completes the proof.

The result obtained yields the following property of symmetric transformations that can readily be verified directly.

The sum of symmetric transformations and also the product of a symmetric transformation by a scalar are again symmetric transformations.

We now prove the following important theorem.

All characteristic roots of a symmetric transformation are real.

Since the characteristic roots of any linear transformation coincide with the characteristic roots of the matrix of this transformation in any basis, and a symmetric transformation is represented in orthonormal bases by symmetric matrices, it suffices to prove the following assertion.

All the characteristic roots of a symmetric matrix are real.

Let λ_0 be a characteristic root (possibly complex) of the symmetric matrix $A = (\alpha_{ij})$,

$$| A - \lambda_0 E | = 0$$

Then the system of homogeneous linear equations with complex coefficients

$$\sum_{j=1}^{n} \alpha_{ij} x_j = \lambda_0 x_i, \qquad i = 1, 2, \ldots, n$$

has a zero determinant, which is to say, it has a nontrivial solution $\beta_1, \beta_2, \ldots, \beta_n$ (generally complex). Thus,

$$\sum_{j=1}^{n} \alpha_{ij} \beta_j = \lambda_0 \beta_i, \qquad i = 1, 2, \ldots, n \tag{3}$$

Multiplying both sides of each ith equation of (3) by a scalar $\bar{\beta}_i$, the conjugate of β_i, and adding separately the left and right members of all the resulting equations, we get the equation

$$\sum_{i,\,j=1}^{n} \alpha_{ij}\beta_j\bar{\beta}_i = \lambda_0 \sum_{i=1}^{n} \beta_i\bar{\beta}_i \tag{4}$$

The coefficient of λ_0 in (4) is a nonzero real number since it is the sum of nonnegative real numbers, of which at least one is strictly positive. The real nature of the number λ_0 will therefore be proved if we prove the real nature of the left-hand side of (4); to do this, it suffices to show that this complex number coincides with its conjugate. Here, for the first time, we make use of the symmetric nature of the (real) matrix A.

$$\overline{\sum_{i,\,j=1}^{n} \alpha_{ij}\beta_j\bar{\beta}_i} = \sum_{i,\,j=1}^{n} \overline{\alpha_{ij}\beta_j\bar{\beta}_i} = \sum_{i,\,j=1}^{n} \alpha_{ij}\bar{\beta}_j\beta_i$$

$$= \sum_{i,\,j=1}^{n} \alpha_{ji}\bar{\beta}_j\beta_i = \sum_{i,\,j=1}^{n} \alpha_{ij}\bar{\beta}_i\beta_j = \sum_{i,\,j=1}^{n} \alpha_{ij}\beta_j\bar{\beta}_i$$

Note that the second last equality is obtained by a simple interchange in the summation indices: j is put in place of i, i in place of j. Hence, the theorem is proved.

A linear transformation φ of the Euclidean space E_n is symmetric if and only if there exists in E_n an orthonormal basis composed of the eigenvectors of the transformation.

In one direction, this assertion is almost obvious: if there exists in E_n an orthonormal basis $e_1, e_2, \ldots, e_n$, and

$$e_i\varphi = \lambda_i e_i, \qquad i = 1, 2, \ldots, n$$

then in the e-basis the transformation φ is represented by the diagonal matrix

$$\begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ 0 & & \lambda_n \end{pmatrix}$$

A diagonal matrix, however, is symmetric, and so the transformation φ is represented in the orthonormal basis e by a symmetric matrix, hence it is symmetric.

The basic inverse assertion of the theorem we prove by induction with respect to the dimension n of the space E_n. Indeed, for $n = 1$, any linear transformation φ of E_1 invariably carries any vector into a proportional vector, whence it follows that any nonzero vector a

is an eigenvector for φ (incidentally, it also follows that any linear transformation of the space E_1 is symmetric). Normalizing the vector a, we obtain the desired orthonormal basis of the space E_1.

Let the assertion of the theorem be proved for an $(n-1)$-dimensional Euclidean space and let a symmetric transformation φ be given in the space E_n. From the above-proved theorem follows the existence, under φ, of a real characteristic root λ_0. Consequently, this number is an eigenvalue of the transformation φ. If a is an eigenvector of the transformation φ corresponding to this eigenvalue, then any nonzero vector proportional to the vector a will (under φ) be an eigenvector corresponding to the same eigenvalue λ_0, since

$$(\alpha a)\,\varphi = \alpha\,(a\varphi) = \alpha\,(\lambda_0 a) = \lambda_0\,(\alpha a)$$

In particular, normalizing the vector a, we obtain a vector e_1 such that

$$e_1\varphi = \lambda_0 e_1,$$
$$(e_1, e_1) = 1$$

As was proved in Sec. 34, the nonzero vector e_1 may be included in the orthogonal basis

$$e_1,\ e_2',\ \ldots,\ e_n' \tag{5}$$

of the space E_n. Those vectors whose first coordinate in the basis (5) is zero, that is, vectors of the form $\alpha_2 e_2' + \ldots + \alpha_n e_n'$ obviously constitute an $(n-1)$-dimensional linear subspace of the space E_n, which we will designate by L. It will even be an $(n-1)$-dimensional Euclidean space, since a scalar product, being defined for all vectors in E_n, is in particular defined for vectors in L and possesses all the requisite properties.

The subspace L consists of all the vectors of E_n which are orthogonal to the vector e_1. Indeed, if

$$a = \alpha_1 e_1 + \alpha_2' e_2' + \ldots + \alpha_n' e_n'$$

then, by the orthogonality of the basis (5) and the normalized character of the vector e_1,

$$(e_1, a) = \alpha_1\,(e_1, e_1) + \alpha_2'\,(e_1, e_2') + \ldots + \alpha_n'\,(e_1, e_n') = \alpha_1$$

that is to say, $(e_1, a) = 0$ if and only if $\alpha_1 = 0$.

If the vector a belongs to the subspace L, i.e., $(e_1, a) = 0$, then the vector $a\varphi$ too lies in L. Indeed, because of the symmetry of the transformation φ,

$$(e_1, a\varphi) = (e_1\varphi, a) = (\lambda_0 e_1, a) = \lambda_0\,(e_1, a) = \lambda_0 \cdot 0 = 0$$

That is, the vector $a\varphi$ is orthogonal to e_1 and therefore lies in L. This property of the subspace L, which is called its *invariance* under the transformation φ, enables us to consider φ (regarded solely with

respect to the vectors in L) as a linear transformation of this $(n-1)$-dimensional Euclidean space. It will even be a symmetric transformation of the space L, since equation (1), which holds for any vectors in E_n, will hold (as a particular case) for vectors lying in L.

By virtue of the induction hypothesis, space L has an orthonormal basis consisting of the eigenvectors of the transformation φ; denote it by $e_2, \ldots, e_n$. All these vectors are orthogonal to the vector e_1, and so $e_1, e_2, \ldots, e_n$ is the desired orthonormal basis of the space E_n consisting of the eigenvectors of the transformation Φ. The theorem is proved.

37. Reducing a Quadratic Form to Principal Axes. Pairs of Forms

Let us apply the last theorem of the preceding section to prove the following matrix theorem.

For every symmetric matrix A it is possible to find an orthogonal matrix Q which diagonalizes matrix A, that is, the matrix $Q^{-1}AQ$ obtained by transforming matrix A by matrix Q will be diagonal.

Let there be given a symmetric matrix A of order n. If $e_1, e_2, \ldots, e_n$ is some orthonormal basis of an n-dimensional Euclidean space E_n, then matrix A represents in this basis a symmetric transformation φ. As has been proved, there is in E_n an orthonormal basis $f_1, f_2, \ldots, f_n$ made up of the eigenvectors of the transformation φ. In this basis, φ is represented by the diagonal matrix B (see Sec. 33). Then, by Sec. 31,

$$B = Q^{-1}AQ \tag{1}$$

where Q is the change-of-basis matrix from the f-basis to the e-basis,

$$e = Qf \tag{2}$$

This matrix, as a matrix for changing from one orthonormal basis to another similar basis, will be orthogonal (see Sec. 35). The theorem is proved.

Since the inverse of orthogonal matrix Q is equal to its transpose, $Q^{-1} = Q'$, equation (1) may be rewritten as

$$B = Q'AQ$$

From Sec. 26, however, we know that such precisely is the transformation of the symmetric matrix A of a quadratic form subject to a linear transformation of the unknowns with the matrix Q. However, taking into account that a linear transformation of unknowns with an orthogonal matrix is an orthogonal transformation (see Sec. 35) and that a quadratic form reduced to canonical form has a diagonal matrix, we arrive, on the basis of the preceding theorem, at the following **theorem on the reduction of a real quadratic form to principal axes.**

Every real quadratic form $f(x_1, x_2, \ldots, x_n)$ *can be reduced to canonical form by an orthogonal transformation of the unknowns.*

Although there may be many different orthogonal transformations of the unknowns which reduce the given quadratic form to canonical form, the canonical form itself is actually determined uniquely.

No matter what the orthogonal transformation that reduces to canonical form the quadratic form $f(x_1, x_2 \ldots, x_n)$ *with matrix* A, *the coefficients of this canonical form are the characteristic roots of the matrix* A (*counting multiplicities*).

Suppose an orthogonal transformation reduces form f to the canonical form

$$f(x_1, x_2, \ldots, x_n) = \mu_1 y_1^2 + \mu_2 y_2^2 + \ldots + \mu_n y_n^2$$

This orthogonal transformation preserves the sum of the square of the unknowns and so, if λ is a new unknown,

$$f(x_1, x_2, \ldots, x_n) - \lambda \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} \mu_i y_i^2 - \lambda \sum_{i=1}^{n} y_i^2$$

Taking determinants of these quadratic forms and taking into account that after completing the linear transformation the determinant of the quadratic form is multiplied by the square of the determinant of the transformation (see Sec. 28), and the square of the determinant of an orthogonal transformation is equal to unity (see Sec. 35), we get the equation

$$|A - \lambda E| = \begin{vmatrix} \mu_1 - \lambda & 0 & \ldots & 0 \\ 0 & \mu_2 - \lambda & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \mu_n - \lambda \end{vmatrix} = \prod_{i=1}^{n} (\mu_i - \lambda)$$

from which follows the assertion of the theorem.

This result may be stated in matrix form as well.

No matter what the orthogonal matrix which diagonalizes the symmetric matrix A, *the principal diagonal of the resulting diagonal matrix will exhibit the characteristic roots of the matrix* A *taken with their multiplicities.*

Finding the orthogonal transformation that reduces a quadratic form to principal axes. In certain problems it is not only necessary to know the canonical form to which a real quadratic form is reduced by an orthogonal transformation, but also the orthogonal transformation itself which accomplishes the reduction. It would be rather difficult to find this transformation by using the principal-axis theorem so we shall point out a different way. Namely, all we need to know is how to find the orthogonal matrix Q which diagonalizes the given symmetric matrix A, or, what is the same thing, to

find its inverse matrix Q^{-1}. By (2), this is the change-of-basis matrix from the e-basis to the f-basis; that is, its rows are coordinate rows (in the e-basis) of an orthonormal system of n eigenvectors of the symmetric transformation φ defined by the matrix A in the e-basis. It remains to find such a system of eigenvectors.

Let λ_0 be any characteristic root of the matrix A and let its multiplicity be equal to k_0. From Sec. 33 we know that the collection of coordinate rows of all eigenvectors of the transformation φ corresponding to the eigenvalue λ_0 coincides with the set of nonzero solutions of the system of homogeneous linear equations

$$(A - \lambda_0 E)\, X = 0 \tag{3}$$

Here, the symmetric nature of the matrix A enables us to write A in place of A'. From the above-proved theorems on the existence of an orthogonal matrix that diagonalizes the symmetric matrix A, and on the uniqueness of this diagonal form, it follows that for system (3) it is at least possible to find k_0 linearly independent solutions. We seek such a system of solutions by the methods taken from Sec. 12, and then we orthogonalize and normalize the resulting system in accord with Sec. 34.

Taking in turn, for λ_0, all the different characteristic roots of the symmetric matrix A and noting that the sum of the multiplicities of these roots is equal to n, we obtain a set of n eigenvectors of the transformation φ represented by their coordinates in the e-basis. To prove that this is the desired orthonormal system of eigenvectors, it remains to prove the following lemma.

The eigenvectors of the symmetric transformation φ which correspond to distinct eigenvalues are mutually orthogonal.

Suppose that

$$b\varphi = \lambda_1 b, \qquad c\varphi = \lambda_2 c$$

and $\lambda_1 \neq \lambda_2$. Since

$$(b\varphi, c) = (\lambda_1 b, c) = \lambda_1 (b, c),$$
$$(b, c\varphi) = (b, \lambda_2 c) = \lambda_2 (b, c)$$

it follows from

$$(b\varphi, c) = (b, c\varphi)$$

that

$$\lambda_1 (b, c) = \lambda_2 (b, c)$$

or, because $\lambda_1 \neq \lambda_2$,

$$(b, c) = 0$$

which is what we set out to prove.

Example. Reduce to principal axes the quadratic form
$f(x_1, x_2, x_3, x_4) = 2x_1x_2 + 2x_1x_3 - 2x_1x_4 - 2x_2x_3 + 2x_2x_4 + 2x_3x_4$
The matrix A of this form looks like

$$A = \begin{pmatrix} 0 & 1 & 1 & -1 \\ 1 & 0 & -1 & 1 \\ 1 & -1 & 0 & 1 \\ -1 & 1 & 1 & 0 \end{pmatrix}$$

Let us find its characteristic polynomial:

$$|A - \lambda E| = \begin{vmatrix} -\lambda & 1 & 1 & -1 \\ 1 & -\lambda & -1 & 1 \\ 1 & -1 & -\lambda & 1 \\ -1 & 1 & 1 & -\lambda \end{vmatrix} = (\lambda - 1)^3 (\lambda + 3)$$

Thus, the matrix A has a triple characteristic root 1 and a simple characteristic root -3. Hence, we can already write the canonical form to which the form f is reduced by an orthogonal transformation:

$$f = y_1^2 + y_2^2 + y_3^2 - 3y_4^2$$

Let us find the orthogonal transformation that accomplishes this reduction. The system of homogeneous linear equations (3) becomes, for $\lambda_0 = 1$,

$$\begin{cases} -x_1 + x_2 + x_3 - x_4 = 0, \\ x_1 - x_2 - x_3 + x_4 = 0, \\ x_1 - x_2 - x_3 + x_4 = 0, \\ -x_1 + x_2 + x_3 - x_4 = 0 \end{cases}$$

The rank of this system is unity and so we can find three linearly independent solutions for it. For example, the vectors

$$b_1 = (1, 1, 0, 0),$$
$$b_2 = (1, 0, 1, 0),$$
$$b_3 = (-1, 0, 0, 1)$$

will be such solutions.

Orthogonalizing this system of vectors, we obtain the following system of vectors:

$$c_1 = b_1 = (1, 1, 0, 0),$$
$$c_2 = -\frac{1}{2} c_1 + b_2 = \left(\frac{1}{2}, -\frac{1}{2}, 1, 0\right),$$
$$c_3 = \frac{1}{2} c_1 + \frac{1}{3} c_2 + b_3 = \left(-\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 1\right)$$

On the other hand, the system of homogeneous linear equations (3) becomes, for $\lambda_0 = -3$,

$$\begin{cases} 3x_1 + x_2 + x_3 - x_4 = 0, \\ x_1 + 3x_2 - x_3 + x_4 = 0, \\ x_1 - x_2 + 3x_3 + x_4 = 0, \\ -x_1 + x_2 + x_3 + 3x_4 = 0 \end{cases}$$

This system has rank 3. Its nontrivial solution is the vector

$$c_4 = (1, -1, -1, 1)$$

The system of vectors c_1, c_2, c_3, c_4 is orthogonal. Normalizing it, we arrive at the orthonormal system of vectors

$$c_1' = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0\right),$$

$$c_2' = \left(\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, \sqrt{\frac{2}{3}}, 0\right)$$

$$c_3' = \left(-\frac{1}{2\sqrt{3}}, \frac{1}{2\sqrt{3}}, \frac{1}{2\sqrt{3}}, \frac{\sqrt{3}}{2}\right),$$

$$c_4' = \left(\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}\right)$$

Thus, the form f is reduced to principal axes by the orthogonal transformation

$$y_1 = \frac{1}{\sqrt{2}} x_1 + \frac{1}{\sqrt{2}} x_2,$$

$$y_2 = \frac{1}{\sqrt{6}} x_1 - \frac{1}{\sqrt{6}} x_2 + \sqrt{\frac{2}{3}} x_3,$$

$$y_3 = -\frac{1}{2\sqrt{3}} x_1 + \frac{1}{2\sqrt{3}} x_2 + \frac{1}{2\sqrt{3}} x_3 + \frac{\sqrt{3}}{2} x_4,$$

$$y_4 = \frac{1}{2} x_1 - \frac{1}{2} x_2 - \frac{1}{2} x_3 + \frac{1}{2} x_4$$

It is well to note that the choice of a system of linearly independent eigenvectors corresponding to a multiple eigenvalue is extremely ambiguous, and so there are many different orthogonal transformations which reduce the form f to canonical form. We found only one of them.

Pairs of forms. Let there be a pair of real quadratic forms in n unknowns, $f(x_1, x_2, \ldots, x_n)$ and $g(x_1, x_2, \ldots, x_n)$. Does there exist a nonsingular linear transformation of the unknowns x_1, x_2, . . ., x_n such that will simultaneously reduce both forms to canonical form?

In the general case, the answer is no. Let us examine the pair of forms

$$f(x_1, x_2) = x_1^2, \qquad g(x_1, x_2) = x_1 x_2$$

Let there be a nonsingular linear transformation

$$\left.\begin{aligned} x_1 &= c_{11} y_1 + c_{12} y_2, \\ x_2 &= c_{21} y_1 + c_{22} y_2 \end{aligned}\right\} \tag{4}$$

which reduces both forms to canonical form. For f to be reduced by transformation (4) to canonical form, one of the coefficients c_{11}, c_{12} must be zero, otherwise the term $2c_{11}c_{12}y_1y_2$ would occur. Renumbering, if necessary, the unknowns y_1, y_2, we can set $c_{12} = 0$ and

so $c_{11} \neq 0$. However, we now find that

$$g(x_1, x_2) = c_{11}y_1(c_{21}y_1 + c_{22}y_2) = c_{11}c_{21}y_1^2 + c_{11}c_{22}y_1y_2$$

Since the form g was also to become canonical, it follows that $c_{11}c_{22} = 0$, that is, $c_{22} = 0$, which, together with $c_{12} = 0$, contradicts the nonsingularity of the linear transformation (4).

The situation is different if we assume that at least one of our forms, say $g(x_1, x_2, \ldots, x_n)$ is positive definite.* Namely, the following theorem holds.

If f and g form a pair of real quadratic forms in n unknowns, and the second one is positive definite, then there exists a nonsingular linear transformation which simultaneously reduces g to normal form and f to canonical form.

For proof, first perform the nonsingular linear transformation of the unknowns $x_1, x_2, \ldots, x_n$,

$$X = TY$$

which reduces the positive definite form g to normal form,

$$g(x_1, x_2, \ldots, x_n) = y_1^2 + y_2^2 + \ldots + y_n^2$$

Then f will go into some form φ in new unknowns,

$$f(x_1, x_2, \ldots, x_n) = \varphi(y_1, y_2, \ldots, y_n)$$

Now perform an orthogonal transformation of the unknowns $y_1, y_2, \ldots, y_n$,

$$Y = QZ$$

which reduces φ to principal axes,

$$\varphi(y_1, y_2, \ldots, y_n) = \lambda_1 z_1^2 + \lambda_2 z_2^2 + \ldots + \lambda_n z_n^2$$

This transformation (see definition in Sec. 35) carries the sum of the squares of the unknowns $y_1, y_2, \ldots, y_n$ into the sum of the squares of the unknowns $z_1, z_2, \ldots, z_n$. As a result we get

$$f(x_1, x_2, \ldots, x_n) = \lambda_1 z_1^2 + \lambda_2 z_2^2 + \ldots + \lambda_n z_n^2,$$
$$g(x_1, x_2, \ldots, x_n) = z_1^2 + z_2^2 + \ldots + z_n^2$$

That is, the linear transformation

$$X = (TQ)\,Z$$

is the required one.

---

* This condition is not of course necessary; thus, both the forms $x_1^2 + x_2^2 - x_3^2$ and $x_1^2 - x_2^2 - x_3^2$ now have canonical form, though none is positive definite.

CHAPTER 9

EVALUATING ROOTS OF POLYNOMIALS

38. Equations of Second, Third, and Fourth Degree

The fundamental theorem proved in Sec. 23 establishes the existence of n complex roots for any polynomial of degree n with numerical coefficients. The proofs (both ours and any other existing proofs) do not however indicate any methods for finding these roots. They are thus pure "existence proofs". The search for such methods began naturally in attempts to derive formulas similar to the one used in the solution of quadratic equations for the case of real coefficients so familiar from school algebra. We will now show that this formula holds true for quadratic equations with complex coefficients as well, and that analogous formulas (though much more involved) can be derived for equations of the third and fourth degree.

Quadratic equations. Suppose we have a quadratic equation

$$x^2 + px + q = 0$$

with arbitrary complex coefficients, the leading coefficient may, without loss of generality, be considered equal to unity. This equation may be written as

$$\left(x + \frac{p}{2}\right)^2 + \left(q - \frac{p^2}{4}\right) = 0$$

As we know, it is possible to take the square root of the complex number $\frac{p^2}{4} - q$ without going outside the complex-number system. The two values of this root which differ in sign alone can be written as $\pm \sqrt{\frac{p^2}{4} - q}$. Therefore,

$$x + \frac{p}{2} = \pm \sqrt{\frac{p^2}{4} - q}$$

That is, the roots of the given equation may be found via the usual formula

$$x = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}$$

Example. Solve

$$x^2 - 3x + (3 - i) = 0$$

Using the formula derived above, we get

$$x = \frac{3}{2} \pm \sqrt{\frac{9}{4} - (3 - i)} = \frac{3}{2} \pm \frac{1}{2}\sqrt{-3 + 4i}$$

Applying the methods of Sec. 19, we find

$$\sqrt{-3 + 4i} = \pm(1 + 2i)$$

and therefore

$$x_1 = 2 + i, \qquad x_2 = 1 - i$$

Cubic equations. Unlike the case of quadratic equations, we have not had a procedure for solving cubic equations even in the case of real coefficients. We will now derive a formula for cubic equations similar to the formula used for quadratic equations, and we will assume from the start that the coefficients can be any complex numbers.

Suppose we have the cubic equation

$$y^3 + ay^2 + by + c = 0 \tag{1}$$

with arbitrary complex coefficients. Replacing in (1) the unknown y by a new unknown x related to y by the equation

$$y = x - \frac{a}{3} \tag{2}$$

we get an equation in the unknown x, which, as can readily be verified, does not contain the square of the unknown; that is, we have an equation of the form

$$x^3 + px + q = 0 \tag{3}$$

If the roots of (3) are found, then, by (2), we will get the roots of the given equation (1) as well. Our job, therefore, is to learn to solve the "incomplete" cubic equation (3) with arbitrary complex coefficients.

By the fundamental theorem, equation (3) has three complex roots. Let x_0 be one of them. We introduce an auxiliary unknown u and consider the polynomial

$$f(u) = u^2 - x_0 u - \frac{p}{3}$$

Its coefficients are complex numbers and therefore it has two complex roots α and β; by Vieta's formulas,

$$\alpha + \beta = x_0 \tag{4}$$

$$\alpha\beta = -\frac{p}{3} \tag{5}$$

Substituting expression (4) of the root x_0 into (3), we get

$$(\alpha + \beta)^3 + p(\alpha + \beta) + q = 0$$

or

$$\alpha^3 + \beta^3 + (3\alpha\beta + p)(\alpha + \beta) + q = 0$$

However, from (5) it follows that $3\alpha\beta + p = 0$, and so we have

$$\alpha^3 + \beta^3 = -q \tag{6}$$

On the other hand, from (5) it follows that

$$\alpha^3\beta^3 = -\frac{p^3}{27} \tag{7}$$

Equations (6) and (7) show that the numbers α^3 and β^3 are roots of the quadratic equation

$$z^2 + qz - \frac{p^3}{27} = 0 \tag{8}$$

with complex coefficients.

Solving (8), we get

$$z = -\frac{q}{2} \pm \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}$$

whence*

$$\alpha = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}, \quad \beta = \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} \tag{9}$$

We arrive at the following formula (*Cardan's formula*) which expresses the roots of equation (3) in terms of its coefficients by means of radicals of index 2 and index 3:

$$x_0 = \alpha + \beta = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}$$

Since a cube root has three values in the field of complex numbers, formulas (9) yield three values for α and three for β. However, when using Cardan's formula, one cannot combine just any value of the root α with any value of the root β; for a given value of α we have to take only that one of the three values of β which satisfies condition (5).

Let α_1 be any one of the three values of the root α. Then the two others may be obtained, as was proved in Sec. 19, by multiplying α_1 by the cube roots ε and ε^2 of unity:

$$\alpha_2 = \alpha_1\varepsilon, \qquad \alpha_3 = \alpha_1\varepsilon^2$$

Denote by β_1 that one of the three values of the root β which corresponds to the value α_1 of the root α on the basis of (5), that is, $\alpha_1\beta_1 =$

* It is immaterial which of the roots of (8) we take for α^3 and which one for β^3 since α and β enter in symmetrical fashion into (6) and (7) and also into the expression (4) for x_0.

$= -\frac{p}{3}$. The two other values of β are

$$\beta_2 = \beta_1\varepsilon, \qquad \beta_3 = \beta_1\varepsilon^2$$

Since, by $\varepsilon^3 = 1$,

$$\alpha_2\beta_3 = \alpha_1\varepsilon\cdot\beta_1\varepsilon^2 = \alpha_1\beta_1\varepsilon^3 = \alpha_1\beta_1 = -\frac{p}{3}$$

it follows that the value α_2 of root α is associated with the value β_3 of root β; similarly, to the value α_3 there corresponds the value β_2. Thus, all three roots of equation (3) can be written as follows:

$$\left.\begin{aligned} x_1 &= \alpha_1 + \beta_1, \\ x_2 &= \alpha_2 + \beta_3 = \alpha_1\varepsilon + \beta_1\varepsilon^2, \\ x_3 &= \alpha_3 + \beta_2 = \alpha_1\varepsilon^2 + \beta_1\varepsilon \end{aligned}\right\} \tag{10}$$

Cubic equations with real coefficients. Let us see what can be said about the roots of the reduced cubic equation

$$x^3 + px + q = 0 \tag{11}$$

if its coefficients are real. It turns out that in this case the main role is played by the sign of the expression $\frac{q^2}{4} + \frac{p^3}{27}$, which in Cardan's formula is under the square-root sign. Notice that the sign of this expression is the opposite of the sign of the expression

$$D = -4p^3 - 27q^2 = -108\left(\frac{q^2}{4} + \frac{p^3}{27}\right)$$

which is called the *discriminant* of equation (11) (see Sec. 54, below). The sign of the discriminant will be used in subsequent statements.

(1) Let $D < 0$. In this case, there is a positive number under each of the square-root signs in Cardan's formula, and so each of the cube roots involves real numbers. However, a cube root of a real number has one real and two conjugate complex values. Let α_1 be the real value of the root α; then the value β_1 of the root β, corresponding to α_1 on the basis of formula (5), will also be real because the number p is real. Thus, the root $x_1 = \alpha_1 + \beta_1$ of equation (11) is real. We find the other two roots by replacing, in formulas (10) of this section, the roots of unity $\varepsilon = \varepsilon_1$ and $\varepsilon^2 = \varepsilon_2$ by their expressions (7), Sec. 19:

$$\begin{aligned} x_2 = \alpha_1\varepsilon + \beta_1\varepsilon^2 &= \alpha_1\left(-\frac{1}{2} + i\frac{\sqrt{3}}{2}\right) + \beta_1\left(-\frac{1}{2} - i\frac{\sqrt{3}}{2}\right) \\ &= -\frac{\alpha_1+\beta_1}{2} + i\sqrt{3}\,\frac{\alpha_1-\beta_1}{2}, \\ x_3 = \alpha_1\varepsilon^2 + \beta_1\varepsilon &= \alpha_1\left(-\frac{1}{2} - i\frac{\sqrt{3}}{2}\right) + \beta_1\left(-\frac{1}{2} + i\frac{\sqrt{3}}{2}\right) \\ &= -\frac{\alpha_1+\beta_1}{2} - i\sqrt{3}\,\frac{\alpha_1-\beta_1}{2} \end{aligned}$$

Since the numbers α_1 and β_1 are real, these two roots turn out to be conjugate complex numbers, the coefficient of the imaginary part being different from zero; since $\alpha_1 \neq \beta_1$, these numbers are the values of distinct cube roots.

Thus, *if* $D < 0$, *then equation* (11) *has one real and two conjugate complex roots.*

(2) Let $D=0$. Then

$$\alpha = \sqrt[3]{-\frac{q}{2}}, \quad \beta = \sqrt[3]{-\frac{q}{2}}$$

Let α_1 be the real value of the root α; then β_1 will also, by (5), be a real number, and $\alpha_1 = \beta_1$. Replacing, in formulas (10), β_1 by α_1 and using the obvious equality $\varepsilon + \varepsilon^2 = -1$, we get

$$x_1 = 2\alpha_1, \quad x_2 = \alpha_1(\varepsilon + \varepsilon^2) = -\alpha_1, \quad x_3 = \alpha_1(\varepsilon^2 + \varepsilon) = -\alpha_1$$

Thus, *if* $D = 0$, *then all roots of* (11) *are real and two of them are equal.*

(3) Finally, let $D > 0$. Then in Cardan's formula there is a negative real number under the square root sign. Therefore, under the signs of the cube roots we have conjugate complex numbers. Thus, all the values of the roots α and β will now be complex numbers. However, there must be at least one real root among the roots of equation (11). Let this root be

$$x_1 = \alpha_0 + \beta_0$$

Since both the sum of the numbers α_0 and β_0 and their product, equal to $-\frac{p}{3}$, are real, it follows that the numbers α_0 and β_0 are conjugate as roots of a quadratic equation with real coefficients. But then the numbers $\alpha_0\varepsilon$ and $\beta_0\varepsilon^2$ and likewise the numbers $\alpha_0\varepsilon^2$ and $\beta_0\varepsilon$ are also conjugate, whence it follows that the roots of equation (11)

$$x_2 = \alpha_0\varepsilon + \beta_0\varepsilon^2, \qquad x_3 = \alpha_0\varepsilon^2 + \beta_0\varepsilon$$

are real numbers too.

We thus see that the three roots of (11) are real, and it is easy to show that they are all distinct, for otherwise the choice of a root x_1 might be accomplished so that we would get the equality $x_2 = x_3$, whence

$$\alpha_0(\varepsilon - \varepsilon^2) = \beta_0(\varepsilon - \varepsilon^2)$$

or $\alpha_0 = \beta_0$, which is clearly impossible.

Thus, *if* $D > 0$, *then equation* (11) *has three distinct real roots.*

The last case that we have just considered shows that Cardan's formula is of slight practical value. Indeed, although for $D > 0$ all roots of (11) with real coefficients are real numbers, to find them using Cardan's formula requires extracting the cube roots of complex numbers, which is only possible if the numbers are represented

in trigonometric form. That is why there is no practical value in writing the roots as radicals. Using methods that go beyond the scope of this book, we could demonstrate that in the case at hand the roots of equation (11) cannot, in general, be expressed in terms of coefficients by means of radicals with real radicands. This case of the solution of (11) is called the *irreducible* case (not to be confused with the irreducibility of polynomials).

Example 1. Solve the equation

$$y^3 + 3y^2 - 3y - 14 = 0$$

The substitution $y = x - 1$ reduces this equation to

$$x^3 - 6x - 9 = 0 \tag{12}$$

Here, $p = -6$, $q = -9$, and so

$$\frac{q^2}{4} + \frac{p^3}{27} = \frac{49}{4} > 0$$

That is, equation (12) has one real and two conjugate complex roots. By (9), $\alpha = \sqrt[3]{\frac{9}{2} + \frac{7}{2}} = \sqrt[3]{8}$, $\beta = \sqrt[3]{\frac{9}{2} - \frac{7}{2}} = \sqrt[3]{1}$. For this reason, $\alpha_1 = 2$, $\beta_1 = 1$, i.e., $x_1 = 3$. The other two roots can be found by using formulas (10): $x_2 = -\frac{3}{2} + i\frac{\sqrt{3}}{2}$, $x_3 = -\frac{3}{2} - i\frac{\sqrt{3}}{2}$.

This implies that the roots of the given equation are the numbers

$$y_1 = 2, \quad y_2 = -\frac{5}{2} + i\frac{\sqrt{3}}{2}, \quad y_3 = -\frac{5}{2} - i\frac{\sqrt{3}}{2}$$

Example 2. Solve

$$x^3 - 12x + 16 = 0$$

Here, $p = -12$, $q = 16$, and so

$$\frac{q^2}{4} + \frac{p^3}{27} = 0$$

Whence $\alpha = \sqrt[3]{-8}$, or $\alpha_1 = -2$. And therefore

$$x_1 = -4, \quad x_2 = x_3 = 2$$

Example 3. Solve

$$x^3 - 19x - 30 = 0$$

Here, $p = -19$, $q = -30$, and so

$$\frac{q^2}{4} + \frac{p^3}{27} = -\frac{784}{27} < 0$$

Thus, Cardan's formula cannot be applied to this equation if we remain in the domain of real numbers, although the roots are the real numbers 2, 3, —5.

Quartic equations. The solution of the quartic equation

$$y^4 + ay^3 + by^2 + cy + d = 0 \tag{13}$$

with arbitrary complex coefficients reduces to a solution of some auxiliary cubic equation. This is achieved by a procedure due to Ferrari.

First, the substitution $y = x - \frac{a}{4}$ reduces equation (13) to the form

$$x^4 + px^2 + qx + r = 0 \tag{14}$$

The left member of this equation is then identically transformed with the aid of the auxiliary parameter α:

$$x^4 + px^2 + qx + r = \left(x^2 + \frac{p}{2} + \alpha\right)^2 + qx + r - \frac{p^2}{4} - \alpha^2 - 2\alpha x^2 - p\alpha$$

or

$$\left(x^2 + \frac{p}{2} + \alpha\right)^2 - \left[2\alpha x^2 - qx + \left(\alpha^2 + p\alpha - r + \frac{p^2}{4}\right)\right] = 0 \tag{15}$$

Now choose α so as to complete the square in the square brackets. This requires that it have one double root; in other words, we must have the equation

$$q^2 - 4 \cdot 2\alpha \left(\alpha^2 + p\alpha - r + \frac{p^2}{4}\right) = 0 \tag{16}$$

Equation (16) is a cubic equation in the unknown α with complex coefficients. As we know, this equation has three complex roots. Let α_0 be one of them; it is expressed, by Cardan's formula, with the aid of radicals in terms of the coefficients of equation (16), that is, in terms of the coefficients of equation (14).

Given this choice of value for α, the polynomial in the square brackets in (15) has the double root $\frac{q}{4\alpha_0}$, and so equation (15) takes the form

$$\left(x^2 + \frac{p}{2} + \alpha_0\right)^2 - 2\alpha_0 \left(x - \frac{q}{4\alpha_0}\right)^2 = 0$$

Hence it decomposes into two quadratic equations:

$$\left.\begin{aligned} x^2 - \sqrt{2\alpha_0}\,x + \left(\frac{p}{2} + \alpha_0 + \frac{q}{2\sqrt{2\alpha_0}}\right) &= 0, \\ x^2 + \sqrt{2\alpha_0}\,x + \left(\frac{p}{2} + \alpha_0 - \frac{q}{2\sqrt{2\alpha_0}}\right) &= 0 \end{aligned}\right\} \tag{17}$$

Since we passed from (14) to (17) by means of identity transformations, the roots of (17) will serve as roots for equation (14) as well. At the same time, it is easy to see that the roots of (14) are expressed in terms of coefficients by means of radicals. We will not write out the appropriate formulas because they are exceedingly unwieldy and of no practical use. Neither will we investigate separately the case when (14) has real coefficients.

Remarks on higher-degree equations. Whereas the ancient Greeks knew the methods for solving quadratic equations, the above-described methods for solving cubic and quartic equations were discovered only in the 16th century. This was followed by almost three

centuries of unsuccessful attempts to find formulas expressing by radicals the roots of any quintic equation (an equation of the fifth degree with literal coefficients) in terms of its coefficients. These attempts ceased only after Abel demonstrated, in the 1820's, that no such formulas can be found for nth-degree equations where $n \geqslant 5$.

This result of Abel's however did not preclude the possibility that the roots of a concrete polynomial with numerical coefficients could, in some way, be expressed in terms of the coefficients by some combination of radicals, or, as we usually say, that any equation is solvable by radicals. In the 1830's, Galois made a complete investigation of the conditions under which a given equation is solvable by radicals. It turned out that for any n equal to or greater than 5 there are nth-degree equations even with integral coefficients that are not solvable by radicals. Such, for instance, is the equation

$$x^5 - 4x - 2 = 0$$

The investigations of Galois exerted a decisive influence on the subsequent development of algebra, but they lie outside the scope of this text.

39. Bounds of Roots

We know that there is no method by which we can find the exact values of the roots of polynomials with numerical coefficients. Nevertheless, a vast range of problems in mechanics, physics and engineering at large reduce to the problem of the roots of polynomials, which at times are of very high degree. This circumstance spurred numerous investigations to find ways of describing the roots of a polynomial with numerical coefficients without actually knowing the roots. For example, studies have been made of the location of roots in the complex plane (the conditions under which all roots lie within the unit circle, that is, are less than unity in absolute value, or the conditions prescribing all roots to lie in the left half-plane, that is, to have negative real parts, etc.). For polynomials with real coefficients, methods have been elaborated for determining the number of their real roots, for finding the bounds within which these roots may be located, etc. Finally, much research has been done into methods of approximation of roots: in engineering situations, it is ordinarily enough to know only certain approximate values of the roots to within a specified accuracy, and if, say, the roots of a polynomial were even written as radicals, the latter would in any case be replaced by their approximations.

At one time, such studies constituted the basic content of higher algebra. We include here only a very small portion of the pertinent results, and taking into account the primary demands of applica-

tions we confine ourselves to the case of polynomials with real coefficients and real roots. In only a few instances will we go farther afield. We will consider the polynomial $f(x)$ with real coefficients as a (continuous) real function of a real variable x and wherever advisable we will take advantage of the results and methods of mathematical analysis.

A good way to begin the study of the real roots of a polynomial $f(x)$ with real coefficients is to examine the graph of the polynomial: *obviously, only the abscissas of the points of intersection of the graph and the x-axis are the real roots of the polynomial.*

To take an example, let us consider the fifth-degree polynomial

$$h(x) = x^5 + 2x^4 - 5x^3 + 8x^2 - 7x - 3$$

On the basis of the results of Sec. 24, we can assert the following concerning the roots of this polynomial: since its degree is odd, $h(x)$ has at least one real root; but if the number of real roots is greater than unity, then it is equal to three or five, since complex roots are pairwise conjugate.

An examination of the graph of the polynomial $h(x)$ enables us to say a good deal more about the roots. We construct the graph (Fig. 9; note that the scale on the x-axis is ten times that on the y-axis), taking only integral values of x and computing the corresponding values of $h(x)$, say by the Horner method:

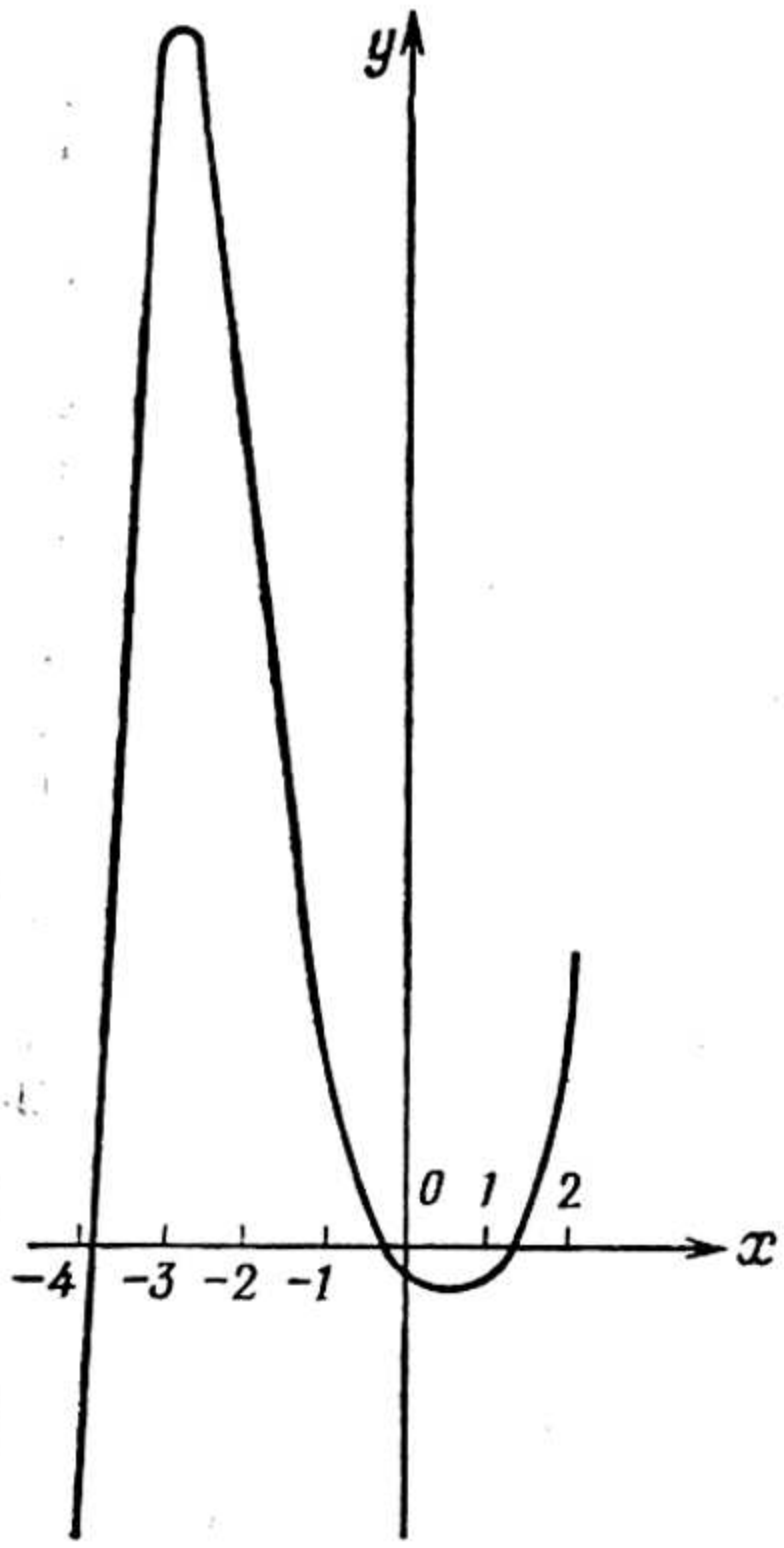

Fig. 9

x	$h(x)$
⋮	⋮
−4	−39
−3	144
−2	83
−1	18
0	−3
1	−4
2	39
⋮	⋮

We see that the polynomial $h(x)$ has in any case three real roots—the positive root α_1 and two negative roots α_2 and α_3,

$$1 < \alpha_1 < 2, \quad -1 < \alpha_2 < 0,$$
$$-4 < \alpha_3 < -3$$

Ordinarily, the information on the (real) roots of a polynomial that we get by examining the graph is very satisfactory in a practical sense. However, the doubt always remains as to whether we have indeed found all the roots. For instance, in the case at hand we did not show that to the right of $x = 2$ and to the left of $x = -4$ there are no roots of the polynomial. What is more, since we only took integral values of x, we can assume that the graph we constructed does not very accurately reflect the true behaviour of the function $h(x)$; it may not, say, take into account the smaller fluctuations and so loses some roots.

True, we could have taken values down to 0.1 or 0.01, in addition to the integral values of x. But then the computations would have been severely complicated and doubts would still remain. On the other hand, we could apply mathematical analysis to test the function $h(x)$ for maxima and minima and thus compare our graph with the true behaviour of the function; but this brings us to the problem of the roots of the derivative $h'(x)$, which is the same kind of problem we are dealing with right now.

The need is evident for more sophisticated procedures enabling us to find the bounds within which lie the real roots of a polynomial with real coefficients and to determine the number of the roots. We shall examine the problem of the bounds of real roots and leave the question of the number of roots to later sections.

The proof of the lemma on the modulus of the highest-degree term (see Sec. 23) already provides a certain bound for the absolute values of the roots of a polynomial. Indeed, setting $k = 1$ in inequality (3), Sec. 23, we find that for

$$|x| \geqslant 1 + \frac{A}{|a_0|} \tag{1}$$

where a_0 is the leading coefficient and A is the maximum of the absolute values of the remaining coefficients, the absolute value of the highest-degree term of the polynomial is greater than the absolute value of the sum of all the other terms, and so no value of x which satisfies inequality (1) can be a root of this polynomial.

Thus, *for a polynomial $f(x)$ with arbitrary numerical coefficients, the number $1 + \frac{A}{|a_0|}$ serves as an upper bound of the moduli (absolute values) of all its roots, real and complex.* For the case above of the polynomial $h(x)$, this bound, since $a_0 = 1$, $A = 8$, is the number 9.

However, this bound is usually too high, particularly if we are only interested in the bounds of the real roots. We now give certain more precise methods. It is well to bear in mind that if the bounds are indicated within which the real roots of a polynomial are to be found, this does not in the least mean that such roots actually exist.

Let us first demonstrate that *it is sufficient to be able to find only the upper bound of the real roots of any polynomial.* Let there be given a polynomial $f(x)$ of degree n and let N_0 be the upper bound of its positive roots. We consider the polynomials

$$\varphi_1(x) = x^n f\left(\frac{1}{x}\right),$$

$$\varphi_2(x) = f(-x),$$

$$\varphi_3(x) = x^n f\left(-\frac{1}{x}\right)$$

and find the upper bounds of their positive roots. Suppose these are the numbers, respectively, N_1, N_2, N_3. Then the *number* $\frac{1}{N_1}$ *will be the lower bound of the positive roots of the polynomial* $f(x)$: if α is a positive root of $f(x)$, then $\frac{1}{\alpha}$ will be a positive root of $\varphi_1(x)$ and from $\frac{1}{\alpha} < N_1$ follows $\alpha > \frac{1}{N_1}$. Similarly, *the numbers* $-N_2$ *and* $-\frac{1}{N_3}$ *serve, respectively, as the upper and lower bounds of the negative roots of the polynomial* $f(x)$. Thus, all positive roots of $f(x)$ satisfy the inequalities $\frac{1}{N_1} < x < N_0$, all negative roots, the inequalities $-N_2 < x < -\frac{1}{N_3}$.

To determine the upper bound of the positive roots we can use the following method. Suppose we have the polynomial

$$f(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n$$

with real coefficients, and $a_0 > 0$. Let a_k, $k \geqslant 1$, be the first of the negative coefficients; if there were no such coefficients, then the polynomial $f(x)$ could not have any positive roots at all. Finally, let B be the greatest of the absolute values of the negative coefficients. Then *the number*

$$1 + \sqrt[k]{\frac{B}{a_0}}$$

serves as the upper bound for the positive roots of the polynomial $f(x)$.

Indeed, setting $x > 1$ and replacing each of the coefficients $a_1, a_2, \ldots, a_{k-1}$ by the number zero, and each of the coefficients $a_k, a_{k+1}, \ldots, a_n$ by the number $-B$, we can only diminish the

value of the polynomial, that is,

$$f(x) \geqslant a_0 x^n - B(x^{n-k} + x^{n-k-1} + \ldots + x + 1) = a_0 x^n - B\frac{x^{n-k+1}-1}{x-1}$$

or, because $x > 1$,

$$f(x) > a_0 x^n - \frac{Bx^{n-k+1}}{x-1} = \frac{x^{n-k+1}}{x-1}[a_0 x^{k-1}(x-1) - B] \tag{2}$$

If

$$x > 1 + \sqrt[k]{\frac{B}{a_0}} \tag{3}$$

then, since

$$a_0 x^{k-1}(x-1) - B \geqslant a_0 (x-1)^k - B$$

the expression in square brackets in formula (2) will prove to be positive; thus, by (2), the value of $f(x)$ will be strictly positive. Thus, the values of x which satisfy the inequality (3) cannot be roots of $f(x)$, which is what we set out to prove.

Taking the above-considered polynomial $h(x)$, this method (since $k = 2$, $B = 7$) yields for the upper bound of the positive roots the number $1 + \sqrt{7}$, which can be replaced by the nearest greater integer 4.

Of the many other methods of finding the upper bound of positive roots, we give *Newton's method*. It is more involved than the one we just gave above, but ordinarily it yields a very good result.

Suppose we have a polynomial $f(x)$ with real coefficients and positive leading coefficient a_0. *If, for* $x = c$, *the polynomial* $f(x)$ *and all its successive derivatives* $f'(x)$, $f''(x)$, . . ., $f^{(n)}(x)$ *take on positive values, then the number* c *serves as the upper bound of the positive roots.*

True enough, by Taylor's formula (see Sec. 23),

$$f(x) = f(c) + (x-c)f'(c) + (x-c)^2\frac{f''(c)}{2!} + \ldots + (x-c)^n\frac{f^{(n)}(c)}{n!}$$

We see that if $x \geqslant c$, then on the right we get a strictly positive number, that is, such values of x cannot be the roots of $f(x)$.

When seeking the appropriate number c for a given polynomial $f(x)$, it is useful to do as follows. The derivative $f^{(n)}(x) = n!a_0$ is a positive number, and so the polynomial $f^{(n-1)}(x)$ is an increasing function of x. Consequently, there is a number c_1 such that for $x \geqslant c_1$ the derivative $f^{(n-1)}(x)$ is positive. Whence it follows that for $x \geqslant c_1$ the derivative $f^{(n-2)}(x)$ will be an increasing function of x and therefore there exists a number c_2, $c_2 \geqslant c_1$, such that for $x \geqslant c_2$ the derivative $f^{(n-2)}(x)$ is also positive. Continuing thus, we finally arrive at the desired number c.

Applying Newton's method to the polynomial $h(x)$ considered above, we have

$$h(x) = x^5 + 2x^4 - 5x^3 + 8x^2 - 7x - 3,$$
$$h'(x) = 5x^4 + 8x^3 - 15x^2 + 16x - 7,$$
$$h''(x) = 20x^3 + 24x^2 - 30x + 16,$$
$$h'''(x) = 60x^2 + 48x - 30,$$
$$h^{IV}(x) = 120x + 48,$$
$$h^{V}(x) = 120$$

It is easy to verify (say, by the Horner method) that all these polynomials are positive for $x = 2$. Thus the number 2 is the upper bound for the positive roots of the polynomial $h(x)$. This result is much more exact than those obtained by other methods.

To find a lower bound for the negative roots of polynomial $h(x)$, let us consider the polynomial $\varphi_2(x) = -h(-x)$ *. Since

$$\varphi_2(x) = x^5 - 2x^4 - 5x^3 - 8x^2 - 7x + 3,$$
$$\varphi_2'(x) = 5x^4 - 8x^3 - 15x^2 - 16x - 7,$$
$$\varphi_2''(x) = 20x^3 - 24x^2 - 30x - 16,$$
$$\varphi_2'''(x) = 60x^2 - 48x - 30,$$
$$\varphi_2^{IV}(x) = 120x - 48,$$
$$\varphi_2^{V}(x) = 120$$

and all these polynomials are positive (as may readily be checked for $x = 4$), the number 4 serves as an upper bound for the positive roots of $\varphi_2(x)$, and so the number -4 will be a lower bound for the negative roots of $h(x)$.

Finally, let us consider the polynomials

$$\varphi_1(x) = -x^5 h\left(\frac{1}{x}\right) = 3x^5 + 7x^4 - 8x^3 + 5x^2 - 2x - 1,$$
$$\varphi_3(x) = -x^5 h\left(-\frac{1}{x}\right) = 3x^5 - 7x^4 - 8x^3 - 5x^2 - 2x + 1$$

For them, again using the Newton method, we find the numbers 1 and 4 as upper bounds for the positive roots and so the number $\frac{1}{1} = 1$ is the lower bound for the positive roots of $h(x)$ and the number $-\frac{1}{4}$ is the upper bound for the negative roots.

---

* $-h(-x)$ in place of $h(-x)$ because Newton's method is applicable only if the leading coefficient is positive. This change of sign of course has no effect whatsoever on the roots of the polynomial $\varphi_2(x)$.

Thus, the positive roots of $h(x)$ lie between 1 and 2 and the negative roots lie between the numbers -4 and $-\frac{1}{4}$. This result is in very good agreement with what we found earlier when we examined the graph.

40. Sturm's Theorem

We now come to the question of the number of real roots of a polynomial $f(x)$ with real coefficients. We will be interested both in the total number of real roots, and, separately, the number of positive and the number of negative roots and the total number of roots in the interval between specified bounds a and b. There are several methods for finding the exact number of roots and all of them are very cumbersome; the most convenient one is the *Sturm method*, which we now discuss.

First let us introduce a definition that will be needed in the next section as well.

Suppose we have a finite ordered sequence of real numbers different from zero, say

$$1, 3, -2, 1, -4, -8, -3, 4, 1 \tag{1}$$

Write down the signs of these numbers in succession:

$$+, +, -, +, -, -, -, +, + \tag{2}$$

We see that there are four variations of sign in (2). We then say that in the ordered sequence (1) there are four *variations in sign*. The number of variations in sign can of course be counted for any finite ordered sequence of nonzero real numbers.

Now let us consider the polynomial $f(x)$ with real coefficients; we will assume that $f(x)$ does not have multiple roots, for then we could divide it by its greatest common divisor and its derivative. The finite ordered sequence of nonzero polynomials with real coefficients

$$f(x) = f_0(x), f_1(x), f_2(x), \ldots, f_s(x) \tag{3}$$

is called the *Sturm sequence* for the polynomial $f(x)$ if the following requirements are met:

(1) Successive polynomials of (3) do not have common roots.

(2) The last polynomial, $f_s(x)$, does not have real roots.

(3) If α is a real root of one of the intermediate polynomials $f_k(x)$ of (3), $1 \leqslant k \leqslant s-1$, then $f_{k-1}(\alpha)$ and $f_{k+1}(\alpha)$ have different signs.

(4) If α is a real root of $f(x)$, then the product $f(x) f_1(x)$ changes sign from minus to plus when x increases and passes through the point α.

The question of whether every polynomial has a Sturm sequence will be considered later on, for the present let us suppose that $f(x)$ does have such a sequence and let us show how it can be used to find the number of real roots.

If a real number c is not a root of the given polynomial $f(x)$ and (3) is a Sturm sequence for this polynomial, then take the set of real numbers

$$f(c),\ f_1(c),\ f_2(c),\ \ldots,\ f_s(c)$$

delete all numbers equal to zero and denote by $W(c)$ the number of variations in sign in the remaining sequence; we call $W(c)$ *the number of variations in sign in the Sturm sequence* (3) *of polynomial* $f(x)$, $x = c$.*

The following theorem holds.

Sturm's theorem. *If the real numbers* a *and* b, $a < b$, *are not the roots of a polynomial* $f(x)$ *which does not have any multiple roots, then* $W(a) \geqslant W(b)$ *and the difference* $W(a) - W(b)$ *is equal to the number of real roots of* $f(x)$ *in the interval between* a *and* b.

Thus, to determine the number of real roots of a polynomial $f(x)$ lying between a and b [recall that $f(x)$ does not, by hypothesis, have multiple roots], it suffices to establish the reduction in the number of variations of sign in the Sturm sequence of this polynomial when moving from a to b.

To prove this theorem, let us see how the number $W(x)$ varies with increasing x. So long as x, as it increases, does not encounter any of the roots of the Sturm sequence (3), the signs of the polynomials of the sequence do not change and so the number $W(x)$ remains unaltered. For this reason, and also because of Condition (2) of the definition of a Sturm sequence, it remains for us to consider two cases: the passage of x through a root of one of the intermediate polynomials $f_k(x)$, $1 \leqslant k \leqslant s-1$, and the passage of x through a root of the polynomial $f(x)$ itself.

Let α be a root of the polynomial $f_k(x)$, $1 \leqslant k \leqslant s-1$. Then, by Condition (1), $f_{k-1}(\alpha)$ and $f_{k+1}(\alpha)$ are different from zero. We can thus find a positive number ε, which may be very small, such that in the interval $(\alpha - \varepsilon,\ \alpha + \varepsilon)$ the polynomials $f_{k-1}(x)$ and $f_{k+1}(x)$ do not have any roots and therefore preserve constant signs: Condition (3) states that these signs are distinct. From this it follows that each of the sequences of numbers

$$f_{k-1}(\alpha - \varepsilon),\ f_k(\alpha - \varepsilon),\ f_{k+1}(\alpha - \varepsilon) \qquad (4)$$

---

* Quite naturally, the variations in sign in the Sturm sequence of the polynomial $f(x)$ have nothing in common with the variation in sign of the polynomial $f(x)$ itself, which variation occurs when x passes through a root of the polynomial.

and

$$f_{k-1}(\alpha+\varepsilon),\ f_k(\alpha+\varepsilon),\ f_{k+1}(\alpha+\varepsilon) \tag{5}$$

has exactly one variation in sign, irrespective of the signs of the numbers $f_k(\alpha-\varepsilon)$ and $f_k(\alpha+\varepsilon)$. Thus, for instance, if the polynomial $f_{k-1}(x)$ is negative on the interval in question and $f_{k+1}(x)$ is positive and if $f_k(\alpha-\varepsilon)>0$, $f_k(\alpha+\varepsilon)<0$, then the sequences (4) and (5) are associated with the sign sequences

$$-,\ +,\ +;\ -,\ -,\ +$$

Thus, when x passes through a root of one of the intermediate polynomials in Sturm's sequence, the variations in sign in the sequence can only shift position, but do not disappear or reappear, and so the number $W(x)$ does not change in such a transition.

On the other hand, let α be a root of the given polynomial $f(x)$. By Condition (1), α will not be a root of $f_1(x)$. Hence, there is a positive number ε such that the interval $(\alpha-\varepsilon, \alpha+\varepsilon)$ does not contain any roots of the polynomial $f_1(x)$, and therefore $f_1(x)$ preserves its sign over this interval. If the sign is positive, then, by Condition (4), the polynomial $f(x)$ itself changes sign from minus to plus when x passes through α, i.e., $f(\alpha-\varepsilon)<0$, $f(\alpha+\varepsilon)>0$. Hence, to the number sequences

$$f(\alpha-\varepsilon),\ f_1(\alpha-\varepsilon) \text{ and } f(\alpha+\varepsilon),\ f_1(\alpha+\varepsilon) \tag{6}$$

there correspond the sign sequences

$$-,\ +\ \text{and}\ +,\ +$$

Thus, the Sturm sequence loses one variation in sign. But if the sign of $f_1(x)$ is negative on the interval $(\alpha-\varepsilon, \alpha+\varepsilon)$, then again, by Condition (4), the polynomial $f(x)$ changes sign from plus to minus as x passes through α, i.e., $f(\alpha-\varepsilon)>0$, $f(\alpha+\varepsilon)<0$. To the number sequences (6) there now correspond the sign sequences

$$+,\ -\ \text{and}\ -,\ -$$

The Sturm sequence again loses one variation in sign.

Thus, *as x increases, the number $W(x)$ changes only when x passes through a root of the polynomial $f(x)$, in this case it is diminished exactly by unity.*

This obviously proves the Sturm theorem. To use it for finding the total number of real roots of a polynomial $f(x)$, it is sufficient to take, for a, the lower limit of the negative roots, and for b, the upper limit of the positive roots. It is simpler however to do as follows. By the lemma proved in Sec. 23 there exists a positive number N, which may be very large, such that for $|x|>N$ the signs of all polynomials of the Sturm sequence will coincide with the signs of their highest-degree terms. In other words, there exists a positive

value of the unknown x which is so large that the signs of the corresponding values of all the polynomials of the Sturm sequence coincide with the signs of their leading coefficients. This value of x, which need not be computed, can be denoted by ∞. On the other hand, there exists a negative value of x which is so large in absolute value that the signs of the corresponding values of the polynomials of the Sturm sequence coincide with the signs of their leading coefficients for polynomials of even degree and are opposite to the signs of the leading coefficients for polynomials of odd degree. Let us agree to denote this value of x by $-\infty$. In the interval $(-\infty, \infty)$ we obviously have all the real roots of all the polynomials of Sturm's sequence and, in particular, all the real roots of the polynomial $f(x)$. Applying the Sturm theorem to this interval, we find the number of these roots; application of the Sturm theorem to the intervals $(-\infty, 0)$ and $(0, \infty)$ yields, respectively, the number of negative and the number of positive roots of the polynomial $f(x)$.

It remains to demonstrate that *any polynomial $f(x)$ with real coefficients and without multiple roots has a Sturm sequence.* Of a variety of methods used for constructing such a sequence, we give the most widely used one. Set $f_1(x) = f'(x)$, thus ensuring fullfilment of Condition (4) of the definition of a Sturm sequence. Indeed, if α is a real root of the polynomial $f(x)$, then $f'(\alpha) \neq 0$. If $f'(\alpha) > 0$, then $f'(x) > 0$ in the neighbourhood of the point α and therefore $f(x)$ changes sign from minus to plus when x passes through α; this is then also true for the product $f(x) f_1(x)$. Similar reasoning is likewise valid for $f'(\alpha) < 0$. Then divide $f(x)$ by $f_1(x)$ and take the remainder (with reversed sign) for $f_2(x)$:

$$f(x) = f_1(x) q_1(x) - f_2(x)$$

Generally, if the polynomials $f_{k-1}(x)$ and $f_k(x)$ have already been found, then $f_{k+1}(x)$ will be the remainder after dividing $f_{k-1}(x)$ by $f_k(x)$ taken with reversed sign:

$$f_{k-1}(x) = f_k(x) q_k(x) - f_{k+1}(x) \qquad (7)$$

This method differs from the Euclidean algorithm as applied to the polynomials $f(x)$ and $f'(x)$ solely in the fact that the sign of the remainder is reversed every time, and the next division is performed by the remainder with reversed sign. Since such a variation in sign is inessential when seeking the greatest common divisor, our process will terminate at some $f_s(x)$, which is the greatest common divisor of the polynomials $f(x)$ and $f'(x)$; since $f(x)$ has no multiple roots [it is prime to $f'(x)$] it will follow that $f_s(x)$ is actually some nonzero real number.

This implies that the sequence of polynomials we have constructed,

$$f(x) = f_0(x),\ f'(x) = f_1(x),\ f_2(x),\ \ldots,\ f_s(x)$$

also satisfies Condition (2) of the definition of a Sturm sequence. To prove that Condition (1) is met, assume that the consecutive polynomials $f_k(x)$ and $f_{k+1}(x)$ have a common root α. Then, by (7), α will also be a root of the polynomial $f_{k-1}(x)$. Passing to the equation

$$f_{k-2}(x) = f_{k-1}(x)\, q_{k-1}(x) - f_k(x)$$

we find that α is a root of $f_{k-2}(x)$ as well. Continuing, we find that α is a common root of $f(x)$ and $f'(x)$, which is in conflict with our assumptions. Finally, fulfillment of Condition (3) follows directly from equation (7); if $f_k(\alpha) = 0$, then $f_{k-1}(\alpha) = -f_{k+1}(\alpha)$.

Let us apply the Sturm method to the polynomial

$$h(x) = x^5 + 2x^4 - 5x^3 + 8x^2 - 7x - 3$$

which we considered in the preceding section. We will not make a preliminary check to see that $h(x)$ does not have any multiple roots, because the method of constructing a Sturm sequence as given above is a simultaneous check on the relative primality of the polynomial and its derivative.

Let us find a Sturm sequence for $h(x)$ by using this method. In the division process, we will (in contrast to the Euclidean algorithm) multiply and divide only by arbitrary *positive* numbers since the signs of the remainders play a fundamental role in the Sturm method. We obtain the following sequence:

$$\begin{aligned}
h(x) &= x^5 + 2x^4 - 5x^3 + 8x^2 - 7x - 3,\\
h_1(x) &= 5x^4 + 8x^3 - 15x^2 + 16x - 7,\\
h_2(x) &= 66x^3 - 150x^2 + 172x + 61,\\
h_3(x) &= -464x^2 + 1135x + 723,\\
h_4(x) &= -32{,}599{,}457x - 8{,}486{,}093,\\
h_5(x) &= -1
\end{aligned}$$

We determine the signs of the polynomials of this sequence for $x = -\infty$ and $x = \infty$; to do this, we (as indicated above) only examine the signs of the leading coefficients and the degrees of the polynomials. We get the following table:

	$h(x)$	$h_1(x)$	$h_2(x)$	$h_3(x)$	$h_4(x)$	$h_5(x)$	Number of variations in sign
$-\infty$	−	+	−	−	+	−	4
∞	+	+	+	−	−	−	1

Thus, when x passes from $-\infty$ to ∞, the Sturm sequence loses three variations in sign and so the polynomial $h(x)$ has exactly three real roots. It will be recalled that when we constructed the graph of this polynomial (in the preceding section) we did not lose a single root.

Let us apply the Sturm method to a simpler polynomial:

$$f(x) = x^3 + 3x^2 - 1$$

Let us find the number of its real roots and also the integral bounds within which each of the roots is located. We shall not construct the graph of this polynomial.

The Sturm sequence associated with the polynomial $f(x)$ is

$$f(x) = x^3 + 3x^2 - 1,$$
$$f_1(x) = 3x^2 + 6x,$$
$$f_2(x) = 2x + 1,$$
$$f_3(x) = 1$$

Let us find the number of variations of sign in this sequence for $x = -\infty$ and $x = \infty$

	$f(x)$	$f_1(x)$	$f_2(x)$	$f_3(x)$	Number of variations in sign
$-\infty$	$-$	$+$	$-$	$+$	3
∞	$+$	$+$	$+$	$+$	0

Consequently, the polynomial $f(x)$ has three real roots. For a more precise location of the roots, continue the above table:

	$f(x)$	$f_1(x)$	$f_2(x)$	$f_3(x)$	Number of variations in sign
$x=-3$	$-$	$+$	$-$	$+$	3
$x=-2$	$+$	0	$-$	$+$	2
$x=-1$	$+$	$-$	$-$	$+$	2
$x=0$	$-$	0	$+$	$+$	1
$x=1$	$+$	$+$	$+$	$+$	0

Thus, the Sturm sequence of the polynomial $f(x)$ loses one variation of sign each time x moves from -3 to -2, from -1 to 0 and from 0 to 1. The roots α_1, α_2 and α_3 of this polynomial thus satisfy the inequalities

$$-3 < \alpha_1 < -2, \quad -1 < \alpha_2 < 0, \quad 0 < \alpha_3 < 1$$

41. Other Theorems on the Number of Real Roots

The Sturm theorem completely resolves the question of the number of real roots of a polynomial, but it has one essential defect and that is the cumbersome computations involved in constructing a Sturm sequence, as the reader could see after performing all the computations of the first example above. We now prove two theorems which do not yield the exact number of real roots but only bound the number from above. These theorems are employed after a graph has been used to bound the number of real roots from below and at times enable us to find the exact number of real roots without resorting to the Sturm method.

Suppose we have an nth-degree polynomial $f(x)$ with real coefficients; we assume it can have multiple roots. Let us consider a sequence of its consecutive derivatives:

$$f(x) = f^{(0)}(x),\ f'(x),\ f''(x),\ \ldots,\ f^{(n-1)}(x),\ f^{(n)}(x) \qquad (1)$$

of which the last one is equal to the leading coefficient a_0 of $f(x)$ multiplied by $n!$ and for this reason preserves sign at all times. If a real number c is not a root of any one of the polynomials of the sequence (1), then by $S(c)$ we denote the number of variations in sign in the ordered sequence of numbers

$$f(c),\ f'(c),\ f''(c),\ \ldots,\ f^{(n-1)}(c),\ f^{(n)}(c)$$

Thus, we can consider the integer-valued function $S(x)$ defined for those values of x which do not make any one of the polynomials in (1) vanish.

Let us see how $S(x)$ varies with increasing x. The number $S(x)$ remains unchanged until x passes through a root of one of the polynomials of (1). We thus have two cases to consider: the passage of x through a root of the polynomial $f(x)$ and the passage of x through a root of one of the derivatives $f^{(k)}(x)$, $1 \leqslant k \leqslant n-1$.

Let α be an l-fold root of the polynomial $f(x)$, $l \geqslant 1$, i.e.,

$$f(\alpha) = f'(\alpha) = \ldots = f^{(l-1)}(\alpha) = 0, \quad f^{(l)}(\alpha) \neq 0$$

Let a positive number ε be so small that the interval $(\alpha - \varepsilon, \alpha + \varepsilon)$ does not contain any roots of the polynomials $f(x)$, $f'(x)$, ... $f^{(l-1)}(x)$, different from α and does not contain any root of the

polynomial $f^{(l)}(x)$ either. We will prove that in the number sequence

$$f(\alpha-\varepsilon),\ f'(\alpha-\varepsilon),\ \ldots,\ f^{(l-1)}(\alpha-\varepsilon),\ f^{(l)}(\alpha-\varepsilon)$$

any two consecutive numbers have opposite signs, whereas all the numbers

$$f(\alpha+\varepsilon),\ f'(\alpha+\varepsilon),\ \ldots,\ f^{(l-1)}(\alpha+\varepsilon),\ f^{(l)}(\alpha+\varepsilon)$$

have the same sign. Since each one of the polynomials of (1) is a derivative of the preceding polynomial, all we have to prove is that if x passes through the root α of polynomial $f(x)$, then, irrespective of the multiplicity of this root, $f(x)$ and $f'(x)$ had different signs prior to the passage and have coincident signs after the passage. If $f(\alpha-\varepsilon)>0$, then $f(x)$ diminishes on the interval $(\alpha-\varepsilon,\ \alpha)$, and so $f'(\alpha-\varepsilon)<0$; but if $f(\alpha-\varepsilon)<0$, then $f(x)$ increases and so $f'(\alpha-\varepsilon)>0$. Hence in both cases the signs differ. On the other hand, if $f(\alpha+\varepsilon)>0$, then $f(x)$ increases on the interval $(\alpha,\ \alpha+\varepsilon)$ and so $f'(\alpha+\varepsilon)>0$; similarly, from $f(\alpha+\varepsilon)<0$ it follows that $f'(\alpha+\varepsilon)<0$. Thus, after the passage through the root α, the signs of $f(x)$ and $f'(x)$ must coincide.

From what has been proved it follows that when x passes through an l-fold root of the polynomial $f(x)$ the sequence

$$f(x),\ f'(x),\ \ldots,\ f^{(l-1)}(x),\ f^{(l)}(x)$$

loses l variations in sign.

Now let α be a root of the derivatives

$$f^{(k)}(x),\ f^{(k+1)}(x),\ \ldots,\ f^{(k+l-1)}(x),\quad 1\leqslant k\leqslant n-1,\quad l\geqslant 1$$

but not a root of $f^{(k-1)}(x)$ or of $f^{(k+l)}(x)$. By what has been proved above, the passage of x through α implies a loss in the sequence

$$f^{(k)}(x),\ f^{(k+1)}(x),\ \ldots,\ f^{(k+l-1)}(x),\ f^{(k+l)}(x)$$

of l variations in sign. True, this passage possibly creates a new variation in sign between $f^{(k-1)}(x)$ and $f^{(k)}(x)$; however, because $l\geqslant 1$, the number of variations in sign, when x passes through α in the sequence

$$f^{(k-1)}(x),\ f^{(k)}(x),\ f^{(k+1)}(x),\ \ldots,\ f^{(k+l-1)}(x),\ f^{(k+l)}(x)$$

either does not change or decreases. It can then decrease only by an even number since the polynomials $f^{(k-1)}(x)$ and $f^{(k+l)}(x)$ do not change sign when x passes through the value α.

These results imply that *if the numbers a and b, $a<b$, are not roots of any one of the polynomials of the sequence* (1), *then the number of real roots of the polynomial f (x) lying between a and b (each counted according to its multiplicity) is equal to the difference S (a) — S (b) or is less than this difference by an even number.*

In order to relax the restrictions imposed on the numbers a and b, let us introduce the following notations. Suppose the real number c is not a root of the polynomial $f(x)$, though it may be a root of some of the other polynomials of the sequence (1). Denote by $S_+(c)$ the number of variations in sign in the number sequence

$$f(c),\ f'(c),\ f''(c),\ \ldots,\ f^{(n-1)}(c),\ f^{(n)}(c) \tag{2}$$

which is computed as follows: if

$$f^{(k)}(c) = f^{(k+1)}(c) = \ldots = f^{(k+l-1)}(c) = 0 \tag{3}$$

but

$$f^{(k-1)}(c) \neq 0, \quad f^{(k+l)}(c) \neq 0 \tag{4}$$

then we take it that $f^{(k)}(c)$, $f^{(k+1)}(c)$, . . ., $f^{(k+l-1)}(c)$ have the same sign as $f^{(k+l)}(c)$; this is obviously the same as deleting the zeros in a count of the number of variations of sign in the sequence (2). On the other hand, by $S_-(c)$ we denote the number of variations of sign in the sequence (2), which is counted as follows: if conditions (3) and (4) hold, then we take it that $f^{(k+i)}(c)$, $0 \leqslant i \leqslant l-1$, has the same sign as $f^{(k+l)}(c)$ if the difference $l - i$ is even, and opposite sign if this difference is odd.

If we now desire to determine the number of real roots of the polynomial $f(x)$ between a and b, $a < b$, and a and b are not roots of $f(x)$ but, possibly, are roots of the other polynomials of the sequence (1), then we do as follows. Let ε be so small that the interval $(a, a + 2\varepsilon)$ does not contain any roots of $f(x)$, or any roots (distinct from a) of the other polynomials of the sequence (1); on the other hand, let η be so small that the interval $(b - 2\eta, b)$ also fails to contain any roots of $f(x)$ and any roots (distinct from b) of the other polynomials of the sequence (1). Then the number we want of real roots of the polynomial $f(x)$ will be equal to the number of the real roots of this polynomial between $a + \varepsilon$ and $b - \eta$, that is, from what has been proved, it will be equal to the difference $S(a + \varepsilon) - S(b - \eta)$ or less than this difference by an even number. However, it is easy to see that

$$S(a + \varepsilon) = S_+(a), \quad S(b - \eta) = S_-(b)$$

This is proof of the following theorem.

Budan-Fourier theorem. *If the real numbers a and b, $a < b$, are not the roots of a polynomial $f(x)$ with real coefficients, then the number of real roots of this polynomial between a and b, each counted according to its multiplicity, is equal to the difference $S_+(a) - S_-(b)$ or is an even number less than this difference.*

Use the symbol ∞ to denote a positive value of the unknown x so large that the signs of the associated values of all the polynomials of the sequence (1) coincide with the signs of their leading coeffi-

cients. Since these coefficients are, sequentially, the numbers a_0, na_0, $n(n-1)a_0$, . . ., $n!a_0$, whose signs coincide, it follows that $S(\infty) = S_-(\infty) = 0$. On the other hand, since

$$f(0) = a_n,\ f'(0) = a_{n-1},\ f''(0) = a_{n-2}2!,$$
$$f'''(0) = a_{n-3}3!,\ \ldots,\ f^{(n)}(0) = a_0 \cdot n!$$

where $a_0, a_1, \ldots, a_n$ are coefficients of the polynomial $f(x)$, then $S_+(0)$ coincides with the number of variations in sign in the sequence of coefficients of $f(x)$, zero coefficients being deleted. Thus, applying the Budan-Fourier theorem to the interval $(0, \infty)$ we arrive at the following theorem.

Descartes' theorem (Descartes' rule of signs). *The number of positive roots of a polynomial $f(x)$, a root of multiplicity m being counted as m roots, is equal to the number of variations in sign in the sequence of coefficients of this polynomial (zero coefficients are not counted) or is less by an even number.*

To determine the number of negative roots of the polynomial $f(x)$ it is obviously sufficient to apply Descartes' theorem to the polynomial $f(-x)$. If none of the coefficients of $f(x)$ is zero, then, obviously, changes of sign in the sequence of coefficients of the polynomial $f(-x)$ will be associated with preservation of signs in the sequence of coefficients of the polynomial $f(x)$, and conversely. Thus, *if the polynomial $f(x)$ does not have zero coefficients, then the number of its negative roots (counting multiplicities) is equal to the number of preservations of signs in the sequence of coefficients or is less by an even number.*

We give another proof of the Descartes theorem that does not rely on the Budan-Fourier theorem. We first prove the following lemma.

If $c > 0$, then the number of variations of sign in the sequence of coefficients of the polynomial $f(x)$ is less than the number of variations of sign in the sequence of coefficients of the product $(x - c) f(x)$ by an odd number.

Indeed, enclosing in parentheses successive terms of the same sign, we can write the polynomial $f(x)$, the leading coefficient a_0 of which can be considered positive, as follows:

$$f(x) = (a_0x^n + \ldots + b_1x^{k_1+1}) - (a_1x^{k_1} + \ldots + b_2x^{k_2+1}) + \ldots + (-1)^s (a_sx^{k_s} + \ldots + b_{s+1}x^t) \quad (5)$$

Here, $a_0 > 0$, $a_1 > 0$, . . ., $a_s > 0$, whereas $b_1, b_2, \ldots, b_s$ are positive or zero, but b_{s+1} is considered strictly positive, that is, x^t, where $t \geqslant 0$, is the smallest power of the unknown x that enters into the polynomial $f(x)$ with a nonzero coefficient. The parenthesis

$$(a_0x^n + \ldots + b_1x^{k_1+1})$$

may accidentally consist of a single term, namely, when $k_1 + 1 = n$. An analogous remark is applicable to the other parentheses of formula (5).

Now write a polynomial equal to the product $(x - c) f(x)$; we will single out only those terms which contain x to the powers $n + 1$, $k_1 + 1, \ldots, k_s + 1$, and t. We obtain

$$(x-c) f(x) = (a_0 x^{n+1} + \ldots) - (a_1' x^{k_1+1} + \ldots) + \ldots + (-1)^s (a_s' x^{k_s+1} + \ldots - c b_{s+1} x^t) \quad (6)$$

where $a_i' = a_i + c b_i$, $i = 1, 2, \ldots, s$, and therefore, since $c > 0$, all the a_i' are strictly positive. Thus, there was one change of sign in the sequence of coefficients of the polynomial $f(x)$ between the terms $a_0 x^n$ and $-a_1 x^{k_1}$ (also between the terms $-a_1 x^{k_1}$ and $a_2 x^{k_2}$, etc.), whereas in the polynomial $(x - c) f(x)$ there will either be one change of sign between the corresponding terms $a_0 x^{n+1}$ and $-a_1' x^{k_1+1}$ (respectively between the terms $-a_1' x^{k_1+1}$ and $a_2' x^{k_2+1}$, etc.) or more changes (but always more by an even number). We are not interested in the exact places of these changes in sign. It may happen, for example, that the coefficient of x^{k_1+2} in (6) is negative, like the coefficient $-a_1'$, and so there is no change of sign between these two successive coefficients; that is to say, the change in sign in the first parenthesis is located at some previous position. Now notice that the last parenthesis in (5) did not have any variation in sign, whereas the last parenthesis in (6) did have variations in sign—an odd number of them: it suffices to note that the last nonzero coefficients of the polynomials $f(x)$ and $(x - c) f(x)$, that is, $(-1)^s b_{s+1}$ and $(-1)^{s+1} b_{s+1} c$ have different signs. Thus, between $f(x)$ and $(x - c) f(x)$ the total number of variations of sign in the sequence of coefficients invariably increases and by an odd number (the sum of several terms, one of which is odd and the others even, will naturally be odd!). The lemma is proved.

To prove Descartes' theorem, denote all the positive roots of the polynomial $f(x)$ by $\alpha_1, \alpha_2, \ldots, \alpha_k$. Then

$$f(x) = (x - \alpha_1)(x - \alpha_2) \ldots (x - \alpha_k) \varphi(x)$$

where $\varphi(x)$ is a polynomial with real coefficients which now has no positive real roots. This implies that the first and the last nonzero coefficients of the polynomial $\varphi(x)$ are of the same sign, which means that the sequence of coefficients of this polynomial contains an even number of variations of sign. Applying the above-proved lemma to the polynomials

$$\varphi(x), \ (x - \alpha_1) \varphi(x), \ (x - \alpha_1)(x - \alpha_2) \varphi(x), \ \ldots, \ f(x)$$

in succession, we find that the number of variations of sign in the sequence of coefficients increases each time by an odd number, that

is to say, by unity plus an even number, and so the number of variations of sign in the sequence of coefficients of the polynomial $f(x)$ is greater than k by an even number.

Let us apply the theorems of Descartes and Budan-Fourier to the earlier considered polynomial

$$h(x) = x^5 + 2x^4 - 5x^3 + 8x^2 - 7x - 3$$

The number of variations of sign in the sequence of coefficients is three, and so by Descartes' theorem, $h(x)$ can have three positive roots or one. On the other hand, $h(x)$ has no zero coefficients, but since the sequence of coefficients has two preservations of sign, $h(x)$ either has two negative roots or none. We compare with the results obtained earlier with the aid of the graph and see that two is the exact number of negative roots of our polynomial.

To determine exactly the number of positive roots, use the Budan-Fourier theorem, applying it to the interval $(1, \infty)$, since in Sec. 39 it was demonstrated that 1 serves as a lower bound to the positive roots of the polynomial $h(x)$. The successive derivatives of $h(x)$ were also written out in Sec. 39. Let us find their signs for $x = 1$ and $x = \infty$:

	$h(x)$	$h'(x)$	$h''(x)$	$h'''(x)$	$h^{IV}(x)$	$h^{V}(x)$	Number of variations in sign
$x=1$	−	+	+	+	+	+	1
$x=\infty$	+	+	+	+	+	+	0

From this it follows that when x moves from 1 to ∞ the sequence of derivatives loses one change of sign, and so $h(x)$ has exactly one positive root.

In connection with this example, it should be noted that, generally speaking, when seeking the number of real roots of a polynomial it is best to begin by constructing a graph and applying the theorems of Descartes and Budan-Fourier, and then only in extreme cases to go on to construct a Sturm sequence.

The Descartes theorem admits of a certain refinement in the special case when we know beforehand that all the roots of the polynomial are real, as for instance in the case of the characteristic polynomial of a symmetric matrix. Namely,

If all the roots of a polynomial $f(x)$ are real, and the constant term is nonzero, then the number k_1 of positive roots of the polynomial is equal to the number s_1 of variations in sign in the sequence of its coefficients, and the number k_2 of negative roots is equal to the number s_2 of variations in sign in the sequence of coefficients of the polynomial $f(-x)$.

Indeed, under our assumptions,

$$k_1 + k_2 = n \tag{7}$$

where n is the degree of the polynomial $f(x)$, and, by Descartes' theorem,

$$k_1 \leqslant s_1, \quad k_2 \leqslant s_2 \tag{8}$$

We will prove that

$$s_1 + s_2 \leqslant n \tag{9}$$

We will prove it by induction with respect to n, since for $n = 1$, due to $a_0 \neq 0$, $a_1 \neq 0$, only one of the polynomials

$$f(x) = a_0 x + a_1, \quad f(-x) = -a_0 x + a_1$$

has a change of sign; that is, for this case, $s_1 + s_2 = 1$. Let formula (9) be proved for polynomials whose degree is less than n. If

$$f(x) = a_0 x^n + a_{n-l} x^l + \ldots + a_n$$

where $l \leqslant n - 1$, $a_{n-l} \neq 0$. we assume

$$g(x) = a_{n-l} x^l + \ldots + a_n$$

Then

$$f(x) = a_0 x^n + g(x), \quad f(-x) = (-1)^n a_0 x^n + g(-x)$$

If s_1' and s_2' are, respectively, the numbers of variations in sign in the sequences of coefficients of the polynomials $g(x)$ and $g(-x)$, then, by the induction hypothesis (it is clear that $l \geqslant 1$),

$$s_1' + s_2' \leqslant l$$

If $l = n - 1$, then the variation in sign in the first place, i.e., for $f(x)$, between a_0 and $a_1 = a_{n-l}$ will occur only in the case of one of the polynomials $f(x)$, $f(-x)$, and so

$$s_1 + s_2 = s_1' + s_2' + 1 \leqslant l + 1 = n$$

But if $l \leqslant n - 2$, then variations of sign are possible in the first places of each of the polynomials $f(x)$, $f(-x)$; however, in this case as well.

$$s_1 + s_2 \leqslant s_1' + s_2' + 2 \leqslant l + 2 \leqslant (n - 2) + 2 = n$$

Comparing (7), (8) and (9), we see that

$$k_1 = s_1, \quad k_2 = s_2$$

The proof is complete.

42. Approximation of Roots

The methods described in the preceding sections enable us to *isolate* the real roots of a polynomial $f(x)$ with real coefficients, that is to say, they permit indicating for each root the interval

containing it alone. If the interval is small enough, then any number in the interval may be taken as an approximation of the desired root. Thus, after it has been demonstrated by the Sturm method (or any other more efficient method) that there is only one root of the polynomial $f(x)$ between the rational numbers a and b, the problem remains of narrowing this interval so that the new limits a' and b' possess a prescribed number of coincident first decimals. The desired root will thus be computed to the needed accuracy.

There are many methods which permit us to speedily approximate the value of a root with any desired accuracy. We will describe two. They are simple theoretically and general enough so that when used in conjunction they quickly yield results. The methods we are about to describe can be applied not only to polynomials but also to the broader classes of continuous functions.

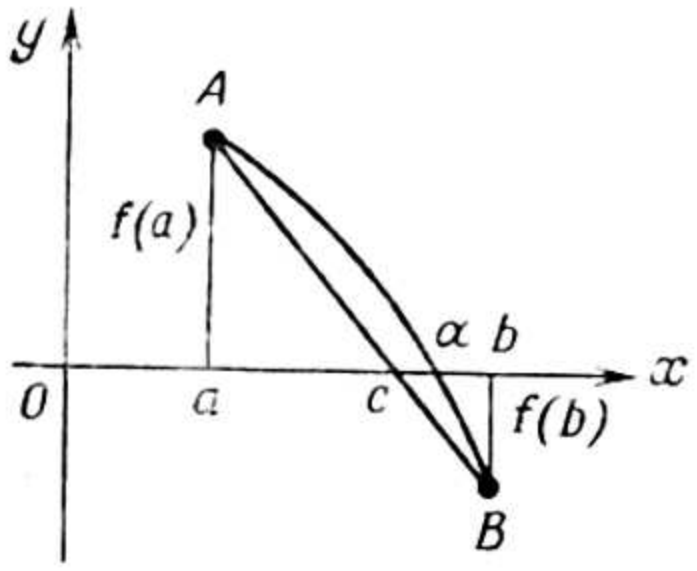

Fig. 10

From here on we assume that α is a simple root of a polynomial $f(x)$, since we can always dispose of multiple roots, and that the root α is isolated between the limits a and b, $a < \alpha < b$; this implies, for one thing, that $f(a)$ and $f(b)$ have different signs.

The method of linear interpolation (also called the method of false position or regula falsi). For an approximate value of the root α we could take, say, the half sum of the limits a and b, $\frac{a+b}{2}$, i.e., the midpoint of the interval from a to b. It is more natural, however, to assume that the root is closer to that endpoint of the interval (a, b) which corresponds to the smallest absolute value of the polynomial. The method of linear interpolation consists in taking a number c for the approximate value of the root α, such that divides the interval (a, b) into parts proportional to the absolute values of the numbers $f(a)$ and $f(b)$; that is,

$$\frac{c-a}{b-c} = -\frac{f(a)}{f(b)}$$

The sign of the right member is minus because $f(a)$ and $f(b)$ have different signs. Whence

$$c = \frac{bf(a) - af(b)}{f(a) - f(b)} \tag{1}$$

Geometrically, as Fig. 10 indicates, the method of linear interpolation consists in replacing the curve $y = f(x)$ on the interval (a, b) by its chord connecting the points $A(a, f(a))$ and $B(b, f(b))$; for the approximate value of the root α we take the abscissa of the point of intersection of the chord and the x-axis.

Newton's method. Since α is a simple root of the polynomial $f(x)$, it follows that $f'(\alpha) \neq 0$. We also assume that $f''(\alpha) \neq 0$ since otherwise the problem would reduce to computing the root of the polynomial $f''(x)$ of lower degree than $f(x)$. We likewise assume that the interval (a, b) does not contain roots of $f(x)$ different from α, neither does it contain any root of the polynomial $f'(x)$ or the polynomial $f''(x)$.* Thus, as follows from mathematical analysis, the curve $y = f(x)$ is either monotonic increasing on the interval

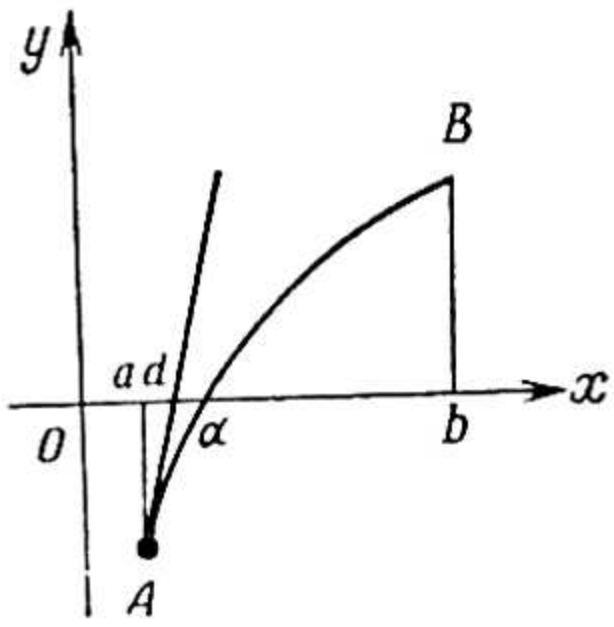

Fig. 11

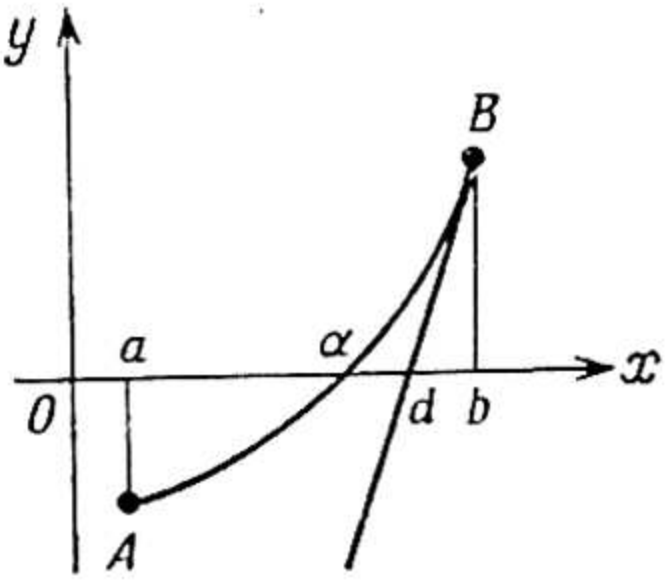

Fig. 12

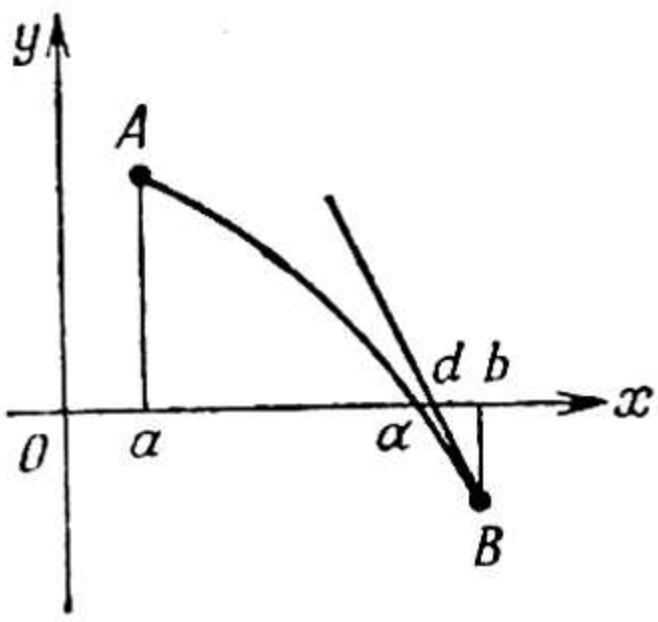

Fig. 13

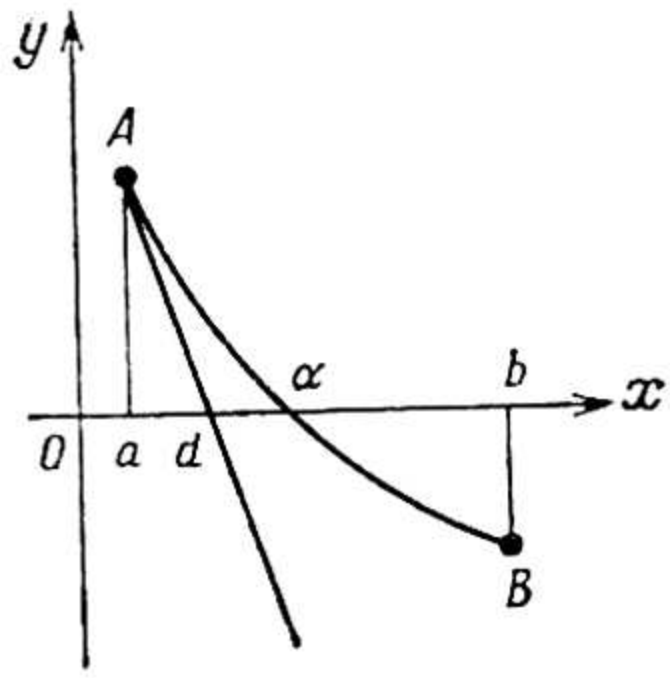

Fig. 14

a, b) or monotonic decreasing; also, it is either convex up at all points of the interval or convex down at all points. Consequently, there are four cases (shown in Figs. 11 to 14) of the location of the curve on the interval (a, b).

Denote by a_0 the endpoint a or b in which the sign of $f(x)$ coincides with the sign of $f''(x)$. Since $f(a)$ and $f(b)$ have different signs, and $f''(x)$ preserves sign throughout the interval (a, b), such an a_0 can be indicated. In the cases given in Figs. 11 and 14, $a_0 = a$, in the other two cases, $a_0 = b$. At the point of the curve $y = f(x)$

* There is usually no difficulty in narrowing the interval so that this condition is satisfied, since the methods given earlier permit establishing the number of roots of polynomials $f'(x)$ and $f''(x)$ in any interval.

with abscissa a_0, that is, at the point with coordinates $(a_0, f(a_0))$, draw a tangent line to this curve and denote by d the abscissa of the intersection point of this tangent with the x-axis. Figs. 11 to 14 show that the number d may be taken as an approximate value of the root α. The Newton method thus consists in replacing the curve $y = f(x)$ on the interval (a, b) by its tangent at one of the endpoints of the interval. The condition imposed on the choice of the point

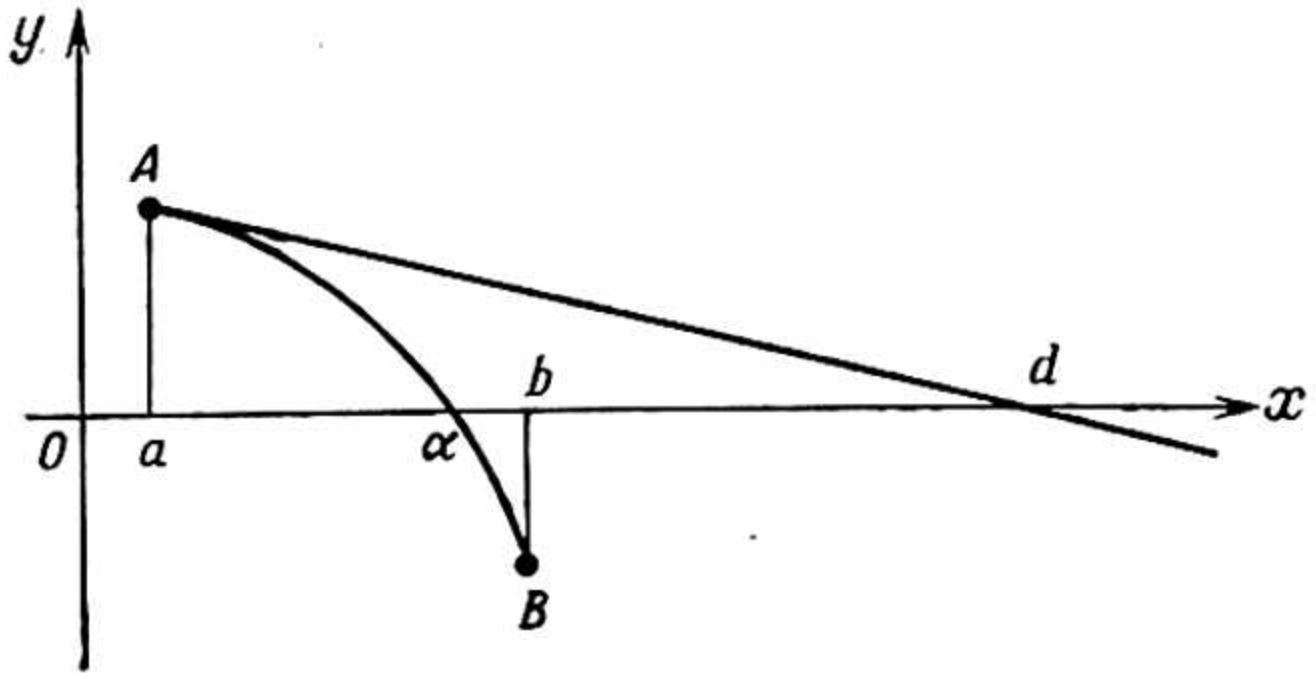

Fig. 15

a_0 is very essential. Fig. 15 shows that if this condition is not observed, the intersection point of the tangent line and x-axis may not at all give an approximation to the desired root.

Let us derive a formula for finding the number d. We recall that the equation of the tangent to the curve $y = f(x)$ at the point $(a_0, f(a_0))$ may be written as

$$y - f(a_0) = f'(a_0)(x - a_0)$$

Substituting the coordinates $(d, 0)$ of the point of intersection of the tangent line with the x-axis, we get

$$-f(a_0) = f'(a_0)(d - a_0)$$

whence

$$d = a_0 - \frac{f(a_0)}{f'(a_0)} \qquad (2)$$

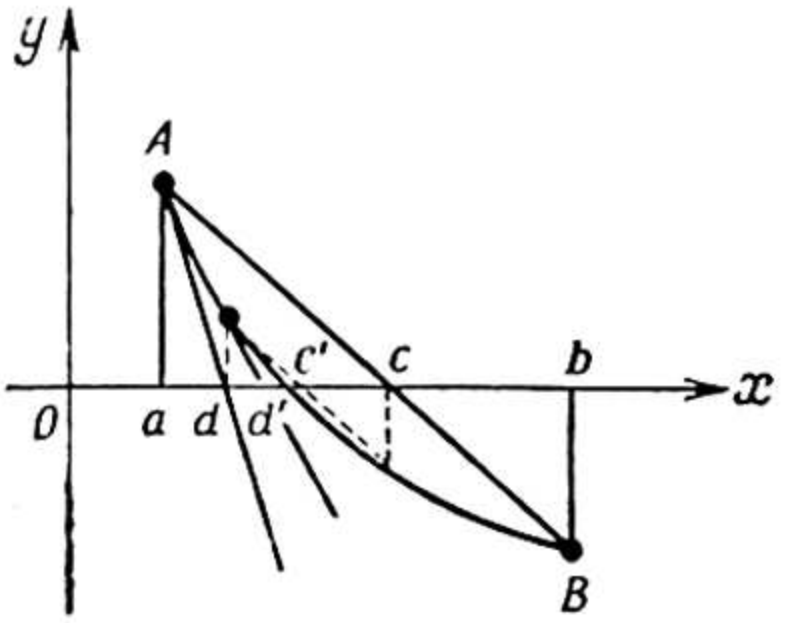

Fig. 16

If in Figs. 11-14 the reader connects A and B by chords, he will see that in all cases *the methods of linear interpolation and of Newton yield approximations to the true value of the root α from different sides.* It is therefore advisable, if the interval (a, b) is such as required by Newton's method, to combine the two methods. In this way we obtain much closer endpoints c and d for the root α. If the accuracy of the approximation is not sufficient, apply both methods (see

Fig. 16) once again to the interval, and so on. We can demonstrate that this process does indeed permit computing the root α to any desired accuracy.

Let us apply these methods to the polynomial

$$h(x) = x^5 + 2x^4 - 5x^3 + 8x^2 - 7x - 3$$

which we dealt with in preceding sections.

We know that this polynomial has a simple root α_1 lying between 1 and 2. We can say right off that these limits are too broad for the methods of linear interpolation and of Newton, used only once each, to yield a decent result. However, let us employ them so as to have one example that does not require involved computations.

As we saw in Sec. 41, for $x = 1$ the derivatives $h'(x)$, $h''(x)$, ..., $h^{V}(x)$ receive positive values. This implies, on the basis of the results of Sec. 39, that the value $x = 1$ serves as an upper bound of the positive roots for $h'(x)$ and also for $h''(x)$. Hence, the interval (1, 2) does not contain any roots of these derivatives and so we can apply the Newton method. Besides, $h''(x)$ is positive everywhere in the interval, and since

$$h(1) = -4, \quad h(2) = 39$$

we have to take $a_0 = 2$. Seeing that $h'(2) = 109$, we get, by formula (2),

$$d = 2 - \frac{39}{109} = \frac{179}{109} = 1.64\ldots$$

On the other hand, formula (1) yields

$$c = \frac{2 \cdot (-4) - 1 \cdot 39}{-4 - 39} = \frac{47}{43} = 1.09\ldots$$

and, consequently, the root α_1 lies within the interval

$$1.09 < \alpha_1 < 1.65$$

This narrowing of the interval that we obtained is too slight to consider the result satisfactory. We could of course apply our methods to the new interval, but it is more advisable from the very beginning to find a sufficiently small interval for α_1, say to within 0.1 or even 0.01, and only then apply the methods. Quite naturally, this at once makes all the computations very cumbersome, but in the solution of concrete problems requiring exact knowledge of the roots of a polynomial, this has to be done.

Let us return to our polynomial $h(x)$ and its root α_1; note that all values of the polynomials given below are computed by the Horner method. Since

$$h(1.3) = -0.13987, \quad h(1.31) = 0.0662923851$$

it follows that

$$1.3 < \alpha_1 < 1.31$$

that is, we have the value of the root α_1 to an accuracy of 0.01. Now let us apply the method of linear interpolation to the new interval:

$$c = \frac{1.31 \cdot (-0.13987) - 1.3 \cdot 0.0662923851}{-0.13987 - 0.0662923851} = \frac{0.26940980063}{0.2061623851} = 1.30678\ldots$$

We also apply Newton's method to this interval, setting $a_0 = 1.31$. Since

$$h'(1.31) = 20.92822405$$

it follows that

$$d = 1.31 - \frac{0.0662923851}{20.92822405} = \frac{27.3496811204}{20.92822405} = 1.30683\ldots$$

Thus,

$$1.30678 < \alpha_1 < 1.30684$$

and therefore, setting $\alpha_1 = 1.30681$, we have an error of less than 0.00003.

We have not yet shown that the foregoing methods actually permit computing a root to any desired accuracy, that is to say we have not proved the convergence of these methods. Let us do so at least with respect to Newton's method.

As above, let the simple root α of the polynomial $f(x)$ lie in the interval (a, b) chosen as required by the Newton method. For one thing, this implies the existence of positive numbers A and B such that everywhere on the interval (a, b),

$$|f'(x)| > A, \quad |f''(x)| < B \tag{3}$$

We introduce the notation

$$C = \frac{B}{2A}$$

and assume that

$$C(b - a) < 1 \tag{4}$$

To fulfil this inequality it may be necessary to replace the interval (a, b) of the root α by a smaller one; but this will not affect the validity of inequalities (3). Let a_0 be the endpoint of the interval (a, b) at which Newton's method is to be applied. On the basis of formula (2) we get a succession of approximate values of the root α: $a_1, a_2, \ldots$ $\ldots, a_k, \ldots$, lying in the interval (a, b) and related by the equalities

$$a_k = a_{k-1} - \frac{f(a_{k-1})}{f'(a_{k-1})}, \quad k = 1, 2, \ldots$$

Let

$$\alpha = a_k + h_k, \quad k = 0, 1, 2, \ldots \tag{6}$$

Then

$$0 = f(\alpha) = f(a_k) + h_k f'(a_k) + \frac{h_k^2}{2} f''(a_k + \theta h_k)$$

where $0 < \theta < 1$. Since $f'(a_k) \neq 0$ due to the condition imposed on the interval (a, b), we get, taking into account (5) and (6),

$$-\frac{h_k^2}{2}\frac{f''(a_k+\theta h_k)}{f'(a_k)} = h_k + \frac{f(a_k)}{f'(a_k)} = \alpha - \left(a_k - \frac{f(a_k)}{f'(a_k)}\right) = \alpha - a_{k+1} = h_{k+1}$$

Whence

$$|h_{k+1}| = h_k^2 \left|\frac{f''(a_k+\theta h_k)}{2f'(a_k)}\right| < h_k^2 \frac{B}{2A} = Ch_k^2, \quad k = 0, 1, 2, \ldots$$

Thus

$$|h_{k+1}| < Ch_k^2 < C^3 h_{k-1}^4 < C^7 h_{k-2}^8 < \ldots < C^{2^{k+1}-1} h_0^{2^{k+1}}$$

or, since $|h_0| = |\alpha - a_0| < b - a$,

$$|h_{k+1}| < C^{-1}[C(b-a)]^{2^{k+1}}, \quad k = 0, 1, 2, \ldots \tag{7}$$

Whence, because of condition (4), it follows that *the difference h_k between the root α and its approximate value a_k obtained by successive application of the Newton method tends to zero with increasing k.* The proof is complete.

Note that (7) gives an *estimate of the error* for the $(k+1)$th step; this is essential if the Newton method is used by itself and not in conjunction with the method of linear interpolation.

Texts dealing with the theory of approximations give procedures with better arranged computations (that simplify their use) than those we have given. Such courses also describe many other methods for approximating roots. These include the *method of Lobachevsky* (sometimes erroneously called the Graeffe method). This method enables one to find at once the approximate values of all roots, including complex roots, and does not require a preliminary isolation of the roots. However, the computations are extremely unwieldy. Underlying this method is the theory of symmetric polynomials, which we describe in Chapter 11 below.

CHAPTER 10

FIELDS AND POLYNOMIALS

43. Number Rings and Fields

In the earlier parts of this book we have frequently been in a position where we investigated complex numbers or only real numbers with the proviso that the results obtained hold true if we restrict ourselves to the real numbers (or, correspondingly, that they carry over word-for-word to the case of any complex numbers). As a rule, in all these cases it might be noted that the theory would hold true completely if we confined ourselves solely within the scope of the rational numbers. The time has now come to indicate the reasons for this parallelism and thus enable us to present the material (which follows) in its natural generality, that is to say, in accepted algebraic language. To do this, we introduce the concept of a *field*, and also the broader concept (which plays a subsidiary role in our course) of a *ring*.

Evidently, the systems of all complex, real and rational numbers, like the system of all integers, *have one property in common: they are all closed not only under addition and multiplication, but under subtraction as well.* This property of the enumerated number systems distinguishes them, say, from the system of positive integers or positive real numbers.

Any system of numbers, complex or (in the particular case) real, containing a sum, a difference and a product of any two of its numbers is termed a *number ring*. Thus, the systems of all integers, and of rational, real and complex numbers are number rings. On the other hand, no system of positive numbers is a ring since if a and b are two different numbers, then either $a - b$, or $b - a$ is negative. Neither is a system of negative numbers a ring because the product of two negative numbers is positive.

The four examples given above do not by any means exhaust the range of number rings. A few more instances will now be given; each time it is left to the reader to verify that the number system is indeed a ring.

The even numbers form a ring; generally, for any natural number n the collection of integers exactly divisible by n is a ring. The odd numbers do not constitute a ring since the sum of two odd numbers is an even number.

Another instance of a ring is the collection of rational numbers whose denominators, in lowest terms, are powers of 2. This collection includes, for example, all integers, since when simplified their denominators are 1, that is, two to the power zero. In this example, in place of 2 we can of course take any prime number p. Generally, taking any (finite or infinite) set of prime numbers and considering the system of rational numbers whose simplified denominators are divisible only by primes belonging to the given set, we again get a ring. On the other hand, the collection of rational numbers whose simplified denominators are not divisible by the square of any prime will not be a ring, since the indicated property of the numbers is not preserved in their multiplication.

Let us now examine number rings that do not lie entirely in the ring of rational numbers. A collection of numbers of the form

$$a + b\sqrt{2} \tag{1}$$

where a and b are any rational numbers, is a ring; in particular, this ring includes all rational numbers (for $b = 0$) and also the number $\sqrt{2}$ itself (for $a = 0$, $b = 1$). We would also have obtained a ring if we had confined ourselves to numbers of the form (1) with integral coefficients a, b. In these examples, we could of course have taken $\sqrt{3}$ or $\sqrt{5}$, etc. in place of $\sqrt{2}$.

The system of numbers of the form

$$a + b\sqrt[3]{2} \tag{2}$$

with rational (or only integral) coefficients a, b is not a ring because the product of $\sqrt[3]{2}$ by itself cannot, as can easily be checked, be written as (2).* However, the system of numbers of the form

$$a + b\sqrt[3]{2} + c\sqrt[3]{4} \tag{3}$$

---

* Indeed, let

$$\sqrt[3]{4} = a + b\sqrt[3]{2} \tag{2'}$$

where the numbers a and b are rational. Multiplying both sides of this equation by $\sqrt[3]{2}$, we get

$$2 = a\sqrt[3]{2} + b\sqrt[3]{4}$$

Substituting the expression (2′) for $\sqrt[3]{4}$, we arrive (after some obvious manipulations) at the equation

$$(a + b^2)\sqrt[3]{2} = 2 - ab \tag{2''}$$

with arbitrary rational coefficients a, b, c, is a ring, and this is also true if we confine ourselves to the case of integral coefficients.

Let us now consider all real numbers obtainable by applying several times the operations of addition, multiplication and subtraction to the familiar number pi (π) and any rational numbers. These will be numbers that can be written as

$$a_0 + a_1\pi + a_2\pi^2 + \ldots + a_n\pi^n \qquad (4)$$

where a_0, a_1, a_2, . . ., a_n are rational numbers, $n \geqslant 0$. Note that no number can have two distinct notations of the type (4), for otherwise, by taking their difference, we would find that the number π satisfies some equation with rational coefficients; now methods of mathematical analysis tell us that actually π cannot satisfy any equation with rational coefficients, which is to say that π is transcendental. Incidentally, even without taking advantage of this result, that is, assuming that the notation of a number in the form (4) is unique, we can show that numbers like (4) constitute a ring.

Another ring is the collection of numbers obtained from π and rational numbers via operations of addition, multiplication, subtraction and division applied several times. To prove this, there is no need to seek a particularly suitable notation for these numbers (though it may possibly be found). If the numbers α and β are obtained from π and some rational numbers by the indicated operations, then quite naturally it will be true of the numbers $\alpha + \beta$, $\alpha - \beta$, $\alpha\beta$ and also (for $\beta \neq 0$) of the number $\frac{\alpha}{\beta}$.

Finally, if we take the collection of complex numbers $a + bi$ with arbitrary rational a, b, we get a ring; this will also be true if we confine ourselves to integral coefficients a, b.

The examples given above do not give a full picture of the great diversity of number rings. But we will not now continue the list of examples and will examine one special and very important type of number ring. We of course know that in the systems of rational, real, and complex numbers, division (except by zero) is unlimited, whereas these number systems are not closed under division of integers. Up to now we paid but slight attention to this difference. Actually, it is very essential and brings us to the following definition.

A number ring is called a *number field* if it contains the quotient of any two of its numbers (the divisor is of course assumed to be

---

If $a + b^2 \neq 0$, then

$$\sqrt[3]{2} = \frac{2 - ab}{a + b^2}$$

which is impossible since the number on the right is rational. But if $a + b^2 = 0$, then, by (2″) we have $2 - ab = 0$. From these two equations follows the fact that $b^3 = -2$ which is again out of the question since the number b is rational.

different from zero). We can thus speak of the field of rational numbers, the field of real numbers, the field of complex numbers, whereas the ring of integers does not constitute a field.

Some of the earlier considered examples of number rings are actually fields. To begin with, notice that there do not exist number fields different from the field of rational numbers and entirely embedded in it (we do not consider the system of zero alone to be a field).

Even the following more general assertion holds true.

The field of rational numbers lies entirely within any number field.

Indeed, let there be some number field, call it P. If a is any number of P different from zero, then P also contains the quotient of the division of a by itself, that is, the number 1. Adding unity to itself several times, we find that all the natural numbers lie in the field P. On the other hand, P must also contain the difference $a - a$, which is the number 0, and so P contains the result of subtracting any natural number from zero, which is to say, any negative integer. Finally, P contains the quotients of all integers, or, generally, all rational numbers.

The field of complex numbers contains many different fields, and the field of rational numbers is only the smallest in it. Thus, the ring, considered above, of numbers like

$$a + b\sqrt{2} \tag{5}$$

with arbitrary rational (and not only integral) coefficients a, b is a field. To see this, consider the quotient of two numbers of the form (5), $a + b\sqrt{2}$ and $c + d\sqrt{2}$; consider the second number to be different from zero, hence the number $c - d\sqrt{2}$ is also nonzero, and so

$$\frac{a+b\sqrt{2}}{c+d\sqrt{2}} = \frac{(a+b\sqrt{2})(c-d\sqrt{2})}{(c+d\sqrt{2})(c-d\sqrt{2})} = \frac{ac-2bd}{c^2-2d^2} + \frac{bc-ad}{c^2-2d^2}\sqrt{2}$$

We again have a number of type (5), and the coefficients remain rational. In this example, the number $\sqrt{2}$ may naturally be replaced by the square root of any rational number whose square root cannot be taken in the field of rational numbers. Thus, the field is made up of numbers of the form $a + bi$ with rational a, b.

44. Rings

In various divisions of mathematics, and also in applications of mathematics to science and engineering, one often has to perform algebraic operations with a variety of nonnumerical entities. The preceding chapters of this book afford numerous examples: the multiplication and addition of matrices, the addition of vectors, operations involving polynomials, operations on linear transforma-

tions. The general definition of an *algebraic operation* that is satisfied by the operations of addition and multiplication in number rings, and also by operations in the indicated examples, consists in the following.

A set M is given that consists either of numbers or of objects of a geometrical nature, or, generally, of certain things which we will call *elements* of the set. We say that *an algebraic operation is defined on the set* M if a law is indicated according to which any two of elements a, b of the set are uniquely associated with some third element c which also belongs to M. This operation may be called *addition*, then c is termed the *sum* of the elements a and b and is denoted by the symbol $c = a + b$; the operation may be called *multiplication*, then c is the *product* of the elements a and b, $c = ab$; finally, it may be that a new terminology and symbolism will be introduced for an operation defined on M.

In each of the number rings are defined two independent operations, addition and multiplication. Subtraction and division will not be considered new operations since they are the inverses of addition and multiplication if we accept the following general definition of an *inverse operation*.

Let an algebraic operation, say addition, be defined on the set M. Then we say that *there is an inverse operation* called subtraction if for any two of elements a, b of M there exists in M an element d that is unique and that satisfies the equation $b + d = a$. The element d is then called the *difference* between the elements a and b and is denoted by the symbol $d = a - b$.

It is obvious that in number fields, both addition and multiplication have inverses. True, there is one restriction relative to multiplication: the divisor must be different from zero. Now in number rings that are not fields (say, in the ring of integers), only addition has an inverse operation.

On the other hand, in the system of all polynomials in the unknown x, whose coefficients belong to a fixed number field P, there are also defined two operations: addition and multiplication, addition having the inverse operation of subtraction.

As we know, both in number rings and in the system of polynomials, the operations of addition and multiplication have the following properties (a, b, c are arbitrary numbers in the number ring under consideration or are arbitrary polynomials in the system at hand):

I. Addition is commutative: $a + b = b + a$.

II. Addition is associative: $a + (b + c) = (a + b) + c$.

III. Multiplication is commutative: $ab = ba$.

IV. Multiplication is associative: $a\ (bc) = (ab)\ c$.

V. Multiplication is distributive over addition:

$$(a + b)\ c = ac + bc.$$

We are now prepared for a general definition of the concept of a ring, one of the most important concepts of algebra.

A set R is called a *ring* if on it are defined two operations: addition and multiplication, both commutative and associative and also related by the distributive law, addition having the inverse operation of subtraction.

We thus have the following examples of rings: number rings, rings of polynomials in the unknown x with coefficients from the given number field or even from the given number ring. Let us take one more example which illustrates the breadth of the ring concept.

The course of mathematical analysis begins with a definition of a function of a real variable x. Let us consider the collection of functions that are defined for all real values of x and that take on real values; let us define algebraic operations in this collection as follows: the *sum* of two functions $f(x)$ and $g(x)$ is a function whose value for any $x = x_0$ is equal to the sum of the values of the given functions, that is, it is equal to $f(x_0) + g(x_0)$. The *product* of these functions is a function whose value for every $x = x_0$ is equal to the product $f(x_0) \cdot g(x_0)$. For any two functions of the collection at hand, there obviously exists a sum and a product. The truth of Properties I to V is verified without any difficulty. The addition and multiplication of functions reduce to the addition and multiplication of their values for any x, which is to say, they reduce to operations on real numbers, for which the Properties I to V hold. Finally, taking for the *difference* of the functions $f(x)$ and $g(x)$ a function whose value for any $x = x_0$ is equal to the difference $f(x_0) - g(x_0)$, we arrive at the operation of subtraction, the inverse of addition. This proves that *the collection of functions defined for all real x becomes a ring as soon as we introduce (as indicated above) the operations of addition and multiplication.*

Other examples of rings of functions may be obtained by considering otherwise defined functions, while preserving the definitions of operations on functions given above: functions defined, say, only for positive values of the unknown x, or functions defined for values of x over the interval $[0, 1]$. Generally, a system of all the functions having some given domain of definition is a ring. We could also obtain rings by regarding not all the functions defined in a given domain, but only the continuous functions studied in the course of mathematical analysis. On the other hand, we could consider the complex functions of a complex variable. Generally speaking, there are very many different function rings, just as there are a great diversity of number rings.

Let us now establish some of the more elementary properties of rings which follow directly from the definition of a ring. For numbers, these properties are quite ordinary, but the reader will

possibly be surprised to find that they are consequences only of the Conditions I to V and the existence of unique subtraction.

First a few remarks regarding the significance of Conditions I to V. The role of the *commutative laws* is evident enough. The significance of the *associative laws* consists in the following: the definition of an algebraic operation speaks of the sum or product of only two elements. If we attempt to define the product of, say, three elements a, b, c, then we have the difficulty that the products au and vc, where $bc = u$, $ab = v$, may, generally speaking, not coincide, that is, $a\,(bc) \neq (ab)\,c$. The associative law demands that these products be equal to one and the same element of the ring: it is natural to take this element for the product abc, written without brackets. What is more, *the associative law permits defining uniquely the product (sum) of any finite number of elements of the ring*; that is, it permits proving that a product of any n elements is independent of the original arrangement of parentheses.

Let us prove this assertion by means of induction with respect to the number n. It has already been proved for $n = 3$, and so let us assume $n > 3$ and also that for all numbers less than n our assertion has already been proved. Let there be elements $a_1, a_2, \ldots, a_n$ and let there be some kind of arrangement of parentheses in this system indicating the order in which multiplication is to be performed. The last step will be the multiplication of the product of the first k elements $a_1 a_2 \ldots a_k$ (where $1 \leqslant k \leqslant n-1$) by the product $a_{k+1} a_{k+2} \ldots a_n$. Since these products consist of a smaller, than n, number of factors and for this reason, by hypothesis, are uniquely defined, it remains to prove the following equation for any k and l:

$$(a_1 a_2 \ldots a_k)\,(a_{k+1} a_{k+2} \ldots a_n) = (a_1 a_2 \ldots a_l)\,(a_{l+1} a_{l+2} \ldots a_n)$$

To do this, it will suffice to consider the case $l = k + 1$. But then, setting

$$a_1 a_2 \ldots a_k = b, \quad a_{k+2} a_{k+3} \ldots a_n = c$$

we get, by the associative law,

$$b\,(a_{k+1} c) = (b a_{k+1})\,c$$

Which proves our assertion.

We can speak, in particular, about the product of n equal elements; that is, we can introduce the concept of a *power*, a^n, of the element a with positive integral exponent n. It is easy to verify that all the ordinary rules for operating with exponents hold true in any ring. Analogously, the associative law of addition leads to the concept of a *multiple*, na, of the element a by a positive integral coefficient n.

The *distributive law*, that is, the usual rule for removing brackets, is the only requirement in the definition of a ring that connects addition and multiplication; it is only through this law that the joint study of the two indicated operations yields more than could be obtained in their separate study. The statement of the distributive law involves the sum of only two terms. However, it can readily be proved that the equality

$$(a_1 + a_2 + \ldots + a_k)\, b = a_1 b + a_2 b + \ldots + a_k b$$

holds for any k and that the general rule of multiplication of a sum by a sum is true.

Also, *the distributive law holds true in any ring for a difference as well.* Indeed, by the definition of a difference, the element $a - b$ satisfies the equality

$$b + (a - b) = a$$

Multiplying both sides of this equation by c and applying the distributive law to the left member, we get

$$bc + (a - b)\, c = ac$$

Element $(a - b)\, c$ is consequently the difference of the elements ac and bc:

$$(a - b)\, c = ac - bc$$

Very important properties of rings follow from the existence of subtraction. If a is an arbitrary element of a ring R, then the difference $a - a$ will be some quite definite element of the ring. Its role is similar to that of zero in number rings, but, by definition, it may depend on the choice of the element a and therefore we will provisionally denote it by 0_a.

We will prove that actually the elements 0_a are equal for all a. Indeed, if b is some other arbitrary element of a ring R, then by adding the element 0_a to both sides of the equation

$$a + (b - a) = b$$

and using the equation $0_a + a = a$, we get

$$0_a + b = 0_a + a + (b - a) = a + (b - a) = b$$

Thus, $0_a = b - b = 0_b$.

We have proved that *any ring R possesses a uniquely defined element which when added to element a of that ring is a.* We call this element the *zero element* of the ring R and we denote it by 0. We believe there is no real danger of confusing it with the number zero. Thus,

$$a + 0 = a \quad \text{for all} \quad a \text{ in } R$$

To continue, *in any ring there exists for any element a uniquely defined inverse element* $-a$ *which satisfies the equation*

$$a + (-a) = 0$$

Namely, this element is the difference $0 - a$; the uniqueness follows from the uniqueness of subtraction. It is obvious that $-(-a) = a$. The difference $b - a$ of any two elements of a ring may now be written as

$$b - a = b + (-a)$$

Indeed,

$$[b + (-a)] + a = b + [(-a) + a] = b + 0 = b$$

For any element a of the ring and for any positive integer n *we have the equality*

$$n(-a) = -(na)$$

And true enough, grouping the terms we get

$$na + n(-a) = n[a + (-a)] = n \cdot 0 = 0$$

We are now in a position to define *negative multiples* of an element of a ring: if $n > 0$, then the equal elements $n(-a)$ and $-(na)$ will be denoted by $(-n)\,a$. Let us finally agree to use the term *zero multiple* $0 \cdot a$ of any element a for the zero element of the ring under consideration.

We have defined zero solely by means of the operation of addition and its inverse, that is to say, without using multiplication. However, in the case of numbers, the number zero has a characteristic and very important property with respect to multiplication too. It turns out that this property is possessed by the zero element of every ring: *in any ring the product of any element by zero is zero.* The proof rests directly on the distributive law: if a is an arbitrary element of a ring R, then no matter what the auxiliary element x of this ring, we get

$$a \cdot 0 = a(x - x) = ax - ax = 0$$

Using this property of zero, we can prove that *in any ring the following equality holds for any elements* a, b:

$$(-a)\,b = -ab$$

True enough,

$$ab + (-a)\,b = [a + (-a)]\,b = 0 \cdot b = 0$$

Which implies that the familiar yet somewhat mysterious rule for the multiplication of negative numbers, "two negatives make a positive", also follows from the definition of a ring, that is, *in*

any ring we have the equality

$$(-a)(-b) = ab$$

Indeed,

$$(-a)(-b) = -[a(-b)] = -(-ab) = ab$$

The reader will not find any difficulty now in proving that in any ring all the rules for operating with the multiples of any number hold true for the multiples (including negative multiples) of any element.

Thus, the algebraic operations in an arbitrary ring have many of the familiar properties of operations on numbers. However, one should not think that every property of addition and multiplication of numbers is preserved in any ring. For instance, the multiplication of numbers has a property which is the converse of the one considered above: if a product of two numbers is equal to zero, then at least one of the factors is zero. This property cannot be carried over to all rings. In some rings we can find pairs of nonzero elements whose product is equal to zero, that is, $a \neq 0$, $b \neq 0$, but $ab = 0$; elements a and b with this property are called *divisors of zero*.

Naturally, among the number rings one cannot find any instances of rings with zero divisors. Likewise there are no zero divisors among the rings of polynomials with numerical coefficients. However, many function rings have zero divisors. First of all, let us note that in any function ring a zero is a function equal to zero for all values of the variable x. Let us now construct the following functions $f(x)$ and $g(x)$ defined for all real values of x:

$$f(x) = 0 \quad \text{for} \quad x \leqslant 0, \quad f(x) = x \quad \text{for} \quad x > 0,$$

$$g(x) = x \quad \text{for} \quad x \leqslant 0, \quad g(x) = 0 \quad \text{for} \quad x > 0$$

Both functions are nonzero since their values are not equal to zero for all values of x, but the product of these functions is zero.

Not all the requirements I to V that enter into the definition of a ring are necessary in equal measure. The development of mathematics shows that whereas the properties I and II of addition and the distributive law V occur in all applications, the inclusion of the multiplication properties III and IV in the definition of a ring is too confining and narrows the sphere of application of this concept. Thus, when the set of square matrices of order n with real elements is regarded with the operations of addition and multiplication of matrices, it satisfies all the requirements in the definition of a ring, with the exception of the commutative law of multiplication. Noncommutative multiplications are encountered so often and in such important instances that the term "ring" is now usually interpreted to mean a *noncommutative ring* (or, more precisely, a not necessarily commutative ring, in the sense of possible noncommuta-

tivity of multiplication), and the special type of ring in which requirement III is fulfilled is termed a *commutative ring.*

There has also been much interest recently in rings with nonassociative multiplication and the general theory of rings under construction is now a theory of nonassociative (that is to say, not necessarily associative) rings. An elementary instance of such a ring is the set of vectors of three-dimensional Euclidean space under the operations of the addition and (taken from the course of analytic geometry) the vector multiplication of vectors.

45. Fields

In the set of number rings, we singled out and gave the name number fields to those rings which admit division (except by zero). It is natural to do this in the general case as well. First note that *no ring admits division by zero* in virtue of the above-proved property of zero under multiplication: to divide an element a by zero means to find, in that ring, an element x such that $0 \cdot x = a$, which for $a \neq 0$ is impossible, since the left-hand side is equal to zero.

Let us introduce the following definition.

A ring P is termed a *field* if it consists of more than zero alone and if division can be performed uniquely in all cases except division by zero; that is to say, for any elements a and b in P, $b \neq 0$, there is in P a unique element q which satisfies the equality $bq = a$. The element q is called the *quotient* of the elements a and b and is denoted by the symbol $q = \frac{a}{b}$.*

Quite naturally, all number fields are instances of fields. A ring of polynomials in the unknown x with real coefficients and, generally, with coefficients taken from some number field, is not a field. The division with a remainder that polynomials have differs of course from exact division, which is assumed in the definition of a field. On the other hand, it is easy to see that *the set of all fractional rational functions with real coefficients* (see Sec. 25) *will be a field* containing the ring of polynomials, just like the field of rational numbers contains the ring of integers.

We could point to certain other instances of fields within the ring of functions, but instead we will examine examples of quite a different sort.

All the number rings, and in general all the rings we have considered so far, contain infinitely many elements. There are, however,

---

* The uniqueness of division in a field, just like the assumed uniqueness of subtraction in the definition of a ring, can actually be proved without any difficulty by means of the requirements that enter into the definition of a field (or ring).

rings and even fields consisting only of a finite number of elements. The simplest examples of *finite rings* and *finite fields* which are essential objects in the theory of numbers are constructed in the following manner.

Take any natural number n different from 1. The integers a and b are called *congruent modulo* n,

$$a \equiv b \pmod{n}$$

if these numbers yield the same remainder when divided by n, that is to say, if their difference is exactly divisible by n. The entire ring of integers is separated into n mutually exclusive (nonintersecting) classes

$$C_0, C_1, \ldots, C_{n-1} \tag{1}$$

of numbers congruent modulo n, the class C_k, $k = 0, 1, \ldots, n-1$, consists of numbers which yield, upon division by n, the remainder k. It turns out that it is possible, in a very natural way, to define the addition and multiplication of these classes.

For this purpose, let us take any (not necessarily distinct) classes C_k and C_l from the system (1). Adding any number of class C_k to any number of class C_l, we obtain numbers lying in one very definite class, namely, in the class C_{k+l}, if $k + l < n$, or in the class C_{k+l-n} if $k + l \geqslant n$. This leads to the following definition of the *addition of classes*:

$$\begin{aligned} C_k + C_l &= C_{k+l} \quad \text{for} \quad k + l < n, \\ C_k + C_l &= C_{k+l-n} \quad \text{for} \quad k + l \geqslant n \end{aligned} \tag{2}$$

On the other hand, multiplying any number of class C_k by any number of class C_l we get numbers lying in a definite class, namely the class C_r, where r is the remainder left after dividing the product kl by n. We thus have the following definition of the *multiplication of classes*:

$$C_k \cdot C_l = C_r, \quad \text{where} \quad kl = nq + r, \qquad 0 \leqslant r < n \tag{3}$$

The system (1) *of classes of integers congruent modulo n is a ring with respect to the operations defined by the conditions* (2) *and* (3). Indeed, the requirements I-V are readily seen to be valid from the definition of a ring, but this validity also follows from the truth of these requirements in the ring of integers and from the relationship, indicated above, between operations on integers and operations on classes. Zero is obviously the class C_0 consisting of numbers exactly divisible by n. The class opposite to C_k, $k = 1, 2, \ldots, n-1$, is the class C_{n-k}. In the system (1) of classes it is thus possible to define subtraction, that is, this system satisfies all the requirements of the definition of a ring. Let us agree to denote the resulting ring by Z_n.

If the number n is a composite number, then the ring Z_n possesses zero divisors and therefore, as will be shown below, it cannot be a field. Indeed, if $n = kl$ where $1 < k < n$, $1 < l < n$, then the classes C_k and C_l are different from the zero class C_0, but by the definition of the multiplication of classes [see (3)], $C_k \cdot C_l = C_0$.

But if the number n is prime, then the ring Z_n is a field.

To see this, let there be classes C_k and C_m, $C_k \neq C_0$, i.e., $1 \leqslant k \leqslant n - 1$. We have to show that it is possible to divide C_m by C_k, or to find a class C_l such that $C_k \cdot C_l = C_m$. If $C_m = C_0$, then $C_l = C_0$ as well. But if $C_m \neq C_0$, then we consider the set of numbers

$$k,\ 2k,\ 3k,\ \ldots,\ (n-1)\,k \tag{4}$$

All these numbers lie outside the zero class C_0, since the product of two natural numbers less than a prime n is not divisible by n. Also, no two numbers sk and tk from (4), $s < t$, can be in one class, for then their difference

$$tk - sk = (t - s)\,k$$

would be divisible by n, which again is in conflict with the primality of the number n. Thus every nonzero class contains exactly one number from the set (4). For instance, in the class C_m there is the number lk, where $1 \leqslant l \leqslant n - 1$, that is, $C_l \cdot C_k = C_m$, and then class C_l will be the desired quotient resulting from the division of C_m by C_k.

We have thus obtained an infinity of distinct finite fields: the field Z_2, consisting of only two elements, and also the fields Z_3, Z_5, Z_7, Z_{11} and so on.

Let us examine some properties of fields that follow from the existence of division. These properties are similar to those of rings based on the existence of subtraction and are demonstrated by the same arguments, and so the proof will be left to the reader.

Every field P has a uniquely defined element whose product by any element a of the field is equal to a. This element, which coincides with equal quotients $\frac{a}{a}$ for all nonzero a is called the *unity* (unit) element of the field P and is denoted by 1. Thus,

$$a \cdot 1 = a \quad \text{for all} \quad a \quad \text{in} \quad P$$

For every nonzero element a, there is, in every field, a unique inverse element a^{-1} which satisfies the equality

$$a \cdot a^{-1} = 1$$

namely, $a^{-1} = \frac{1}{a}$. It is obvious that $(a^{-1})^{-1} = a$. The quotient $\frac{b}{a}$ may now be written in the form

$$\frac{b}{a} = b \cdot a^{-1}$$

For any element a different from zero and for any positive integer n we have the equality

$$(a^{-1})^n = (a^n)^{-1}$$

Denoting these equal elements by a^{-n}, we arrive at *negative powers* of an element of the field for which the ordinary operating rules hold. Let us finally agree that $a^0 = 1$ for all a.

The existence of a unit element is not a characteristic property of fields: the ring of integers, for instance, has a unit element. Yet the example of the ring of even numbers shows that not all rings possess a unit element. On the other hand, *any ring possessing a unit element and an inverse for every nonzero element is a field.* Indeed, in this case for the quotient $\frac{b}{a}$, $a \neq 0$, we have the product ba^{-1}. It is easy to prove the uniqueness of this quotient.

Notice that *no field has zero divisors.* Let $ab = 0$, but $a \neq 0$. Multiplying both sides of the equality by the element a^{-1}, we get $(a^{-1}a)\, b = 1 \cdot b = b$ on the left and $a^{-1} \cdot 0 = 0$ on the right, or $b = 0$. From this it follows that *in any field any equality may be divided by a common nonzero factor.* This is so, since if $ac = bc$ and $c \neq 0$, then $(a - b)\, c = 0$, whence $a - b = 0$, or $a = b$.

From the definition of the quotient $\frac{a}{b}$ (where $b \neq 0$) and from the above-proved possibility of writing it as the product ab^{-1}, it is easy to see that *all the ordinary rules for handling fractions hold true in any field*, namely:

$$\frac{a}{b} = \frac{c}{d} \text{ if and only if } ad = bc,$$

$$\frac{a}{b} \pm \frac{c}{d} = \frac{ad \pm bc}{bd},$$

$$\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd},$$

$$\frac{-a}{b} = -\frac{a}{b}$$

The characteristic of a field. Not all properties of number fields hold true in the case of arbitrary fields. Say, if we take 1 and add 1 to it several times, that is, if we take any positive integral multiple of one, we will never get zero, and, generally, all these multiples (that is, all natural numbers) are distinct. But if we take integral multiples of unity in some finite field, then there will invariably be equal integral multiples, since the field has only a finite number of distinct elements. If all the integral multiples of unity of a field P are distinct elements of P, that is, $k \cdot 1 \neq l \cdot 1$ for $k \neq l$ then we say that the field P has *characteristic zero.* Such for example are all the number fields. But if there exist integers k and l such that $k > l$,

but in P we have the equality $k \cdot 1 = l \cdot 1$, then $(k - l) \cdot 1 = 0$, i.e., there exists in P a positive multiple of unity which is equal to zero. In this case P is called a field of a *finite characteristic*, namely of the *characteristic* p, if p is the first positive coefficient with which the unit element of the field P vanishes. All finite fields are examples of fields of a finite characteristic. Incidentally, there also exist infinite fields having a finite characteristic.

If a field P has a characteristic p, then the number p is prime.

Indeed, from the equality $p = st$, where $s < p$, $t < p$, would follow the equality $(s \cdot 1)(t \cdot 1) = p \cdot 1 = 0$, that is to say, since a field cannot have zero divisors, then either $s \cdot 1 = 0$ or $t \cdot 1 = 0$, which, however, runs counter to the definition of a characteristic as the least positive coefficient which makes the unit element of the field vanish.

If the characteristic of a field P is equal to p, then for any element a of the field we have the equality $pa = 0$. But if the characteristic of the field P is 0 and a is an element of the field, n an integer, then from $a \neq 0$ and $n \neq 0$ it follows that $na \neq 0$.

Indeed, in the first case the element pa (that is, the sum of p terms equal to a) can, by factoring out a, be represented as

$$pa = a(p \cdot 1) = a \cdot 0 = 0$$

In the second case, from the equality $na = 0$, that is, $a(n \cdot 1) = 0$, we would get $n \cdot 1 = 0$, $a \neq 0$; that is, since the characteristic of the field is zero, $n = 0$.

Subfields, extensions. Suppose in the field P a portion of the elements (some set P') is itself a field with respect to the operations defined in P; that is to say, for any two elements a, b in P', the elements (in the field P) $a + b$, ab, $a - b$, and, for $b \neq 0$, $\frac{a}{b}$ belong to P' (the laws I to V will of course hold in P' since they hold in P). Then P' is a *subfield* of the field P, and P is an *extension* of the field P'. Quite naturally, the zero and unity of P will lie in P' as well and will also serve in P' as zero and unity. Thus, the field of rational numbers is a subfield of the field of real numbers; all number fields are subfields of the field of complex numbers.

Let there be given in the field P a subfield P' and an element c exterior to P' and suppose we have a minimum subfield P'' of P which contains both P' and c. There can only be one such minimum subfield, since if P''' were one more subfield with these properties, then the intersection of subfields P'' and P''' (i.e., the collection of elements common to both subfields) would contain P' and the element c and, together with any two of its elements, it would contain their sum (this sum must lie both in P'' and in P''', and so also in their intersection) and likewise their product, difference and quotient; in other words the intersection would itself be a subfield,

but this contradicts the minimality of the subfield P''. We will say that the field P'' *is obtained by adjoining an element c to the field* P'; symbolically, we write $P'' = P'(c)$.

The field $P'(c)$ naturally contains, besides the element c and all the elements of the field P', also all the elements which are derived from them by the operations of addition, multiplication, subtraction and division. By way of illustration, recall the extension (considered in Sec. 43) of the field of rational numbers consisting of numbers of the form $a + b\sqrt{2}$ with rational a, b; this extension results from adjoining the number $\sqrt{2}$ to the field of rational numbers.

46. Isomorphisms of Rings (Fields). The Uniqueness of the Field of Complex Numbers

The concept of an isomorphism plays an important role in the theory of rings. Namely, the rings L and L' are called *isomorphic* if a one-to-one correspondence can be set up between them such that for any elements a, b in L and for the corresponding elements a', b' in L', the sum $a + b$ corresponds to the sum $a' + b'$, and the product ab corresponds to the product $a'b'$.

Suppose an isomorphic correspondence exists between the rings L and L'. *In this correspondence, the zero* 0 *of* L *corresponds to the zero* $0'$ *of* L'. Indeed, suppose the element 0 is associated with an element c' of L'. Take an arbitrary element a of L and the associated element a' of L'. Then to the element $a + 0$ there has to correspond the element $a' + c'$; but $a + 0 = a$, and so $a' + c' = a'$, whence $c' = 0'$. Furthermore, *the element* $-a$ *is associated with the element* $-a'$. Indeed, let the element $-a$ be associated with the element d'. Then to the element $a + (-a) = 0$ there will have to correspond the element $a' + d'$, that is, $a' + d' = 0'$, whence $d' = -a'$. This implies that *to a difference of elements in* L *there corresponds a difference of the corresponding elements of* L'. By similar arguments it may be shown that if the ring L has a unit element, then the image of this element (i.e., the element corresponding to it in L' under the given isomorphism) will be the unit element of the ring L', and if the element a from L has the inverse a^{-1}, then in L' the image of a^{-1} is the inverse element of a'.

This implies that *a ring isomorphic to a field is itself a field.* It is also easy to see that the property of a ring not to have zero divisors also holds in an isomorphic correspondence. Generally speaking, isomorphic rings can differ as to the nature of their elements, but they are identical with respect to their algebraic properties. Any theorem which has been proved relative to some ring will hold true for all rings isomorphic to that ring, provided that the proof does not involve any individual properties of the elements of the ring

but only the properties of the operations. For this reason *we will not consider isomorphic rings or fields to be distinct*; for us they will simply be different copies of one and the same ring or field.

Let us apply this concept to the problem of constructing the field of complex numbers. The construction, given in Sec. 17, of the field of complex numbers was based on the use of points in the plane. This is not the only possible construction. In place of points, we could have taken line segments (vectors) in the plane that emanate from the coordinate origin, and by specifying these vectors via their components a, b on the coordinate axes, we could have defined addition and multiplication of the vectors with the aid of the same formulas (2) and (3) of Sec. 17, as in the case of points in the plane. We could have gone further still and dispensed with geometrical material altogether: noting that points in a plane and also vectors in a plane can be represented by ordered pairs of real numbers (a, b), we could simply take the collection of all such pairs and introduce addition and multiplication via formulas (2) and (3) of that section.

With respect to their algebraic properties, all these fields would be indistinguishable, as witness the following theorem.

All extensions of the field D of real numbers derived by adjoining to D a root of the equation

$$x^2 + 1 = 0 \tag{1}$$

are isomorphic among themselves.

Indeed, suppose we have a field P which is an extension of the field D and contains an element satisfying equation (1). The choice of denoting this element is up to us, and we use the letter i. We thus get the equation $i^2 + 1 = 0$ (whence $i^2 = -1$), where involution and addition are to be understood in the sense of the operations defined in the field P. We now want to find the field $D(i)$ obtained by adjoining the element i to the field D, that is, we wish to find the minimal subfield of the field P containing both D and the element i.

For this purpose, let us examine all the elements α of the field P which can be written in the form

$$\alpha = a + bi \tag{2}$$

where a and b are arbitrary real numbers, and the product of the number b by element i and the sum of the number a and this product are to be understood in the sense of the operations defined in the field P. No element α of P can possess two different representations of that form: from

$$\alpha = a + bi = \bar{a} + \bar{b}i$$

and $b \neq \bar{b}$ there would follow

$$i = \frac{\bar{a} - a}{b - \bar{b}}$$

That is, i would be a real number, but if $b = \bar{b}$, then $a = \bar{a}$. In particular, the elements of P written as (2) include all real numbers (the case $b = 0$) and also the element i (the case $a = 0$, $b = 1$).

We will now show that *the collection of all elements of type* (2) *constitutes a subfield of the field* P. This will then be the desired field $D(i)$. Suppose we have the elements $\alpha = a + bi$ and $\beta = c + di$. Then, using the commutativity and associativity of addition and the distributive law, all of which hold in P, we get

$$\alpha + \beta = (a + bi) + (c + di) = (a + c) + (bi + di)$$

whence

$$\alpha + \beta = (a + c) + (b + d)\, i \tag{3}$$

Thus, this sum again belongs to the set of elements under consideration. Furthermore,

$$-\beta = (-c) + (-d)\, i$$

since, by (3), the equality $\beta + (-\beta) = 0 + 0i = 0$ holds true. Therefore

$$\alpha - \beta = \alpha + (-\beta) = (a - c) + (b - d)\, i \tag{3'}$$

That is to say, this set is also closed under subtraction. Again using properties from I to V, which hold for operations in the field P (see Sec. 44), and relying on the equality $i^2 = -1$, we get

$$\alpha\beta = (a + bi)(c + di) = ac + adi + bci + bdi^2$$

that is,

$$\alpha\beta = (ac - bd) + (ad + bc)\, i \tag{4}$$

Thus the product of any two elements of the type (2) is again an element of this type. Finally, suppose that $\beta \neq 0$, i.e., at least one of the numbers c, d is nonzero. Then we will also have $c - di \neq 0$ and

$$(c + di)(c - di) = c^2 - (di)^2 = c^2 - d^2 i^2 = c^2 + d^2$$

and $c^2 + d^2 \neq 0$. Therefore, using the assertion (stated in the preceding section) that all the ordinary rules of handling fractions hold true in any field, and thus, in particular, that a fraction remains unchanged when the numerator and denominator are multiplied by the same nonzero element, we obtain

$$\frac{\alpha}{\beta} = \frac{a + bi}{c + di} = \frac{(a + bi)(c - di)}{(c + di)(c - di)} = \frac{(ac + bd) + (bc - ad)\, i}{c^2 + d^2}$$

That is to say, the element

$$\frac{\alpha}{\beta} = \frac{ac + bd}{c^2 + d^2} + \frac{bc - ad}{c^2 + d^2}\, i \tag{4'}$$

again has the form (2).

We will now show that *the subfield D (i) which we have derived from the field P is isomorphic to that field of points in a plane that was constructed in* Sec. 17. Associating with the element $a + bi$ of the field $D(i)$ a point (a, b), we obtain [due to the uniqueness—just proved—of the notation (2) for elements of the field $D(i)$] a one-to-one correspondence between the elements of this field and all the points in the plane. In this correspondence, the real number a is associated with the point $(a, 0)$ because of the equality $a = a + 0i$, and the element $i = 0 + 1 \cdot i$ is associated with the point $(0, 1)$. On the other hand, comparing formulas (3) and (4) of this section with formulas (2) and (3) of Sec. 17, we find that the sum and product of the elements α and β of the field $D(i)$ are correlated with the points which are the sum and, respectively, the product of points associated with the elements α and β.

This completes the proof of the theorem, since all fields that are isomorphic to some given field are isomorphic among themselves. For one thing, we see that the choice (in Sec. 17) of formulas (2) and (3) for determining operations involving points was not accidental and cannot be altered.

There are many other ways of constructing the field of complex numbers. Let us examine one which uses the addition and multiplication of matrices.

We consider a noncommutative ring of second-order matrices over the field of real numbers. It is obvious that the scalar matrices

$$\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}$$

constitute in this ring a subfield that is isomorphic to the field of real numbers. It turns out, however, that *in the ring of second-order matrices over the field of reals, we can also find a subfield that is isomorphic to the field of complex numbers.* Indeed, associate with every complex number $a + bi$ the *matrix*

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$$

In this way, the entire field of complex numbers is mapped one-to-one onto a part of the ring of second-order matrices, and from the equations

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix} + \begin{pmatrix} c & d \\ -d & c \end{pmatrix} = \begin{pmatrix} a+c & b+d \\ -(b+d) & a+c \end{pmatrix},$$

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix} \cdot \begin{pmatrix} c & d \\ -d & c \end{pmatrix} = \begin{pmatrix} ac-bd & ad+bc \\ -(ad+bc) & ac-bd \end{pmatrix}$$

it follows that this mapping is isomorphic, since the matrices in the right-hand members correspond to the complex numbers $(a + c) + (b + d)\, i = (a + bi) + (c + di)$ and $(ac - bd) + (ad + bc)\, i = (a + bi)(c + di)$. In particular, the role of the imaginary unit i is played by the matrix

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

The foregoing result indicates yet another possible way of constructing the field of complex numbers that is just as satisfactory as those considered earlier.

47. Linear Algebra and the Algebra of Polynomials over an Arbitrary Field

In the earlier chapters of this book devoted to linear algebra, the base field was the field of real numbers. It is easy to verify, however, that much of what was written in those chapters can be carried over word for word to the case of an arbitrary base field.

Thus, *for an arbitrary base field P, the Gaussian method for solving systems of linear equations, the theory of determinants and Cramer's rule, which were given in Chapter 1, all hold true.* It is only the remark concerning skew-symmetric determinants (at the end of Sec. 4) which requires the assumption that the characteristic of the field P is different from two. Incidentally, the proof of Property 4 (same section) also breaks down if the characteristic of the field P is equal to two, though the property itself holds true.

It is also useful to note that the assertion (mentioned repeatedly in Chapter 1) on the existence of an infinity of distinct solutions to an indeterminate system of linear equations holds true in the case of any infinite base field P, but ceases to hold if P is finite.

The following carry over completely to the case of an arbitrary base field: the theory of linear dependence of vectors, the theory of the rank of a matrix and the general theory of systems of linear equations (see Chapter 2), and also the algebra of matrices (Chapter 3).

The general theory of quadratic forms constructed in Sec. 26 *is carried over to the case of any base field P whose characteristic is different from two.* As can be readily demonstrated, the fundamental theorem of this section ceases to hold without this restriction.

For example, let $P = Z_2$, that is, let P be a field consisting of two elements 0 and 1; let $1 + 1 = 0$, whence $-1 = 1$, and let there be a quadratic form $f = x_1x_2$ over this field. If there exists a linear transformation

$$x_1 = b_{11}y_1 + b_{12}y_2,$$
$$x_2 = b_{21}y_1 + b_{22}y_2$$

which reduces f to canonical form, then in the equation

$$f = (b_{11}y_1 + b_{12}y_2)(b_{21}y_1 + b_{22}y_2)$$
$$= b_{11}b_{21}y_1^2 + (b_{11}b_{22} + b_{12}b_{21})\, y_1y_2 + b_{12}b_{22}y_2^2$$

the coefficient $b_{11}b_{22} + b_{12}b_{21}$ of the product y_1y_2 must be equal to zero. But this coefficient is equal to the determinant of the linear transformation that we took, since irrespective of whether $b_{12}b_{21} = 1$ or $b_{12}b_{21} = 0$, we have $b_{12}b_{21} = -b_{12}b_{21}$ in both cases. Our linear transformation turned out to be singular.

The rest of Chapter 6 is largely devoted to quadratic forms with complex or real coefficients.

Finally, *the entire theory of linear spaces and their linear transformations which was constructed in Chapter 7 holds true for the case of an arbitrary base field P.* Incidentally, the concept of a characteristic root is connected with the theory of polynomials over an arbitrary field (this will be discussed below). Notice that the theorem, in Sec. 33, on the relationship between characteristic roots and eigenvalues will now be formulated as follows: the characteristic roots of a linear transformation φ which lie in the base field P, and they alone, serve as the eigenvalues of this transformation.

Now the theory of Euclidean spaces (Chapter 8) is essentially connected with the field of real numbers.

We can also extend to the case of an arbitrary base field P certain of the above-discussed sections of the algebra of polynomials. However, it is first necessary to make precise the meaning of the concept of a polynomial over an arbitrary field.

In Sec. 20 we indicated two viewpoints concerning the concept of a polynomial: the formal-algebraic view and the function-theoretic view. Both can be transferred to the case of an arbitrary base field. However, though they are equivalent in the case of number fields (see Sec. 24), and, as can readily be verified, of infinite fields in general, *they cease to be equivalent in the case of finite fields.*

Consider, for instance, the field Z_2 introduced in Sec. 45 and consisting of two elements 0 and 1 with $1 + 1 = 0$. The polynomials $x + 1$ and $x^2 + 1$ with coefficients from this field are distinct; that is to say, they do not satisfy the algebraic definition of equality of polynomials. Yet, for $x = 0$, both these polynomials become 1, and for $x = 1$ they have the value 0, that is to say, they must be considered equal as "functions" of the "variable" x, which takes on values in the field Z_2. In the field Z_3, consisting of three elements: 0, 1, 2, with $1 + 2 = 0$, the situation is the same relative to the polynomials $x^3 + x + 1$ and $2x + 1$. Examples of this type can, generally, be indicated for all finite fields.

Thus, in the theory of an arbitrary field P, one cannot accept the function-theoretic view of polynomials. It consequently becomes

necessary to make explicit the formal-algebraic definition of a polynomial. For this purpose, we will construct a ring of polynomials over an arbitrary field P such that dispenses, from the very start, with the ordinary notation of polynomials in terms of an "unknown" x.

Consider all possible ordered finite systems of elements of the field P having the form

$$(a_0, a_1, \ldots, a_{n-1}, a_n) \tag{1}$$

Here, n is arbitrary, $n \geqslant 0$, but for $n > 0$ it must be true that $a_n \neq 0$. Defining addition and multiplication for systems of the form (1) in accord with formulas (3) and (4), Sec. 20, we convert the collection of these systems into a commutative ring; the necessary proofs of the properties repeat word for word what was accomplished for number polynomials in Sec. 20.

In the ring we have constructed, systems of the form (a) (the case $n = 0$) constitute a subfield isomorphic to the field P. This permits identifying such systems with corresponding elements a of the field P, that is, setting

$$(a) = a \quad \text{for all} \quad a \text{ in } P \tag{2}$$

On the other hand, denote the system (0, 1) by the letter x,

$$x = (0, 1)$$

Then, applying the above-indicated definition of multiplication, we find that $x^2 = (0, 0, 1)$ and, generally,

$$x^k = (\underbrace{0, 0, \ldots, 0}_{k \text{ times}}, 1) \tag{3}$$

Now using the definitions of addition and multiplication of ordered systems, and also equalities (2) and (3), we get

$$\begin{aligned}
&(a_0, a_1, a_2, \ldots, a_{n-1}, a_n) \\
&\quad = (a_0) + (0, a_1) + (0, 0, a_2) \\
&\quad + \ldots + (\underbrace{0, 0, \ldots, 0}_{n-1 \text{ times}}, a_{n-1}) + (\underbrace{0, 0, \ldots, 0}_{n \text{ times}}, a_n) \\
&\quad = (a_0) + (a_1)(0, 1) + (a_2)(0, 0, 1) \\
&\quad + \ldots + (a_{n-1})(\underbrace{0, 0, \ldots, 0}_{n-1 \text{ times}}, 1) + (a_n)(\underbrace{0, 0, \ldots, 0}_{n \text{ times}}, 1) \\
&\quad = a_0 + a_1 x + a_2 x^2 + \ldots + a_{n-1} x^{n-1} + a_n x^n
\end{aligned}$$

Thus, any ordered system of type (1) can be written as a polynomial in x with coefficients from the field P, and this notation will evidently be unique. Finally, starting with the already proved commutativity of addition, we can go over to the notation in descending powers of x.

Consequently, we construct a commutative ring which it is natural to call *a ring of polynomials in the unknown x over the field P*. This ring is symbolized as $P[x]$.

The ring $P[x]$ contains the field P itself, as was demonstrated above. Now, as in the case of rings of polynomials over number fields (see Sec. 20), *the ring $P[x]$ has a unit element, does not have zero divisors and is not a field.*

If the field P is contained in a greater field $\overline{P}$, then the ring $P[x]$ is a subring of the ring $\overline{P}[x]$: any polynomial with coefficients from P can of course be considered a polynomial over the field $\overline{P}$ too; now the sum and product of polynomials depend solely on their coefficients, and for this reason they do not change when passing to a larger field.

To get a still better picture of the true extent of the concept of a "ring of polynomials over a field P", let us examine it from yet another angle.

Let the field P be contained as a subring in some commutative ring L. The element α of ring L is called *algebraic over the field P* if there exists an equation of degree n, $n \geqslant 1$, with coefficients from the field P that is satisfied by the element α. If there is no such equation, then the element α is termed *transcendental over the field P*. Naturally, the element x of the ring $P[x]$ is transcendental over the field P.

The following theorem holds true.

If the element α of ring L is transcendental over the field P, then the subring L' obtained by adjoining the element α to the field P (i.e., the minimal subring of the ring L containing the field P and the element α) *is isomorphic to the ring $P[x]$ of polynomials.*

Indeed, any element β of the ring L which can be written as

$$\beta = a_0\alpha^n + a_1\alpha^{n-1} + \ldots + a_{n-1}\alpha + a_n, \qquad n \geqslant 0 \qquad (4)$$

with coefficients $a_0, a_1, \ldots, a_{n-1}, a_n$ from the field P will be contained in the subring L'. The element β cannot have two distinct notations of the form (4), since by subtracting one from the other we would find that there exists an equation over the field P satisfied by the element α, but this is in conflict with the transcendental nature of this element. Combining the elements of type (4) by the rules of addition in the ring L, it is of course possible to combine coefficients of like powers of α; but this coincides with the rule for

adding polynomials. On the other hand, by multiplying elements of form (4) by the rules for multiplication in the ring L, we can, using the distributive law, perform termwise multiplication and then collect like terms. This evidently leads to the familiar law of multiplication of polynomials. This proves that elements of the type (4) constitute, in the ring L, a subring containing the field P and the element α (that is, a subring coinciding with L'), and that this subring is isomorphic to the polynomial ring $P[x]$.

We see that the choice of definitions for operations on polynomials we made above was not accidental; it is fully determined by the fact that the element x of the ring $P[x]$ must be transcendental over the field P.

Note that in constructing the polynomial ring $P[x]$ we never used the division of elements of the field P and only once (namely, in proving the assertion on the degree of a product of polynomials) had to refer to the absence of zero divisors in the field P. It is therefore possible to take an arbitrary commutative ring L and, repeating the foregoing construction, derive a *polynomial ring $L[x]$ over the ring L*; if in this case the ring L does not contain divisors of zero, the power of the product of the polynomials will be equal to the sum of the powers of the factors and therefore the polynomial ring $L[x]$ will not contain divisors of zero either.

Returning to polynomials with coefficients from an arbitrary field P, notice that actually the entire theory of divisibility of polynomials (described in Secs. 20-22 of this book) is carried over to this case. Namely, *in the ring $P[x]$ we have the division algorithm*, and both the quotient and the remainder will themselves belong to the ring $P[x]$. Also, *the concept of a divisor is meaningful in the ring $P[x]$ and all its basic properties are preserved.* The fact that the division algorithm does not take us outside the base field P, permits us to assert that *the property of a polynomial $\varphi(x)$ to be a divisor of $f(x)$ does not depend upon whether we consider the field P or any extension of it.*

Also preserved in the ring $P[x]$ are the definition and all the properties of a greatest common divisor, together with the Euclidean algorithm and the theorem proved in Sec. 21 *with the aid of this algorithm.* Notice that since the division algorithm is, as we know, independent of the choice of the base field, we can assert that *the greatest common divisor of two given polynomials is likewise independent of whether we consider the field P or an arbitrary extension of it, $\bar{P}$.*

Finally, *for polynomials over the field P, the concept of a root is meaningful and the basic properties of roots hold true.* Likewise preserved is the theory of multiple roots. Incidentally, we will return to this question at the end of the next section.

These remarks will enable us, in our subsequent study of polynomials over any field P, to refer to Secs. 20-22.

48. Factorization of Polynomials into Irreducible Factors

On the basis of the theorem on the existence of a root, we proved in Sec. 24 the existence and uniqueness of factorization of a polynomial into irreducible factors for fields of complex and real numbers. These results are particular cases of general theorems referring to polynomials over an arbitrary field P. The present section is devoted to this general theory, which parallels the theory of the prime factorization of integers.

First let us define those polynomials which play the same role in the polynomial ring as primes play in the ring of integers. We stress from the start that in this definition we deal solely with polynomials whose degree is greater than or equal to unity. This is in full accord with the fact that in the definition of prime numbers and in the study of the factorization of integers into prime factors, the numbers 1 and -1 are ruled out.

Suppose we have a polynomial $f(x)$ of degree n, $n \geqslant 1$, with coefficients from the field P. By Property V, Sec. 21, all polynomials of zero degree are divisors of $f(x)$. On the other hand, by Property VII, all polynomials $cf(x)$, where c is a nonzero element of P, will also be divisors of $f(x)$; note that these polynomials exhaust all the divisors (with degree n) of the polynomial $f(x)$. As to divisors (of $f(x)$) whose degree is greater than 0 but less than n, it will be seen that they may or may not be in the ring $P[x]$. In the former case, the polynomial $f(x)$ is called *reducible* in the field P (or over the field P), in the latter case, *irreducible* over this field.

Recalling the definition of a divisor, we may say that *a polynomial $f(x)$ of degree n is reducible over the field P if it can be factored over this field* (*i. e., in the ring $P[x]$*) *into a product of two factors of degree less than n:*

$$f(x) = \varphi(x)\,\psi(x) \tag{1}$$

and $f(x)$ is irreducible over the field P if in any factorization of the type (1), *one of its factors is of degree 0 and the other is of degree n.*

Note particularly that one can speak of reducibility or irreducibility of a polynomial only as regards a given field P, since a polynomial that is irreducible over one field may prove to be reducible over some extension $\overline{P}$ of that field. Thus, the polynomial $x^2 - 2$ with integral coefficients is irreducible over the field of rational numbers: it cannot be factored into a product of two linear factors with rational coefficients. However, this polynomial is reducible over the field of real numbers, as the following equation shows:

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

The polynomial $x^2 + 1$ is irreducible not only over the field of rational numbers but also over the field of real numbers. It becomes reducible however in the field of complex numbers, since

$$x^2 + 1 = (x - i)(x + i)$$

Let us point to certain basic properties of irreducible polynomials, bearing in mind that we will be speaking of polynomials irreducible over the field P.

(a) *Any polynomial of degree one is irreducible.*

This is rather evident since if the polynomial could be factored into a product of factors of lower degree, then they would have to be of degree 0. But the product of any polynomials of zero degree is again a polynomial of zero degree and not first degree.

(b) *If a polynomial $p(x)$ is irreducible, then any polynomial $cp(x)$, where c is a nonzero element of P, is also irreducible.*

This property follows from Properties I and VII of Sec. 21. It will permit us, where necessary, to confine our consideration to irreducible polynomials whose leading coefficients are unity.

(c) *If $f(x)$ is an arbitrary polynomial and $p(x)$ is an irreducible polynomial, then either $f(x)$ is divisible by $p(x)$ or the polynomials are coprime (relatively prime).*

If $(f(x), p(x)) = d(x)$, then $d(x)$, being a divisor of the irreducible polynomial $p(x)$ is either of degree 0 or is a polynomial of the form $cp(x)$, $c \neq 0$. In the former case, $f(x)$ and $p(x)$ are coprime, in the latter, $f(x)$ is divisible by $p(x)$.

(d) *If the product of the polynomials $f(x)$ and $g(x)$ is divisible by an irreducible polynomial $p(x)$, then at least one of these polynomials is divisible by $p(x)$.*

Indeed, if $f(x)$ is not divisible by $p(x)$, then, by (c), $f(x)$ and $p(x)$ are coprime, and then, by Property (b) of Sec. 21, the polynomial $g(x)$ must be divisible by $p(x)$.

Property (d) is readily carried over to the case of a product of any finite number of factors.

The two theorems which follow are the main purpose of this whole section.

Any polynomial $f(x)$ in the ring $P[x]$ having degree n, $n \geqslant 1$, can be factored into a product of irreducible factors.

Indeed, if a polynomial $f(x)$ is itself irreducible, then the indicated product consists of only one polynomial. But if it is reducible, then it can be factored into a product of factors of lower degree. If, among these factors, we again find irreducibles, then we decompose them into factors again, etc. This process will cease after a finite number of steps, since in any factorization of $f(x)$ into factors, the sum of the degrees of the factors must be equal to n and therefore the number of factors dependent on x cannot exceed n.

The factorization of integers into prime factors is unique if we confine our consideration to positive integers. However, in the ring of all integers, uniqueness only occurs to within sign: thus, $-6 = 2\cdot(-3) = (-2)\cdot 3$, $10 = 2\cdot 5 = (-2)\cdot(-5)$ and so on. A similar situation obtains in the polynomial ring as well. If

$$f(x) = p_1(x)\, p_2(x) \ldots p_s(x)$$

is a factorization of the polynomial $f(x)$ into a product of irreducible factors and if the elements $c_1, c_2, \ldots, c_s$ from the field P are such that their product is equal to 1, then

$$f(x) = [c_1 p_1(x)]\cdot[c_2 p_2(x)] \ldots [c_s p_s(x)]$$

will also, by (b), be a factorization of $f(x)$ into a product of irreducible factors. It turns out that this exhausts all factorizations of $f(x)$.

If a polynomial $f(x)$ from a ring $P[x]$ can be decomposed in two ways into a product of irreducible factors:

$$f(x) = p_1(x)\, p_2(x) \ldots p_s(x) = q_1(x)\, q_2(x) \ldots q_t(x) \qquad (2)$$

then, $s = t$, and, with appropriate numbering, we have the equalities

$$q_i(x) = c_i p_i(x), \qquad i = 1, 2, \ldots, s \qquad (3)$$

where c_i are nonzero elements from the field P.

This theorem holds for polynomials of degree one, since they are irreducible. We will therefore argue by induction with respect to the degree of the polynomial, that is, we will prove the theorem for $f(x)$, assuming that for polynomials of lower degree it is already proved.

Since $q_1(x)$ is a divisor of $f(x)$, it follows, by Property (d) and equality (2), that $q_1(x)$ will be a divisor of at least one of the polynomials $p_i(x)$, say of $p_1(x)$. However, since the polynomial $p_1(x)$ is irreducible and the degree of $q_1(x)$ is greater than zero, there exists an element c_1 such that

$$q_1(x) = c_1 p_1(x) \qquad (4)$$

Substituting this expression of $q_1(x)$ into (2) and cancelling $p_1(x)$ (which is permissible since there are no zero divisors in the ring $P[x]$), we obtain the equation

$$p_2(x)\, p_3(x) \ldots p_s(x) = [c_1 q_2(x)]\, q_3(x) \ldots q_t(x)$$

Since the degree of the polynomial equal to these products is lower than that of $f(x)$, then it is already proved that $s - 1 = t - 1$, whence $s = t$, and there exist elements $c_2', c_3, \ldots, c_s$ such that $c_2' p_2(x) = c_1 q_2(x)$, whence $q_2(x) = (c_1^{-1} c_2')\, p_2(x)$ and $c_i p_i(x) = q_i(x)$, $i = 3, \ldots, s$. Assuming $c_1^{-1} c_2' = c_2$ and taking into account (4), we get the equations (3) completely.

The theorem we have just proved may be stated more succinctly: *every polynomial may be uniquely decomposed into irreducible factors to within zero-degree factors.*

Incidentally, it is always possible to consider the following special type of factorization *which will be quite unique for every polynomial*: take any factorization of the polynomial $f(x)$ into irreducible factors and factor out of each the leading coefficient. We get the factorization

$$f(x) = a_0 p_1(x)\, p_2(x) \ldots p_s(x) \tag{5}$$

where all the $p_i(x)$, $i = 1, 2, \ldots, s$, are irreducible polynomials with leading coefficients equal to unity. The factor a_0 will be equal to the leading coefficient of the polynomial $f(x)$, as can readily be verified by multiplying out the right member of (5).

The irreducible factors in (5) do not necessarily have to be distinct. If an irreducible polynomial $p(x)$ appears several times in the factorization (5), it is called a *multiple factor* of $f(x)$, namely a *k-fold* (double, triple, etc.) *factor* if (5) contains exactly k factors equal to $p(x)$. But if the factor $p(x)$ appears in (5) only once, then it is called a *simple* (or *single*) *factor* of $f(x)$.

If in the factorization (5) the factors $p_1(x)$, $p_2(x)$, ..., $p_l(x)$ are distinct and any other factor is equal to one of them and if $p_i(x)$, $i = 1, 2, \ldots, l$, is a k_i-fold factor of the polynomial $f(x)$, then (5) may be rewritten as

$$f(x) = a_0 p_1^{k_1}(x)\, p_2^{k_2}(x) \ldots p_l^{k_l}(x) \tag{6}$$

This is the notation that we will ordinarily make use of without specifying that the exponents are equal to the multiplicities of the corresponding factors, i.e., that $p_i(x) \neq p_j(x)$ for $i \neq j$.

If we are given the factorizations of the polynomials $f(x)$ and $g(x)$ into irreducible factors, then the greatest common divisor $d(x)$ of these polynomials is equal to the product of the factors appearing in both factorizations at the same time, and each factor is taken to the power equal to the least of its multiplicities in the two given polynomials.

Indeed, the indicated product will be a divisor of each of the polynomials $f(x)$, $g(x)$ and therefore also of $d(x)$. If this product were different from $d(x)$, then the factorization of $d(x)$ into irreducible factors would either contain a factor that does not appear in the factorization of at least one of the polynomials $f(x)$ and $g(x)$, which is impossible, or one of the factors would have a higher power than it has in the factorization of one of the polynomials $f(x)$ and $g(x)$, which is again impossible.

This theorem is similar to the rule ordinarily used to find the greatest common divisor of integers. However, in the case of polynomials, it cannot replace the Euclidean algorithm, for, since there is only a finite number of primes less than a given positive integer,

the factorization of an integer into prime factors is attained by a finite number of trials. This is not the case in a polynomial ring over an infinite base field, and, in the general case, one cannot offer a method for factoring polynomials into irreducible factors. What is more, it is very hard even to decide in the general case the question of whether a polynomial $f(x)$ is irreducible over a given field P. Thus, the description of all irreducible polynomials for the case of the fields of complex and real numbers was obtained in Sec. 24 as a corollary to a very profound theorem on the existence of a root. As to the field of rational numbers, only a few assertions of a specific nature concerning polynomials that are irreducible over this field will be made in Sec. 56.

We have shown that in the polynomial ring (as in the ring of integers) we have a factorization into "prime" (irreducible) factors and that this factorization is in a certain sense unique. The question arises as to whether it is possible to carry over these results to broader classes of rings. We confine ourselves here to the case of such commutative rings as have a unit element and do not have divisors of zero.

We will use the term *divisor of unity* for an element a of the ring such that in this ring there exists an inverse element a^{-1}:

$$aa^{-1} = 1$$

In the ring of integers, these are the numbers 1 and -1, in the ring $P[x]$ of polynomials, all the polynomials of zero degree (that is, nonzero numbers from the field P). The element c, which is nonzero and is not a divisor of unity, will be called a *prime* element of the ring if in any decomposition of it into a product of two factors, $c = ab$, one of the factors is invariably a divisor of unity. In the ring of integers, the prime elements are prime numbers, in the polynomial ring they are irreducible polynomials.

Will every element of the ring under consideration that is nonzero and is not a divisor of unity be decomposable into a product of prime factors? If it is, will the factorization be unique? This is to be understood as follows: if

$$a = p_1 p_2 \cdots p_k = q_1 q_2 \cdots q_l$$

are two factorizations of the element a into prime factors, then $k = l$ and (possibly after a change in the numbering)

$$q_i = p_i c_i, \qquad i = 1, 2, \ldots, k$$

where c_i is a divisor of unity.

It turns out that in both instances the answer is no. We give one example, namely, we indicate a ring in which factorization into prime factors is possible but not unique.

Consider complex numbers of the form

$$\alpha = a + b\sqrt{-3} \tag{7}$$

where a and b are integers. All such numbers form a ring without divisors of zero and containing a unit element; indeed,

$$(a+b\sqrt{-3})(c+d\sqrt{-3})=(ac-3bd)+(bc+ad)\sqrt{-3} \quad (8)$$

We use the term *norm* of a number $\alpha = a + b\sqrt{-3}$ for the positive integer

$$N(\alpha) = a^2 + 3b^2$$

By (8), *the norm of a product is equal to the product of the norms*

$$N(\alpha\beta) = N(\alpha)\,N(\beta) \quad (9)$$

Indeed,

$$(ac - 3bd)^2 + 3(bc + ad)^2 = a^2c^2 + 9b^2d^2 + 3b^2c^2 + 3a^2d^2$$
$$= (a^2 + 3b^2)(c^2 + 3d^2)$$

If in our ring the number α is a divisor of unity, that is the number α^{-1} is also of the form (7), then, by (9),

$$N(\alpha)\cdot N(\alpha^{-1}) = N(\alpha\alpha^{-1}) = N(1) = 1$$

and therefore $N(\alpha) = 1$, since the numbers $N(\alpha)$ and $N(\alpha^{-1})$ are integers and are positive. If $\alpha = a + b\sqrt{-3}$, then from $N(\alpha) = 1$ it follows that

$$N(\alpha) = a^2 + 3b^2 = 1$$

which, however, is possible only when $b=0$, $a = \pm 1$. Thus, *in our ring, as in the ring of integers, only the numbers* 1 *and* -1 *will be divisors of unity, and only these numbers have a norm equal to unity.*

The equation (9) for the norm of a product can naturally be extended to the case of any finite number of factors. It is thus easy to conclude that *any number* α *in our ring can be factored into a product of a finite number of prime factors.* We leave the proof to the reader.

However, we cannot assert that the factorization into prime factors is unique. For example, the following equations hold true:

$$4 = 2\cdot 2 = (1 + \sqrt{-3})(1 - \sqrt{-3})$$

In our ring there are no other divisors of unity except 1 and -1, and so the number $1 + \sqrt{-3}$ (like the number $1 - \sqrt{-3}$) cannot differ from the number 2 solely by a factor which is a divisor of unity. It remains to show that *each one of the numbers* 2, $1 + \sqrt{-3}$, $1 - \sqrt{-3}$ *will be prime in the ring under consideration.* Indeed, the norm of each of these three numbers is equal to 4. Let α be any one of these numbers and let

$$\alpha = \beta\gamma$$

Then, by (9), one of the following three cases is possible:

(1) $N(\beta) = 4$, $N(\gamma) = 1$; (2) $N(\beta) = 1$, $N(\gamma) = 4$; (3) $N(\beta) = N(\gamma) = 2$. In the first case, the number γ will, as we know, be a divisor of unity; in the second case, β will be a divisor of unity. The third case is impossible due to the impossibility of the equality

$$a^2 + 3b^2 = 2$$

where a and b are integers.

Multiple factors. Although, as has been demonstrated above, we are not able to decompose polynomials into irreducible factors, there exist methods which enable us to determine whether a given polynomial has multiple factors or not and, if it does, to reduce the study of that polynomial to the study of polynomials that do not contain multiple factors. True, these methods require that we impose certain restrictions on the base field. In the rest of this section we will assume that the field P has characteristic 0. Without this restriction, the theorems on multiple factors that will be proved below break down. At the same time, the case of fields of characteristic zero is the most important one from the viewpoint of applications since, for one thing, all number fields are included here.

To begin with, notice that we can extend to this case both the concept of a derivative of a polynomial (introduced in Sec. 22 for polynomials with complex coefficients) and the basic properties of this concept.* Let us now prove the following theorem.

If $p(x)$ is a k-fold irreducible factor of the polynomial $f(x)$, $k \geqslant 1$, then it will be the $(k-1)$-fold factor of the derivative of this polynomial. In particular, a prime factor of the polynomial does not enter into the factorization of the derivative.

Indeed, let

$$f(x) = p^k(x)\, g(x) \tag{10}$$

$g(x)$ is no longer divisible by $p(x)$. Differentiating (10), we get

$$\begin{aligned} f'(x) &= p^k(x)\, g'(x) + kp^{k-1}(x)\, p'(x)\, g(x) \\ &= p^{k-1}(x)\, [p(x)\, g'(x) + kp'(x)\, g(x)] \end{aligned}$$

The second term in the brackets is not divisible by $p(x)$; indeed, $g(x)$ is not divisible by $p(x)$ by hypothesis, $p'(x)$ is of lower degree, i.e., it is not divisible by $p(x)$ either; hence, due to the irreducibility of the polynomial $p(x)$ and Property (d) of this section and Property IX of Sec. 21, our assertion follows. On the other hand, the first term in the sum in the square brackets is divisible by $p(x)$ and so the entire sum cannot be divisible by $p(x)$; which is to say that the factor $p(x)$ does indeed appear in $f'(x)$ with a multiplicity of $k-1$.

---

* For fields of a finite characteristic, the assertion that the derivative of a polynomial of degree n is of degree $n-1$ fails.

From our theorem and from the above-indicated method of finding the greatest common divisor of two polynomials it follows that if a factorization of the polynomial $f(x)$ into irreducible factors is given,

$$f(x) = a_0 p_1^{k_1}(x)\, p_2^{k_2}(x) \ldots p_l^{k_l}(x) \tag{11}$$

then *the greatest common divisor of* $f(x)$ *and of its derivative has the following factorization into irreducible factors*:

$$(f(x), f'(x)) = p_1^{k_1-1}(x)\, p_2^{k_2-1}(x) \ldots p_l^{k_l-1}(x) \tag{12}$$

where the factor $p_i^{k_i-1}(x)$ should naturally be replaced by unity for $k_i = 1$. In particular, *a polynomial* $f(x)$ *does not contain multiple factors if and only if it is relatively prime to its derivative.*

We now know how to answer the question of the existence of multiple factors in a given polynomial. What is more, since neither the derivative of a polynomial nor the greatest common divisor of two polynomials depend on whether we are considering the field P or any extension $\overline{P}$ of it, we obtain the following corollary to the result that has just been proved.

If a polynomial $f(x)$ *with coefficients in a field* P *of characteristic zero does not have multiple factors over this field, then neither will there be any multiple factors over any extension* $\overline{P}$ *of the field* P.

In particular, if $f(x)$ is irreducible over P and $\overline{P}$ is some extension of P, then, although $f(x)$ can be reducible over $\overline{P}$, it will definitely not be divisible by the square of an irreducible (over $\overline{P}$) polynomial.

Isolating multiple factors. If we have a polynomial $f(x)$ with the factorization (11) and if by $d_1(x)$ we denote the greatest common divisor of $f(x)$ and of its derivative $f'(x)$, then (12) will be a factorization of $d_1(x)$. Dividing (11) by (12), we get

$$v_1(x) = \frac{f(x)}{d_1(x)} = a_0 p_1(x)\, p_2(x) \ldots p_l(x)$$

That is, we obtain a polynomial without multiple factors, and any irreducible factor of $v_1(x)$ will also be a factor of $f(x)$. In this way, finding the irreducible factors of $f(x)$ is reduced to finding them for the polynomial $v_1(x)$ which, generally speaking, is of lower degree and, at any rate, contains only prime factors. If the problem is solved for $v_1(x)$, then it only remains to determine the multiplicity of the irreducible factors found in $f(x)$; this is done by means of the division algorithm.

A more sophisticated variant of this method enables us to consider several polynomials without multiple factors; also, having found the irreducible factors of these polynomials, we not only find all the irreducible factors of $f(x)$, but also their multiplicities.

Let (11) be a factorization of $f(x)$ into irreducible factors, the greatest multiplicity of the factors being s, $s \geqslant 1$. Denote by $F_1(x)$ the product of all single factors of $f(x)$, by $F_2(x)$ the product of all double factors, but taken only once at a time, and so forth; finally, denote by $F_s(x)$ the product of all s-fold factors taken once at a time, as before. If under these conditions, for some j in $f(x)$, there are no j-fold factors, set $F_j(x) = 1$. Then $f(x)$ will be divisible by the kth degree of the polynomial $F_k(x)$, $k = 1, 2, \ldots, s$, and the factorization (11) becomes

$$f(x) = a_0 F_1(x) F_2^2(x) F_3^3(x) \ldots F_s^s(x)$$

and the factorization (12) for $d_1(x) = (f(x), f'(x))$ will be rewritten as

$$d_1(x) = F_2(x) F_3^2(x) \ldots F_s^{s-1}(x)$$

Denoting by $d_2(x)$ the greatest common divisor of the polynomial $d_1(x)$ and of its derivative, and generally by $d_k(x)$ the greatest common divisor of the polynomials $d_{k-1}(x)$ and $d'_{k-1}(x)$, we obtain in the same fashion

$$d_2(x) = F_3(x) F_4^2(x) \ldots F_s^{s-2}(x),$$
$$d_3(x) = F_4(x) F_5^2(x) \ldots F_s^{s-3}(x),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$d_{s-1}(x) = F_s(x),$$
$$d_s(x) = 1$$

Whence

$$v_1(x) = \frac{f(x)}{d_1(x)} = a_0 F_1(x) F_2(x) F_3(x) \ldots F_s(x),$$
$$v_2(x) = \frac{d_1(x)}{d_2(x)} = F_2(x) F_3(x) \ldots F_s(x),$$
$$v_3(x) = \frac{d_2(x)}{d_3(x)} = F_3(x) \ldots F_s(x)$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$v_s(x) = \frac{d_{s-1}(x)}{d_s(x)} = F_s(x)$$

and, therefore, finally,

$$F_1(x) = \frac{v_1(x)}{a_0 v_2(x)}, \quad F_2(x) = \frac{v_2(x)}{v_3(x)}, \ldots, F_s(x) = v_s(x)$$

Thus, using only procedures that do not require a knowledge of the irreducible factors of the polynomial $f(x)$, namely, taking the derivative, using the Euclidean algorithm and the division algorithm, we can find the polynomials $F_1(x), F_2(x), \ldots, F_s(x)$ without multiple factors; every irreducible factor of the polynomial $F_k(x)$, $k = 1, 2, \ldots, s$, will be k-fold for $f(x)$.

This method cannot, of course, be regarded as a procedure for factoring a polynomial into irreducible factors, since for the case of $s = 1$ (that is, for a polynomial without multiple factors) we only get $f(x) = F_1(x)$.

49. Theorem on the Existence of a Root

Quite naturally, the fundamental theorem (proved in Sec. 23) on the existence, for every numerical polynomial, of a root in the field of complex numbers cannot be extended to the case of an arbitrary field. In this section we will prove a theorem which in the general theory of fields replaces to some extent the afore-mentioned fundamental theorem of the algebra of complex numbers.

Let there be given a polynomial $f(x)$ over a field P. A natural question arises: if the polynomial $f(x)$ does not have any roots at all in the field P, then does there exist an extension $\overline{P}$ of P in which there will be at least one root of $f(x)$? We can assume that the degree of the polynomial $f(x)$ is greater than unity: the question is meaningless for a zero-degree polynomial, and every polynomial of degree one, $ax + b$, has the root $-\frac{b}{a}$ in the field P itself. On the other hand, we can evidently confine ourselves to the case of $f(x)$ being irreducible: if it is reducible over P, then the root of any one of its irreducible factors will be a root of $f(x)$ itself.

The answer to the question that interests us is given by the following **theorem on the existence of a root.**

For every polynomial $f(x)$ that is irreducible over the field P there is an extension of the field such that contains a root of $f(x)$. All minimal fields containing the field P and a root of this polynomial are isomorphic among themselves.

Let us first prove the second part of the theorem.

Suppose we have a polynomial irreducible over P:

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n \qquad (1)$$

and $n \geqslant 2$, that is, $f(x)$ has no roots in the field P itself. Suppose that there is an extension $\overline{P}$ of P which contains a root α of $f(x)$. Let us prove the following lemma which will be needed later on but which is of interest in itself.

If a root α, in $\overline{P}$, of a polynomial $f(x)$ which is irreducible over P serves also as a root of some polynomial $g(x)$ in the ring $P[x]$ then $f(x)$ will be a divisor of $g(x)$.

Indeed, the polynomials $f(x)$ and $g(x)$ over the field $\overline{P}$ have a common divisor $x - \alpha$ and so are not relatively prime. The property of polynomials not to be relatively prime does not, however, depend on the choice of the field. It is therefore possible to pass to the field P and apply Property (c) of Sec. 48.

Now let us find the minimal subfield $P(\alpha)$ of $\overline{P}$ which contains the field P and the element α. It definitely includes all elements of the form

$$\beta = b_0 + b_1\alpha + b_2\alpha^2 + \ldots + b_{n-1}\alpha^{n-1} \qquad (2)$$

where $b_0, b_1, b_2, \ldots, b_{n-1}$ are elements of P. No element of $\overline{P}$ can have two distinct notations of the form (2); if it is also true that

$$\beta = c_0 + c_1\alpha + c_2\alpha^2 + \ldots + c_{n-1}\alpha^{n-1}$$

and for at least one k, $c_k \neq b_k$, then α will be a root of the polynomial

$$g(x) = (b_0 - c_0) + (b_1 - c_1)x + (b_2 - c_2)x^2 + \ldots + (b_{n-1} - c_{n-1})x^{n-1}$$

which runs counter to the lemma proved above since the degree of $g(x)$ is lower than the degree of $f(x)$.

The elements of the field $\overline{P}$ having the form (2) include all the elements of the field P (for $b_1 = b_2 = \ldots = b_{n-1} = 0$), and also the element α itself (for $b_1 = 1$, $b_0 = b_2 = \ldots = b_{n-1} = 0$). We now prove that *elements of the form* (2) *constitute the entire sought-for subfield* $P(\alpha)$. Indeed, if we are given elements β [with notation (2)] and

$$\gamma = c_0 + c_1\alpha + c_2\alpha^2 + \ldots + c_{n-1}\alpha^{n-1}$$

then, on the basis of the properties of operations in the field $\overline{P}$,

$$\beta \pm \gamma = (b_0 \pm c_0) + (b_1 \pm c_1)\alpha + (b_2 \pm c_2)\alpha^2 + \ldots + (b_{n-1} \pm c_{n-1})\alpha^{n-1}$$

That is to say, the sum and difference of any two elements of the type (2) are again elements of that type.

If we multiply β and γ, we get an expression containing α^n and other higher powers of α. However, it follows from (1) and the equality $f(\alpha) = 0$ that α^n and therefore α^{n+1}, α^{n+2} and so on can be expressed in terms of lower powers of the element α. The simplest way of finding an expression for $\beta\gamma$ is this: let

$$\varphi(x) = b_0 + b_1x + \ldots + b_{n-1}x^{n-1},$$

$$\psi(x) = c_0 + c_1x + \ldots + c_{n-1}x^{n-1}$$

whence $\varphi(\alpha) = \beta$, $\psi(\alpha) = \gamma$. Multiply the polynomials $\varphi(x)$ and $\psi(x)$ and divide the product by $f(x)$. This yields

$$\varphi(x)\,\psi(x) = f(x)\,q(x) + r(x) \qquad (3)$$

where

$$r(x) = d_0 + d_1x + \ldots + d_{n-1}x^{n-1}$$

Taking the values of both sides of (3) for $x = \alpha$ we find that

$$\varphi(\alpha)\,\psi(\alpha) = f(\alpha)\,q(\alpha) + r(\alpha)$$

That is to say, by $f(\alpha) = 0$,

$$\beta\gamma = d_0 + d_1\alpha + \ldots + d_{n-1}\alpha^{n-1}$$

Thus, the product of two elements of the type (2) will again be an element of this type.

Finally, we will show that if element β is of the type (2), $\beta \neq 0$, then the element β^{-1} existing in the field $\overline{P}$ can also be written as (2). To do this, take the polynomial

$$\varphi(x) = b_0 + b_1x + \ldots + b_{n-1}x^{n-1}$$

in the ring $P[x]$. Since the degree of $\varphi(x)$ is lower than the degree of $f(x)$, and the polynomial $f(x)$ is irreducible over P, it follows that $\varphi(x)$ and $f(x)$ are relatively prime and therefore, by Secs. 21 and 47, there exist in the ring $P[x]$ polynomials $u(x)$ and $v(x)$ such that

$$\varphi(x)\,u(x) + f(x)\,v(x) = 1$$

We can assume here that the degree of $u(x)$ is less than n:

$$u(x) = s_0 + s_1x + \ldots + s_{n-1}x^{n-1}$$

Whence, by $f(\alpha) = 0$, it follows that

$$\varphi(\alpha)\,u(\alpha) = 1$$

and therefore, by the equality $\varphi(\alpha) = \beta$, we have

$$\beta^{-1} = u(\alpha) = s_0 + s_1\alpha + \ldots + s_{n-1}\alpha^{n-1}$$

Thus, the collection of elements of the field $\overline{P}$ having the form (2) constitutes a subfield of $\overline{P}$, which is the desired field $P(\alpha)$. Furthermore, since we saw that in seeking the sum and product of the elements β and γ of the type (2) we need only know the coefficients of the expressions of these elements in terms of powers of α, we can assert the truth of the following result. If besides $\overline{P}$ there is another extension $\overline{P}'$ of the field P, which also contains a root α' of the polynomial $f(x)$, and if $P(\alpha')$ is a minimal subfield of the field $\overline{P}'$ containing P and α' then the *fields* $P(\alpha)$ *and* $P(\alpha')$ *are isomorphic.* To obtain the isomorphic correspondence between them, it is necessary to associate with the element β of type (2) in $P(\alpha)$ an element

$$\beta' = b_0 + b_1\alpha' + b_2\alpha'^2 + \ldots + b_{n-1}\alpha'^{n-1}$$

in $P(\alpha')$ having the same coefficients. This completes the proof of the second part of the theorem.

Let us now prove the basic first part of this theorem. The foregoing will help to point the way. We have a polynomial $f(x)$ of degree $n \geqslant 2$ that is irreducible over the field P and it is required to construct an extension of P containing a root of $f(x)$. To do this,

let us take the entire polynomial ring $P[x]$ and partition it into disjoint classes, combining in one class the polynomials which yield the same remainders upon division by the given polynomial $f(x)$. In other words, the polynomials $\varphi(x)$ and $\psi(x)$ belong to the same class if their difference is exactly divisible by $f(x)$.

We agree to denote the resulting classes by the letters A, B, C and so on and to define the sum and product of classes in the following natural manner. Take any two classes A and B; choose in A a polynomial $\varphi_1(x)$, in B a polynomial $\psi_1(x)$ and denote by $\chi_1(x)$ the sum of these polynomials:

$$\chi_1(x) = \varphi_1(x) + \psi_1(x)$$

and by $\Theta_1(x)$ their product:

$$\Theta_1(x) = \varphi_1(x) \cdot \psi_1(x)$$

Now choose any other polynomial $\varphi_2(x)$ in A and any polynomial $\psi_2(x)$ in B and denote by $\chi_2(x)$ and $\Theta_2(x)$ their sum and product, respectively:

$$\chi_2(x) = \varphi_2(x) + \psi_2(x),$$
$$\Theta_2(x) = \varphi_2(x) \cdot \psi_2(x)$$

By hypothesis, the polynomials $\varphi_1(x)$ and $\varphi_2(x)$ are in the same class A and therefore their difference $\varphi_1(x) - \varphi_2(x)$ is exactly divisible by $f(x)$; the difference $\psi_1(x) - \psi_2(x)$ has the same property. From this it follows that the difference

$$\begin{aligned}\chi_1(x) - \chi_2(x) &= [\varphi_1(x) + \psi_1(x)] - [\varphi_2(x) + \psi_2(x)] \\ &= [\varphi_1(x) - \varphi_2(x)] + [\psi_1(x) - \psi_2(x)] \qquad (4)\end{aligned}$$

is also exactly divisible by the polynomial $f(x)$. This is also true of the difference $\Theta_1(x) - \Theta_2(x)$ since

$$\begin{aligned}\Theta_1(x) - \Theta_2(x) &= \varphi_1(x)\,\psi_1(x) - \varphi_2(x)\,\psi_2(x) \\ &= \varphi_1(x)\,\psi_1(x) - \varphi_1(x)\,\psi_2(x) + \varphi_1(x)\,\psi_2(x) - \varphi_2(x)\,\psi_2(x) \\ &= \varphi_1(x)\,[\psi_1(x) - \psi_2(x)] + [\varphi_1(x) - \varphi_2(x)]\,\psi_2(x) \qquad (5)\end{aligned}$$

Equation (4) shows that the polynomials $\chi_1(x)$ and $\chi_2(x)$ lie in the same class. In other words, the sum of any polynomial from class A and any polynomial from class B belongs to a very definite class C, which does not depend on what polynomials are chosen as "representatives" in classes A and B. We call this class C the *sum* of the classes A and B:

$$C = A + B$$

Similarly, because of (5), there is a class D which is independent of the choice of representatives in classes A and B and in which lies the product of any polynomial of A by any polynomial of B. We

call this class the *product* of the classes A and B:

$$D = AB$$

We shall show that the collection of classes into which we have partitioned the ring $P[x]$ of polynomials is converted into a field after the indicated introduction of the operations of addition and multiplication. Indeed, the validity of the associative and commutative laws for both operations and of the distributive law follows from the validity of these laws in the ring $P[x]$, since operations on classes reduce to operations on the polynomials lying in these classes. The role of zero is evidently played by the class composed of polynomials divisible exactly by the polynomial $f(x)$. We call this the *zero class* and denote it by the symbol 0. The *opposite* of class A, which is made up of polynomials that yield the remainder $\varphi(x)$ upon division by $f(x)$, is the class made up of polynomials which yield the remainder $-\varphi(x)$ upon division by $f(x)$, whence it follows that *subtraction* is unique on the set of classes.

To prove that *division* is possible on the set of classes, we have to show that there exists a class playing the role of unity and that for any class different from zero there is an inverse class. The class of polynomials which upon division by $f(x)$ yields a remainder 1 will obviously be *unity*. We call this the *unit class* and denote it by the symbol E.

Now suppose we have a class A different from zero. A polynomial $\varphi(x)$ chosen in A as a representative will thus not be exactly divisible by $f(x)$ and therefore, because of the irreducibility of $f(x)$, these two polynomials are relatively prime. Thus, in the ring $P[x]$ there exist polynomials $u(x)$ and $v(x)$ that satisfy the equation

$$\varphi(x)\, u(x) + f(x)\, v(x) = 1$$

whence

$$\varphi(x)\, u(x) = 1 - f(x)\, v(x) \tag{6}$$

Upon division by $f(x)$, the right member of (6) yields a remainder 1, which means it belongs to the unit class E. If the class to which the polynomial $u(x)$ belongs is denoted by B, then (6) shows that

$$AB = E$$

whence $B = A^{-1}$. This is proof of the existence of an inverse class for every nonzero class; in other words, this completes the proof that classes form a field.

We will denote this field by $\overline{P}$ and will show that *it is an extension of the field* P. With every element a of the field P is associated a class composed of polynomials which upon division by $f(x)$ yield a remainder a; the element a itself, regarded as a zero-degree polynomial,

belongs to this class. All classes of this special type constitute, in the field $\overline{P}$, a subfield that is isomorphic to the field P. Indeed, the one-to-one nature of the correspondence is obvious; on the other hand, for representatives in these classes we can choose elements of the field P and therefore with the sum (product) of elements of P is associated a sum (product) of corresponding classes. Consequently, in the future we will not need to distinguish between the elements of a field P and the classes corresponding to them.

Finally, use X to denote the class made up of polynomials which upon division by $f(x)$ yield the remainder x. This class is a definite element of the field $\overline{P}$, and we wish to demonstrate that *it is a root of the polynomial* $f(x)$. Let

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n$$

We denote by A_i the class corresponding, in the foregoing sense, to the element a_i of the field P, $i = 0, 1, \ldots, n$, and will find out what the element

$$A_0X^n + A_1X^{n-1} + \ldots + A_{n-1}X + A_n \qquad (7)$$

of the field $\overline{P}$ is equal to. Assuming elements a_i, $i = 0, 1, \ldots, n$, to be representatives of the classes A_i and the polynomial x to be a representative of the class X, and using the definition of addition and multiplication of classes, we find that the polynomial $f(x)$ is itself contained in class (7). However, $f(x)$ is exactly divisible by itself and therefore class (7) turns out to be the zero class. Thus, by replacing in (7) the classes A_i by the elements a_i of P corresponding to them, we find that the following equality holds in the field $\overline{P}$:

$$a_0X^n + a_1X^{n-1} + \ldots + a_{n-1}X + a_n = 0$$

That is to say, the class X is indeed a root of the polynomial $f(x)$.

This completes the proof of the theorem on the existence of a root. Note that by taking the field of real numbers for P and setting $f(x) = x^2 + 1$, we obtain yet another method for constructing the field of complex numbers.

Certain corollaries can be derived from the theorem on the existence of a root similar to those derived in Sec. 24 from the fundamental theorem of the algebra of complex numbers. One remark is in order first, however. Since any linear factor $x - c$ of a polynomial $f(x)$ is irreducible, it must appear in the unique factorization of $f(x)$ into irreducible factors.

However, the number of linear factors in the factorization of $f(x)$ into irreducible factors cannot exceed the degree of the polynomial. We get the following result.

A polynomial $f(x)$ of degree n cannot have more than n roots in the field P, even if each of the roots is counted with its multiplicity.

We use the term *splitting field* of a polynomial $f(x)$ of degree n over the field P for an extension Q of P such that contains n roots of $f(x)$ (counting multiplicity in the case of multiple roots). Consequently, over the field Q the polynomial $f(x)$ will decompose into linear factors, and no further extension of the field Q can make new roots appear for $f(x)$.

For every polynomial $f(x)$ in the ring $P[x]$ there is a splitting field over the field P.

Indeed, if a polynomial $f(x)$ of degree n, $n \geqslant 1$, has n roots in the field P itself, then P will be the desired splitting field. But if $f(x)$ does not decompose into linear factors over P, then we take one of its nonlinear irreducible factors $\varphi(x)$ and, on the basis of the theorem of the existence of a root, we extend P to the field P', which contains a root of $\varphi(x)$. If the polynomial $f(x)$ still does not break up into linear factors over P', we again extend the field, thus creating a root for one more of the remaining nonlinear irreducible factors. In a finite number of steps we will obviously arrive at the splitting field for $f(x)$.

Quite naturally, $f(x)$ can have many different splitting fields. One can prove that all the minimal fields containing the field P and n roots of the polynomial $f(x)$ (where n is the degree of the polynomial) are isomorphic. However, we will not make use of this assertion and will therefore not give the proof.

Multiple roots. In the previous section we proved that a polynomial $f(x)$ over a field P of characteristic 0 does not have multiple factors if and only if it is relatively prime to its derivative; it was also noted that the absence, in $f(x)$, of multiple factors over P implies the absence of such factors over any extension $\overline{P}$ of the field P. Let us apply this to the case when $\overline{P}$ is a splitting field for $f(x)$; recalling the definition of a miltiple root, we arrive at the following result.

If a polynomial $f(x)$ over a field P of characteristic 0 does not have multiple roots in the given splitting field, then it is relatively prime to its derivative $f'(x)$. Conversely, if $f(x)$ is relatively prime to its derivative, then it does not have multiple roots in any one of its splitting fields.

Whence, in particular, it follows that *a polynomial $f(x)$ which is irreducible over a field P of characteristic* 0, *cannot have multiple roots in any extension of the field.* This assertion does not hold in fields of a finite characteristic. This circumstance plays a perceptible role in the general theory of fields.

Note in conclusion that *for an arbitrary field, the Vieta formulas hold too* (see Sec. 24); here, the roots of the polynomial are taken in some splitting field of this polynomial.

50. The Field of Rational Fractions

The theory of rational fractions described in Sec. 25 *holds in full for the case of an arbitrary base field as well.* However, when passing from the field of real numbers to an arbitrary field P, the view taken of the expression $\frac{f(x)}{g(x)}$ as a function of the variable x must be rejected, for, as we know, it is not applicable to polynomials. Our job here is to figure out the meaning of these expressions for the case when the coefficients belong to an arbitrary field P. More precisely, we want to construct a field containing the polynomial ring $P[x]$ and in such a way that the operations of addition and multiplication defined in the new field coincide, as applied to polynomials, with the operations in the ring $P[x]$; in short, the ring $P[x]$ must be a subring of this new field. On the other hand, any element of the new field must be representable (in the sense of division as defined in this field) in the form of a quotient of two polynomials. As will now be shown, such a field can be constructed for any P. We denote it by $P(x)$ (the unknown is in the parentheses) and call it *the field of rational fractions* over the field P.

First assume that the ring $P[x]$ is already a subring of some field Q. If $f(x)$ and $g(x)$ are arbitrary polynomials from $P[x]$, and $g(x) \neq 0$, then there is, in the field Q, a uniquely defined element equal to the quotient obtained by the division of $f(x)$ by $g(x)$. Denoting this element by $\frac{f(x)}{g(x)}$, as is the usual way in the case of a field, we can write the following equation on the basis of the definition of a quotient:

$$f(x) = g(x) \cdot \frac{f(x)}{g(x)} \tag{1}$$

where the product is to be understood in the sense of multiplication in the field Q. It may happen that some quotients $\frac{f(x)}{g(x)}$ and $\frac{\varphi(x)}{\psi(x)}$ are one and the same element of Q. The condition for this is the ordinary condition of equality of fractions:

$\frac{f(x)}{g(x)} = \frac{\varphi(x)}{\psi(x)}$ *if and only if* $f(x)\psi(x) = \varphi(x)g(x)$.

Indeed, if $\frac{f(x)}{g(x)} = \frac{\varphi(x)}{\psi(x)} = \alpha$, then, by (1),

$$f(x) = g(x)\alpha, \quad \varphi(x) = \psi(x)\alpha$$

whence

$$f(x)\psi(x) = g(x)\psi(x)\alpha = g(x)\varphi(x)$$

Conversely, if $f(x)\psi(x) = g(x)\varphi(x) = u(x)$ in the sense of multiplication in the ring $P[x]$, then, passing to the field Q, we obtain

the equalities

$$\frac{f(x)}{g(x)} = \frac{u(x)}{g(x)\psi(x)} = \frac{\varphi(x)}{\psi(x)}$$

Furthermore, it is easy to see that the sum and product of any elements of Q, which are quotients of polynomials in $P[x]$, can again be represented in the form of such quotients, and the ordinary rules of addition and multiplication of fractions hold true:

$$\frac{f(x)}{g(x)} + \frac{\varphi(x)}{\psi(x)} = \frac{f(x)\psi(x) + g(x)\varphi(x)}{g(x)\psi(x)}, \tag{2}$$

$$\frac{f(x)}{g(x)} \cdot \frac{\varphi(x)}{\psi(x)} = \frac{f(x)\cdot\varphi(x)}{g(x)\cdot\psi(x)} \tag{3}$$

Indeed, multiplying both sides of these equations by the product $g(x)\psi(x)$ and applying (1), we get equalities which hold true in the ring $P[x]$. The validity of (2) and (3) now follows from the fact that, thanks to the absence of zero divisors in the field Q, both sides of each of the resulting equalities may be reduced by a nonzero element $g(x)\psi(x)$ without spoiling the equalities.

These preliminary remarks suggest the path we should take in constructing the field $P(x)$. Suppose we have an arbitrary field P and over it a polynomial ring $P[x]$. With every ordered pair of polynomials $f(x)$, $g(x)$, where $g(x) \neq 0$, we associate the symbol $\frac{f(x)}{g(x)}$, called a *rational fraction with numerator* $f(x)$ *and denominator* $g(x)$. We stress the fact that this is only a symbol corresponding to the given pair of polynomials, since, generally speaking, division of polynomials in the ring $P[x]$ itself is impossible, and so far the ring $P[x]$ is not contained in any field. Even if $g(x)$ is a divisor of $f(x)$, the new symbol $\frac{f(x)}{g(x)}$ should for the time being be distinguished from the polynomial obtained as the quotient in the division of $f(x)$ by $g(x)$.

We now call the rational fractions $\frac{f(x)}{g(x)}$ and $\frac{\varphi(x)}{\psi(x)}$ *equal*,

$$\frac{f(x)}{g(x)} = \frac{\varphi(x)}{\psi(x)} \tag{4}$$

if in the ring $P[x]$ we have the equality $f(x)\psi(x) = g(x)\varphi(x)$. It is obvious that any fraction is equal to itself and that if one fraction is equal to another, then the second one is equal to the first. Let us prove the *transitive* property of this concept of equality. We are given equalities (4) and

$$\frac{\varphi(x)}{\psi(x)} = \frac{u(x)}{v(x)} \tag{5}$$

From the equalities

$$f(x)\psi(x) = g(x)\varphi(x), \qquad \varphi(x)v(x) = \psi(x)u(x)$$

equivalent to them in the ring $P[x]$ it follows that

$$f(x)\, v(x)\, \psi(x) = g(x)\, \varphi(x)\, v(x) = g(x)\, u(x)\, \psi(x)$$

and therefore, after cancelling out the nonzero (as the denominator of one of the fractions) polynomial $\psi(x)$, we get

$$f(x)\, v(x) = g(x)\, u(x)$$

whence, by the definition of the equality of fractions,

$$\frac{f(x)}{g(x)} = \frac{u(x)}{v(x)}$$

This completes the proof.

Now let us combine into one class all fractions equal to some one given fraction, and therefore (by virtue of the transitivity of the equality) equal among themselves. If one class has even a single fraction not contained in another class, then, as follows from the transitivity of the equality, these two classes do not have a single element in common.

Thus, the collection of all rational fractions written by means of polynomials from the ring $P[x]$ breaks up into disjoint classes of fractions equal among themselves. We would now like to define algebraic operations in this set of classes of equal fractions so that it becomes a field. To do this, we will define operations on rational fractions and will each time verify that the replacement of summands (or factors) by fractions equal to them replaces the sum (or product) also by an equal fraction. This will enable us to speak of the sum and product of classes of equal fractions.

First, let us make the following remark which will be used repeatedly in what follows. *A rational fraction becomes an equal fraction if its numerator and denominator are multiplied by one and the same nonzero polynomial, or reduced by any common factor.* Indeed,

$$\frac{f(x)}{g(x)} = \frac{f(x)\, h(x)}{g(x)\, h(x)}$$

since in the ring $P[x]$

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

We define the *addition* of rational fractions by formula (2), since from $g(x) \neq 0$ and $\psi(x) \neq 0$ it follows that $g(x)\,\psi(x) \neq 0$, the right member of this formula is indeed a rational fraction. Furthermore, if it is given that

$$\frac{f(x)}{g(x)} = \frac{f_0(x)}{g_0(x)}, \quad \frac{\varphi(x)}{\psi(x)} = \frac{\varphi_0(x)}{\psi_0(x)}$$

that is,

$$f(x)\, g_0(x) = g(x)\, f_0(x), \quad \varphi(x)\, \psi_0(x) = \psi(x)\, \varphi_0(x) \qquad (6)$$

then, by multiplying both members of the first of the equalities (6) by $\psi(x)\,\psi_0(x)$, both members of the second equality by $g(x)\,g_0(x)$ and then adding these equalities termwise, we obtain

$$[f(x)\,\psi(x)+g(x)\,\varphi(x)]\,g_0(x)\,\psi_0(x) = [f_0(x)\,\psi_0(x)+g_0(x)\,\varphi_0(x)]\,g(x)\,\psi(x)$$

which is equivalent to the equation

$$\frac{f(x)\,\psi(x)+g(x)\,\varphi(x)}{g(x)\,\psi(x)}=\frac{f_0(x)\,\psi_0(x)+g_0(x)\,\varphi_0(x)}{g_0(x)\,\psi_0(x)}$$

Thus, if we have two classes of equal fractions, the sum of any fraction of one class and any fraction of the other class is equal to any other such sum, that is to say, such sums lie in some definite third class. This class is called the *sum* of the two given classes.

The commutativity of this addition follows directly from (2); the associativity is proved as follows:

$$\left[\frac{f(x)}{g(x)}+\frac{\varphi(x)}{\psi(x)}\right]+\frac{u(x)}{v(x)}=\frac{f(x)\,\psi(x)+g(x)\,\varphi(x)}{g(x)\,\psi(x)}+\frac{u(x)}{v(x)}$$

$$=\frac{f(x)\,\psi(x)\,v(x)+g(x)\,\varphi(x)\,v(x)+g(x)\,\psi(x)\,u(x)}{g(x)\,\psi(x)\,v(x)}$$

$$=\frac{f(x)}{g(x)}+\frac{\varphi(x)\,v(x)+\psi(x)\,u(x)}{\psi(x)\,v(x)}=\frac{f(x)}{g(x)}+\left[\frac{\varphi(x)}{\psi(x)}+\frac{u(x)}{v(x)}\right]$$

From the definition of equality of fractions it is easy to derive that all fractions of the form $\frac{0}{g(x)}$, i.e., fractions with zero numerator, are equal and that they form a complete class of equal fractions. We call this class the *zero class* and we will prove that in our addition it plays the part of zero. Indeed, if we have an arbitrary fraction $\frac{\varphi(x)}{\psi(x)}$, then

$$\frac{0}{g(x)}+\frac{\varphi(x)}{\psi(x)}=\frac{0\cdot\psi(x)+g(x)\,\varphi(x)}{g(x)\,\psi(x)}=\frac{g(x)\,\varphi(x)}{g(x)\,\psi(x)}=\frac{\varphi(x)}{\psi(x)}$$

From the equation

$$\frac{f(x)}{g(x)}+\frac{-f(x)}{g(x)}=\frac{0}{g^2(x)}$$

the right side of which belongs to the zero class, it now follows that the class of fractions equal to the fraction $\frac{-f(x)}{g(x)}$ will be *opposite* to the class of fractions equal to the fraction $\frac{f(x)}{g(x)}$. From this, as we know, follows the validity of unique *subtraction*.

We define *multiplication* of rational fractions by formula (3): since $g(x)\,\psi(x)\neq 0$, the right member of this formula will indeed

be a rational fraction. Furthermore, if

$$\frac{f(x)}{g(x)}=\frac{f_0(x)}{g_0(x)}, \quad \frac{\varphi(x)}{\psi(x)}=\frac{\varphi_0(x)}{\psi_0(x)}$$

that is,

$$f(x)\,g_0(x)=g(x)\,f_0(x), \quad \varphi(x)\,\psi_0(x)=\psi(x)\,\varphi_0(x)$$

then, by multiplying out these latter equations termwise, we get

$$f(x)\,g_0(x)\,\varphi(x)\,\psi_0(x)=g(x)\,f_0(x)\,\psi(x)\,\varphi_0(x)$$

which is equivalent to the equation

$$\frac{f(x)\,\varphi(x)}{g(x)\,\psi(x)}=\frac{f_0(x)\,\varphi_0(x)}{g_0(x)\,\psi_0(x)}$$

Thus, by analogy with the above-defined sum of classes, we can speak of a *product* of classes of equal fractions.

The commutativity and associativity of this multiplication follow immediately from (3) and the validity of the distributive law is proved as follows:

$$\left[\frac{f(x)}{g(x)}+\frac{\varphi(x)}{\psi(x)}\right]\frac{u(x)}{v(x)}=\frac{f(x)\,\psi(x)+g(x)\,\varphi(x)}{g(x)\,\psi(x)}\cdot\frac{u(x)}{v(x)}$$

$$=\frac{[f(x)\,\psi(x)+g(x)\,\varphi(x)]\,u(x)}{g(x)\,\psi(x)\,v(x)}=\frac{f(x)\,\psi(x)\,u(x)+g(x)\,\varphi(x)\,u(x)}{g(x)\,\psi(x)\,v(x)}$$

$$=\frac{f(x)\,\psi(x)\,u(x)\,v(x)+g(x)\,\varphi(x)\,u(x)\,v(x)}{g(x)\,\psi(x)\,v^2(x)}=\frac{f(x)\,u(x)}{g(x)\,v(x)}+\frac{\varphi(x)\,u(x)}{\psi(x)\,v(x)}$$

$$=\frac{f(x)}{g(x)}\cdot\frac{u(x)}{v(x)}+\frac{\varphi(x)}{\psi(x)}\cdot\frac{u(x)}{v(x)}$$

It is easy to see that fractions of the type $\frac{f(x)}{f(x)}$, i.e., fractions whose numerators are equal to the denominators, are equal and constitute a separate class. This class is termed the *unit class* and in our multiplication plays the role of unity:

$$\frac{f(x)}{f(x)}\cdot\frac{\varphi(x)}{\psi(x)}=\frac{f(x)\,\varphi(x)}{f(x)\,\psi(x)}=\frac{\varphi(x)}{\psi(x)}$$

Finally, if the fraction $\frac{f(x)}{g(x)}$ does not belong to the zero class, i.e., $f(x)\neq 0$, then there is a fraction $\frac{g(x)}{f(x)}$. Since

$$\frac{f(x)}{g(x)}\cdot\frac{g(x)}{f(x)}=\frac{f(x)\,g(x)}{g(x)\,f(x)}$$

and the right member of this equality belongs to the unit class, the class of fractions equal to the fraction $\frac{g(x)}{f(x)}$ will be *inverse* to the class

of fractions equal to $\frac{f(x)}{g(x)}$. Whence follows the validity of unique *division.*

Thus, *the classes of equal rational fractions with coefficients from the field P constitute, in our definition of operations, a commutative field.* This is the desired field $P(x)$. Incidentally, we still have to prove that this field which we have constructed contains a subring isomorphic to the ring $P[x]$ and that every element of the field can be represented as a quotient of two elements of this subring.

If we associate with an arbitrary polynomial $f(x)$ from the ring $P[x]$ a class of rational fractions equal to the fraction $\frac{f(x)}{1}$ (among all these fractions there are of course fractions whose denominators are equal to unity), we obtain a one-to-one mapping of the ring $P[x]$ into the field we have constructed. Indeed, from the equality

$$\frac{f(x)}{1} = \frac{\varphi(x)}{1}$$

it would follow that $f(x) \cdot 1 = 1 \cdot \varphi(x)$, that is to say, $f(x) = \varphi(x)$. This mapping will even be isomorphic, as the following equations show:

$$\frac{f(x)}{1} + \frac{g(x)}{1} = \frac{f(x) \cdot 1 + g(x) \cdot 1}{1^2} = \frac{f(x) + g(x)}{1},$$

$$\frac{f(x)}{1} \cdot \frac{g(x)}{1} = \frac{f(x) \cdot g(x)}{1}$$

Thus, *the classes of fractions equal to fractions of form* $\frac{f(x)}{1}$ *constitute, in our field, a subring that is isomorphic to the ring* $P[x]$. The fraction $\frac{f(x)}{1}$ can therefore be denoted simply as $f(x)$. And finally, since for $g(x) \neq 0$, the class of fractions equal to the fraction $\frac{1}{g(x)}$ is the inverse of the class of fractions equal to the fraction $\frac{g(x)}{1}$, it follows from the equality

$$\frac{f(x)}{1} \cdot \frac{1}{g(x)} = \frac{f(x)}{g(x)}$$

that *all elements of our field may be considered* (in the sense of operations defined in this field) *to be quotients of polynomials of the ring* $P[x]$.

Over an arbitrary field P we constructed the field of rational fractions $P(x)$. Using this same method, we can construct the field of rational numbers by taking the ring of integers in place of the ring of polynomials. Combining these two cases and using the same kind of method, we could prove a theorem asserting that, generally, any commutative ring without divisors of zero is a subring of some field.

CHAPTER 11

POLYNOMIALS IN SEVERAL UNKNOWNS

51. The Ring of Polynomials in Several Unknowns

One often has to consider polynomials that depend on two, three, and, generally, several unknowns. In the first chapters of this book we studied linear and quadratic forms, which are examples of such polynomials. Generally speaking, *a polynomial* $f(x_1, x_2, \ldots, x_n)$ *in* n *unknowns* $x_1, x_2, \ldots, x_n$ *over some field* P is the sum of a finite number of terms of the form $x_1^{k_1}, x_2^{k_2}, \ldots, x_n^{k_n}$, where all $k_i \geqslant 0$, with coefficients from the field P. It is assumed, quite naturally, that the polynomial $f(x_1, x_2, \ldots, x_n)$ does not contain like terms and that only terms with nonzero coefficients are considered. Two polynomials in n unknowns, $f(x_1, x_2, \ldots, x_n)$ and $g(x_1, x_2, \ldots, x_n)$ are called *equal* (or *identically equal*) if the coefficients of like terms are equal.

If a polynomial $f(x_1, x_2, \ldots, x_n)$ is given over a field P, then *its degree with respect to the unknown* x_i, $i = 1, 2, \ldots, n$, is the highest exponent with which x_i appears in the terms of the polynomial. By chance, the power may be 0, which means that although f is considered a polynomial in n unknowns $x_1, x_2, \ldots, x_i, \ldots, x_n$, the unknown x_i does not actually appear in the notation.

On the other hand, if we call the number $k_1 + k_2 + \ldots + k_n$ (that is, the sum of the exponents of the unknowns) the *degree of the term*

$$x_1^{k_1} x_2^{k_2} \ldots x_n^{k_n}$$

then the *degree of the polynomial* $f(x_1, x_2, \ldots, x_n)$ (that is, the degree of the unknowns taken together) is the highest degree of its terms. In particular, as in the case of one unknown, only nonzero elements from the field P will be polynomials of degree zero. On the other hand, as in the case of polynomials in one unknown, zero will be the only polynomial in n unknowns whose degree is not defined. Of course, a polynomial can in the general case contain several highest-

degree terms and therefore one cannot speak of the highest-degree term of a polynomial.

The operations of addition and multiplication are defined as follows for polynomials in n unknowns over a field P. The *sum* of the polynomials $f(x_1, x_2, \ldots, x_n)$ and $g(x_1, x_2, \ldots, x_n)$ is a polynomial whose coefficients are obtained by adding the corresponding coefficients of the polynomials f and g; if some term occurs in only one of the polynomials f, g, then its coefficient in the other polynomial is naturally taken to be zero. The product of two "monomials" is defined by the equation

$$ax_1^{k_1}x_2^{k_2}\ldots x_n^{k_n}\cdot bx_1^{l_1}x_2^{l_2}\ldots x_n^{l_n}=(ab)\,x_1^{k_1+l_1}x_2^{k_2+l_2}\ldots x_n^{k_n+l_n}$$

after which the *product* of the polynomials $f(x_1, x_2, \ldots, x_n)$ and $g(x_1, x_2, \ldots, x_n)$ is defined as the result of a termwise multiplication and subsequent collecting of like terms.

Given this definition of operations, the collection of polynomials in n unknowns over the field P becomes a commutative ring, which does not contain divisors of zero. Indeed, for $n = 1$ our definitions coincide with those which were given in Sec. 20 for the case of polynomials in one unknown. Let it already be proved that the polynomials in $n-1$ unknowns $x_1, x_2, \ldots, x_{n-1}$ with coefficients from the field P constitute a ring without divisors of zero. Any polynomial in n unknowns $x_1, x_2, \ldots, x_{n-1}, x_n$ may be uniquely represented as a polynomial in the unknown x_n with coefficients which are polynomials in $x_1, x_2, \ldots, x_{n-1}$; conversely, any polynomial in x_n with coefficients from the ring of polynomials in $x_1, x_2, \ldots, x_{n-1}$ over the field P may of course be regarded as a polynomial over this same field P with respect to the entire collection of unknowns $x_1, x_2, \ldots, x_{n-1}, x_n$. It may readily be verified that the one-to-one correspondence we have obtained between the polynomials in n unknowns and the polynomials in one unknown over the ring of polynomials in $n-1$ unknowns is isomorphic with respect to the operations of addition and multiplication. The assertion being proved follows now from the fact that polynomials in one unknown over the ring of polynomials in $n-1$ unknowns themselves constitute a ring, and, as a ring of polynomials in one unknown over the ring without zero divisors, it does not itself contain any divisors of zero (see Sec. 47).

Consequently, we have proved *the existence of a ring of polynomials in n unknowns over the field P*. This ring is denoted by the symbol $P[x_1, x_2, \ldots, x_n]$.

The following considerations permit regarding the ring of polynomials in n unknowns from a somewhat different angle. Let a field P be contained in some commutative ring L as a subring. In L take n elements $\alpha_1, \alpha_2, \ldots, \alpha_n$ and find the minimal subring L' of the ring L which contains these elements and also the entire field P, that is, the subring obtained by *adjoining* the elements $\alpha_1, \alpha_2, \ldots, \alpha_n$

to the field P. The subring L' consists of all elements of the ring L which are expressed in terms of the elements $\alpha_1, \alpha_2, \ldots, \alpha_n$ and the elements of the field P by means of addition, subtraction and multiplication. It is easy to see that what we have are precisely those elements of the ring L which may be written (with the aid of the operations occurring in L) in the form of polynomials in $\alpha_1, \alpha_2, \ldots, \alpha_n$ with coefficients from P; these elements, being elements of the ring L, will add and multiply precisely in accord with the rules of addition and multiplication of polynomials in n unknowns.

Of course, speaking generally, a given element β of the subring L' will possess many different notations in the form of a polynomial in $\alpha_1, \alpha_2, \ldots, \alpha_n$ with coefficients from the field P. If for any β in L' such a notation is unique, i.e., if the different polynomials in $\alpha_1, \alpha_2, \ldots, \alpha_n$ are distinct elements of the ring L' (and, hence, of the ring L), then the system of elements $\alpha_1, \alpha_2, \ldots, \alpha_n$ is called *algebraically independent* over the field P, otherwise it is *algebraically dependent.** From this we can draw the following conclusion.

If the field P is a subring of a commutative ring L and if the system of elements $\alpha_1, \alpha_2, \ldots, \alpha_n$ of L is algebraically independent over P, then the subring L' of the ring L generated by adjoining to P the elements $\alpha_1, \alpha_2, \ldots, \alpha_n$ is isomorphic to the polynomial ring $P[x_1, x_2, \ldots, x_n]$.

Of the other properties of the ring $P[x_1, x_2, \ldots, x_n]$ of polynomials in n unknowns we indicate the following: this ring may be included in the *field $P(x_1, x_2, \ldots, x_n)$ of rational fractions in n unknowns over the field P*. Every element of this field can be written as $\frac{f}{g}$, where f and g are polynomials of the ring $P[x_1, x_2, \ldots, x_n]$; then $\frac{f}{g} = \frac{\varphi}{\psi}$ if and only if $f\psi = g\varphi$. Addition and multiplication of these rational fractions is performed by the rules which, as indicated in Sec. 45, hold true for quotients in any field. The existence proof of the field $P(x_1, x_2, \ldots, x_n)$ is carried out just as it was in Sec. 50 for the case $n = 1$.

We can construct a theory of divisibility for polynomials in several unknowns that generalizes the theory of divisibility for polynomials in one unknown, which we studied in Chapters 5 and 10. However, since we do not intend to go into a detailed study of the ring of polynomials in several unknowns, we will confine ourselves to the problem of factoring a polynomial into irreducible factors.

First let us introduce the following concept: if all terms of a polynomial $f(x_1, x_2, \ldots, x_n)$ have one and the same degree s, then

---

* The appropriate concepts for the case of $n = 1$ were introduced in Sec. 47: there, an element α, algebraically independent over the field P in the sense of the foregoing definition, was called transcendental over P, otherwise it was algebraic over P.

it is called a *homogeneous polynomial* or, briefly, a *form* of degree s; we are acquainted with *linear* and *quadratic forms*, and we could consider *cubic forms*, all terms of which are of degree 3 in the unknowns taken together, etc. Any polynomial in n unknowns can be uniquely represented as a sum of several forms in these unknowns, the latter having various degrees. To obtain the desired representation, all we need to do is combine all terms of the same degree. For example, a polynomial of degree four $f(x_1, x_2, x_3) = 3x_1x_2^2 - 7x_1^2x_3^2 + x_2 - 5x_1x_2x_3 + + x_1^4 - 2x_3 - 6 + x_3^3$ is the sum of the quartic form $x_1^4 - 7x_1^2x_3^2$, the cubic form $3x_1x_2^2 - 5x_1x_2x_3 + x_3^3$, the linear form $x_2 - 2x_3$ and the constant term (a form of degree zero) -6.

Let us now prove the following theorem.

The degree of a product of two nonzero polynomials in n unknowns is equal to the sum of the degrees of the polynomials.

First suppose that we have the forms $\varphi(x_1, x_2, \ldots, x_n)$ of degree s and $\psi(x_1, x_2, \ldots, x_n)$ of degree t. The product of any term of the form φ by any term of the form ψ will obviously have the degree $s + t$, and so the product $\varphi\psi$ will be a form of degree $s + t$, since collecting like terms cannot make all the coefficients of this product vanish due to the absence of divisors of zero in the ring $P[x_1, x_2, \ldots \ldots, x_n]$.

If we are now given arbitrary polynomials $f(x_1, x_2, \ldots, x_n)$ and $g(x_1, x_2, \ldots, x_n)$ of degrees s and t, respectively, then, by representing each of them as a sum of forms of different degrees, we get

$$f(x_1, x_2, \ldots, x_n) = \varphi(x_1, x_2, \ldots, x_n) + \ldots,$$
$$g(x_1, x_2, \ldots, x_n) = \psi(x_1, x_2, \ldots, x_n) + \ldots$$

where φ and ψ are, respectively, forms of degrees s and t, and the dots stand for sums of forms of lower degrees. Then

$$fg = \varphi\psi + \ldots$$

By what has been proved, the form $\varphi\psi$ is of degree $s + t$, and since all terms replaced by dots are of lower degree, the degree of the product fg will be equal to $s + t$. The theorem is proved.

The polynomial φ is called the *divisor* of the polynomial f, and f is the *dividend which is divided by* φ, if in the ring $P[x_1, x_2, \ldots, x_n]$ there is a polynomial ψ such that $f = \varphi\psi$. It is easy to see that the divisibility properties I-IX (Sec. 21) are preserved in this general case as well. A polynomial f of degree k, $k \geqslant 1$ is called *reducible* over a field P if it can be decomposed into a product of polynomials from the ring $P[x_1, x_2, \ldots, x_n]$ whose degrees are less than k. Otherwise it is an *irreducible* polynomial.

Any polynomial in the ring $P[x_1, x_2, \ldots, x_n]$ having a nonzero degree can be decomposed into a product of irreducible factors. This decomposition (factorization) is unique to within factors of degree zero.

This theorem generalizes the corresponding results of Sec. 48 which refer to polynomials in one unknown. The first assertion is proved by repeating exactly the reasoning of Sec. 48. The proof of the second assertion is much more difficult. Before attempting it, we note that from the second assertion of this theorem there follows a corollary: *if the product of two polynomials f and g from the ring* $P[x_1, x_2, \ldots, x_n]$ *is divisible by an irreducible polynomial p, then at least one of these polynomials is divisible by p.* This is so, for otherwise we would have, for the product fg, two decompositions into irreducible factors, one of which contains p and the other does not.

Suppose the theorem has been proved for polynomials in n unknowns and we wish to prove it for a polynomial in $n + 1$ unknowns $x, x_1, x_2, \ldots, x_n$. Write this polynomial as $\varphi(x)$. Its coefficients will consequently be polynomials in $x_1, x_2, \ldots, x_n$. For these coefficients the theorem has already been proved, that is to say, each of them can be uniquely decomposed into a product of irreducible factors. Let us call $\varphi(x)$ a *primitive* polynomial (more exactly, *primitive over the ring* $P[x_1, x_2, \ldots, x_n]$), if its coefficients do not contain a single common irreducible factor, that is to say, are all relatively prime, and let us prove the following lemma (Gauss' lemma).

The product of two primitive polynomials is itself primitive.

Indeed, suppose we have the primitive polynomials

$$f(x) = a_0x^k + a_1x^{k-1} + \ldots + a_ix^{k-i} + \ldots + a_k,$$

$$g(x) = b_0x^l + b_1x^{l-1} + \ldots + b_jx^{l-j} + \ldots + b_l$$

with coefficients from the ring $P[x_1, x_2, \ldots, x_n]$ and let

$$f(x)\,g(x) = c_0x^{k+l} + c_1x^{k+l-1} + \ldots + c_{i+j}x^{k+l-(i+j)} + \ldots + c_{k+l}$$

If this product is not primitive, then the coefficients $c_0, c_1, \ldots, c_{k+l}$ will have a common irreducible factor $p = p(x_1, x_2, \ldots, x_n)$. Since all the coefficients of the primitive polynomial $f(x)$ cannot be divisible by p, let the coefficient a_i be the first that is not divisible by p; similarly, by b_j denote the first coefficient of the polynomial $g(x)$ that is not divisible by p. Multiplying $f(x)$ and $g(x)$ termwise and collecting terms in $x^{k+l-(i+j)}$, we get

$$c_{i+j} = a_ib_j + a_{i-1}b_{j+1} + a_{i-2}b_{j+2} + \ldots + a_{i+1}b_{j-1} + a_{i+2}b_{j-2} + \ldots$$

The left member is divisible by the irreducible polynomial p. All terms of the right member (except the first) are also definitely divisible by p. Indeed, by the conditions imposed on the choice of i and j, all coefficients $a_{i-1}, a_{i-2}, \ldots$, and also $b_{j-1}, b_{j-2}, \ldots$ are divisible by p. From this it follows that the product a_ib_j is also divisible by p and therefore, as noted above, at least one of the polynomials a_i, b_j must be divisible by p, which however is not the case. This

completes the proof of the lemma, under the assumption that the fundamental theorem for polynomials in n unknowns holds true.

As we know, the ring $P[x_1, x_2, \ldots, x_n]$ is contained in the field of rational fractions $P(x_1, x_2, \ldots, x_n)$ which we will denote by Q:

$$Q = P(x_1, x_2, \ldots, x_n)$$

Let us consider the polynomial ring $Q[x]$. If the polynomial $\varphi(x)$ belongs to this ring, then each coefficient of it can be represented as a quotient of polynomials from the ring $P[x_1, x_2, \ldots, x_n]$. Taking out the common denominator of these quotients and then removing the common factors from the numerators, we can represent $\varphi(x)$ as

$$\varphi(x) = \frac{a}{b} f(x)$$

Here, a and b are polynomials of the ring $P[x_1, x_2, \ldots, x_n]$ and $f(x)$ is a polynomial in x with coefficients from $P[x_1, x_2, \ldots, x_n]$; it is even a primitive polynomial since its coefficients do not have common factors.

In this way, we associate with every polynomial $\varphi(x)$ of the ring $Q[x]$ a primitive polynomial $f(x)$. *For the given polynomial* $\varphi(x)$, *the polynomial* $f(x)$ *is defined uniquely to within a nonzero factor in the field* P. Indeed, let

$$\varphi(x) = \frac{a}{b} f(x) = \frac{c}{d} g(x)$$

where $g(x)$ is again a primitive polynomial. Then

$$adf(x) = bcg(x)$$

Thus, ad and bc are obtained by taking out all common factors from the coefficients of one and the same polynomial over the ring $P[x_1, x_2, \ldots, x_n]$. Whence it follows, due to the validity, in this ring (on the induction hypothesis), of the unique factorization theorem, that ad and bc can differ only by a factor of degree zero. Hence, the primitive polynomials $f(x)$ and $g(x)$ differ by the same factor.

The product of two polynomials from the ring $Q[x]$ *is associated with the product of the primitive polynomials corresponding to them.* Indeed, if

$$\varphi(x) = \frac{a}{b} f(x), \quad \psi(x) = \frac{c}{d} g(x)$$

where $f(x)$ and $g(x)$ are primitive polynomials, then

$$\varphi(x)\psi(x) = \frac{ac}{bd} f(x) g(x)$$

But, as was proved above, the product $f(x) g(x)$ is a primitive polynomial.

Furthermore, note that *if the polynomial* $\varphi(x)$ *from the ring* $Q[x]$ *is irreducible over the field* Q, *then the corresponding primitive polynomial* $f(x)$, *regarded as a polynomial in* $x, x_1, x_2, \ldots, x_n$, *is also irreducible, and conversely.* Indeed, if the polynomial f is reducible, $f = f_1 f_2$, then both factors must contain the unknown x, since otherwise the polynomial f would not be primitive, whence follows the decomposition of the polynomial $\varphi(x)$ over the field Q:

$$\varphi(x) = \frac{a}{b} f(x) = \left(\frac{a}{b} f_1\right) f_2$$

Conversely, if the polynomial $\varphi(x)$ is reducible over Q, $\varphi(x) = \varphi_1(x)\,\varphi_2(x)$, then the primitive polynomials $f_1(x)$ and $f_2(x)$, corresponding to the polynomials $\varphi_1(x)$ and $\varphi_2(x)$, will both contain x, but their product, as was proved above, is equal to $f(x)$ (to within a factor from the field P).

Now let us take a primitive polynomial f and factor it into irreducible factors, $f = f_1 \cdot f_2 \ldots f_k$. Not only must all these factors contain the unknown x, they will even be primitive polynomials, for otherwise the polynomial f would not be primitive. *This factorization of the primitive polynomial* f *is unique to within factors from the field* P. True enough, due to the preceding lemma, we can regard this factorization as a factorization of $f(x)$ into irreducible factors over the field Q, but we already know of the uniqueness of factorization of polynomials in one unknown over some field; this uniqueness occurs to within factors from Q. However, in our case, due to the primitivity of all factors f_i, it will be to within factors from P.

After these lemmas, proved by induction, the proof of our fundamental theorem does not present any difficulties. Indeed, any irreducible polynomial in the ring $P[x, x_1, x_2, \ldots, x_n]$ will either be an irreducible polynomial from the ring $P[x_1, x_2, \ldots, x_n]$ or an irreducible primitive polynomial. From this it follows that if we have some factorization of the polynomial $\varphi(x, x_1, x_2, \ldots, x_n)$ into irreducible factors, then, by combining factors, we can represent φ as

$$\varphi(x, x_1, x_2, \ldots, x_n) = a(x_1, x_2, \ldots, x_n) f(x, x_1, x_2, \ldots, x_n)$$

where a is independent of x, and f is a primitive polynomial. However, we know that this factorization of φ is unique to within factors from P. On the other hand, since for the polynomial a in n unknowns the uniqueness of factorization into irreducible factors holds by the induction hypothesis, and, for the primitive polynomial f, was proved in the preceding lemma, the proof of our theorem for the case of $n + 1$ unknowns is also complete.

An interesting corollary stems from the lemmas proved above: *if a polynomial* $\varphi(x)$ *with coefficients in* $P[x_1, x_2, \ldots, x_n]$ *is reducible over the field* $Q = P(x_1, x_2, \ldots, x_n)$ *then it can be factored into factors*

dependent on x and having, as coefficients, polynomials from the ring $P[x_1, x_2, \ldots, x_n]$. Indeed, if to the polynomial $\varphi(x)$ there corresponds a primitive polynomial $f(x)$, that is, $\varphi(x) = af(x)$, then, as we know, the factorability of $f(x)$ follows from the factorability of $\varphi(x)$. But this latter fact leads to the factorization of $\varphi(x)$ over the ring $P[x_1, x_2, \ldots, x_n]$.

In contrast to the case of polynomials in one unknown, which, as we know from Sec. 49, can be factored into linear factors over an appropriately chosen extension of the base field under consideration, *there exist over any field P absolutely irreducible polynomials of arbitrary degree in several* (two or more) *unknowns*, that is to say, polynomials that remain irreducible under any extension of the field.

Such, for instance, is the polynomial

$$f(x, y) = \varphi(x) + y$$

where $\varphi(x)$ is an arbitrary polynomial in one unknown over the field P. Indeed, if there were a factorization

$$f(x, y) = g(x, y)\, h(x, y)$$

in some extension $\overline{P}$ of the field P, then, by writing g and h in terms of powers of y, we would have, say,

$$g(x, y) = a_0(x)\, y + a_1(x), \quad h(x, y) = b_0(x)$$

that is, h is not dependent on y; and then, because $a_0(x)\, b_0(x) = 1$, we would have that $b_0(x)$ has degree 0, i.e., h is not dependent on x either.

Alphabetical order of the terms of a polynomial. For polynomials in one unknown, we have two natural ways of arranging the terms — as descending and ascending powers of the unknown. This is not possible for polynomials in several unknowns. If we have a polynomial of degree five in three unknowns,

$$f(x_1, x_2, x_3) = x_1x_2^2x_3^2 + x_1^4x_3 + x_2^3x_3^2 + x_1^2x_2x_3^2$$

it may also be written as

$$f(x_1, x_2, x_3) = x_1^4x_3 + x_1^2x_2x_3^2 + x_1x_2^2x_3^2 + x_2^3x_3^2$$

and there is no reason to prefer one notation to the other. There is, however, a very definite way of ordering the terms of a polynomial in several unknowns; it depends incidentally on the manner in which the unknowns are numbered. For polynomials in one unknown it reduces to ordering the terms in descending powers of the unknown. It is known as the *alphabetical* method.

Suppose we have a polynomial $f(x_1, x_2, \ldots, x_n)$ in the ring $P[x_1, x_2, \ldots, x_n]$ and two distinct terms of the polynomial

$$x_1^{k_1}x_2^{k_2} \ldots x_n^{k_n} \tag{1}$$

$$x_1^{l_1}x_2^{l_2} \ldots x_n^{l_n} \tag{2}$$

whose coefficients are certain nonzero elements of P. Since the terms (1) and (2) are distinct, at least one of the differences of the exponents on the unknowns

$$k_i - l_i, \qquad i = 1, 2, \ldots, n$$

is nonzero. Term (1) will be considered *higher* than term (2) [and term (2) *lower* than term (1)] if the first of these differences (nonzero) is positive, that is, if there is an i, $1 \leqslant i \leqslant n$, such that

$$k_1 = l_1, \quad k_2 = l_2, \ldots, k_{i-1} = l_{i-1}, \quad \text{but} \quad k_i > l_i$$

In other words, term (1) will be higher than term (2) if the exponent on x_1 in (1) is greater than in (2), or if these exponents are equal but the exponent on x_2 in (1) is greater than in (2), and so forth. It will readily be seen that from the fact that term (1) is higher than term (2) it does not follow that the degree of the former (all unknowns taken together) is greater than that of the latter: of the terms

$$x_1^3 x_2 x_3, \quad x_1 x_2^5 x_3^2$$

the first is higher though it is of lower degree.

It is obvious that of any two distinct terms of the polynomial $f(x_1, x_2, \ldots, x_n)$, one will be higher than the other. It is also easy to verify that if term (1) is higher than term (2), and (2), in turn, is higher than the term

$$x_1^{m_1} x_2^{m_2} \ldots x_n^{m_n} \tag{3}$$

that is, there exists a j, $1 \leqslant j \leqslant n$, such that

$$l_1 = m_1, \quad l_2 = m_2, \ldots, l_{j-1} = m_{j-1}, \quad \text{but} \quad l_j > m_j$$

then, irrespective of whether i is greater than, equal to, or less than j, term (1) will be higher than term (3). Thus, placing first that term which is higher, we get a definite ordering of the terms of the polynomial $f(x_1, x_2, \ldots, x_n)$, which is called alphabetical.

Thus, the polynomial

$$f(x_1, x_2, x_3, x_4) = x_1^4 + 3x_1^2 x_2^3 x_3 - x_1^2 x_2^3 x_4^2 + 5x_1 x_3 x_4^2 + 2x_2 + x_3^3 x_4 - 4$$

is arranged in alphabetical order.

In the alphabetical notation of the polynomial $f(x_1, x_2, \ldots, x_n)$ one of its terms will occupy first place, that is, will be higher than any of the others. This term is called the *highest term of the polynomial*, in the example given above, x_1^4 is the highest term. We will now prove a lemma concerning highest terms; it will be used in the proof of the fundamental theorem of the next section.

The highest term of a product of two polynomials in n unknowns is equal to the product of the highest terms of the factors.

Indeed, suppose we are multiplying the polynomials $f(x_1, x_2, \ldots, x_n)$ and $g(x_1, x_2, \ldots, x_n)$. If

$$ax_1^{k_1}x_2^{k_2}\ldots x_n^{k_n} \tag{4}$$

is the highest term of the polynomial $f(x_1, x_2, \ldots, x_n)$, and

$$a'x_1^{s_1}x_2^{s_2}\ldots x_n^{s_n} \tag{5}$$

is any other term of this polynomial, then there is an i, $1 \leqslant i \leqslant n$, such that

$$k_1 = s_1, \ldots, k_{i-1} = s_{i-1}, \quad k_i > s_i$$

If, on the other hand,

$$bx_1^{l_1}x_2^{l_2}\ldots x_n^{l_n} \tag{6}$$

$$b'x_1^{t_1}x_2^{t_2}\ldots x_n^{t_n} \tag{7}$$

are the highest term and any other term of the polynomial $g(x_1, x_2, \ldots, x_n)$, then there is a j, $1 \leqslant j \leqslant n$, such that

$$l_1 = t_1, \ldots, l_{j-1} = t_{j-1}, \quad l_j > t_j$$

Multiplying the terms (4) and (6) and also the terms (5) and (7), we get

$$abx_1^{k_1+l_1}x_2^{k_2+l_2}\ldots x_n^{k_n+l_n}, \tag{8}$$

$$a'b'x_1^{s_1+l_1}x_2^{s_2+l_2}\ldots x_n^{s_n+l_n} \tag{9}$$

It is easy to see, however, that term (8) is higher than term (9); if, say, $i \leqslant j$, then

$$k_1 + l_1 = s_1 + t_1, \ldots, k_{i-1} + l_{i-1} = s_{i-1} + t_{i-1} \text{ but}$$

$$k_i + l_i > s_i + t_i$$

since $k_i > s_i$, $l_i \geqslant t_i$. In the same way, we see that term (8) is higher than the product of the terms (4) and (7), and also higher than the product of the terms (5) and (6). Thus, term (8)—the product of the highest terms of the polynomials f and g—will be higher than all other terms obtained by termwise multiplication of the polynomials f and g, and so this term does not vanish when we collect terms; that is to say, it remains the highest term in the product fg.

52. Symmetric Polynomials

Conspicuous among polynomials in several unknowns are those that remain unchanged no matter what rearrangements of the unknowns occur. Thus, all unknowns appear in these polynomials in symmetric fashion, whence the name *symmetric polynomials* (or *symmetric functions*). Among the simplest examples are the sum of

all unknowns $x_1 + x_2 + \ldots + x_n$, the sum of the squares of the unknowns $x_1^2 + x_2^2 + \ldots + x_n^2$, the product of the unknowns $x_1 x_2 \ldots x_n$, and so on. Since any permutation on n symbols can be represented in the form of a product of transpositions (see Sec. 3), it is sufficient, when proving the symmetry of a polynomial, to verify that it remains unchanged under any transposition of two unknowns.

We shall now consider symmetric polynomials in n unknowns with coefficients from some field P. It is easy to see that *the sum, difference and product of two symmetric polynomials are symmetric*; that is to say, symmetric polynomials form a subring in the ring $P[x_1, x_2, \ldots, x_n]$ of all polynomials in n unknowns over the field P; this is called *the ring of symmetric polynomials in n unknowns over the field P*. It includes all elements of P (that is, all polynomials of degree zero and also zero), since they definitely do not change under any rearrangement of the unknowns. Any other symmetric polynomial invariably contains all n unknowns and even has one and the same degree with respect to them: if a symmetric polynomial $f(x_1, x_2, \ldots, x_n)$ has a term in which the unknown x_i appears with an exponent k, then it also has a term obtained from the first one by a transposition of the unknowns x_i and x_j, that is, one containing the unknown x_j to the same power k.

The following n symmetric polynomials in n unknowns are called *elementary symmetric polynomials*:

$$\left.\begin{aligned}
\sigma_1 &= x_1 + x_2 + \ldots + x_n,\\
\sigma_2 &= x_1x_2 + x_1x_3 + \ldots + x_{n-1}x_n,\\
\sigma_3 &= x_1x_2x_3 + x_1x_2x_4 + \ldots + x_{n-2}x_{n-1}x_n,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
\sigma_{n-1} &= x_1x_2\ldots x_{n-1} + x_1x_2\ldots x_{n-2}x_n + \ldots + x_2x_3\ldots x_n,\\
\sigma_n &= x_1x_2\ldots x_n
\end{aligned}\right\} \quad (1)$$

These polynomials, whose symmetry is obvious, play a very great role in the theory of symmetric polynomials. They are suggested by the Vieta formulas (see Sec. 24) and so we can say that *the coefficients of a polynomial in one unknown, the leading coefficient of which is unity, will, to within sign, be elementary symmetric polynomials with respect to its roots.* This relationship between elementary symmetric polynomials and the Vieta formulas will be very essential for those applications of symmetric polynomials to the theory of polynomials in one unknown which justify their study.

Since symmetric polynomials in n unknowns $x_1, x_2, \ldots, x_n$ over the field P constitute a ring, the following assertions are obvious: we have a symmetric polynomial in the case of any positive integer power of any one of the elementary symmetric polynomials, also in the case of a product of such powers (taken with any coefficient of P),

and, finally, in the case of any sum of these products. In other words, *any polynomial in the elementary symmetric polynomials* $\sigma_1, \sigma_2, \ldots, \sigma_n$ *with coefficients from* P, *which polynomial is regarded as a polynomial in the unknowns* $x_1, x_2, \ldots, x_n$, *will be symmetric.* For example, set $n = 3$ and take the polynomial $\sigma_1\sigma_2 + 2\sigma_3$. Replacing σ_1, σ_2 and σ_3 by their expressions, we get

$$\sigma_1\sigma_2 + 2\sigma_3 = x_1^2x_2 + x_1^2x_3 + x_1x_2^2 + x_2^2x_3 + x_1x_3^2 + x_2x_3^2 + 5x_1x_2x_3$$

What we have on the right is obviously a symmetric polynomial in x_1, x_2, x_3.

An inversion of this result is the following **fundamental theorem on symmetric polynomials.**

Any symmetric polynomial in the unknowns $x_1, x_2, \ldots, x_n$ *over the field* P *is a polynomial in the elementary symmetric polynomials* $\sigma_1, \sigma_2, \ldots, \sigma_n$ *with coefficients belonging to* P.

Indeed, suppose we have the symmetric polynomial

$$f(x_1, x_2, \ldots, x_n)$$

and, in the alphabetical notation, let the highest term be

$$a_0x_1^{k_1}x_2^{k_2} \ldots x_n^{k_n} \tag{2}$$

The exponents on the unknowns in this term must satisfy the inequalities

$$k_1 \geqslant k_2 \geqslant \ldots \geqslant k_n \tag{3}$$

Indeed, suppose, for some i, we have $k_i < k_{i+1}$. However, since the polynomial $f(x_1, x_2, \ldots, x_n)$ is symmetric, it must contain the term

$$a_0x_1^{k_1}x_2^{k_2} \ldots x_i^{k_{i+1}}x_{i+1}^{k_i} \ldots x_n^{k_n} \tag{4}$$

which is obtained from term (2) by a transposition of the unknowns x_i and x_{i+1}. This is a contradiction, since term (4) is higher than term (2) alphabetically: the exponents on $x_1, x_2, \ldots, x_{i-1}$ coincide in both terms, but the exponent on x_i in term (4) is greater than in term (2).

Let us now take the following product of elementary symmetric polynomials [all exponents will be nonnegative because of inequalities (3)]:

$$\varphi_1 = a_0\sigma_1^{k_1-k_2}\sigma_2^{k_2-k_3} \ldots \sigma_{n-1}^{k_{n-1}-k_n}\sigma_n^{k_n} \tag{5}$$

This is a symmetric polynomial in the unknowns $x_1, x_2, \ldots, x_n$, and its highest term is equal to term (2). Indeed, the highest terms of the polynomials $\sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_n$ are equal, respectively, to $x_1, x_1x_2, x_1x_2x_3, \ldots, x_1x_2 \ldots x_n$, and since it was proved at the end of the preceding section that the highest term of a product is equal to the product of the highest terms of the factors, it follows

that the highest term of the polynomial φ_1 is

$$a_0 x_1^{k_1-k_2}(x_1x_2)^{k_2-k_3}(x_1x_2x_3)^{k_3-k_4}\ldots(x_1x_2\ldots x_{n-1})^{k_{n-1}-k_n}(x_1x_2\ldots x_n)^{k_n} = a_0 x_1^{k_1}x_2^{k_2}\ldots x_n^{k_n}$$

From this it follows that when we subtract φ_1 from f, the highest terms of these polynomials cancel out, that is, the highest term of the symmetric polynomial $f - \varphi_1 = f_1$ will be lower than the term (2). which is the highest one in f. Repeating this same procedure for the polynomial f_1, whose coefficients obviously belong to the field P. we get the equality

$$f_1 = \varphi_2 + f_2$$

where φ_2 is the product of the powers of elementary symmetric polynomials with a coefficient in P, and f_2 is a symmetric polynomial whose highest term is lower than the highest term in f_1, whence the equality

$$f = \varphi_1 + \varphi_2 + f_2$$

Continuing this process, we get $f_s = 0$ for some s and therefore arrive at an expression of f in the form of a polynomial in $\sigma_1, \sigma_2, \ldots, \sigma_n$ with coefficients in P:

$$f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{s} \varphi_i = \varphi(\sigma_1, \sigma_2, \ldots, \sigma_n)$$

Indeed, if this process were endless,* we would obtain an infinite sequence of symmetric polynomials:

$$f_1, f_2, \ldots, f_s, \ldots \tag{6}$$

and the highest term of each would be lower than the highest terms of the preceding polynomials, and all the more so lower than (2). However, if

$$bx_1^{l_1}x_2^{l_2}\ldots x_n^{l_n} \tag{7}$$

is the highest term of the polynomial f_s, then from the symmetry of this polynomial there follow the inequalities

$$l_1 \geqslant l_2 \geqslant \ldots \geqslant l_n \tag{8}$$

which are similar to the inequalities (3). On the other hand, since term (2) is higher than term (7), it follows that

$$k_1 \geqslant l_1 \tag{9}$$

* One must bear in mind that, generally speaking, the polynomial φ_s also contains terms not found in the polynomial f_{s-1} and therefore the transition from f_{s-1} to $f_s = f_{s-1} - \varphi_s$ is connected not only with eliminating certain terms from f_{s-1} but also with the appearance of new terms. Here, $s = 1, 2, \ldots$.

It is readily seen, however, that the systems of nonnegative integers $l_1, l_2, \ldots, l_n$ which satisfy the inequalities (8) and (9), may be chosen in only a finite number of ways. Indeed, even if we give up the requirement (8) and only assume that all l_i, $i = 1, 2, \ldots, n$, do not exceed k_1, then the choice of numbers l_i will be possible in only $(k_1 + 1)^n$ ways. Whence it follows that the sequence of polynomials (6) with strictly descending highest terms cannot be infinite.

This completes the proof of the theorem.

The above-indicated relationship between elementary symmetric polynomials and the Vieta formulas permits deriving the following important corollary from the fundamental theorem on symmetric polynomials.

Let $f(x)$ be a polynomial in one unknown over the field P having the leading coefficient unity. Then any symmetric polynomial (with coefficients from P) in the roots of the polynomial $f(x)$, which roots belong to some splitting field of the polynomial $f(x)$ over P, will be a polynomial (with coefficients from P) in the coefficients of the polynomial $f(x)$ and therefore will be an element of P.

The foregoing proof of the fundamental theorem also provides us with a practical method for finding the expressions of symmetric polynomials in terms of elementary polynomials. Let us first introduce the following notation: if

$$ax_1^{k_1}x_2^{k_2}\ldots x_n^{k_n} \tag{10}$$

is some product of powers of the unknowns $x_1, x_2, \ldots, x_n$ (some of the exponents may be equal to zero), then

$$S(ax_1^{k_1}x_2^{k_2}\ldots x_n^{k_n}) \tag{11}$$

will denote the sum of all terms obtained from (10) by all possible rearrangements of the unknowns. It is obvious that this will be a symmetric polynomial and homogeneous too, and that any symmetric polynomial in n unknowns containing the term (10) will also contain all the other terms of the polynomial (11). For example, $S(x_1) = \sigma_1$, $S(x_1x_2) = \sigma_2$, $S(x_1^2)$ is the sum of the squares of all the unknowns, etc.

Example. Express the symmetric polynomial $f = S(x_1^2x_2)$ in n unknowns in terms of the elementary symmetric polynomials.

Here, the highest term is $x_1^2x_2$ and therefore $\varphi_1 = \sigma_1^{2-1}\sigma_2 = \sigma_1\sigma_2$, that is,

$$\begin{aligned}\varphi_1 &= (x_1 + x_2 + \ldots + x_n)(x_1x_2 + x_1x_3 + \ldots + x_{n-1}x_n)\\ &= S(x_1^2x_2) + 3S(x_1x_2x_3)\end{aligned}$$

whence

$$f_1 = f - \varphi_1 = -3S(x_1x_2x_3) = -3\sigma_3$$

Therefore, $f = \varphi_1 + f_1 = \sigma_1\sigma_2 - 3\sigma_3$.

In more involved cases, it is advisable first to determine which terms can enter into the expression of the given polynomial via elementary polynomials,

and then to find the coefficients of these terms by the method of undetermined coefficients.

Example 1. Find an expression for the symmetric polynomial $f = S(x_1^2 x_2^2)$.

We know (see the proof of the fundamental theorem) that the terms of the desired polynomial $\varphi(\sigma_1, \sigma_2, \ldots, \sigma_n)$ are determined via the highest terms of the symmetric polynomials $f_1, f_2, \ldots$, these highest terms being lower than the highest term of the given polynomial f, that is, lower than $x_1^2 x_2^2$. We find all the products $x_1^{l_1} x_2^{l_2} \ldots x_n^{l_n}$ that satisfy the following conditions: (1) they are lower than the term $x_1^2 x_2^2$, (2) they can serve as the highest terms of symmetric polynomials, i.e., they satisfy the inequalities $l_1 \geqslant l_2 \geqslant \ldots \geqslant l_n$, (3) with respect to all unknowns taken together they have the degree 4 (since, as we know, all the polynomials $f_1, f_2, \ldots$ have the same degree as the homogeneous polynomial f). Writing out only appropriate combinations of exponents and indicating, alongside, those products of powers of σ which products are determined by them, we get the following table:

$$22000 \ldots \sigma_1^{2-2}\sigma_2^{2-0} = \sigma_2^2,$$

$$21100 \ldots \sigma_1^{2-1}\sigma_2^{1-1}\sigma_3^{1-0} = \sigma_1\sigma_3,$$

$$11110 \ldots \sigma_1^{1-1}\sigma_2^{1-1}\sigma_3^{1-1}\sigma_4^{1-0} = \sigma_4$$

Thus, the polynomial f has the form

$$f = \sigma_2^2 + A\sigma_1\sigma_3 + B\sigma_4$$

We set the coefficient of σ_2 equal to unity, since this term is determined by the highest term of the polynomial f and, as we know from the proof of the fundamental theorem, has the same coefficient. The coefficients A and B are found as follows.

Set $x_1 = x_2 = x_3 = 1$, $x_4 = \ldots = x_n = 0$. It is easy to see that for these values of the unknowns the polynomial f has the value 3, and the polynomials σ_1, σ_2, σ_3 and σ_4, the values of 3, 3, 1, and 0, respectively. Therefore,

$$3 = 9 + A \cdot 3 \cdot 1 + B \cdot 0$$

whence $A = -2$. Now put $x_1 = x_2 = x_3 = x_4 = 1$, $x_5 = \ldots = x_n = 0$. The values of the polynomials f, σ_1, σ_2, σ_3 and σ_4 will be 6, 4, 6, 4, 1, respectively. Therefore,

$$6 = 36 - 2 \cdot 4 \cdot 4 + B \cdot 1$$

whence $B = 2$. Thus, for f the desired expression is

$$f = \sigma_2^2 - 2\sigma_1\sigma_3 + 2\sigma_4$$

Example 2. Find the sum of the cubes of the roots of the polynomial

$$f(x) = x^4 + x^3 + 2x^2 + x + 1$$

To solve this problem, let us find the expression for the symmetric polynomial $S(x_1^3)$ in terms of the elementary symmetric polynomials. Applying the same method as in the preceding example, we get the table

$$3000 \ldots \sigma_1^3,$$

$$2100 \ldots \sigma_1\sigma_2,$$

$$1110 \ldots \sigma_3$$

and therefore

$$S(x_1^3) = \sigma_1^3 + A\sigma_1\sigma_2 + B\sigma_3$$

First assuming $x_1 = x_2 = 1$, $x_3 = \ldots = x_n = 0$, and then $x_1 = x_2 = x_3 = 1$, $x_4 = \ldots = x_n = 0$, we get $A = -3$, $B = 3$, that is,

$$S(x_1^3) = \sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3 \qquad (12)$$

To find the sum of the cubes of the roots of the given polynomial $f(x)$, it is necessary (because of the Vieta formulas) to replace, in the above-found expression, σ_1 by the coefficient of x^3 with sign reversed, that is, by -1, then to replace σ_2 by the coefficient of x^2, that is, by 2, and, finally, to replace σ_3 by the coefficient of x with sign reversed, i.e., by -1. Thus, the sum we are interested in (the sum of the cubes of the roots) is equal to

$$(-1)^3 - 3\cdot(-1)\cdot 2 + 3\cdot(-1) = 2$$

The reader can verify this result if he takes into account that $f(x)$ has as roots the numbers i, $-i$, $-\frac{1}{2} + i\frac{\sqrt{3}}{2}$ and $-\frac{1}{2} - i\frac{\sqrt{3}}{2}$. It is also obvious that the formula (12) does not depend on the given polynomial $f(x)$ and enables us to find the sum of the cubes of the roots of any polynomial.

The method, obtained in the proof of the fundamental theorem, for expressing a symmetric polynomial f in terms of the elementary polynomials leads to a very definite polynomial in $\sigma_1, \sigma_2, \ldots, \sigma_n$. It turns out that there is no way of obtaining a different expression for f in terms of $\sigma_1, \sigma_2, \ldots, \sigma_n$. This is indicated by the following uniqueness theorem.

Every symmetric polynomial has only a unique expression in the form of a polynomial in the elementary symmetric polynomials.

Here is the proof. If a symmetric polynomial $f(x_1, x_2, \ldots, x_n)$ over a field P had two distinct expressions in terms of $\sigma_1, \sigma_2, \ldots, \sigma_n$

$$f(x_1, x_2, \ldots, x_n) = \varphi(\sigma_1, \sigma_2, \ldots, \sigma_n) = \psi(\sigma_1, \sigma_2, \ldots, \sigma_n)$$

then the difference

$$\chi(\sigma_1, \sigma_2, \ldots, \sigma_n) = \varphi(\sigma_1, \sigma_2, \ldots, \sigma_n) - \psi(\sigma_1, \sigma_2, \ldots, \sigma_n)$$

would be a nonzero polynomial in $\sigma_1, \sigma_2, \ldots, \sigma_n$; that is, not all its coefficients would be zero, whereas replacing $\sigma_1, \sigma_2, \ldots, \sigma_n$ in this polynomial by their expressions in terms of $x_1, x_2, \ldots, x_n$ would lead to the zero of the ring $P[x_1, x_2, \ldots, x_n]$. It therefore remains to prove that if a polynomial $\chi(\sigma_1, \sigma_2, \ldots, \sigma_n)$ is different from zero, that is, has at least one nonzero coefficient, then the polynomial $g(x_1, x_2, \ldots, x_n)$ obtained from χ by replacing $\sigma_1, \sigma_2, \ldots, \sigma_n$ by their expressions in terms of $x_1, x_2, \ldots, x_n$,

$$\chi(\sigma_1, \sigma_2, \ldots, \sigma_n) = g(x_1, x_2, \ldots, x_n) \qquad (13)$$

is also nonzero.

If $a\sigma_1^{k_1}\sigma_2^{k_2}\ldots\sigma_n^{k_n}$ is one of the terms of the polynomial χ, $a \neq 0$, then after replacing all σ by their expressions (1), we get a polynomial in $x_1, x_2, \ldots, x_n$ whose highest term (in the sense of alphabetical ordering) is, as we already know from the proof of the fundamental theorem, the term

$$ax_1^{k_1}(x_1x_2)^{k_2}\ldots(x_1x_2\ldots x_n)^{k_n} = ax_1^{l_1}x_2^{l_2}\ldots x_n^{l_n}$$

where

$$\begin{aligned} l_1 &= k_1 + k_2 + \ldots + k_n, \\ l_2 &= \quad k_2 + \ldots + k_n, \\ &\ldots\ldots\ldots\ldots \\ l_n &= \quad k_n \end{aligned}$$

Whence

$$k_i = l_i - l_{i+1}, \quad k_n = l_n, \quad i = 1, 2, \ldots, n-1$$

That is to say, using the exponents $l_1, l_2, \ldots, l_n$, we can restore the exponents $k_1, k_2, \ldots, k_n$ of the initial term of the polynomial χ. Thus, distinct terms of the polynomial χ, which are regarded as polynomials in $x_1, x_2, \ldots, x_n$, have distinct highest terms.

Let us now consider all the terms of the polynomial χ: for each one of them let us find the highest term of its representation in the form of a polynomial in $x_1, x_2, \ldots, x_n$ and select that highest term which is highest in the alphabetical-ordering sense. As has been pointed out above, this term does not have any similar ones among the highest terms obtained from the other terms of the polynomial χ, and since, by hypothesis, it is higher than each of these highest terms, it is all the more so higher than the other terms obtained when replacing in the terms of the polynomial χ the elements $\sigma_1, \sigma_2, \ldots, \sigma_n$ by their expressions (1). We have thus found a term which, when passing from $\chi(\sigma_1, \sigma_2, \ldots, \sigma_n)$ to $g(x_1, x_2, \ldots, x_n)$, appears (with nonzero coefficient) only once and for this reason cannot be cancelled out with anything in any way. Whence it follows that not all coefficients of the polynomial $g(x_1, x_2, \ldots, x_n)$ are equal to zero, that is, this polynomial is not a zero element of the ring $P[x_1, x_2, \ldots, x_n]$. The proof is complete.

Evidently, this theorem could also be stated in the following manner.

A system of elementary symmetric polynomials $\sigma_1, \sigma_2, \ldots, \sigma_n$ regarded as elements of the polynomial ring $P[x_1, x_2, \ldots, x_n]$ is algebraically independent over the field P.

53. Symmetric Polynomials Continued

Remarks on the fundamental theorem. The proof of the fundamental theorem on symmetric polynomials given in the preceding section admits of a number of essential supplements to the statement of the theorem. We will make use of them in what follows. First of all, the coefficients of the polynomial $\varphi(\sigma_1, \sigma_2, \ldots, \sigma_n)$ which we found as an expression for the symmetric polynomial $f(x_1, x_2, \ldots, x_n)$ in terms of the elementary symmetric polynomials not only belong to the field P, but *are even expressed in terms of the coefficients of the*

polynomial f by means of addition and subtraction, i.e., they belong to the ring L generated by the coefficients of the polynomial f inside the field P.

True enough, all coefficients of the polynomial φ_1 [see formula (5) of the preceding section] in the unknowns $x_1, x_2, \ldots, x_n$ are, as will readily be seen, integral multiples of the coefficient a_0 of the highest term of the polynomial f and for this reason belong to the ring L. Let it be already proved that L contains all coefficients (in $x_1, x_2, \ldots, x_n$) of the polynomials $\varphi_1, \varphi_2, \ldots, \varphi_l$. Then the coefficients of the polynomial $f_l = f - \varphi_1 - \varphi_2 - \ldots - \varphi_l$ will also belong to L, and therefore L also contains all coefficients of the polynomial φ_{l+1} in $x_1, x_2, \ldots, x_n$.

On the other hand, *the degree of the polynomial* $\varphi(\sigma_1, \sigma_2, \ldots, \sigma_n)$ *with respect to* $\sigma_1, \sigma_2, \ldots, \sigma_n$ *taken together is equal to the degree of the polynomial* $f(x_1, x_2, \ldots, x_n)$ *with respect to each of the unknowns* x_i. Indeed, since (2) of Sec. 52, is the highest term of polynomial f, it follows that k_1 will be the degree of f in the unknown x_1, and therefore, by symmetry, in any other of the unknowns x_i as well. However, the degree of φ_1 with respect to σ jointly is, by (5) of Sec. 52, equal to the number

$$(k_1 - k_2) + (k_2 - k_3) + \ldots + (k_{n-1} - k_n) + k_n = k_1$$

Furthermore, since the leading term of the polynomial f_1 is lower than the leading term of the polynomial f, it follows that the degree of f_1 with respect to each one of the x_i will not exceed the degree of f with respect to each one of these unknowns. However, for f_1 the polynomial φ_2 plays the same role as φ_1 for f, and so the degree of φ_2 with respect to σ jointly is equal to the degree of f_1 with respect to each one of x_i; that is, it does not exceed k_1 and so on. Thus, likewise, the degree of $\varphi(\sigma_1, \sigma_2, \ldots, \sigma_n)$ does not exceed k_1. But since no φ_i with $i > 1$ can contain all $\sigma_1, \sigma_2, \ldots, \sigma_n$ to the same powers as φ_1, the degree of $\varphi(\sigma_1, \sigma_2, \ldots, \sigma_n)$ is exactly equal to k_1. Our assertion is thus proved.

Finally, let $a\sigma_1^{l_1}\sigma_2^{l_2} \ldots \sigma_n^{l_n}$ be one of the terms of the polynomial $\varphi(\sigma_1, \sigma_2, \ldots, \sigma_n)$. We give the name "*weight*" of this term to the number

$$l_1 + 2l_2 + \ldots + nl_n$$

that is, to the sum of the exponents multiplied by the indices of the corresponding σ_i. In other words, this is the degree of our term with respect to the unknowns $x_1, x_2, \ldots, x_n$ taken together, as follows from the theorem (proved in Sec. 51) on the degree of a product of polynomials. Then the following assertion holds true.

If, with respect to the totality of unknowns, a homogeneous symmetric polynomial $f(x_1, x_2, \ldots, x_n)$ *has degree s, then all terms of its expression* $\varphi(\sigma_1, \sigma_2, \ldots, \sigma_n)$ *via* σ *will have the same weight equal to s.*

Indeed, if (2) of Sec. 52 is the highest term of the homogeneous polynomial f, then

$$s = k_1 + k_2 + \ldots + k_n$$

However, the weight of the term φ_1 is, by (5) of Sec. 52, equal to

$$(k_1 - k_2) + 2(k_2 - k_3) + \ldots + (n-1)(k_{n-1} - k_n) + nk_n = k_1 + k_2 + k_3 + \ldots + k_n$$

That is, it is also equal to s. Furthermore, the polynomial $f_1 = f - \varphi_1$, being the difference of two homogeneous polynomials of degree s, will itself be homogeneous of degree s, and therefore the term φ_2 of the polynomial φ will have weight s, etc.

Symmetric rational fractions. The fundamental theorem on symmetric polynomials can be extended to the case of rational fractions. Let us call the rational fraction $\frac{f}{g}$ in n unknowns $x_1, x_2, \ldots, x_n$ *symmetric* if it remains equal to itself under any rearrangement of the unknowns. It is easy to demonstrate that *this definition does not depend on whether we take the fraction* $\frac{f}{g}$ *or an equivalent fraction* $\frac{f_0}{g_0}$. Indeed, if ω is some arrangement of our unknowns, and φ is an arbitrary polynomial in these unknowns, then let us agree to use φ^ω to denote the polynomial into which φ is carried by the arrangement ω. By hypothesis, for any ω,

$$\frac{f}{g} = \frac{f^\omega}{g^\omega}$$

That is, $fg^\omega = gf^\omega$. On the other hand, from

$$\frac{f}{g} = \frac{f_0}{g_0}$$

it follows that $fg_0 = gf_0$, whence $f^\omega g_0^\omega = g^\omega f_0^\omega$. Multiplying both sides by f, we get

$$ff^\omega g_0^\omega = fg^\omega f_0^\omega = gf^\omega f_0^\omega$$

whence, by cancelling out f^ω, it follows that $fg_0^\omega = gf_0^\omega$ or

$$\frac{f_0^\omega}{g_0^\omega} = \frac{f}{g} = \frac{f_0}{g_0}$$

The following theorem is valid.

Any symmetric rational fraction in the unknowns $x_1, x_2, \ldots, x_n$ *with coefficients from the field* P *can be represented as a rational fraction in the elementary symmetric polynomials* $\sigma_1, \sigma_2, \ldots, \sigma_n$ *with coefficients which again belong to* P.

Indeed, suppose we have the symmetric rational fraction

$$\frac{f(x_1, x_2, \ldots, x_n)}{g(x_1, x_2, \ldots, x_n)}$$

Assuming it to be in lowest terms, we could prove that both f and g are symmetric polynomials. However, a simpler way is the following. If the polynomial g is not symmetric, multiply the numerator and the denominator by the product of all $n! - 1$ polynomials obtained from g under all possible nonidentical permutations of the unknowns. It is easy to check that the denominator will now be a symmetric polynomial. From this it follows, by the symmetry of the entire fraction, that the numerator will now also be symmetric, and so to prove the theorem all we have to do is express the numerator and the denominator in terms of the elementary symmetric polynomials.

Power sums. In applications we often encounter the symmetric polynomials

$$s_k = x_1^k + x_2^k + \ldots + x_n^k, \quad k = 1, 2, \ldots$$

which are sums of the kth powers of the unknowns $x_1, x_2, \ldots, x_n$. These polynomials, called *power sums*, must be expressed (by the fundamental theorem) in terms of elementary symmetric polynomials. However, for large k, it is extremely difficult to find these expressions, and so of interest is the relationship between the polynomials $s_1, s_2, \ldots$ and $\sigma_1, \sigma_2, \ldots, \sigma_n$, which we will now establish.

First of all, $s_1 = \sigma_1$. Next, if $k \leqslant n$, then it is easy to verify the truth of the following equalities:

$$\left.\begin{aligned}
&s_{k-1}\sigma_1 = s_k + S(x_1^{k-1}x_2),^* \\
&s_{k-2}\sigma_2 = S(x_1^{k-1}x_2) + S(x_1^{k-2}x_2x_3), \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&s_{k-i}\sigma_i = S(x_1^{k-i+1}x_2 \ldots x_i) + S(x_1^{k-j}x_2 \ldots x_ix_{i+1}) \\
&\qquad\qquad\qquad\qquad\qquad\qquad 2 \leqslant i \leqslant k-2 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&s_1\sigma_{k-1} = S(x_1^2x_2 \ldots x_{k-1}) + k\sigma_k
\end{aligned}\right| \qquad (1)$$

Taking the alternating sum of these equalities (that is, the sum with alternating signs), and then transposing all terms to one side, we get the following formula:

$$s_k - s_{k-1}\sigma_1 + s_{k-2}\sigma_2 - \ldots + (-1)^{k-1}s_1\sigma_{k-1} + (-1)^k k\sigma_k = 0 \qquad (2)$$

$$(k \leqslant n)$$

---

* See (11) of Sec. 52.

But if $k > n$, then the system (1) of equations takes the form

$$s_{k-1}\sigma_1 = s_k + S(x_1^{k-1}x_2),$$
$$s_{k-2}\sigma_2 = S(x_1^{k-1}x_2) + S(x_1^{k-2}x_2x_3),$$
$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$
$$s_{k-i}\sigma_i = S(x_1^{k-i+1}x_2 \dots x_i) + S(x_1^{k-i}x_2 \dots x_ix_{i+1}), \quad 2 \leqslant i \leqslant n-1,$$
$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$
$$s_{k-n}\sigma_n = S(x_1^{k-n+1}x_2 \dots x_n)$$

whence follows the formula

$$s_k - s_{k-1}\sigma_1 + s_{k-2}\sigma_2 - \dots + (-1)^n s_{k-n}\sigma_n = 0 \quad (k > n) \qquad (3)$$

Formulas (2) and (3) are called *Newton's formulas*. They connect power sums with elementary symmetric polynomials and permit one to find, successively, the expressions for s_1, s_2, s_3 . . . in terms of $\sigma_1, \sigma_2, \dots, \sigma_n$. Thus, we know that $s_1 = \sigma_1$, which also follows from formula (2). Furthermore, if $k = 2 \leqslant n$, then, by (2), $s_2 - s_1\sigma_1 + 2\sigma_2 = 0$, whence

$$s_2 = \sigma_1^2 - 2\sigma_2$$

For $k = 3 \leqslant n$ we have $s_3 - s_2\sigma_1 + s_1\sigma_2 - 3\sigma_3 = 0$, whence, using the expressions already found for s_1 and s_2, we get

$$s_3 = \sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3$$

which is already familiar to us [see (12) of Sec. 52]. Now if $k = 3$ but $n = 2$, then, by (3), $s_3 - s_2\sigma_1 + s_1\sigma_2 = 0$, whence $s_3 = \sigma_1^3 - 3\sigma_1\sigma_2$. Using the Newton formulas, we can obtain a general formula expressing s_k in terms of $\sigma_1, \sigma_2, \dots, \sigma_n$. True, this formula is very unwieldy and so we will not give it.

If the base field P has characteristic 0 and for this reason division by any natural number n is meaningful*, then formula (2) permits successively expressing the elementary symmetric polynomials $\sigma_1, \sigma_2, \dots, \sigma_n$ in terms of the first n power sums $s_1, s_2, \dots, s_n$. Thus, $\sigma_1 = s_1$ and therefore

$$\sigma_2 = \frac{1}{2}(s_1\sigma_1 - s_2) = \frac{1}{2}(s_1^2 - s_2),$$
$$\sigma_3 = \frac{1}{3}(s_3 - s_2\sigma_1 + s_1\sigma_2) = \frac{1}{6}(s_1^3 - 3s_1s_2 + 2s_3)$$

and so forth. From the foregoing and from the fundamental theorem follows the result that

---

* In a field of characteristic p, the expression $\frac{a}{p}$ is meaningless for $a \neq 0$ since in this field $px = 0$ for any x.

Any symmetric polynomial in n unknowns $x_1, x_2, \ldots, x_n$ over a field P of characteristic zero can be represented as a polynomial in the power sums $s_1, s_2, \ldots, s_n$ with coefficients belonging to the field P.

Polynomials symmetric in two systems of unknowns. In the next section, and also in Sec. 58, use will be made of a generalization of the concept of a symmetric polynomial. Suppose we have two systems of unknowns $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_r$, and suppose their union

$$x_1, x_2, \ldots, x_n, \quad y_1, y_2, \ldots, y_r \tag{4}$$

is algebraically independent over the field P. The polynomial $f(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_r)$ over the field P is called *symmetric in two systems of unknowns* if it remains unchanged under any arrangements of the unknowns $x_1, x_2, \ldots, x_n$ among themselves and of the unknowns $y_1, y_2, \ldots, y_r$ among themselves. If we denote the elementary symmetric polynomials in $x_1, x_2, \ldots, x_n$ by $\sigma_1, \sigma_2, \ldots, \sigma_n$ and the elementary symmetric polynomials in $y_1, y_2, \ldots, y_r$ by $\tau_1, \tau_2, \ldots, \tau_r$ then the fundamental theorem is generalized as follows.

Any polynomial $f(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_r)$ over the field P, which polynomial is symmetric with respect to the systems of unknowns $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_r$, can be represented as a polynomial (with coefficients from P) in the elementary symmetric polynomials with respect to these two systems of unknowns:

$$f(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_r) = \varphi(\sigma_1, \sigma_2, \ldots, \sigma_n, \tau_1, \tau_2, \ldots \tau_r)$$

Indeed, the polynomial f may be regarded as a polynomial $\overline{f}(y_1, y_2, \ldots, y_r)$ with coefficients which are polynomials in $x_1, x_2, \ldots, x_n$. Since f remains unchanged under rearrangements of the unknowns $x_1, x_2, \ldots, x_n$, it follows that the coefficients of the polynomial $\overline{f}$ will be symmetric polynomials in $x_1, x_2, \ldots, x_n$ and therefore, by the fundamental theorem, can be represented as polynomials (with coefficients from P) in $\sigma_1, \sigma_2, \ldots, \sigma_n$. On the other hand, the polynomial $\overline{f}(y_1, y_2, \ldots, y_r)$ regarded over the field $P(x_1, x_2, \ldots, x_n)$ will be symmetric with respect to $y_1, y_2, \ldots, y_r$ and therefore can be represented as the polynomial $\overline{\varphi}(\tau_1, \tau_2, \ldots, \tau_r)$. The coefficients of the polynomial $\overline{\varphi}$ will, as was demonstrated at the beginning of this section, be expressed in terms of the coefficients of $\overline{f}$ by means of addition and subtraction, and so they too will be polynomials in $\sigma_1, \sigma_2, \ldots, \sigma_n$. This obviously leads us to the desired expression for f in terms of $\sigma_1, \sigma_2, \ldots, \sigma_n, \tau_1, \tau_2, \ldots, \tau_r$.

Example. The polynomial

$$f(x_1, x_2, x_3, y_1, y_2) = x_1x_2x_3 - x_1x_2y_1 - x_1x_2y_2 - x_1x_3y_1 - x_1x_3y_2 - x_2x_3y_1 - x_2x_3y_2 + x_1y_1y_2 + x_2y_1y_2 + x_3y_1y_2$$

is symmetric both with respect to the unknowns x_1, x_2, x_3 and to the unknowns y_1, y_2, but is not symmetric with respect to the five unknowns taken together, as is evident from, say, a transposition of the unknowns x_1 and y_1. Let us find the expression for f in terms of σ_1, σ_2, σ_3, τ_1, τ_2:

$$f = x_1x_2x_3 - (x_1x_2 + x_1x_3 + x_2x_3)\, y_1 - (x_1x_2 + x_1x_3 + x_2x_3)\, y_2$$
$$+ (x_1 + x_2 + x_3)\, y_1y_2 = \sigma_3 - \sigma_2y_1 - \sigma_2y_2 + \sigma_1y_1y_2 = \sigma_3 - \sigma_2\tau_1 + \sigma_1\tau_2$$

The theorem just proved can naturally be extended to the case of three or more systems of unknowns.

For polynomials symmetric with respect to two systems of unknowns, the theorem of unique representation in terms of elementary symmetric polynomials also holds true. In other words, the following theorem is valid.

The combined system

$$\sigma_1,\ \sigma_2,\ \ldots,\ \sigma_n,\quad \tau_1,\ \tau_2,\ \ldots,\ \tau_r$$

of elementary symmetric polynomials in the given systems of unknowns $x_1, x_2, \ldots, x_n$ *and* $y_1, y_2, \ldots, y_r$ *is algebraically independent over the field* P.

Indeed, suppose over the field P there is a polynomial

$$\varphi\,(\sigma_1,\ \sigma_2,\ \ldots,\ \sigma_n,\ \tau_1,\ \tau_2,\ \ldots,\ \tau_r)$$

equal to zero although not all its coefficients are zeros. This polynomial may be regarded as a polynomial $\psi\,(\tau_1,\ \tau_2,\ \ldots,\ \tau_r)$ with coefficients which are polynomials in $\sigma_1, \sigma_2, \ldots, \sigma_n$. We can, consequently, take it that ψ is a polynomial in $\tau_1, \tau_2, \ldots, \tau_r$ over the field of rational fractions

$$Q = P\,(x_1,\ x_2,\ \ldots,\ x_n)$$

The system $y_1, y_2, \ldots, y_r$ remains algebraically independent over the field Q: if, in this system, there were algebraic dependence with coefficients from Q, then, by eliminating the denominators, we would obtain an algebraic dependence in system (4), which contradicts the assumption. Proceeding from the uniqueness theorem of the preceding section, we now find that the system $\tau_1, \tau_2, \ldots, \tau_r$ must also be algebraically independent over the field Q, and therefore all coefficients of the polynomial ψ are equal to zero. However, these coefficients are polynomials in $\sigma_1, \sigma_2, \ldots, \sigma_n$ and therefore, again on the basis of the uniqueness theorem for the case of one system of unknowns (this time, the system $x_1, x_2, \ldots, x_n$), all coefficients of these latter polynomials are themselves zero. This proves that, in contradiction with the hypothesis, all coefficients of the polynomial φ must be zero.

54. Resultant. Elimination of Unknown. Discriminant

If we have a polynomial $f(x_1, x_2, \ldots, x_n)$ from the ring $P[x_1, x_2, \ldots, x_n]$, then its *solution* is a set of values of the unknowns

$$x_1 = \alpha_1, \; x_2 = \alpha_2, \; \ldots, \; x_n = \alpha_n$$

taken in the field P or in some extension $\overline{P}$ of this field, a set that makes the polynomial f vanish:

$$f(\alpha_1, \alpha_2, \ldots, \alpha_n) = 0$$

Every polynomial f of degree greater than zero has solutions: if the unknown x_1 occurs in the notation of this polynomial, then for $\alpha_2, \ldots, \alpha_n$ we can actually take any elements of the field P, provided only that the degree of the polynomial $f(x_1, \alpha_2, \ldots, \alpha_n)$ is strictly positive, and then, using the theorem on the existence of a root (Sec. 49), take an extension $\overline{P}$ of the field P in which the polynomial $f(x_1, \alpha_2, \ldots, \alpha_n)$ in the single unknown x_1 has the root α_1. At the same time, we see that the property of a polynomial of degree n in one unknown to have, in any field, not more than n roots ceases to hold true for polynomials in several unknowns.

If we have several polynomials in n unknowns, we can pose the question of finding solutions that are common to all these polynomials; that is, solutions of the system of equations which is obtained by equating the given polynomials to zero. A particular case of this problem, namely the case of systems of linear equations, was considered in detail in Chapter 2. However, concerning the opposite case of one equation in one unknown but of arbitrary degree, we know nothing about the roots except that they exist in some extension of the base field. Finding and studying solutions of an arbitrary nonlinear system of equations in several unknowns is, quite understandably, a still more involved problem that goes beyond the scope of our present course and constitutes a special branch of mathematics known as algebraic geometry. Here, we confine ourselves to a system of two equations of arbitrary degree in two unknowns; we will show that this case can be reduced to that of one equation in one unknown.

Let us first take up the question of the existence of common roots of two polynomials in one unknown. Suppose we have the polynomials

$$\left.\begin{aligned} f(x) &= a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n, \\ g(x) &= b_0x^s + b_1x^{s-1} + \ldots + b_{s-1}x + b_s \end{aligned}\right\} \tag{1}$$

over the field P, $a_0 \neq 0$, $b_0 \neq 0$.

From the results of the preceding chapter, it readily follows that *polynomials $f(x)$ and $g(x)$ have a common root in some extension of the field P if and only if they are not relatively prime.* Thus, the question

of the existence of common roots of the given polynomials can be resolved by applying the Euclidean algorithm.

We will now give another method. Let $\overline{P}$ be some extension of the field P in which $f(x)$ has n roots $\alpha_1, \alpha_2, \ldots, \alpha_n$ and $g(x)$ has s roots $\beta_1, \beta_2, \ldots, \beta_s$; for $\overline{P}$ we can take the splitting field for the product $f(x)\,g(x)$. The element

$$R(f, g) = a_0^s b_0^n \prod_{i=1}^{n} \prod_{j=1}^{s} (\alpha_i - \beta_j) \tag{2}$$

of the field $\overline{P}$ is called the *resultant* of the polynomials $f(x)$ and $g(x)$. It is obvious that $f(x)$ and $g(x)$ *have a common root in* $\overline{P}$ *if and only if* $R(f, g) = 0$. Since

$$g(x) = b_0 \prod_{j=1}^{s} (x - \beta_j)$$

and therefore

$$g(\alpha_i) = b_0 \prod_{j=1}^{s} (\alpha_i - \beta_j)$$

it follows that the resultant $R(f, g)$ can also be written as

$$R(f, g) = a_0^s \prod_{i=1}^{n} g(\alpha_i) \tag{3}$$

The polynomials $f(x)$ and $g(x)$ are utilized in nonsymmetric fashion in determining the resultant. Indeed,

$$R(g, f) = b_0^n a_0^s \prod_{j=1}^{s} \prod_{i=1}^{n} (\beta_j - \alpha_i) = (-1)^{ns} R(f, g) \tag{4}$$

In accordance with (3), $R(g, f)$ may be written as

$$R(g, f) = b_0^n \prod_{j=1}^{s} f(\beta_j) \tag{5}$$

Expression (2) for a resultant requires a knowledge of the roots of the polynomials $f(x)$ and $g(x)$ and therefore is, in a practical sense, useless for solving the problem of the existence of a common root of these two polynomials. However, it turns out that *the resultant* $R(f, g)$ *may be represented in the form of a polynomial in the coefficients* $a_0, a_1, \ldots, a_n, b_0, b_1, \ldots, b_s$ *of the polynomials* $f(x)$ *and* $g(x)$.

The possibility of such a representation follows readily from the results of the preceding section. Indeed, formula (2) shows that the resultant $R(f, g)$ is a symmetric polynomial in two sets of unknowns: the set $\alpha_1, \alpha_2, \ldots, \alpha_n$ and the set $\beta_1, \beta_2, \ldots, \beta_s$. Therefore, as proved at the end of the preceding section, it can be represented in the form of a polynomial in the elementary symmetric polynomials with respect to these two systems of unknowns, that is, by the Vieta

formulas, as a polynomial in the quotients $\frac{a_i}{a_0}$, $i = 1, 2, \ldots, n$, and $\frac{b_j}{b_0}$, $j = 1, 2, \ldots, s$; the factor $a_0^s b_0^n$ included in (2) eliminates a_0 and b_0 from the denominators of the resulting expression. Incidentally, it would be an arduous task to find the expression of the resultant in terms of the coefficients by means of methods described in the preceding sections, and so we will proceed differently.

The expression for the resultant of the polynomials (1) that we will find will suit any pair of such polynomials. To be more precise, we will take it that the set of roots

$$\alpha_1, \alpha_2, \ldots, \alpha_n, \quad \beta_1, \beta_2, \ldots, \beta_s \tag{6}$$

of the polynomials (1) is a set of $n + s$ independent unknowns, that is, a set of $n + s$ elements which are algebraically independent over the field P in the sense of Sec. 51.

We will get an expression for the resultant, which expression, regarded as a polynomial in the unknowns (6) (after replacement of the coefficients by the roots via the Vieta formulas), will be equal to the right member of (2); this member is also regarded as a polynomial in the unknowns (6).

Regarding the equality precisely in the sense of an identity in the set of unknowns (6), we will prove that *the resultant* $R(f, g)$ *of the polynomials* (1) *is equal to the following determinant of order* $n + s$:

$$D = \left|\begin{array}{ccccc} a_0 & a_1 & \ldots & a_n & \\ & a_0 & a_1 & \ldots & a_n \\ & \cdots & \cdots & \cdots & \cdots \\ & & a_0 & a_1 & \ldots \ a_n \\ b_0 & b_1 & \ldots & b_s & \\ & b_0 & b_1 & \ldots & b_s \\ & \cdots & \cdots & \cdots & \cdots \\ & & b_0 & b_1 & \ldots \ b_s \end{array}\right| \begin{array}{l} \Big\} s \text{ rows} \\ \\ \\ \\ \Big\} n \text{ rows} \\ \\ \\ \end{array} \tag{7}$$

(all vacancies are occupied by zeros). The structure of this determinant is clear enough; it need only be noted that the coefficient a_0 appears s times on the principal diagonal and the coefficient b_s occurs n times.

To prove our assertion, we compute in two ways the product $a_0^s b_0^n DM$, where M is the auxiliary determinant of order $n + s$

$$M = \left|\begin{array}{cccccccc} \beta_1^{n+s-1} & \beta_2^{n+s-1} & \ldots & \beta_s^{n+s-1} & \alpha_1^{n+s-1} & \alpha_2^{n+s-1} & \ldots & \alpha_n^{n+s-1} \\ \beta_1^{n+s-2} & \beta_2^{n+s-2} & \ldots & \beta_s^{n+s-2} & \alpha_1^{n+s-2} & \alpha_2^{n+s-2} & \ldots & \alpha_n^{n+s-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \beta_1^2 & \beta_2^2 & \ldots & \beta_s^2 & \alpha_1^2 & \alpha_2^2 & \ldots & \alpha_n^2 \\ \beta_1 & \beta_2 & \ldots & \beta_s & \alpha_1 & \alpha_2 & \ldots & \alpha_n \\ 1 & 1 & \ldots & 1 & 1 & 1 & \ldots & 1 \end{array}\right|$$

M is the Vandermonde determinant and so it is equal (see Sec. 6) to the product of the differences of the elements of its second last row, any succeeding element being subtracted from any preceding element. Thus,

$$M = \prod_{1 \leqslant i < j \leqslant s} (\beta_i - \beta_j) \cdot \prod_{j=1}^{s} \prod_{i=1}^{n} (\beta_j - \alpha_i) \cdot \prod_{1 \leqslant i < j \leqslant n} (\alpha_i - \alpha_j)$$

and therefore, by (4),

$$a_0^s b_0^n DM = D \cdot R(g, f) \cdot \prod_{1 \leqslant i < j \leqslant s} (\beta_i - \beta_j) \cdot \prod_{1 \leqslant i < j \leqslant n} (\alpha_i - \alpha_j) \qquad (8)$$

On the other hand, let us compute the product DM on the basis of the theorem on the determinant of a product of matrices. Multiplying out the appropriate matrices and taking into account that all α are roots of $f(x)$ and all β are roots of $g(x)$, we get

$$DM = \begin{vmatrix}
\beta_1^{s-1} f(\beta_1) & \beta_2^{s-1} f(\beta_2) & \dots & \beta_s^{s-1} f(\beta_s) & 0 & 0 & \dots & 0 \\
\beta_1^{s-2} f(\beta_1) & \beta_2^{s-2} f(\beta_2) & \dots & \beta_s^{s-2} f(\beta_s) & 0 & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
\beta_1 f(\beta_1) & \beta_2 f(\beta_2) & \dots & \beta_s f(\beta_s) & 0 & 0 & \dots & 0 \\
f(\beta_1) & f(\beta_2) & \dots & f(\beta_s) & 0 & 0 & \dots & 0 \\
0 & 0 & \dots & 0 & \alpha_1^{n-1} g(\alpha_1) & \alpha_2^{n-1} g(\alpha_2) & \dots & \alpha_n^{n-1} g(\alpha_n) \\
0 & 0 & \dots & 0 & \alpha_1^{n-2} g(\alpha_1) & \alpha_2^{n-2} g(\alpha_2) & \dots & \alpha_n^{n-2} g(\alpha_n) \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & \dots & 0 & \alpha_1 g(\alpha_1) & \alpha_2 g(\alpha_2) & \dots & \alpha_n g(\alpha_n) \\
0 & 0 & \dots & 0 & g(\alpha_1) & g(\alpha_2) & \dots & g(\alpha_n)
\end{vmatrix}$$

Applying the Laplace theorem, then taking common factors out of the columns of the determinants and computing the remaining determinants as Vandermonde determinants, we obtain

$$a_0^s b_0^n DM = a_0^s b_0^n \prod_{j=1}^{s} f(\beta_j) \cdot \prod_{1 \leqslant i < j \leqslant s} (\beta_i - \beta_j) \cdot \prod_{i=1}^{n} g(\alpha_i) \cdot \prod_{1 \leqslant i < j \leqslant n} (\alpha_i - \alpha_j)$$

or, using (3) and (5),

$$a_0^s b_0^n DM = R(f, g) R(g, f) \cdot \prod_{1 \leqslant i < j \leqslant s} (\beta_i - \beta_j) \cdot \prod_{1 \leqslant i < j \leqslant n} (\alpha_i - \alpha_j) \qquad (9)$$

We find that the right sides of (8) and (9), considered as polynomials in the unknowns (6), are equal. Both sides of the resulting equation can be reduced by common factors not identically zero. The common factor $R(g, f)$ is not equal to zero: since $a_0 \neq 0$ and $b_0 \neq 0$ by hypothesis, it suffices to select for the unknowns (6) nonequal values (in the base field or in some extension of it) in order to obtain from

(4) a nonzero value for the polynomial $R(g, f)$. In the same way, we prove that the other two common factors are also different from zero. Cancelling out common factors, we arrive at the equality

$$R(f, g) = D \tag{10}$$

which is what we set out to prove.

Let us now give up the requirement that the leading coefficients of the polynomials (1) be different from zero*. Concerning the true degrees of these polynomials, it is thus possible to assert only that they do not exceed their "formal" degrees n and, respectively, s. For the resultant, the expression (2) is now meaningless, since it may be that the polynomials in question have fewer roots than n or s. On the other hand, determinant (7) can be written now as well, and since it is already proved that for $a_0 \neq 0$, $b_0 \neq 0$ this determinant is equal to the resultant, it follows that in our general case too we can call it the *resultant* of the polynomials $f(x)$ and $g(x)$ and denote it by $R(f, g)$.

However, we can no longer hope that the fact that the resultant is zero is equivalent to our polynomials having a root in common. Indeed, if $a_0 = 0$ and $b_0 = 0$, then $R(f, g) = 0$, irrespective of whether the polynomials f and g have common roots or not. It turns out, however, that this case is the only case when one cannot conclude that if the resultant is zero, the given polynomials have common roots**. Namely, the following theorem is valid.

If we have polynomials (1) *with arbitrary leading coefficients, then the resultant* (7) *of these polynomials is zero if and only if the polynomials have a common root or if their leading coefficients are both zero.*

Proof. The case of $a_0 \neq 0$, $b_0 \neq 0$ has already been considered, and the case of $a_0 = b_0 = 0$ is covered in the statement of the theorem. It remains to consider the case when one of the leading coefficients of the polynomials (1), say a_0, is nonzero and b_0 is equal to zero.

If $b_i = 0$ for all i, $i = 0, 1, \ldots, s$, then $R(f, g) = 0$ since the determinant (7) contains zero rows. In this case, however, the polynomial $g(x)$ is identically zero and therefore has common roots with $f(x)$. However, if

$$b_0 = b_1 = \ldots = b_{k-1} = 0, \quad \text{but} \quad b_k \neq 0, \quad k \leqslant s$$

and if

$$\bar{g}(x) = b_k x^{s-k} + b_{k+1} x^{s-k-1} + \ldots + b_{s-1} x + b_s$$

---

* This temporary rejection of the condition on the leading coefficient of the polynomial, which was valid up to now, is due to subsequent applications: we want to consider systems of polynomials in two unknowns and we want to regard one of the unknowns as a coefficient. Thus, the leading coefficient can vanish for particular values of this unknown.

** The determinant (7) is of course also equal to zero when $a_n = b_s = 0$. However, in this case the polynomials (1) have a common root 0.

then, replacing the elements $b_0, b_1, \ldots, b_{k-1}$ in (7) with zeros and applying the Laplace theorem, we obviously get

$$R(f, g) = a_0^k R(f, \bar{g}) \tag{11}$$

But since the leading coefficients of both polynomials f and $\bar{g}$ are different from zero, it follows, from what was proved above, that the equality $R(f, \bar{g}) = 0$ is necessary and sufficient for the polynomials f and $\bar{g}$ to have a root in common. On the other hand, by (11), the equalities $R(f, g) = 0$ and $R(f, \bar{g}) = 0$ are equivalent, and since the polynomials g and $\bar{g}$ of course have the same roots, we find that in the case at hand as well the fact that the resultant $R(f, g)$ is zero is equivalent to the polynomials $f(x)$ and $g(x)$ having a common root. This proves the theorem.

Let us find the resultant of the two quadratic polynomials

$$f(x) = a_0x^2 + a_1x + a_2, \quad g(x) = b_0x^2 + b_1x + b_2$$

By (7),

$$R(f, g) = \begin{vmatrix} a_0 & a_1 & a_2 & 0 \\ 0 & a_0 & a_1 & a_2 \\ b_0 & b_1 & b_2 & 0 \\ 0 & b_0 & b_1 & b_2 \end{vmatrix}$$

or, computing the determinant via expansion by the first and third rows,

$$R(f, g) = (a_0b_2 - a_2b_0)^2 - (a_0b_1 - a_1b_0)(a_1b_2 - a_2b_1) \tag{12}$$

Thus, if we have the polynomials

$$f(x) = x^2 - 6x + 2, \quad g(x) = x^2 + x + 5$$

then, by (12), $R(f, g) = 233$ and so these polynomials do not have any roots in common. But if we have the polynomials

$$f(x) = x^2 - 4x - 5, \quad g(x) = x^2 - 7x + 10$$

then $R(f, g) = 0$, which means that they have a common root, the number 5.

Eliminating an unknown from a system of two equations in two unknowns. Suppose we have two polynomials f and g in two unknowns x and y with coefficients from some field P. We write the polynomials in descending powers of x:

$$\left.\begin{aligned} f(x, y) &= a_0(y)\,x^k + a_1(y)\,x^{k-1} + \ldots + a_{k-1}(y)\,x + a_k(y), \\ g(x, y) &= b_0(y)\,x^l + b_1(y)\,x^{l-1} + \ldots + b_{l-1}(y)\,x + b_l(y) \end{aligned}\right\} \tag{13}$$

The coefficients will be polynomials from the ring $P[y]$. We find the resultant of f and g, which are regarded as polynomials in x, and denote it by $R_x(f, g)$. By (7) it will be a polynomial in the single unknown y with coefficients from the field P:

$$R_x(f, g) = F(y) \tag{14}$$

Let the system of polynomials (13) have, in some extension of the field P, the common solution $x = \alpha$, $y = \beta$. Substituting the value β in place of y in (13), we get two polynomials $f(x, \beta)$ and $g(x, \beta)$ in the one unknown x. These polynomials have the common root α and therefore their resultant, which by (14) is equal to $F(\beta)$, must be equal to zero, that is, β *must be a root of the resultant* $R_x(f, g)$. Conversely, if the resultant $R_x(f, g)$ of the polynomials (13) has the root β, then the resultant of the polynomials $f(x, \beta)$ and $g(x, \beta)$ is zero. That is to say, *either these polynomials have a common root or both their leading coefficients are zero*,

$$a_0(\beta) = b_0(\beta) = 0$$

The finding of common solutions of the system (13) of polynomials is reduced to the finding of roots of the single polynomial (14) in the single unknown y. We say that *the unknown* x *has been eliminated from the system* (13) *of polynomials.*

The next theorem relates to the question of the degree of the polynomial which we obtain after eliminating one unknown from the system of two polynomials in two unknowns.

If, taking the unknowns together, the polynomials $f(x, y)$ *and* $g(x, y)$ *are respectively of degrees* n *and* s, *then the degree of the polynomial* $R_x(f, g)$ *in the unknown* y *does not exceed the product* ns, *if, of course, this polynomial is not identically zero.*

First of all, if we regard two polynomials in one unknown with leading coefficients equal to unity, then, by (2), their resultant $R(f, g)$ is a homogeneous polynomial in $\alpha_1, \alpha_2, \ldots, \alpha_n, \beta_1, \beta_2, \ldots, \beta_s$ of degree ns. From this it follows that if the term

$$a_1^{k_1} a_2^{k_2} \ldots a_n^{k_n} b_1^{l_1} b_2^{l_2} \ldots b_s^{l_s}$$

enters into the expression of the resultant via the coefficients $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_s$ and if the *weight* of this term is the number

$$k_1 + 2k_2 + \ldots + nk_n + l_1 + 2l_2 + \ldots + sl_s$$

then *all terms of* $R(f, g)$ *expressed via the coefficients have the same weight equal to* ns. This assertion also holds true in the general case for terms of the resultant (7) if the number

$$0 \cdot k_0 + 1 \cdot k_1 + \ldots + nk_n + 0 \cdot l_0 + 1 \cdot l_1 + \ldots + sl_s \tag{15}$$

is given as the *weight* of the term $a_0^{k_0}a_1^{k_1}\ldots a_n^{k_n}b_0^{l_0}b_1^{l_1},\ \ldots\ b_s^{l_s}$. Indeed, replacing the factors a_0 and b_0 by unity in the terms of determinant (7), we arrive at the case that has already been considered; however, the exponents on these factors enter into (15) with coefficients 0.

Now write the polynomials f and g as follows:

$$f(x, y) = a_0(y)\,x^n + a_1(y)\,x^{n-1} + \ldots + a_n(y),$$
$$g(x, y) = b_0(y)\,x^s + b_1(y)\,x^{s-1} + \ldots + b_s(y)$$

Since n is the degree of $f(x, y)$ in the unknowns jointly, the power of the coefficient $a_r(y)$, $r = 0, 1, 2, \ldots, n$, cannot exceed its index r; this holds true for $b_r(y)$ as well. Whence it follows that the degree of each term of the resultant $R_x(f, g)$ does not exceed the weight of this term, which is to say it is not greater than the number ns. This completes the proof.

Example 1. Find the common solutions to the following system of polynomials:

$$f(x, y) = x^2y + 3xy + 2y + 3,$$
$$g(x, y) = 2xy - 2x + 2y + 3$$

Eliminate x from this system; to do this, rewrite it as

$$\left.\begin{aligned} f(x, y) &= y\cdot x^2 + (3y)\cdot x + (2y + 3),\\ g(x, y) &= (2y - 2)\,x + (2y + 3) \end{aligned}\right\} \qquad (16)$$

then

$$R_x(f, g) = \begin{vmatrix} y & 3y & 2y+3 \\ 2y-2 & 2y+3 & 0 \\ 0 & 2y-2 & 2y+3 \end{vmatrix} = 2y^2 + 11y + 12$$

The numbers $\beta_1 = -4$, $\beta_2 = -\frac{3}{2}$ will be the roots of the resultant. The leading coefficients of the polynomials (16) do not vanish for these values of the unknown y, and so each of them, together with some value for x, constitutes a solution of the given system of polynomials. The polynomials

$$f(x, -4) = -4x^2 - 12x - 5,$$
$$g(x, -4) = -10x - 5$$

have the common root $\alpha_1 = -\frac{1}{2}$. The polynomials

$$f\left(x, -\frac{3}{2}\right) = -\frac{3}{2}x^2 - \frac{9}{2}x,$$
$$g\left(x, -\frac{3}{2}\right) = -5x$$

have the common root $\alpha_2 = 0$. Thus, the given system of polynomials has two solutions:

$$\alpha_1 = -\frac{1}{2},\quad \beta_1 = -4 \quad \text{and} \quad \alpha_2 = 0,\quad \beta_2 = -\frac{3}{2}$$

Example 2. Eliminate one unknown from the system of polynomials

$$f(x, y) = 2x^3y - xy^2 + x + 5,$$
$$g(x, y) = x^2y^2 + 2xy^2 - 5y + 1$$

Since both polynomials are of degree 2 in the unknown y, whereas one of them is of degree 3 in x, it is advisable to eliminate y. Rewrite the system as

$$\left.\begin{aligned} f(x, y) &= (-x)\cdot y^2 + (2x^3)\cdot y + (x + 5), \\ g(x, y) &= (x^2 + 2x)\, y^2 - 5y + 1 \end{aligned}\right\} \tag{17}$$

and find its resultant, applying formula (12):

$$\begin{aligned} R_y(f, g) &= [(-x)\cdot 1 - (x + 5)(x^2 + 2x)]^2 \\ &\quad - [(-x)(-5) - 2x^3(x^2 + 2x)]\,[2x^3\cdot 1 - (x + 5)(-5)] \\ &= 4x^8 + 8x^7 + 11x^6 + 84x^5 + 161x^4 + 154x^3 + 96x^2 - 125x \end{aligned}$$

One of the roots of the resultant is 0. However, for this value of the unknown x, both leading coefficients of the polynomials (17) vanish; and, as is readily seen, the polynomials $f(0, y)$ and $g(0, y)$ do not have any common roots. We do not have any method for finding the other roots of the resultant. We can only assert that if we found them [say in the splitting field for $R_y(f, g)$], then not one of them would make both leading coefficients of the polynomials (17) vanish, and therefore each of these roots, together with some value for y (one or even several), would constitute a solution of the given system of polynomials.

There are also methods for successively eliminating the unknowns from systems with an arbitrary number of polynomials and unknowns. They are too involved however to be included in this course.

Discriminant. By analogy with the question that led us to the concept of a resultant, we can ask about the conditions under which a polynomial $f(x)$ of degree n from the ring $P[x]$ has multiple roots. Let

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n, \quad a_0 \neq 0$$

and suppose that in some extension of the field P this polynomial has the roots $\alpha_1, \alpha_2, \ldots, \alpha_n$. It is obvious that *there will be equal roots among them if and only if the following product is zero:*

$$\begin{aligned} \Delta = (\alpha_2 - \alpha_1)(\alpha_3 - \alpha_1) \ldots (\alpha_n - \alpha_1)(\alpha_3 - \alpha_2)(\alpha_4 - \alpha_2) \ldots (\alpha_n - \alpha_2) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \times (\alpha_n - \alpha_{n-1}) = \prod_{n \geqslant i > j \geqslant 1} (\alpha_i - \alpha_j) \end{aligned}$$

or, equivalently, if the product

$$D = a_0^{2n-2} \prod_{n \geqslant i > j \geqslant 1} (\alpha_i - \alpha_j)^2$$

called the discriminant of the polynomial $f(x)$ is zero.

Unlike the product Δ, which can change sign upon a rearrangement of the roots, the discriminant D is symmetric with respect to

$\alpha_1, \alpha_2, \ldots, \alpha_n$ and can therefore be expressed in terms of the coefficients of the polynomial $f(x)$. To find this expression, under the assumption that the field P has characteristic zero, we can take advantage of the connection between the discriminant of the polynomial $f(x)$ and the resultant of this polynomial and its derivative. It is natural to expect such a connection: we know from Sec. 49 that a polynomial has multiple roots if and only if it has roots in common with the derivative $f'(x)$ and therefore $D = 0$ if and only if $R(f, f') = 0$.

By formula (3) of this section,

$$R(f, f') = a_0^{n-1} \prod_{i=1}^{n} f'(\alpha_i)$$

Differentiating

$$f(x) = a_0 \prod_{k=1}^{n} (x - \alpha_k)$$

we get

$$f'(x) = a_0 \sum_{k=1}^{n} \prod_{j \neq k} (x - \alpha_j)$$

After substitution of α_i instead of x, all terms, except the ith, vanish and so

$$f'(\alpha_i) = a_0 \prod_{j \neq i} (\alpha_i - \alpha_j)$$

whence

$$R(f, f') = a_0^{n-1} \cdot a_0^n \prod_{i=1}^{n} \prod_{j \neq i} (\alpha_i - \alpha_j)$$

For any i and j, $i > j$, two factors enter into this product: $\alpha_i - \alpha_j$ and $\alpha_j - \alpha_i$. Their product is equal to $(-1) \cdot (\alpha_i - \alpha_j)^2$ and since there are $\frac{n(n-1)}{2}$ pairs of indices i, j satisfying the inequalities $n \geqslant i > j \geqslant 1$, it follows that

$$R(f, f') = (-1)^{\frac{n(n-1)}{2}} a_0^{2n-1} \prod_{n \geqslant i > j \geqslant 1} (\alpha_i - \alpha_j)^2 = (-1)^{\frac{n(n-1)}{2}} a_0 D$$

Example. Find the discriminant of the quadratic trinomial

$$f(x) = ax^2 + bx + c$$

Since $f'(x) = 2ax + b$, it follows that

$$R(f, f') = \begin{vmatrix} a & b & c \\ 2a & b & 0 \\ 0 & 2a & b \end{vmatrix} = a(-b^2 + 4ac)$$

In our case, $\frac{n(n-1)}{2} = 1$ and so

$$D = -a^{-1} R(f, f') = b^2 - 4ac$$

This coincides with what school algebra calls the discriminant of a quadratic equation.

Another way of finding the discriminant is the following. Form a Vandermonde determinant from the powers of the roots $\alpha_1, \alpha_2, \ldots, \alpha_n$. As indicated in Sec. 6,

$$\begin{vmatrix} 1 & 1 & \ldots & 1 \\ \alpha_1 & \alpha_2 & \ldots & \alpha_n \\ \alpha_1^2 & \alpha_2^2 & \ldots & \alpha_n^2 \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_1^{n-1} & \alpha_2^{n-1} & \ldots & \alpha_n^{n-1} \end{vmatrix} = \prod_{n \geqslant i > j \geqslant 1} (\alpha_i - \alpha_j) = \Delta$$

and so the discriminant is equal to the square of this determinant multiplied by a_0^{2n-2}. Multiplying this determinant by its transpose by the rule for matrix multiplication and recalling the power sums defined in the preceding section, we get

$$D = a_0^{2n-2} \begin{vmatrix} n & s_1 & s_2 & \ldots & s_{n-1} \\ s_1 & s_2 & s_3 & \ldots & s_n \\ s_2 & s_3 & s_4 & \ldots & s_{n+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ s_{n-1} & s_n & s_{n+1} & \ldots & s_{2n-2} \end{vmatrix} \tag{18}$$

where s_k is the sum of the kth powers of the roots $\alpha_1, \alpha_2, \ldots, \alpha_n$.

Example. Find the discriminant of the cubic polynomial $f(x) = x^3 + ax^2 + bx + c$. By (18)

$$D = \begin{vmatrix} 3 & s_1 & s_2 \\ s_1 & s_2 & s_3 \\ s_2 & s_3 & s_4 \end{vmatrix}$$

As we know from the preceding section,

$$s_1 = \sigma_1 = -a,$$
$$s_2 = \sigma_1^2 - 2\sigma_2 = a^2 - 2b,$$
$$s_3 = \sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3 = -a^3 + 3ab - 3c$$

Using Newton's formula, we will also find that (because $\sigma_4 = 0$)

$$s_4 = \sigma_1^4 - 4\sigma_1^2\sigma_2 + 4\sigma_1\sigma_3 + 2\sigma_2^2 = a^4 - 4a^2b + 4ac + 2b^2$$

Whence

$$\begin{aligned} D &= 3s_2s_4 + 2s_1s_2s_3 - s_2^3 - s_1^2s_4 - 3s_3^2 \\ &= a^2b^2 - 4b^3 - 4a^3c + 18abc - 27c^2 \end{aligned} \tag{19}$$

In particular, for $a = 0$, i.e., for an incomplete cubic polynomial, we obtain

$$D = -4b^3 - 27c^2$$

in complete accordance with what was said in Sec. 38.

55. Alternative Proof of the Fundamental Theorem of the Algebra of Complex Numbers

The proof of the fundamental theorem given in Sec. 23 was completely nonalgebraic. Here we give another proof, which takes advantage of an extensive algebraic apparatus: essential use is made of the fundamental theorem on symmetric polynomials (Sec. 52) and also of the theorem on the existence of a splitting field for any polynomial (Sec. 49). At the same time, the nonalgebraic portion of the proof is minimal and is reduced to a single simple assertion.

First note that in Sec. 23 we proved a lemma on the modulus of the highest-degree term of a polynomial. Taking the coefficients of a polynomial $f(x)$ to be real and putting $k = 1$, we obtain the following corollary of this lemma.

For real values of x sufficiently large in absolute value the sign of a polynomial $f(x)$ with real coefficients coincides with the sign of the highest-degree term.

From this follows the result that

A polynomial of odd degree with real coefficients has at least one real root.

Indeed, let

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n$$

and all coefficients be real. Because of the oddness of n, the highest-degree term a_0x^n has different signs for positive and negative values of x, and therefore, as was proved above, the polynomial $f(x)$ will also have different signs for positive and negative values of x sufficiently large in absolute value. There consequently exist real values of x, say a and b, such that

$$f(a) < 0, \quad f(b) > 0$$

However, from the course of analysis we know that a polynomial (a rational integral function, that is) $f(x)$ is a continuous function and for this reason, because of one of the basic properties of continuous functions, $f(x)$ takes on any given value intermediate between $f(a)$ and $f(b)$ for certain real values of x between a and b. For example, there is an α between a and b such that $f(\alpha) = 0$.

Using this result, we will prove the following assertion.

Every polynomial of arbitrary degree with real coefficients has at least one complex root.

Indeed, suppose we have a polynomial $f(x)$ with real coefficients having degree $n = 2^kq$, where q is an odd number. Since the case $k = 0$ has already been considered (see above), we shall assume $k > 0$, that is, we consider n an even number and we will argue by induction with respect to k, on the assumption that our assertion has

been proved for all polynomials with real coefficients whose degrees are divisible by 2^{k-1} but not divisible by 2^k *.

Let P be a splitting field for the polynomial $f(x)$ over the field of complex numbers (see Sec. 49), and let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be the roots of $f(x)$ in P. Choose an arbitrary real number c and take the elements of the field P having the form

$$\beta_{ij} = \alpha_i \alpha_j + c(\alpha_i + \alpha_j), \quad i < j \tag{1}$$

The number of elements β_{ij} is obviously equal to

$$\frac{n(n-1)}{2} = \frac{2^k q(2^k q - 1)}{2} = 2^{k-1} q(2^k q - 1) = 2^{k-1} q' \tag{2}$$

where q' is an odd number.

Let us now construct from the ring $P[x]$ a polynomial $g(x)$ having for its roots all the elements β_{ij} and only these elements:

$$g(x) = \prod_{i,\, j,\, i<j} (x - \beta_{ij})$$

The coefficients of this polynomial are elementary symmetric polynomials in β_{ij}. Consequently, by (1), they will be polynomials in $\alpha_1, \alpha_2, \ldots, \alpha_n$ with real coefficients (since the number c is real), they will even be symmetric polynomials. True enough, a transposition of any two α, say α_k and α_l, implies merely a rearrangement in the set of all β_{ij}: every β_{kj}, where j is different from k and from l, is converted into β_{lj}, and conversely, whereas β_{kl} and all β_{ij}, for i and j different from k and l, remain fixed. But the coefficients of the polynomial $g(x)$ remain unchanged under a rearrangement of its roots.

From this it follows, by the fundamental theorem on symmetric polynomials, that the coefficients of the polynomial $g(x)$ will be polynomials (with real coefficients) in the coefficients of the given polynomial $f(x)$ and for this reason will themselves be real numbers. The degree of this polynomial, which is equal to the number of the roots β_{ij}, is divisible, according to (2), by 2^{k-1}, but is not divisible by 2^k. And so, by the induction hypothesis, at least one of the roots β_{ij} of the polynomial $g(x)$ must be a complex number.

Thus, for any choice of the real number c there is a pair of indices, i, j, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant n$, such that the element $\alpha_i \alpha_j + c(\alpha_i + \alpha_j)$ is a complex number (recall that the field P contains the field of complex numbers as a subfield). Quite naturally, for any other choice of the number c there will, generally speaking, correspond to it (in the indicated sense) another pair of indices. However, there exist an infinitude of distinct real numbers c, whereas we have at our disposal only a finite number of distinct pairs i, j. Whence it

---

* Consequently, this degree can even be greater than n.

follows that we can choose two distinct real numbers c_1 and c_2, $c_1 \neq c_2$, such that they are associated with one and the same pair of indices i, j, for which

$$\left.\begin{aligned} \alpha_i\alpha_j + c_1(\alpha_i + \alpha_j) &= a, \\ \alpha_i\alpha_j + c_2(\alpha_i + \alpha_j) &= b \end{aligned}\right\} \tag{3}$$

are complex numbers.

From equality (3) it follows that

$$(c_1 - c_2)(\alpha_i + \alpha_j) = a - b$$

whence

$$\alpha_i + \alpha_j = \frac{a-b}{c_1 - c_2}$$

That is to say, this sum is a complex number. From this and at least from the first of the equalities (3) it follows that the product $\alpha_i\alpha_j$ will also be a complex number. Thus, the elements α_i and α_j are roots of the quadratic equation

$$x^2 - (\alpha_i + \alpha_j)x + \alpha_i\alpha_j = 0,$$

with complex coefficients and therefore, as follows from the formula (derived in Sec. 38) for solving quadratic equations with complex coefficients, they must themselves be complex numbers. Thus, among the roots of the polynomial $f(x)$ we have even found two complex roots and the proof of our assertion is complete.

For complete proof of the fundamental theorem, we have yet to consider the case of a polynomial with arbitrary complex coefficients. Let

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n$$

be such a polynomial. Take the polynomial

$$\bar{f}(x) = \bar{a}_0x^n + \bar{a}_1x^{n-1} + \ldots + \bar{a}_n$$

obtained from $f(x)$ by replacing all coefficients with conjugate complex numbers and then consider the product

$$F(x) = f(x)\bar{f}(x) = b_0x^{2n} + b_1x^{2n-1} + \ldots + b_kx^{2n-k} + \ldots + b_{2n}$$

where, evidently,

$$b_k = \sum_{i+j=k} a_i\bar{a}_j, \qquad k = 0, 1, 2, \ldots, 2n$$

Using the familiar properties of conjugate complex numbers (see Sec. 18), we find that

$$\bar{b}_k = \sum_{i+j=k} \bar{a}_ia_j = b_k$$

That is, all coefficients of the polynomial $F(x)$ prove to be real.

It then follows, as proved above, that the polynomial $F(x)$ has at least one complex root β,

$$F(\beta) = f(\beta)\,\bar{f}(\beta) = 0$$

That is, either $f(\beta) = 0$ or $\bar{f}(\beta) = 0$. In the former case, the theorem is proved. But if the latter case occurs, that is,

$$\bar{a}_0\beta^n + \bar{a}_1\beta^{n-1} + \ldots + \bar{a}_n = 0$$

then, replacing all complex numbers here by their conjugates (which, as we know, does not affect the equality), we get

$$f(\bar{\beta}) = a_0\bar{\beta}^n + \alpha_1\bar{\beta}^{n-1} + \ldots + a_n = 0$$

Thus, $f(x)$ has the complex number $\bar{\beta}$ for its root. This completes the proof of the fundamental theorem.

CHAPTER 12

POLYNOMIALS WITH RATIONAL COEFFICIENTS

56. Reducibility of Polynomials over the Field of Rationals

The field of rational numbers, R, is the third number field of particular interest to us, along with the fields of real and complex numbers. It is the smallest of the number fields: as proved in Sec. 43, the field R is contained in its entirety in any number field. We will now investigate the question of the reducibility of polynomials over the field of rationals, in the next section we deal with the rational (integral and fractional) roots of polynomials with rational coefficients. We stress once again that these are two different things: the polynomial

$$x^4 + 2x^2 + 1 = (x^2 + 1)^2$$

is reducible over the field of rational numbers, though it does not have a single rational root.

What can be said about the reducibility of polynomials over the field R? First of all, note that if we have a polynomial $f(x)$ whose coefficients are rational but are not all integral, then, reducing the coefficients to a common denominator and multiplying $f(x)$ by this denominator (equal, say, to k), we obtain a polynomial $kf(x)$, all the coefficients of which will now be integers. It is evident that the polynomials $f(x)$ and $kf(x)$ have the same roots; on the other hand, they will at the same time be reducible or irreducible over the field R.

However, we are not yet entitled to confine ourselves to a consideration of polynomials with integral coefficients. Indeed, let the integral polynomial $g(x)$ (i.e., a polynomial with integral coefficients) be reducible over the field of rationals, i.e., factorable into lower-degree factors with rational (in the general case, fractional) coefficients. Does factorability of $g(x)$ into factors with integral coefficients follow from this? In other words, might it not be true that a polynomial with integral coefficients that is reducible over the field of rational numbers is irreducible over the ring of integers?

The answers may be obtained via considerations similar to those carried out in Sec. 51. Let us call a polynomial $f(x)$ with integral coefficients *primitive* if its coefficients are jointly relatively prime, that is, if they do not have any common divisors different from 1 and -1. If we have an arbitrary polynomial $\varphi(x)$ with rational coefficients, it may be uniquely represented in the form of a product of a lowest-terms fraction by some primitive polynomial:

$$\varphi(x) = \frac{a}{b} f(x) \tag{1}$$

To do this, factor out the common denominator of all coefficients of the polynomial $\varphi(x)$ and then also the common factors of the numerators of these coefficients; note that the degree of $f(x)$ is the same as that of $\varphi(x)$. The uniqueness (to within sign) of the representation (1) is proved as follows. Let

$$\varphi(x) = \frac{a}{b} f(x) = \frac{c}{d} g(x)$$

where $g(x)$ is again a primitive polynomial. Then

$$adf(x) = bcg(x)$$

Thus, ad and bc are obtained by taking all the common factors out of the coefficients of one and the same integral polynomial, and therefore they can differ in sign alone. Whence it follows that the primitive polynomials $f(x)$ and $g(x)$ can likewise differ only in sign.

The Gaussian lemma holds true for integral primitive polynomials.

The product of two integral primitive polynomials is a primitive polynomial.

Indeed, suppose we have the integral primitive polynomials

$$f(x) = a_0x^k + a_1x^{k-1} + \ldots + a_ix^{k-i} + \ldots + a_k,$$
$$g(x) = b_0x^l + b_1x^{l-1} + \ldots + b_jx^{l-j} + \ldots + b_l$$

and let

$$f(x)\,g(x) = c_0x^{k+l} + c_1x^{k+l-1} + \ldots + c_{i+j}x^{(k+l)-(i+j)} + \ldots + c_{k+l}$$

If this product is not primitive, then there is a prime p such that serves as a common divisor of all coefficients $c_0, c_1, \ldots, c_{k+l}$. Since all the coefficients of the primitive polynomial $f(x)$ cannot be divisible by p, let the coefficient a_i be the first one not divisible by p. Similarly, denote by b_j the first coefficient of the polynomial $g(x)$ not divisible by p. Multiplying $f(x)$ and $g(x)$ termwise and collecting terms in $x^{(k+l)-(i+j)}$, we obtain

$$c_{i+j} = a_ib_j + a_{i-1}b_{j+1} + a_{i-2}b_{j+2} + \ldots + a_{i+1}b_{j-1} + a_{i+2}b_{j-2} + \ldots$$

The left side is divisible by p. Also, all terms on the right are certainly divisible by p, except the first. Indeed, by the conditions imposed on the choice of i and j, all the coefficients $a_{i-1}, a_{i-2}, \ldots$, and also $b_{j-1}, b_{j-2}, \ldots$ are divisible by p. It then follows that the product $a_i b_j$ is also divisible by p and therefore, due to the primality of the number p, p should divide at least one of the coefficients a_i, b_j, which, however, is not the case. The lemma is proved.

Let us now answer the questions posed above. Let a polynomial $g(x)$ of degree n with integral coefficients be reducible over the field of rational numbers:

$$g(x) = \varphi_1(x)\,\varphi_2(x)$$

where $\varphi_1(x)$ and $\varphi_2(x)$ are polynomials with rational coefficients and of degree less than n. Then

$$\varphi_i(x) = \frac{a_i}{b_i} f_i(x), \qquad i = 1, 2$$

where $\frac{a_i}{b_i}$ is in lowest terms and $f_i(x)$ is a primitive polynomial. Then

$$g(x) = \frac{a_1 a_2}{b_1 b_2} [f_1(x)\, f_2(x)]$$

The left member is an integral polynomial and so the denominator $b_1 b_2$ in the right member must be reducible. But the polynomial in brackets will, by the Gaussian lemma, be primitive, and so any prime factor from $b_1 b_2$ can cancel out only with some prime factor from $a_1 a_2$, and since a_i and b_i are relatively prime, $i = 1, 2$, the number a_2 must be exactly divisible by b_1, and a_1 by b_2;

$$a_2 = b_1 a_2', \quad a_1 = b_2 a_1'$$

Whence

$$g(x) = a_1' a_2' f_1(x)\, f_2(x)$$

Adjoining the coefficient $a_1' a_2'$ to any one of the factors $f_1(x)$, $f_2(x)$, we obtain a factorization of the polynomial $g(x)$ into factors of lower degree with integral coefficients. This is the proof of the following theorem.

A polynomial with integral coefficients that is irreducible over the ring of integers will also be irreducible over the field of rational numbers.

We can now restrict ourselves, in questions relating to the reducibility of polynomials over the field of rationals, to a consideration of factorizations of integral polynomials into factors whose coefficients are all likewise integral.

We know that any polynomial of degree greater than unity is reducible over the field of complex numbers, and any polynomial (with real coefficients) of degree greater than two is reducible over the field of real numbers. The situation regarding the field of ratio-

nal numbers is quite different: *for any n there is a polynomial of degree n with rational (even integral) coefficients that is irreducible over the field of rational numbers.* The proof of this assertion is based on the following sufficient criterion of the irreducibility of a polynomial over the field R, called the **Eisenstein criterion.**

Suppose we have the polynomial

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n$$

with integral coefficients. If there is at least one way in which we can choose the prime number p that satisfies the following requirements:

(1) *the leading coefficient a_0 is not divisible by p,*

(2) *all the other coefficients are divisible by p,*

(3) *the constant term is divisible by p but not by p^2,*

then the polynomial $f(x)$ is irreducible over the field of rational numbers.

Indeed, if the polynomial $f(x)$ is reducible over the field R, then it can be factored into two factors of lower degree with integral coefficients:

$$f(x) = (b_0x^k + b_1x^{k-1} + \ldots + b_k)(c_0x^l + c_1x^{l-1} + \ldots + c_l)$$

where $k < n$, $l < n$, $k + l = n$. From this, comparing coefficients in both members of the equation, we obtain

$$\left.\begin{aligned} a_n &= b_kc_l, \\ a_{n-1} &= b_kc_{l-1} + b_{k-1}c_l, \\ a_{n-2} &= b_kc_{l-2} + b_{k-1}c_{l-1} + b_{k-2}c_l, \\ &\cdots\cdots\cdots\cdots \\ a_0 &= b_0c_0 \end{aligned}\right\} \qquad (2)$$

From the first of the equalities (2) it follows that, since a_n is divisible by p and p is prime, one of the factors b_k, c_l must be divisible by p. Both cannot be divisible by p at the same time since a_n, by hypothesis, is not divisible by p^2. For instance, let p divide b_k; therefore c_l is prime to p. We now go over to the second of the equalities (2). Its left member and also the first term in the right member are divisible by p, and so p divides the product $b_{k-1}c_l$ as well, but since p does not divide c_l, p does divide b_{k-1}. In the same fashion, we find from the third equality of (2) that p divides b_{k-2}, and so on. Finally, from the $(k+1)$th equality it will be found that p divides b_0; but then from the last equality of (2) it follows that p divides a_0, which contradicts our assumption.

It is extremely easy to write, for any n, integral polynomials of degree n that satisfy the conditions of the Eisenstein criterion and, hence, are irreducible over the field of rational numbers. Such, for example, is the polynomial $x^n + 2$; the Eisenstein criterion is applicable for $p = 2$.

The Eisenstein criterion is only a sufficient condition for irreducibility over the field R, but by no means is it a necessary condition: if it is not possible, for a given polynomial $f(x)$, to find a prime number p such that the conditions of the Eisenstein criterion are fulfilled, it may be reducible, like $x^2 - 5x + 6$, but it may also be irreducible, like $x^2 + 1$. There are a large number of other sufficient criteria besides the Eisenstein criterion (though less important) for irreducibility of polynomials over the field R. There is also a method, due to Kronecker, which permits one to decide whether any polynomial with integral coefficients is reducible or not over R. However, it is very unwieldy and hardly at all applicable in a practical sense.

Example. Consider the polynomial

$$f_p(x) = \frac{x^p - 1}{x - 1} = x^{p-1} + x^{p-2} + \ldots + x + 1$$

where p is a prime number. The roots of this polynomial are pth roots of unity different from unity itself; since these roots, together with 1, divide the unit circle of the complex plane into p equal parts, the polynomial $f_p(x)$ is called a *cyclotomic polynomial.*

The Eisenstein criterion cannot be directly applied to this polynomial. But by changing the variable, setting $x = y + 1$, we get

$$g(y) = f_p(y+1) = \frac{(y+1)^p - 1}{(y+1) - 1} = \frac{1}{y}\left[y^p + py^{p-1} + \frac{p(p-1)}{2!}y^{p-2} + \ldots + py\right]$$

$$= y^{p-1} + py^{p-2} + \frac{p(p-1)}{2!}y^{p-3} + \ldots + p$$

The coefficients of the polynomial $g(y)$ are binomial coefficients and so all, except the leading coefficient, are divisible by p; the constant term is not divisible by p^2. Thus, by the Eisenstein criterion, the polynomial $g(y)$ is irreducible over the field R, whence follows *the irreducibility over R of the cyclotomic polynomial* $f_p(x)$. Indeed, if

$$f_p(x) = \varphi(x)\,\psi(x)$$

then

$$g(y) = \varphi(y+1)\,\psi(y+1)$$

57. Rational Roots of Integral Polynomials

It was pointed out above that the question of the factorization of a given polynomial over the field of rational numbers into irreducible factors has no really satisfactory practical solution. However, the particular case referring to the isolation of linear factors of a polynomial with rational coefficients, that is, to the finding of its rational roots, is very simple and may be solved without excessive computations. Quite naturally, the problem of finding rational roots of a polynomial with rational coefficients does not in the least exhaust the general problem of the real roots of these polynomials;

that is to say, the methods and results given in Chapter 9 are valid in toto when applied to polynomials with rational coefficients.

As we take up the question of finding the rational roots of polynomials with rational coefficients, it is well to note that, as indicated in the preceding section, we can confine ourselves to polynomials with integral coefficients. We shall consider separately the case of integral and that of fractional roots.

If an integer α is a root of a polynomial $f(x)$ with integral coefficients, then α is a divisor of the constant term of the polynomial.

Indeed, let

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n$$

Divide $f(x)$ by $x - \alpha$:

$$f(x) = (x - \alpha)(b_0x^{n-1} + b_1x^{n-2} + \ldots + b_{n-1})$$

Performing the division by the Horner method (see Sec. 22), we find that *all coefficients of the quotient, including b_{n-1}, are integers,* and since

$$a_n = -\alpha b_{n-1} = \alpha(-b_{n-1})$$

our assertion is proved.*

Thus, if an integral polynomial $f(x)$ has integral roots, they will be found among the divisors of the constant term. It is thus necessary to test all possible divisors (both positive and negative) of the constant term. If none is a root of the polynomial, then the polynomial has no integral roots at all.

To test all the divisors of the constant term may turn out to be extremely complicated even if the values of the polynomial have been computed by the Horner method and not via direct substitution of each of the divisors in place of the unknown. The following remarks somewhat simplify computations. First of all, since both 1 and -1 are always divisors of the constant term, we compute $f(1)$ and $f(-1)$. This presents no difficulties. Now if the integer α is a root of $f(x)$,

$$f(x) = (x - \alpha)\,q(x)$$

then, as indicated above, all the coefficients of the quotient $q(x)$ will be integers, and therefore the quotients

$$\frac{f(1)}{\alpha - 1} = -q(1), \quad \frac{f(-1)}{\alpha + 1} = -q(-1)$$

---

* It would be wrong to attempt to prove this theorem by referring to the fact that the constant term a_n is (to within sign) a product of the roots of the polynomial $f(x)$: these roots can include fractional, irrational, and complex roots, and one cannot, therefore, assert beforehand that the product of all these roots (except α) will be integral.

must be integers. Thus, *only such divisors α of the constant term (from among those which differ from 1 and -1) have to be tested, relative to which each of the quotients* $\frac{f(1)}{\alpha-1}$, $\frac{f(-1)}{\alpha+1}$ *is an integer.*

Example 1. Find the integral roots of the polynomial

$$f(x)=x^3-2x^2-x-6$$

The numbers ±1, ±2, 3, ±6 are divisors of the constant term. Since $f(1)=-8$, $f(-1)=-8$, it follows that 1 and -1 are not roots. Furthermore, the numbers

$$\frac{-8}{2+1},\quad \frac{-8}{-2-1},\quad \frac{-8}{6-1},\quad \frac{-8}{-6-1}$$

are fractions and so the divisors 2, -2, 6, -6 have to be rejected, whereas the numbers

$$\frac{-8}{3-1},\quad \frac{-8}{3+1},\quad \frac{-8}{-3-1},\quad \frac{-8}{-3+1}$$

are integers and so the divisors 3 and -3 have yet to be tested. We apply the Horner method:

$$\begin{array}{r|cccc} & 1 & -2 & -1 & -6 \\ \hline -3 & 1 & -5 & 14 & -48 \end{array}$$

That is, $f(-3)=-48$ and so -3 is not a root of $f(x)$. Finally,

$$\begin{array}{r|cccc} & 1 & -2 & -1 & -6 \\ \hline 3 & 1 & 1 & 2 & 0 \end{array}$$

That is, $f(3)=0$: the number 3 is a root of $f(x)$. At the same time we found the coefficients of the quotient obtained by dividing $f(x)$ by $x-3$:

$$f(x)=(x-3)(x^2+x+2)$$

It is readily seen that the quotient x^2+x+2 does not have 3 as its root, which means that this number is not a multiple root of $f(x)$.

Example 2. Find integral roots of the polynomial

$$f(x)=3x^4+x^3-5x^2-2x+2$$

Here, ±1 and ±2 are divisors of the constant term. Furthermore $f(1)=-1$, $f(-1)=1$, i.e., 1 and -1 do not serve as roots. Finally, since the numbers

$$\frac{1}{2+1} \quad \text{and} \quad \frac{-1}{-2-1}$$

are fractions, it follows that 2 and -2 will not be roots either and so the polynomial $f(x)$ does not have any integral roots at all.

Let us now examine the question of fractional roots.

If an integral polynomial whose leading coefficient is unity has a rational root, then this root is an integer.

Indeed, let the polynomial

$$f(x)=x^n+a_1x^{n-1}+a_2x^{n-2}+\ldots+a_n$$

with integral coefficients have for a root the fraction $\frac{b}{c}$ in lowest terms, i.e.,

$$\frac{b^n}{c^n} + a_1 \frac{b^{n-1}}{c^{n-1}} + a_2 \frac{b^{n-2}}{c^{n-2}} + \ldots + a_n = 0$$

From this it follows that

$$\frac{b^n}{c} = -a_1 b^{n-1} - a_2 b^{n-2} c - \ldots - a_n c^{n-1}$$

Thus the simplified fraction is equal to an integer, which is impossible.

To obtain all the rational (fractional and integral) roots of an integral polynomial

$$f(x) = a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \ldots + a_{n-1} x + a_n$$

it is necessary to find all the integral roots of the polynomial

$$\varphi(y) = y^n + a_1 y^{n-1} + a_0 a_2 y^{n-2} + \ldots + a_0^{n-2} a_{n-1} y + a_0^{n-1} a_n$$

and divide them by a_0.

Multiply $f(x)$ by a_0^{n-1} and then change the unknown, putting $y = a_0 x$. Clearly,

$$\varphi(y) = \varphi(a_0 x) = a_0^{n-1} f(x)$$

whence it follows that the roots of the polynomial $f(x)$ are equal to the roots of the polynomial $\varphi(y)$ divided by a_0. In particular, to rational roots of $f(x)$ there will correspond rational roots of $\varphi(y)$; however, since the leading coefficient of $\varphi(y)$ is equal to unity, these roots can only be integral, and we already have a procedure for finding them.

Example. Find rational roots of the polynomial

$$f(x) = 3x^4 + 5x^3 + x^2 + 5x - 2$$

Multiplying $f(x)$ by 3^3 and setting $y = 3x$, we get

$$\varphi(y) = y^4 + 5y^3 + 3y^2 + 45y - 54$$

We seek integral roots of the polynomial $\varphi(y)$.

Let us find $\varphi(1)$ by the Horner method:

	1	5	3	45	−54
1	1	6	9	54	0

Thus, $\varphi(1) = 0$, that is, 1 is a root of $\varphi(y)$, and

$$\varphi(y) = (y-1)\, q(y)$$

where

$$q(y) = y^3 + 6y^2 + 9y + 54$$

Let us find the integral roots of the polynomial $q(y)$. The numbers ± 1, ± 2, ± 3, ± 6, ± 9, ± 18, ± 27, ± 54 are divisors of the constant term. Here,

$$q(1) = 70, \quad q(-1) = 50$$

Computing $\frac{q(1)}{\alpha-1}$ and $\frac{q(-1)}{\alpha+1}$ for every divisor α we find that all divisors, except $\alpha = -6$, must be rejected. Test this divisor:

$$\begin{array}{r|l} -6 & \begin{array}{cccc} 1 & 6 & 9 & 54 \end{array} \\ \hline & \begin{array}{cccc} 1 & 0 & 9 & 0 \end{array} \end{array}$$

Thus, $q(-6) = 0$, or -6 is a root of $q(y)$ and therefore also of $\varphi(y)$.

Consequently, the polynomial $\varphi(y)$ has integral roots 1 and -6. Thus the numbers $\frac{1}{3}$ and -2, and only these numbers, are rational roots of the polynomial $f(x)$.

It must be stressed once again that the above-described methods are applicable only to polynomials with integral coefficients and only for finding their rational roots.

58. Algebraic Numbers

Every polynomial of degree n with rational coefficients has n roots in the field of complex numbers; some of these roots (or even all of them) can lie outside the field of rational numbers. However, not every complex or real number serves as a root of some polynomial with rational coefficients. The complex (or, in particular, real) numbers which are roots of such polynomials are called *algebraic* numbers in contrast to *transcendental* numbers. Algebraic numbers include all rational numbers (as the roots of first-degree polynomials with rational coefficients) and also any radical of the form $\sqrt[n]{a}$ with rational radicand a (as a root of the binomial $x^n - a$). On the other hand, the more comprehensive courses of mathematical analysis offer proof of the transcendence of the number e (the base of the system of natural logarithms) and also of the familiar number π of elementary geometry.

If a number α is algebraic, then it will even be a root of some polynomial with integral coefficients and therefore a root of one of the irreducible divisors of this polynomial, also with integral coefficients. *The irreducible integral polynomial, of which α is a root, is determined uniquely to within a constant factor, that is to say, quite uniquely if we require that the coefficients of the polynomial be relatively prime jointly* (i.e., that the polynomial be primitive). Indeed, if α serves as a root of two irreducible polynomials $f(x)$ and $g(x)$, then the greatest common divisor of these polynomials will be different from unity, and therefore the polynomials, due to their irreducibility, can differ from one another by a zero-degree factor only.

Algebraic numbers which are roots of one and the same irreducible (over the field R) polynomial are termed *conjugate.** Thus, the whole

* Not to be confused with the concept of the conjugacy of complex numbers.

set of algebraic numbers breaks up into disjoint finite classes of conjugate numbers. No rational number, as a root of a first-degree polynomial, has conjugate numbers different from itself; this property is characteristic of rational numbers: every algebraic number which is not rational is a root of an irreducible polynomial of degree greater than unity, and for this reason it has conjugate numbers different from itself.

The set of all algebraic numbers is a subfield of the field of complex numbers. In other words, *the sum, difference, product and quotient of algebraic numbers are algebraic numbers.*

In fact, suppose we have the algebraic numbers α and β. Denote by $\alpha_1 = \alpha, \alpha_2, \ldots, \alpha_n$ all numbers conjugate to α, by $\beta_1 = \beta, \beta_2, \ldots, \beta_s$ all numbers conjugate to β, by $f(x)$ and $g(x)$, irreducible polynomials with rational coefficients having for roots α and β respectively. Write a polynomial whose roots are all possible sums $\alpha_i + \beta_j$; this is

$$\varphi(x) = \prod_{i=1}^{n} \prod_{j=1}^{s} [x - (\alpha_i + \beta_j)]$$

It is obvious that the coefficients of this polynomial will not change under rearrangements of all α_i and also of all β_j. Hence, on the basis of the theorem on polynomials symmetric with respect to two systems of unknowns (see end of Sec. 53), they are polynomials in the coefficients of the polynomials $f(x)$ and $g(x)$. In other words, the coefficients of the polynomial $\varphi(x)$ prove to be rational numbers, and therefore the number $\alpha + \beta = \alpha_1 + \beta_1$, which is one of its roots, will be algebraic.

The algebraic nature of the numbers $\alpha - \beta$ and $\alpha\beta$ is proved in similar fashion with the aid of the polynomials

$$\psi(x) = \prod_{i=1}^{n} \prod_{j=1}^{s} [x - (\alpha_i - \beta_j)]$$

and

$$\chi(x) = \prod_{i=1}^{n} \prod_{j=1}^{s} (x - \alpha_i \beta_j)$$

To prove the algebraic nature of a quotient, it suffices to demonstrate that if a number α is algebraic and different from zero, then α^{-1} will also be an algebraic number. Let α be a root of the polynomial

$$f(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n$$

with rational coefficients. Then, evidently, the polynomial

$$g(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0$$

which also has rational coefficients, will have for a root the number α^{-1}, which is what we set out to prove.

It follows, from the theorem just proved, that any sum of a rational number and a radical, say $1 + \sqrt[3]{2}$, and also any sum of radicals, say $\sqrt{3} + \sqrt[7]{5}$, will be algebraic numbers. However, we cannot as yet assert that numbers written as radicals within radicals, say $\sqrt{1 + \sqrt{2}}$, are algebraic. This will be a consequence of the following theorem.

If the number ω *is a root of the polynomial*

$$\varphi(x) = x^n + \alpha x^{n-1} + \beta x^{n-2} + \ldots + \lambda x + \mu$$

whose coefficients are algebraic numbers, then ω *is also an algebraic number.*

Let $\alpha_i, \beta_j, \ldots, \lambda_s, \mu_t$ run through numbers which are respectively conjugate to $\alpha, \beta, \ldots, \lambda, \mu$, it being true that $\alpha_1 = \alpha$, $\beta_1 = \beta, \ldots, \lambda_1 = \lambda, \mu_1 = \mu$. Consider all possible polynomials of the form

$$\varphi_{i, j, \ldots, s, t}(x) = x^n + \alpha_i x^{n-1} + \beta_j x^{n-2} + \ldots + \lambda_s x + \mu_t$$

so that $\varphi_{1, 1, \ldots, 1, 1}(x) = \varphi(x)$ and take the product of all these polynomials:

$$F(x) = \prod_{i, j, \ldots, s, t} \varphi_{i, j, \ldots, s, t}(x)$$

The coefficients of the polynomial $F(x)$ are obviously symmetric with respect to each of the systems $\alpha_i, \beta_j, \ldots, \lambda_s, \mu_t$ and therefore (again by the theorem of Sec. 53) are polynomials in the coefficients of those irreducible polynomials (with rational coefficients) whose roots are, respectively, $\alpha, \beta, \ldots, \lambda, \mu$; that is to say, they are themselves rational numbers. The number ω, being a root of $\varphi(x)$, will, consequently, be a root of the polynomial $F(x)$ with rational coefficients, i.e., it will be an algebraic number.

Let us apply this theorem to the number $\omega = \sqrt{1 + \sqrt{2}}$. The number $\alpha = 1 + \sqrt{2}$ is algebraic by the previous theorem and therefore the number ω is a root of the polynomial $x^2 - \alpha$ with algebraic coefficients; that is, it is itself algebraic. Generally, applying several times both theorems that have just been proved, the reader will easily arrive at the following result.

Any number written in radicals over the field of rational numbers (that is to say, a number expressed in terms of some arbitrarily complicated combination of radicals—radicals within radicals, in the general case) is an algebraic number.

Obviously, algebraic numbers written as radicals constitute a field. One must bear in mind, however, that this field, as follows from the remark made (without proof) at the end of Sec. 38, will only be a part of the field of all algebraic numbers.

We have already mentioned the transcendence of two numbers: e and π. Actually, however, there are an infinity of transcendental numbers. What is more, using the concepts and methods of set theory, we will show that there are, so to say, even more transcendental numbers than algebraic numbers. The exact meaning of this sentence will be clear from what follows.

An infinite set M is called *countable* (*denumerable*), if it can be put into one-to-one correspondence with the set of natural numbers, that is to say, if its elements can be enumerated with the aid of the natural numbers, otherwise it is *noncountable*.

Lemma 1. *Every infinite set M contains a countable subset.*

Indeed, take an arbitrary element a_1 in M and then an element a_2 different from a_1. Generally, let there be chosen n distinct elements $a_1, a_2, \ldots, a_n$ in M. Since the set M is infinite, it cannot be exhausted by these elements, and so we can find an element a_{n+1} different from them. Continuing this process, we will find in M an infinite subset composed of the elements

$$a_1, a_2, \ldots, a_n, \ldots$$

The countability of this subset is obvious.

Lemma 2. *Every infinite subset B of a countable set A is itself countable.*

Because of its countability, the set A can be written as

$$a_1, a_2, \ldots, a_n, \ldots \tag{1}$$

Let a_{k_1} be the first element of the sequence (1) belonging to B, a_{k_2} the second element with this same property, etc. Assuming $a_{k_n} = b_n$. $n = 1, 2, \ldots$, we find that the elements of the subset B constitute a sequence

$$b_1, b_2, \ldots, b_n, \ldots$$

It is clear that this subset is countable.

Lemma 3. *The union of a countable set of finite sets which pairwise do not have any common elements is a countable set.*

Indeed, suppose we have the finite sets

$$A_1, A_2, \ldots, A_n, \ldots$$

Let their union be B. We will obviously enumerate all elements of the set B if, in arbitrary fashion, we number the elements of the finite set A_1, then continue the numbering by passing to the elements of the set A_2, and so on.

Lemma 4. *The union of two countable sets which are devoid of common elements is a countable set.*

Let there be given a countable set A with elements

$$a_1, a_2, \ldots, a_n, \ldots$$

and a countable set B with elements

$$b_1, b_2, \ldots, b_n, \ldots$$

and let the union of these sets be C. If we put

$$a_n = c_{2n-1}, \; b_n = c_{2n}, \qquad n = 1, 2, \ldots$$

then all elements of C will be represented as the sequence

$$c_1, c_2, \ldots, c_{2n-1}, c_{2n}, \ldots$$

This completes the proof of the countability of this set.

Now let us prove the following theorem.

The set of all algebraic numbers is countable.

First let us prove *the countability of the set of all polynomials in one unknown with integral coefficients.* If

$$f(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n$$

is such a polynomial (different from zero), let us use the term *height* of the polynomial for the natural number

$$h_f = n + |a_0| + |a_1| + \ldots + |a_{n-1}| + |a_n|$$

It is obvious that there is only a finite number of integral polynomials with a given height h; denote this set by M_h. Denote the set consisting of zero alone by M_0. The set of all integral polynomials will be the union of the countable set of the finite sets M_0, M_1, $M_2, \ldots, M_h, \ldots$; that is to say, by Lemma 3, it is countable.

From this, by Lemma 2, it follows that *the set of all integral primitive irreducible polynomials is also countable.* At the same time, we know that every algebraic number is a root of one and only one integral primitive irreducible polynomial. Consequently, collecting the roots of all such polynomials, that is, taking the union of the countable set of finite sets, we obtain the set of all algebraic numbers. This set will thus, by Lemma 3, be countable.

Finally, let us prove the following theorem.

The set of all transcendental numbers is noncountable.

Let us first consider the set F of all real numbers x between zero and unity, $0 < x < 1$, and let us prove that *this set is noncountable.* We know that each of the indicated numbers x may be written as a regular infinite decimal fraction

$$x = 0, \quad \alpha_1 \alpha_2 \ldots \alpha_n \ldots$$

and that this notation is unique if we do not allow for fractions in which for all n beyond some $n = N$ all $\alpha_n = 9$; conversely, any fraction of this form is equal to some number x from the set F. Now suppose that the set F is countable, that is, that all the numbers x can be written as the sequence

$$x_1, x_2, \ldots, x_k, \ldots \tag{2}$$

Let

$$x_k = 0, \quad \alpha_{k1}\alpha_{k2} \ldots \alpha_{kn} \ldots$$

be the notation of the number x_k in the form of an infinite decimal. Now write the infinite decimal fraction

$$0, \beta_1\beta_2 \ldots \beta_n \ldots \tag{3}$$

assuming β_1 to be different from the first decimal digit of the fraction x_1, that is, $\beta_1 \neq \alpha_{11}$, β_2 to be different from the second decimal digit of the fraction x_2, i.e., $\beta_2 \neq \alpha_{22}$ and, generally, $\beta_n \neq \alpha_{nn}$. Besides, assume that among the digits β_n there are infinitely many that are different from the digit 9. It is clear that there is a fraction (3) which satisfies all these requirements. It is thus a number in the set F, but it is different, by its construction, from all the numbers of the sequence (2). This contradiction proves the noncountability of the set F.

Whence follows the *noncountability of the set of all complex numbers*: if the set were countable, then, by Lemma 2, it could not contain the noncountable subset F. The noncountability of the set of all transcendental numbers is now, by Lemma 4, obvious, since the union of this set with the countable set of all algebraic numbers is the set of all complex numbers, that is to say, it is noncountable.

The two theorems we have proved show us, due to Lemma 1, that the set of the transcendental numbers is indeed much richer in elements (that is to say, more "potent") than the set of algebraic numbers.

CHAPTER 13

NORMAL FORM OF A MATRIX

59. Equivalence of λ-Matrices

We return now to problems of linear algebra. Chapter 7 demonstrated the important role of the concept of similarity of matrices. Namely, two square matrices of order n are similar if and only if they represent (in different bases) the same linear transformation of n-dimensional linear space. However, we are not yet able to tell whether two given specific matrices are similar or not. On the other hand, among all matrices similar to a given matrix A, we are not able to indicate a matrix of elementary form (in one sense or another); even the question of the conditions under which a matrix A is similar to a diagonal matrix was considered in Sec. 33 only for a particular case. These are the questions we will take up in this chapter. (Note that they are discussed straight off for the case of an arbitrary base field P.)

Let us first investigate square matrices of order n whose elements are polynomials of arbitrary degree in a single unknown λ with coefficients from the field P. These are called *polynomial matrices* or, briefly, *λ-matrices*. An example of a λ-matrix is the characteristic matrix $A - \lambda E$ of an arbitrary square matrix A with elements in P. The principal diagonal of this matrix contains first-degree polynomials, all off-diagonal elements are zero-degree polynomials or zeros. Every matrix with elements from the field P (for brevity, we call them *numerical matrices*) is also a special case of a λ-matrix: its elements are polynomials of degree zero or zeros.

Suppose we have a λ-matrix

$$A(\lambda) = \begin{pmatrix} a_{11}(\lambda) & \dots & a_{1n}(\lambda) \\ \dots & \dots & \dots \\ a_{n1}(\lambda) & \dots & a_{nn}(\lambda) \end{pmatrix}$$

We use the term *elementary transformations* of this matrix for the following four types of transformation:

(1) multiplication of any row of the matrix $A(\lambda)$ by any scalar α in P different from zero;

(2) multiplication of any column of $A(\lambda)$ by any scalar α in P different from zero;

(3) addition, to any ith row of matrix $A(\lambda)$, of any jth row of it, $j \neq i$, multiplied by any polynomial $\varphi(\lambda)$ in the ring $P[\lambda]$;

(4) addition, to any ith column of matrix $A(\lambda)$, of any jth column of it, $j \neq i$, multiplied by any polynomial $\varphi(\lambda)$ in the ring $P[\lambda]$.

It is readily seen that *for every elementary transformation of the λ-matrix there is an inverse transformation which is also elementary.* Thus, the inverse of (1) is an elementary transformation consisting in the multiplication of that row by the number α^{-1}, which exists due to the condition $\alpha \neq 0$; the inverse of (3) is a transformation which consists in adding to the ith row the jth row multiplied by $-\varphi(\lambda)$.

It is possible to interchange any two rows or any two columns in a matrix $A(\lambda)$ by a number of elementary transformations.

Suppose we wish to interchange the ith and jth rows of $A(\lambda)$. This can be accomplished by means of four elementary transformations as the scheme below illustrates:

$$\begin{pmatrix} i \\ j \end{pmatrix} \to \begin{pmatrix} i+j \\ j \end{pmatrix} \to \begin{pmatrix} i+j \\ -i \end{pmatrix} \to \begin{pmatrix} j \\ -i \end{pmatrix} \to \begin{pmatrix} j \\ i \end{pmatrix}$$

The sequence of transformations is: (a) add jth row to ith row; (b) subtract the new ith row from the jth row; (c) add the new jth row to the new ith row; (d) multiply the new jth row by -1.

We will say that the λ-matrices $A(\lambda)$ and $B(\lambda)$ *are equivalent* and we will write $A(\lambda) \sim B(\lambda)$ if the matrix $A(\lambda)$ can be carried into the matrix $B(\lambda)$ by means of a finite number of elementary transformations. This equivalence relation is obviously reflexive and transitive and also symmetric, due to the existence of an inverse elementary transformation for every elementary transformation. In other words, *all square λ-matrices of order n over the field P break up into disjoint classes of equivalent matrices.*

Our immediate aim will be to find the simplest kind of matrices among all the λ-matrices equivalent to the given matrix $A(\lambda)$. To do this, we introduce the following concept. A *canonical λ-matrix* is a λ-matrix with the following three properties:

(a) the matrix is diagonal, that is, of the form

$$\begin{pmatrix} e_1(\lambda) & & & 0 \\ & e_2(\lambda) & & \\ & & \ddots & \\ 0 & & & e_n(\lambda) \end{pmatrix} \tag{1}$$

(b) any polynomial $e_i(\lambda)$, $i = 2, 3, \ldots, n$, is exactly divisible by the polynomial $e_{i-1}(\lambda)$;

(c) the leading coefficient of every polynomial $e_i(\lambda)$, $i = 1, 2, \ldots, n$, is equal to unity if the polynomial is nonzero.

Note that if among the polynomials $e_i(\lambda)$ on the principal diagonal of the canonical λ-matrix (1) there are some equal to zero, then, by property (b), they invariably occupy the last positions on the principal diagonal. On the other hand, if there are zero-degree polynomials among the polynomials $e_i(\lambda)$, then, by Property (c), they are all equal to unity, and, by Property (b), they occupy the first positions on the principal diagonal of the matrix (1).

The canonical λ-matrices embrace, among others, the numerical matrices, including the unit and zero matrices.

Any λ-matrix is equivalent to some canonical λ-matrix, that is to say, it can be reduced to canonical form via elementary transformations.

We will prove this theorem by induction with respect to the order n of the λ-matrices at hand. Indeed, for $n = 1$ we have

$$A(\lambda) = (a(\lambda))$$

If $a(\lambda) = 0$, then our matrix is already canonical. But if $a(\lambda) \neq 0$, then it suffices to divide the polynomial $a(\lambda)$ by its leading coefficient (this is an elementary matrix transformation) in order to get a canonical matrix.

Suppose the theorem has been proved for λ-matrices of order $n-1$. We consider an arbitrary λ-matrix $A(\lambda)$ of order n. If it is a zero matrix, it is already canonical and no proof is needed. We therefore take it that there are nonzero elements among the elements of matrix $A(\lambda)$.

Interchanging (if necessary) rows and columns of $A(\lambda)$, we can move one of the nonzero elements into the upper left-hand corner. Thus, of the λ-matrices equivalent to $A(\lambda)$, there are some with a nonzero polynomial in the upper left corner. Let us consider all such matrices. The polynomials in the upper left corner of these matrices may have different degrees. But the degree of a polynomial is a natural number, and in any nonempty set of natural numbers there is a least number. It is thus possible to find, from among all the λ-matrices equivalent to $A(\lambda)$ and having a nonzero element in the upper left corner, one matrix such that the polynomial in the upper left corner is of the lowest possible degree. Finally, dividing the first row of this matrix by the leading coefficient of the indicated polynomial, we get a λ-matrix equivalent to $A(\lambda)$,

$$A(\lambda) \sim \begin{pmatrix} e_1(\lambda) & b_{12}(\lambda) & \ldots & b_{1n}(\lambda) \\ b_{21}(\lambda) & b_{22}(\lambda) & \ldots & b_{2n}(\lambda) \\ \ldots & \ldots & \ldots & \ldots \\ b_{n1}(\lambda) & b_{n2}(\lambda) & \ldots & b_{nn}(\lambda) \end{pmatrix}$$

such that $e_1(\lambda) \neq 0$, the leading coefficient of this polynomial is equal to unity, and no combination of elementary transformations can carry the resulting matrix into a matrix in which the upper left-hand corner would be occupied by a nonzero polynomial of lower degree.

We now prove that *all elements of the first row and first column of the matrix obtained are exactly divisible by* $e_1(\lambda)$. Suppose, for example, for $2 \leqslant j \leqslant n$,

$$b_{1j}(\lambda) = e_1(\lambda)\, q(\lambda) + r(\lambda)$$

where the degree of $r(\lambda)$ is less than the degree of $e_1(\lambda)$ if $r(\lambda)$ is different from zero. Then, subtracting from the jth column of our matrix the first column multiplied by $q(\lambda)$ and interchanging the first and jth columns, we obtain a matrix equivalent to $A(\lambda)$ in the upper left corner of which is the polynomial $r(\lambda)$, that is to say, a polynomial of lower degree than $e_1(\lambda)$, which contradicts the choice of this polynomial, whence it follows that $r(\lambda) = 0$. The proof is complete.

Now subtracting from the jth column of our matrix the first column multiplied by $q(\lambda)$, we replace the element $b_{1j}(\lambda)$ by zero. Performing such transformations for $j = 2, 3, \ldots, n$, we substitute zeros for all elements $b_{1j}(\lambda)$. In similar fashion we substitute zeros for all elements $b_{i1}(\lambda)$, $i = 2, 3, \ldots, n$. *We thus arrive at a matrix, equivalent to* $A(\lambda)$, *in the upper left corner of which is the polynomial* $e_1(\lambda)$, *all other elements of the first row and the first column being zero:*

$$A(\lambda) \sim \begin{pmatrix} e_1(\lambda) & 0 & \ldots & 0 \\ 0 & c_{22}(\lambda) & \ldots & c_{2n}(\lambda) \\ \cdot & \cdot & \cdot & \cdot \\ 0 & c_{n2}(\lambda) & \ldots & c_{nn}(\lambda) \end{pmatrix} \tag{2}$$

By the induction hypothesis, the matrix of order $n - 1$ in the lower right corner of the matrix (2) that we have obtained can be reduced to canonical form by elementary transformations:

$$\begin{pmatrix} c_{22}(\lambda) & \ldots & c_{2n}(\lambda) \\ \cdot & \cdot & \cdot \\ c_{n2}(\lambda) & \ldots & c_{nn}(\lambda) \end{pmatrix} \sim \begin{pmatrix} e_2(\lambda) & & 0 \\ & \ddots & \\ 0 & & e_n(\lambda) \end{pmatrix}$$

Having performed these same transformations on the corresponding rows and columns of matrix (2) (in the process, the first row and first column will obviously remain unchanged), we find that

$$A(\lambda) \sim \begin{pmatrix} e_1(\lambda) & & & 0 \\ & e_2(\lambda) & & \\ & & \ddots & \\ 0 & & & e_n(\lambda) \end{pmatrix} \tag{3}$$

To prove that the matrix (3) is canonical, it remains to demonstrate that $e_2(\lambda)$ is exactly divisible by $e_1(\lambda)$. Suppose

$$e_2(\lambda) = e_1(\lambda)\, q(\lambda) + r(\lambda)$$

where $r(\lambda) \not\equiv 0$ and the degree of $r(\lambda)$ is less than that of $e_1(\lambda)$. However, by adding to the second column of (3) the first column multiplied by $q(\lambda)$ and then subtracting the first row from the second, we replace the element $e_2(\lambda)$ by the element $r(\lambda)$. Then, by interchanging the first two rows and the first two columns, we transfer the polynomial $r(\lambda)$ to the upper left corner of the matrix, but this contradicts the choice of the polynomial $e_1(\lambda)$.

The theorem on the reduction of a λ-matrix to canonical form is proved. We have to supplement it with the following uniqueness theorem.

Every λ-matrix is equivalent to one canonical matrix only.

Suppose we have an arbitrary λ-matrix $A(\lambda)$ of order n. Take some natural number k, $1 \leqslant k \leqslant n$, and consider all kth-order minors of $A(\lambda)$. Computing these minors, we obtain a finite system of polynomials in λ; we denote the greatest common divisor of this system of polynomials with leading coefficient 1 by $d_k(\lambda)$.

We thus have the polynomials

$$d_1(\lambda),\ d_2(\lambda),\ \ldots,\ d_n(\lambda) \tag{4}$$

which are uniquely defined by the matrix $A(\lambda)$ itself. Here, $d_1(\lambda)$ is the greatest common divisor of all elements of $A(\lambda)$ with coefficient 1, and $d_n(\lambda)$ is equal to the determinant of the matrix $A(\lambda)$ divided by its leading coefficient. Also note that if the matrix $A(\lambda)$ has rank r, then

$$d_{r+1}(\lambda) = \ldots = d_n(\lambda) = 0$$

whereas all the remaining polynomials of system (4) are different from zero.

The greatest common divisor $d_k(\lambda)$ of all minors of order k of the λ-matrix $A(\lambda)$, $k = 1, 2, \ldots, n$, remains unchanged under elementary transformations of $A(\lambda)$.

This assertion is almost obvious when an elementary transformation of type (1) or (2) is performed in matrix $A(\lambda)$. For instance, if the ith row of the matrix is multiplied by a number α in the field P, $\alpha \neq 0$, then the kth-order minors through which the ith row passes will be multiples of α, whereas all the other kth-order minors will remain unchanged. But when seeking the greatest common divisor of several polynomials, any one of the polynomials can be multiplied with impunity by nonzero numbers from P.

Let us now consider elementary transformations of type (3) or (4). Let us, say, add to the ith row of $A(\lambda)$ the jth row, $j \neq i$,

multiplied by the polynomial $\varphi(\lambda)$; denote the resulting matrix by $\overline{A}(\lambda)$ and denote by $\overline{d}_k(\lambda)$ the greatest common divisor of all its kth-order minors taken with leading coefficient 1. Let us see what happens to the kth-order minors of $A(\lambda)$ under this transformation.

It is clear that minors through which the ith row does not pass remain unchanged. Likewise, there is no change in those minors through which both the ith and jth rows pass, since a determinant is unaltered by adding a multiple of one row to another row. Finally, let us take any kth-order minor with the ith row passing through it, but not the jth row; denote it by M. The corresponding minor of the matrix $\overline{A}(\lambda)$ can evidently be represented by the sum of the minor M and the minor M', multiplied by $\varphi(\lambda)$, of the matrix $A(\lambda)$, which M' is obtained from M by replacing the elements of the ith row af $A(\lambda)$ by the corresponding elements of its jth row. Since both M and M' are divisible by $d_k(\lambda)$, it follows that $M + \varphi(\lambda) M'$ will also be divisible by $d_k(\lambda)$.

From the foregoing it follows that all the kth-order minors of matrix $\overline{A}(\lambda)$ are exactly divisible by $d_k(\lambda)$ and therefore $\overline{d}_k(\lambda)$ too is divisible by $d_k(\lambda)$. But since the elementary transformation at hand has an inverse of the same type, it follows that $d_k(\lambda)$ is likewise divisible by $\overline{d}_k(\lambda)$. But if one takes into account that the leading coefficients of both these polynomials are equal to unity, then $\overline{d}_k(\lambda) = d_k(\lambda)$, which completes the proof.

Thus, *all λ-matrices equivalent to the matrix $A(\lambda)$ are associated with one and the same set of polynomials* (4). Specifically, this refers to any one (if there are several) canonical matrix equivalent to $A(\lambda)$. Let (3) be such a matrix.

Let us compute the polynomial $d_k(\lambda)$, $k = 1, 2, \ldots, n$, using matrix (3). Clearly, the kth-order minor in the upper left corner of this matrix is equal to the product

$$e_1(\lambda)\, e_2(\lambda) \ldots e_k(\lambda) \tag{5}$$

Furthermore, if we take, in matrix (3), the kth-order minor in the rows with indices $i_1, i_2, \ldots, i_k$, where $i_1 < i_2 \ldots < i_k$ and in columns with the same indices, then this minor is equal to the product $e_{i_1}(\lambda)\, e_{i_2}(\lambda) \ldots e_{i_k}(\lambda)$ which is divisible by (5). Indeed, $1 \leqslant i_1$ and so $e_{i_1}(\lambda)$ is divisible by $e_1(\lambda)$, $2 \leqslant i_2$ and therefore $e_{i_2}(\lambda)$ is divisible by $e_2(\lambda)$, and so on. Finally, if in matrix (3) we take the kth-order minor, through which the ith row of this matrix passes for at least one i but does not pass its ith column, then this minor contains a zero row and is therefore equal to zero.

It follows from the foregoing that the product (5) will be the greatest common divisor of all kth-order minors of matrix (3) and,

therefore, of the original matrix $A(\lambda)$,

$$d_k(\lambda) = e_1(\lambda)\, e_2(\lambda) \ldots e_k(\lambda), \qquad k = 1, 2, \ldots, n \tag{6}$$

It is now easy to show that *the polynomials* $e_k(\lambda)$, $k = 1, 2, \ldots, n$, *are uniquely determined by the matrix* $A(\lambda)$ *itself*. Let the rank of this matrix be r. Then, as we know, $d_r(\lambda) \neq 0$, but $d_{r+1}(\lambda) = 0$, and therefore, by (6), $e_{r+1}(\lambda) = 0$. Whence, because of the properties of a canonical matrix, it follows generally that if the rank r of matrix $A(\lambda)$ is less than n, then

$$e_{r+1}(\lambda) = e_{r+2}(\lambda) = \ldots = e_n(\lambda) = 0 \tag{7}$$

On the other hand, for $k \leqslant r$, it follows from (6), because $d_{k-1}\lambda \neq 0$, that

$$e_k(\lambda) = \frac{d_k(\lambda)}{d_{k-1}(\lambda)} \tag{8}$$

This completes the proof of the uniqueness of the canonical form of the λ-matrix. At the same time we have obtained a direct procedure for finding polynomials $e_k(\lambda)$, which are called *invariant factors* of the matrix $A(\lambda)$.

Example. Reduce the λ-matrix

$$A(\lambda) = \begin{pmatrix} \lambda^3 - \lambda & 2\lambda^2 \\ \lambda^2 + 5\lambda & 3\lambda \end{pmatrix}$$

to canonical form. Performing a series of elementary transformations, we get

$$A(\lambda) \sim \begin{pmatrix} \lambda^3 - \lambda & \frac{2}{3}\lambda^2 \\ \lambda^2 + 5\lambda & \lambda \end{pmatrix} \sim \begin{pmatrix} \frac{1}{3}\lambda^3 - \frac{10}{3}\lambda^2 - \lambda & 0 \\ \lambda^2 + 5\lambda & \lambda \end{pmatrix}$$

$$\sim \begin{pmatrix} \frac{1}{3}\lambda^3 - \frac{10}{3}\lambda^2 - \lambda & 0 \\ 0 & \lambda \end{pmatrix} \sim \begin{pmatrix} \lambda^3 - 10\lambda^2 - 3\lambda & 0 \\ 0 & \lambda \end{pmatrix} \sim \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^3 - 10\lambda^2 - 3\lambda \end{pmatrix}$$

On the other hand, it might be possible to compute the invariant factors of the matrix $A(\lambda)$ directly. Namely, computing the greatest common divisor of the elements of this matrix, we obtain

$$d_1(\lambda) = e_1(\lambda) = \lambda$$

Now, computing the determinant of $A(\lambda)$ and noting that its leading coefficient is equal to 1, we obtain

$$d_2(\lambda) = \lambda^4 - 10\lambda^3 - 3\lambda^2$$

and so

$$e_2(\lambda) = \frac{d_2(\lambda)}{d_1(\lambda)} = \lambda^3 - 10\lambda^2 - 3\lambda$$

60. Unimodular λ-matrices. Relationship Between Similarity of Numerical Matrices and the Equivalence of Their Characteristic Matrices

From the results of the preceding section there follows a criterion of equivalence of λ-matrices, which may be stated in either of two almost identical formulations.

Two λ-matrices are equivalent if and only if they can be reduced to one and the same canonical form.

Two λ-matrices are equivalent if and only if they have the same invariant factors.

Let us derive another criterion of a different nature.

We know that the unit matrix E is a canonical λ-matrix. We call the λ-matrix $U(\lambda)$ *unimodular* if it has the matrix E for its canonical form; that is to say, if all its invariant factors are equal to unity.

The λ-matrix $U(\lambda)$ is unimodular if and only if its determinant is nonzero but does not depend on λ; that is, it is a nonzero number of the base field P.

Indeed, if $U(\lambda) \sim E$, then these two matrices are associated with one and the same polynomial $d_n(\lambda)$. However, $d_n(\lambda) = 1$ for the unit matrix. From this it follows that the determinant of the matrix $U(\lambda)$, which determinant differs from $d_n(\lambda)$ only by a nonzero numerical factor, will be a nonzero number of the field P. Conversely, if the determinant of the matrix $U(\lambda)$ is different from zero and is not dependent on λ, then for this matrix the polynomial $d_n(\lambda)$ will be equal to unity and therefore, by (6) of Sec. 59, all invariant factors $e_i(\lambda)$ of $U(\lambda)$, $i = 1, 2, \ldots, n$, are equal to unity.

This implies that *any nonsingular numerical matrix is a unimodular λ-matrix.* However, a unimodular λ-matrix can be very complicated. Thus, the λ-matrix

$$\begin{pmatrix} \lambda & \lambda^3 + 5 \\ \lambda^2 - \lambda - 4 & \lambda^4 - \lambda^3 - 4\lambda^2 + 5\lambda - 5 \end{pmatrix}$$

is unimodular, since its determinant is equal to 20; that is to say, it is different from zero and is not dependent on λ.

From the theorem proved above it follows that a *product of unimodular λ-matrices is unimodular*: it suffices to recall that in matrix multiplication the determinants are multiplied together.

The λ-matrix $U(\lambda)$ is unimodular if and only if there is an inverse matrix which is also a λ-matrix.

Indeed, if we have a nonsingular λ-matrix, then in seeking the inverse matrix in ordinary fashion we will have to divide the cofactors of the elements of the given matrix by the determinant of the matrix, i.e., by some polynomial in λ. Therefore, in the general

case, the elements of the inverse matrix will be rational fractions in λ and not polynomials in λ; that is, this matrix is not a λ-matrix. But if a unimodular matrix is given, then we will have to divide the cofactors only by a nonzero number from the field P; i.e., the elements of the inverse matrix will be polynomials in λ and therefore the inverse matrix will itself be a λ-matrix. Conversely, if the λ-matrix $U(\lambda)$ has an inverse λ-matrix $U^{-1}(\lambda)$, then the determinants of both matrices are polynomials in λ, their product is equal to 1, and therefore both determinants must be zero-degree polynomials.

There follows from this last remark a supplement to the theorem just proved: *A λ-matrix inverse to a unimodular λ-matrix is unimodular.*

The concept of a unimodular matrix is used in the statement of the following new equivalence criterion of λ-matrices: *Two λ-matrices $A(\lambda)$ and $B(\lambda)$ of order n are equivalent if and only if there exist unimodular λ-matrices $U(\lambda)$ and $V(\lambda)$ of the same order n such that*

$$B(\lambda) = U(\lambda)\, A(\lambda)\, V(\lambda) \tag{1}$$

First, we introduce the following concept used in the proof of this criterion. We use the term *elementary matrix* to denote a numerical (and, hence, λ-) matrix

$$\begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ \cdots & \cdots & \alpha & \cdots & \cdots \\ & & & 1 & \\ & & & & \ddots \\ 0 & & & & 1 \end{pmatrix} (i) \tag{2}$$

that differs from the unit matrix in only one way: there is an arbitrary nonzero number α from the field P in some ith position of the principal diagonal, $1 \leqslant i \leqslant n$. On the other hand, we will use the term *elementary matrix* for the λ-matrix

$$\begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & \vdots & & \\ \cdots & 1 & \cdots & \varphi(\lambda) & \cdots \\ & & \ddots & \vdots & \\ & & & 1 & \\ & & & \vdots & \ddots \\ & & & \vdots & & 1 \\ & & & (j) & & \end{pmatrix} (i) \tag{3}$$

which differs from the unit matrix in only one way: an arbitrary polynomial $\varphi(\lambda)$ from the ring $P[\lambda]$ occupies the position at the intersection of the ith row and the jth column, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant n$, $i \neq j$.

Every elementary matrix is unimodular. This is quite obvious since the determinant of matrix (2) is equal to α, but, by hypothesis, $\alpha \neq 0$; however, the determinant of the matrix (3) is equal to 1.

Performance of any elementary transformation in the λ-matrix $A(\lambda)$ is equivalent to multiplying this matrix on the left or on the right by some elementary matrix.

It will be easy for the reader to verify the truth of the following four assertions: (1) multiplication of the matrix $A(\lambda)$ on the left by the matrix (2) is equivalent to multiplication of the ith row of $A(\lambda)$ by the scalar α; (2) multiplication of $A(\lambda)$ on the right by matrix (2) is equivalent to multiplication of the ith column of the matrix $A(\lambda)$ by the scalar α; (3) multiplication of matrix $A(\lambda)$ on the left by matrix (3) is equivalent to adding to the ith row of $A(\lambda)$ its jth row multiplied by $\varphi(\lambda)$; (4) multiplication of the matrix $A(\lambda)$ on the right by matrix (3) is equivalent to adding to the jth column of $A(\lambda)$ its ith column multiplied by $\varphi(\lambda)$.

Let us now take up the proof of our criterion of the equivalence of λ-matrices. If $A(\lambda) \sim B(\lambda)$, then we can proceed from $A(\lambda)$ to $B(\lambda)$ by means of a finite number of elementary transformations. Replacing each of these transformations by multiplication on the left or on the right by an elementary matrix, we arrive at the equation

$$B(\lambda) = U_1(\lambda) \ldots U_k(\lambda) A(\lambda) V_1(\lambda) \ldots V_l(\lambda) \qquad (4)$$

where all the matrices $U_1(\lambda), \ldots, U_k(\lambda), V_1(\lambda), \ldots, V_l(\lambda)$ are elementary and, hence, unimodular. Hence, the matrices

$$U(\lambda) = U_1(\lambda) \ldots U_k(\lambda), \quad V(\lambda) = V_1(\lambda) \ldots V_l(\lambda) \qquad (5)$$

which are products of unimodular matrices will also be unimodular, and equation (4) will be rewritten as (1). Notice that if, say, $k = 0$, i.e., elementary transformations are performed on columns only, then we simply put $U(\lambda) = E$.

This portion of the proof already allows us to make the following assertion.

A λ-matrix is unimodular if and only if it is representable as a product of elementary matrices.

True enough, for we have already taken advantage of the fact that a product of elementary matrices is unimodular. Conversely, if we have an arbitrary unimodular matrix $W(\lambda)$ then it is equivalent to the unit matrix E. Applying the foregoing proof to matrices E and $W(\lambda)$ instead of $A(\lambda)$ and $B(\lambda)$, we get from (4) the equation

$$W(\lambda) = U_1(\lambda) \ldots U_k(\lambda) V_1(\lambda) \ldots V_l(\lambda)$$

which is to say that the matrix $W(\lambda)$ is represented as a product of elementary matrices.

It is now easy to prove the converse assertion of our criterion. Suppose that for the matrices $A(\lambda)$ and $B(\lambda)$ there are unimodular matrices $U(\lambda)$ and $V(\lambda)$ such that (1) holds. From what has been proved, the matrices $U(\lambda)$ and $V(\lambda)$ may be represented as products of elementary matrices; let these be the representations (5). Then (1) can be rewritten as (4) and, substituting the corresponding elementary transformation for each multiplication by an elementary matrix, we finally obtain $A(\lambda) \sim B(\lambda)$.

Matrix polynomials. We can take an entirely different view of the λ-matrix concept and use the term *matrix λ-polynomial of order n over the field P* for a polynomial in λ whose coefficients are square matrices of the same order n with elements from the field P. Its general aspect is

$$A_0\lambda^k + A_1\lambda^{k-1} + \ldots + A_{k-1}\lambda + A_k \qquad (6)$$

Regarding (in accordance with Sec. 15) the multiplication of matrix A_i by λ^{k-i}, $i = 0, 1, \ldots, k$, as the multiplication by λ^{k-i} of all elements of the matrix A_i, and then performing matrix addition in accord with that same Sec. 15, we find that *any matrix λ-polynomial of order n may be written as a λ-matrix of order n.* Thus,

$$\begin{pmatrix} 4 & 0 \\ -1 & 1 \end{pmatrix}\lambda^3 + \begin{pmatrix} 0 & -3 \\ 0 & 1 \end{pmatrix}\lambda^2 + \begin{pmatrix} 1 & 2 \\ 0 & -2 \end{pmatrix}\lambda + \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$
$$= \begin{pmatrix} 4\lambda^3 + \lambda & -3\lambda^2 + 2\lambda + 1 \\ -\lambda^3 & \lambda^3 + \lambda^2 - 2\lambda \end{pmatrix}$$

Conversely, *any λ-matrix of order n may be written in the form of a matrix λ-polynomial of order n.* Thus,

$$\begin{pmatrix} 3\lambda^2 - 5 & \lambda + 1 \\ \lambda^4 + 2\lambda & -3 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\lambda^4 + \begin{pmatrix} 3 & 0 \\ 0 & 0 \end{pmatrix}\lambda^2 + \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}\lambda + \begin{pmatrix} -5 & 1 \\ 0 & -3 \end{pmatrix}$$

The correspondence between λ-matrices and matrix λ-polynomials is one-to-one and isomorphic in the meaning of Sec. 46. Indeed, the equality of λ-polynomials of the form (6) as matrices is equivalent to the equality of matrix coefficients of identical powers of λ, and the multiplication of a matrix by λ is equivalent to its multiplication by a scalar matrix with λ on the principal diagonal.

Suppose we have a λ-matrix $A(\lambda)$, and

$$A(\lambda) = A_0\lambda^k + A_1\lambda^{k-1} + \ldots + A_{k-1}\lambda + A_k$$

where the matrix A_0 is not a zero matrix. We call the number k the *degree* of the λ-matrix $A(\lambda)$; clearly, this is the highest power (in λ) of the elements of the matrix $A(\lambda)$.

The view taken of λ-matrices as matrix polynomials permits developing for λ-matrices a theory of divisibility similar to the

theory of divisibility for numerical polynomials, made more complicated, true, by the noncommutativity of matrix multiplication and the presence of divisors of zero. We restrict ourselves to the sole problem of the division algorithm (with remainder).

Given, over the field P, the nth-order λ-matrices

$$A(\lambda) = A_0\lambda^k + A_1\lambda^{k-1} + \ldots + A_{k-1}\lambda + A_k,$$
$$B(\lambda) = B_0\lambda^l + B_1\lambda^{l-1} + \ldots + B_{l-1}\lambda + B_l$$

Assume that the matrix B_0 is nonsingular, i.e., there exists a matrix B_0^{-1}. Then, over the field P it is possible to find λ-matrices $Q_1(\lambda)$ and $R_1(\lambda)$ of the same order n such that

$$A(\lambda) = B(\lambda)\,Q_1(\lambda) + R_1(\lambda) \tag{7}$$

The degree of $R_1(\lambda)$ is less than the degree of $B(\lambda)$ or $R_1(\lambda) = 0$. On the other hand, there are, over P, λ-matrices $Q_2(\lambda)$ and $R_2(\lambda)$ of order n such that

$$A(\lambda) = Q_2(\lambda)\,B(\lambda) + R_2(\lambda) \tag{8}$$

The degree of $R_2(\lambda)$ is less than the degree of $B(\lambda)$ or $R_2(\lambda) = 0$. The matrices $Q_1(\lambda)$ and $R_1(\lambda)$ and also $Q_2(\lambda)$ and $R_2(\lambda)$ which satisfy these conditions are uniquely determined.

The proof of this theorem follows the same lines as that of the corresponding theorem for numerical polynomials (see Sec. 20). For instance, let condition (7) be satisfied also by the matrices $\overline{Q}_1(\lambda)$ and $\overline{R}_1(\lambda)$ and the degree of $\overline{R}_1(\lambda)$ is less than the degree of $B(\lambda)$. Then

$$B(\lambda)\,[Q_1(\lambda) - \overline{Q}_1(\lambda)] = \overline{R}_1(\lambda) - R_1(\lambda)$$

The degree of the right side is less than l, but the degree of the left side (if the square bracket is nonzero) is greater than or equal to l, since the matrix B_0 is nonsingular. Whence follows the uniqueness of the matrices $Q_1(\lambda)$ and $R_1(\lambda)$.

To prove the existence of such matrices, notice that for $k \geqslant l$ the degree of the difference

$$A(\lambda) - B(\lambda)\cdot B_0^{-1}A_0\lambda^{k-l}$$

will be strictly less than k; therefore $B_0^{-1}A_0\lambda^{k-l}$ will be the highest-degree term of the matrix λ-polynomial $Q_1(\lambda)$. The continuation is the same as in Sec. 20. On the other hand, the degree of the difference

$$A(\lambda) - A_0B_0^{-1}\lambda^{k-l}\cdot B(\lambda)$$

is also strictly less than k, that is, $A_0B_0^{-1}\lambda^{k-l}$ will be the highest-degree term of the matrix λ-polynomial $Q_2(\lambda)$. We see that the λ-matrices $Q_1(\lambda)$ and $Q_2(\lambda)$ [and also $R_1(\lambda)$ and $R_2(\lambda)$] which satisfy

the conditions of the theorem, will indeed be distinct in the general case.

Fundamental theorem on the similarity of matrices. Earlier we mentioned the fact that as yet we have no way of deciding whether two numerical matrices A and B (that is, matrices with elements in the base field P) are similar or not. On the other hand, their characteristic matrices $A - \lambda E$ and $B - \lambda E$ are λ-matrices and the question of the equivalence of these matrices is something that can be resolved effectively. It is therefore clear why the following theorem is of such great importance.

The matrices A and B with elements in the field P are similar if and only if their characteristic matrices $A - \lambda E$ and $B - \lambda E$ are equivalent.

Indeed, let the matrices A and B be similar, i.e., there is, over the field P, a nonsingular matrix C such that

$$B = C^{-1}AC$$

Then

$$C^{-1}(A - \lambda E)C = C^{-1}AC - \lambda(C^{-1}EC) = B - \lambda E$$

The nonsingular numerical matrices C^{-1} and C are, however, unimodular λ-matrices. We see that the matrix $B - \lambda E$ is obtained by multiplying the matrix $A - \lambda E$ on the left and on the right by unimodular matrices, that is, $A - \lambda E \sim B - \lambda E$.

Proof of the converse assertion is more complicated. Let

$$A - \lambda E \sim B - \lambda E$$

Then there exist unimodular matrices $U(\lambda)$ and $V(\lambda)$ such that

$$U(\lambda)(A - \lambda E)V(\lambda) = B - \lambda E \tag{9}$$

Taking into account that unimodular matrices have inverse matrices which are λ-matrices, we derive from (9) the following equalities which will be used in the sequel:

$$\left.\begin{aligned} U(\lambda)(A - \lambda E) &= (B - \lambda E)V^{-1}(\lambda) \\ (A - \lambda E)V(\lambda) &= U^{-1}(\lambda)(B - \lambda E) \end{aligned}\right\} \tag{10}$$

Since the λ-matrix $B - \lambda E$ has degree 1 in λ, the nonsingular matrix $-E$ serving as the leading coefficient of the corresponding matrix polynomial, it follows that we can apply the division algorithm to the matrices $U(\lambda)$ and $B - \lambda E$: there are matrices $Q_1(\lambda)$ and R_1 (the latter, if nonzero, must have degree 0 in λ, i.e., it is independent of λ) such that

$$U(\lambda) = (B - \lambda E)Q_1(\lambda) + R_1 \tag{11}$$

Similarly,

$$V(\lambda) = Q_2(\lambda)(B - \lambda E) + R_2 \tag{12}$$

Using (11) and (12), we get, from (9),

$$R_1 (A - \lambda E) R_2 = (B - \lambda E) - U(\lambda)(A - \lambda E) Q_2(\lambda)(B - \lambda E) - (B - \lambda E) Q_1(\lambda)(A - \lambda E) V(\lambda) + (B - \lambda E) Q_1(\lambda)(A - \lambda E) Q_2(\lambda)(B - \lambda E)$$

or, by (10),

$$R_1 (A - \lambda E) R_2 = (B - \lambda E) - (B - \lambda E) V^{-1}(\lambda) Q_2(\lambda)(B - \lambda E) - (B - \lambda E) Q_1(\lambda) U^{-1}(\lambda)(B - \lambda E) + (B - \lambda E) Q_1(\lambda)(A - \lambda E) Q_2(\lambda)(B - \lambda E) = (B - \lambda E)\{E - [V^{-1}(\lambda) Q_2(\lambda) + Q_1(\lambda) U^{-1}(\lambda) - Q_1(\lambda)(A - \lambda E) Q_2(\lambda)](B - \lambda E)\}$$

The square bracket on the right is actually zero, for otherwise, being a λ-matrix [since both $V^{-1}(\lambda)$ and $U^{-1}(\lambda)$ are λ-matrices], it would at least be of degree 0, but then the degree of the curly brackets would not be less than 1 and, hence, the degree of the entire right member would not be less than 2. But this is impossible since on the left-hand side we have a λ-matrix of degree 1.

Thus,

$$R_1 (A - \lambda E) R_2 = B - \lambda E$$

whence, equating the matrix coefficients of identical powers of λ we get

$$R_1 A R_2 = B, \tag{13}$$

$$R_1 R_2 = E \tag{14}$$

Equation (14) shows that the numerical matrix R_2 is not only nonzero but is even nonsingular, and

$$R_2^{-1} = R_1$$

But then equation (13) takes the form

$$R_2^{-1} A R_2 = B$$

which proves the similarity of the matrices A and B.

We have at the same time learned to find the nonsingular matrix R_2 which transforms matrix A into matrix B. Namely, if the matrices $A - \lambda E$ and $B - \lambda E$ are equivalent, then a finite number of elementary transformations carries the first into the second. Take those transformations which refer to columns; denote by $V(\lambda)$ the product of the corresponding elementary matrices taken in the same order. Then divide $V(\lambda)$ by $B - \lambda E$ and perform the division so that the quotient is on the left of the divisor [see (8)]. The remainder of this division will be just the matrix R_2.

Actually, this division need not be performed; one can take advantage of the following lemma, which will also be of use in Sec. 62.

Lemma. *Let*

$$V(\lambda) = V_0\lambda^s + V_1\lambda^{s-1} + \ldots + V_{s-1}\lambda + V_s, \quad V_0 \neq 0 \tag{15}$$

If

$$V(\lambda) = (\lambda E - B)\, Q_1(\lambda) + R_1, \tag{16}$$

$$V(\lambda) = Q_2(\lambda)\,(\lambda E - B) + R_2$$

then

$$R_1 = B^s V_0 + B^{s-1} V_1 + \ldots + B V_{s-1} + V_s, \tag{17}$$

$$R_2 = V_0 B^s + V_1 B^{s-1} + \ldots + V_{s-1} B + V_s$$

It suffices to prove the first of these two assertions, because the second is proved similarly. The proof consists in direct verification of the validity of (16) if the polynomial $V(\lambda)$ is replaced by its notation (15), if (17) is substituted for R_1, and if in place of $Q_1(\lambda)$ we take the polynomial

$$Q_1(\lambda) = V_0\lambda^{s-1} + (BV_0 + V_1)\lambda^{s-2} + (B^2V_0 + BV_1 + V_2)\lambda^{s-3} + \ldots + (B^{s-1}V_0 + B^{s-2}V_1 + \ldots + V_{s-1})$$

This verification is left to the reader.

Example. Given the matrices

$$A = \begin{pmatrix} -2 & 1 \\ 0 & 3 \end{pmatrix}, \qquad B = \begin{pmatrix} -10 & -4 \\ 26 & 11 \end{pmatrix}$$

Their characteristic matrices are equivalent since they can be reduced to one and the same canonical form

$$\begin{pmatrix} 1 & 0 \\ 0 & \lambda^2 - \lambda - 6 \end{pmatrix}.$$

The matrices A and B are thus similar.

To find the matrix R_2 that transforms A into B, let us find some chain of elementary transformations that carries $A - \lambda E$ into $B - \lambda E$. Thus,

$$A - \lambda E = \begin{pmatrix} -2-\lambda & 1 \\ 0 & 3-\lambda \end{pmatrix} \sim \begin{pmatrix} -2-\lambda & 1 \\ -16-8\lambda & 11-\lambda \end{pmatrix} \sim \begin{pmatrix} 8+4\lambda & -4 \\ -16-8\lambda & 11-\lambda \end{pmatrix}$$

$$\sim \begin{pmatrix} 40+4\lambda & -4 \\ -104 & 11-\lambda \end{pmatrix} \sim \begin{pmatrix} -10-\lambda & -4 \\ 26 & 11-\lambda \end{pmatrix} = B - \lambda E$$

The last two transformations refer to columns: to the first column we add the second multiplied by —8 and then we multiply the first column by $-\frac{1}{4}$. The product of the corresponding elementary matrices will be

$$V(\lambda) = \begin{pmatrix} 1 & 0 \\ -8 & 1 \end{pmatrix} \begin{pmatrix} -\frac{1}{4} & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -\frac{1}{4} & 0 \\ 2 & 1 \end{pmatrix}$$

This matrix does not depend on λ and therefore it is the sought-for matrix R_2.

Of course, the matrix that transforms A into B is not by far determined uniquely. For example, the matrix

$$\begin{pmatrix} 3 & 1 \\ 2 & 1 \end{pmatrix}$$

will also be of that kind.

61. Jordan Normal Form

We will now consider nth-order square matrices with elements in the field P. We will isolate a special type called Jordan matrices, and it will be shown that these matrices serve as a normal form for a very broad class of matrices. Namely, *matrices, all the characteristic roots of which lie in the base field P (and only such matrices) are similar to certain Jordan matrices; we say that they can be reduced to a Jordan normal form.* It will then follow, if for the field P we take the field of complex numbers, that *any matrix with complex elements can be reduced to a Jordan normal form in the field of complex numbers.*

We will need some definitions. A kth-order *Jordan submatrix* referring to the number λ_0 is a matrix of order k, $1 \leqslant k \leqslant n$, of the form

$$\begin{pmatrix} \lambda_0 & 1 & & 0 \\ & \lambda_0 & 1 & \\ & & \ddots & \ddots \\ & & & \ddots & 1 \\ & & & & \lambda_0 \\ 0 & & & & \end{pmatrix} \tag{1}$$

In other words, one and the same number λ_0 from the field P occupies the principal diagonal, with unity along the diagonal immediately above and zero elsewhere. Thus,

$$(\lambda_0), \quad \begin{pmatrix} \lambda_0 & 1 \\ 0 & \lambda_0 \end{pmatrix}, \quad \begin{pmatrix} \lambda_0 & 1 & 0 \\ 0 & \lambda_0 & 1 \\ 0 & 0 & \lambda_0 \end{pmatrix}$$

are, respectively, Jordan submatrices of first, second and third order.

A *Jordan matrix* of order n is a matrix of order n having the form

$$J = \begin{pmatrix} \boxed{J_1} & & & 0 \\ & \boxed{J_2} & & \\ & & \ddots & \\ & & & \ddots \\ 0 & & & \boxed{J_s} \end{pmatrix} \tag{2}$$

The elements along the principal diagonal are Jordan submatrices $J_1, J_2, \ldots, J_s$ of certain orders, not necessarily distinct, referring to certain numbers (not necessarily distinct either) lying in the field P. All other positions have zeros. Here, $s \geqslant 1$, that is to say,

one Jordan submatrix of order n belongs to Jordan matrices of this order, and, naturally, $s \leqslant n$.

It may be noted (though this will not be used in what follows) that the structure of the Jordan matrix can be described without resorting to the concept of the Jordan submatrix. It is obvious, namely, that the matrix J is a Jordan matrix if and only if it has the form

$$\begin{pmatrix} \lambda_1 & \varepsilon_1 & & & 0 \\ & \lambda_2 & \varepsilon_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \varepsilon_{n-1} \\ 0 & & & & \lambda_n \end{pmatrix}$$

where λ_i, $i = 1, 2, \ldots, n$, are arbitrary numbers in P and every ε_j, $j = 1, 2, \ldots, n-1$, is equal to unity or zero; note that if $\varepsilon_j = 1$, then $\lambda_j = \lambda_{j+1}$.

Diagonal matrices are a special case of Jordan matrices. These are Jordan matrices whose Jordan submatrices are of order 1.

Our immediate aim is to find the canonical form of the characteristic matrix $J - \lambda E$ of an arbitrary Jordan matrix J of order n. We will first find the canonical form of the characteristic matrix

$$\begin{pmatrix} \lambda_0 - \lambda & 1 & & & 0 \\ & \lambda_0 - \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ 0 & & & & \lambda_0 - \lambda \end{pmatrix} \tag{3}$$

of a single Jordan submatrix (1) of order k. Computing the determinant of this matrix and recalling that the leading coefficient of the polynomial $d_k(\lambda)$ must be equal to 1, we find that

$$d_k(\lambda) = (\lambda - \lambda_0)^k$$

On the other hand, among the $(k-1)$th-order minors of the matrix (3) there is a minor equal to unity; this is the minor obtained by deleting the first column and the last row of the matrix. Therefore

$$d_{k-1}(\lambda) = 1$$

From this it follows that the *following* kth-*order* λ-*matrix*

$$\begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ 0 & & & (\lambda - \lambda_0)^k \end{pmatrix} \tag{4}$$

is the canonical form of the matrix (3).

We now prove the following lemma.

If the polynomials $\varphi_1(\lambda), \varphi_2(\lambda), \ldots, \varphi_t(\lambda)$ *from the ring* $P[\lambda]$ *are pairwise prime, the following equivalence holds true:*

$$\begin{pmatrix} \varphi_1(\lambda) & & & 0 \\ & \varphi_2(\lambda) & & \\ & & \ddots & \\ 0 & & & \varphi_t(\lambda) \end{pmatrix} \sim \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ 0 & & & \prod_{i=1}^{t} \varphi_i(\lambda) \end{pmatrix}.$$

It is evidently sufficient to consider the case of $t = 2$. Since the polynomials $\varphi_1(\lambda)$ and $\varphi_2(\lambda)$ are relatively prime, there are polynomials $u_1(\lambda)$ and $u_2(\lambda)$ in the ring $P[\lambda]$ such that

$$\varphi_1(\lambda)\, u_1(\lambda) + \varphi_2(\lambda)\, u_2(\lambda) = 1$$

Therefore

$$\begin{pmatrix} \varphi_1(\lambda) & 0 \\ 0 & \varphi_2(\lambda) \end{pmatrix} \sim \begin{pmatrix} \varphi_1(\lambda) & \varphi_1(\lambda)\, u_1(\lambda) \\ 0 & \varphi_2(\lambda) \end{pmatrix}$$

$$\sim \begin{pmatrix} \varphi_1(\lambda) & \varphi_1(\lambda)\, u_1(\lambda) + \varphi_2(\lambda)\, u_2(\lambda) \\ 0 & \varphi_2(\lambda) \end{pmatrix} = \begin{pmatrix} \varphi_1(\lambda) & 1 \\ 0 & \varphi_2(\lambda) \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & \varphi_1(\lambda) \\ \varphi_2(\lambda) & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & \varphi_1(\lambda) \\ 0 & -\varphi_1(\lambda)\, \varphi_2(\lambda) \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 0 \\ 0 & -\varphi_1(\lambda)\, \varphi_2(\lambda) \end{pmatrix} \sim \begin{pmatrix} 1 & 0 \\ 0 & \varphi_1(\lambda)\, \varphi_2(\lambda) \end{pmatrix}$$

which is what we set out to prove.

Let us now consider the characteristic matrix

$$J - \lambda E = \begin{pmatrix} \boxed{J_1 - \lambda E_1} & & & 0 \\ & \boxed{J_2 - \lambda E_2} & & \\ & & \ddots & \\ 0 & & & \boxed{J_s - \lambda E_s} \end{pmatrix} \tag{5}$$

of the Jordan matrix J of type (2); here, E_i, $i = 1, 2, \ldots, s$, is a unit matrix of the same order as the submatrix J_i. Let the Jordan submatrices of the matrix J refer to the following distinct numbers: $\lambda_1, \lambda_2, \ldots, \lambda_t$, where $t \leqslant s$. Furthermore, let there refer to the number λ_i, $i = 1, 2, \ldots, t$, q_i Jordan submatrices, $q_i \geqslant 1$, and let the orders of the submatrices (arranged in nonincreasing order) be

$$k_{i1} \geqslant k_{i2} \geqslant \ldots \geqslant k_{iq_i} \tag{6}$$

Let it be noted (though we will not make use of this fact) that

$$\sum_{i=1}^{t} q_i = s,$$

$$\sum_{i=1}^{t} \sum_{j=1}^{q_i} k_{ij} = n$$

Applying elementary transformations to the rows and columns of matrix (5) which pass through the submatrix $J_i - \lambda E_i$ of this matrix, we will quite obviously not involve the other diagonal submatrices, whence it follows that it is possible, in matrix (5), to replace by means of elementary transformations every submatrix $J_i - \lambda E_i$, $i = 1, 2, \ldots, s$, by a corresponding submatrix of the type (4). In other words, *the matrix $J - \lambda E$ is equivalent to a diagonal matrix, the diagonal elements of which consist (aside from a certain number of units) of the following polynomials which correspond to all Jordan submatrices of the matrix J:*

$$\left.\begin{array}{l} (\lambda-\lambda_1)^{k_{11}},\ (\lambda-\lambda_1)^{k_{12}},\ \ldots,\ (\lambda-\lambda_1)^{k_{1q_1}}, \\ (\lambda-\lambda_2)^{k_{21}},\ (\lambda-\lambda_2)^{k_{22}},\ \ldots,\ (\lambda-\lambda_2)^{k_{2q_2}}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ (\lambda-\lambda_t)^{k_{t1}},\ (\lambda-\lambda_t)^{k_{t2}},\ \ldots,\ (\lambda-\lambda_t)^{k_{tq_t}} \end{array}\right\} \qquad (7)$$

We do not indicate the positions of the polynomials (7) on the principal diagonal, since the diagonal elements of any diagonal λ-matrix can be arbitrarily rearranged by interchanging rows and like columns. This is worth bearing in mind for the future.

Let q be the largest of the numbers q_i, $i = 1, 2, \ldots, t$. Denote by $e_{n-j+1}(\lambda)$ the product of polynomials in the jth column of array (7), $j = 1, 2, \ldots, q$, that is,

$$e_{n-j+1}(\lambda) = \prod_{i=1}^{t} (\lambda - \lambda_i)^{k_{ij}} \qquad (8)$$

If there are certain vacancies in the jth column—it may happen that $q_i < j$ for certain i—then the corresponding factors in (8) are considered to be unity. Since, by hypothesis, the numbers $\lambda_1, \lambda_2, \ldots, \lambda_t$ are distinct, the powers of the linear binomials in the jth column of array (7) are pairwise relatively prime. Therefore, on the basis of the lemma proved above, they can, by means of elementary transformations, be replaced in the diagonal matrix at hand by their product $e_{n-j+1}(\lambda)$ and by a certain number of units.

Doing this for $j = 1, 2, \ldots, q$, we find that

$$J - \lambda E \sim \begin{Vmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & 1 & & & \\ & & & e_{n-q+1}(\lambda) & & \\ & & & & \ddots & \\ & & & & & e_{n-1}(\lambda) \\ 0 & & & & & e_n\lambda \end{Vmatrix} \tag{9}$$

This is the desired canonical form of the matrix $J - \lambda E$. Indeed, the leading coefficients of all polynomials on the principal diagonal of (9) are equal to unity and each of the polynomials is exactly divisible by the preceding one, by Condition (6).

Example. Let

$$J = \begin{pmatrix} \boxed{\begin{matrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{matrix}} & & & & 0 \\ & \boxed{2} & & & \\ & & \boxed{\begin{matrix} 5 & 1 \\ 0 & 5 \end{matrix}} & & \\ & & & \boxed{\begin{matrix} 5 & 1 \\ 0 & 5 \end{matrix}} & \\ 0 & & & & \boxed{2} \end{pmatrix}$$

For this Jordan matrix of order 9, the polynomial array (7) is of the form

$$\begin{matrix} (\lambda - 2)^3, & \lambda - 2, & \lambda - 2, \\ (\lambda - 5)^2, & (\lambda - 5)^2 \end{matrix}$$

Therefore, the invariant factors of the J matrix are the polynomials

$$e_9(\lambda) = (\lambda - 2)^3 (\lambda - 5)^2,$$
$$e_8(\lambda) = (\lambda - 2)(\lambda - 5)^2,$$
$$e_7(\lambda) = (\lambda - 2)$$

whereas $e_6(\lambda) = \ldots = e_1(\lambda) = 1$.

Now that we have learned how, judging by the form of a given Jordan matrix J, to write down the canonical form of its characteristic matrix straightaway, we can prove the following theorem.

Two Jordan matrices are similar if and only if they consist of the same Jordan submatrices, that is to say, if they differ at most solely in the order of these submatrices on the principal diagonal.

Actually, the polynomial array (7) was completely determined by the set of Jordan submatrices of the Jordan matrix J and did not in the least reflect the arrangement of the Jordan submatrices along the principal diagonal of the matrix. It then follows that if Jordan matrices J and J' have the same set of Jordan submatrices, then they are associated with one and the same array (7) of polynomials and therefore the same polynomials (8). Thus, the characteristic matrices $J - \lambda E$ and $J' - \lambda E$ have the same invariant factors, that is to say, they are equivalent, and therefore the matrices J and J' are similar.

Conversely, if the Jordan matrices J and J' are similar, then their characteristic matrices have the same invariant factors. Let the polynomials (8) for $j = 1, 2, \ldots, q$, be those invariant factors which are different from unity. But the polynomial array (7) can be restored from the polynomials (8). Namely, the polynomials (8) can be factored into a product of powers of linear factors, since, as has already been proved, this property is possessed by the invariant factors of the characteristic matrix of any Jordan matrix. Array (7) just consists of all those maximal powers of the linear factors into which the polynomials (8) are factored. Finally, using array (7) we can restore the Jordan submatrices of the original Jordan matrices: to every polynomial $(\lambda - \lambda_i)^{k_{ij}}$ of (7) there corresponds a Jordan submatrix of order k_{ij} that refers to the number λ_i. This proves that the matrices J and J' consist of the same Jordan submatrices and differ at most in their order alone.

One consequence of this theorem is that *a Jordan matrix similar to a diagonal matrix is diagonal and that two diagonal matrices are similar if and only if they can be obtained from one another by permuting the numbers on the principal diagonal.*

Reducing a matrix to Jordan normal form. If a matrix A with elements from the field P *can be reduced to a Jordan normal form*, i.e., is similar to a Jordan matrix, then, as follows from the theorem that was proved above, *the Jordan normal form is determined uniquely for matrix A to within the order of the Jordan submatrices on the principal diagonal.* The condition that allows a matrix A to be so reduced is given in the following theorem, the proof of which offers a practical procedure for finding a Jordan matrix similar to A if such a Jordan matrix exists. Note that reducibility over the field P means that all the elements of the matrix undergoing transformation are in P.

Matrix A with elements in the field P can be reduced over P to the Jordan normal form if and only if all the characteristic roots of A lie in the base field P itself.

Indeed, if matrix A is similar to the Jordan matrix J, these two matrices have the same characteristic roots. However, the characteristic roots of J are easily found: since the determinant of the

matrix $J - \lambda E$ is equal to the product of its elements on the principal diagonal, the polynomial $|J - \lambda E|$ can be factored over P into linear factors and its roots are numbers (and only these numbers) on the principal diagonal of J.

Conversely, let all characteristic roots of matrix A be in the field P. If the different-from-unity invariant factors of the matrix $A - \lambda E$ are

$$e_{n-q+1}(\lambda), \ldots, e_{n-1}(\lambda), e_n(\lambda) \tag{10}$$

then

$$|A - \lambda E| = (-1)^n e_{n-q+1}(\lambda) \ldots e_{n-1}(\lambda) e_n(\lambda)$$

Indeed, the determinants of the matrix $A - \lambda E$ and its canonical matrix can only differ in a constant factor that is actually equal to $(-1)^n$, since such, precisely, is the leading coefficient of the characteristic polynomial $|A - \lambda E|$. Thus, among the polynomials (10) there are none equal to zero, the sum of the degrees of these polynomials is equal to n, and all can be factored over the field P into linear factors, which is due to the fact that, by hypothesis, the polynomial $|A - \lambda E|$ has such a factorization.

Let (8) be factorizations of the polynomials (10) into products of the powers of the linear factors. We use the term *elementary divisors of the polynomial* e_{n-j+1}, $j = 1, 2, \ldots, q$, for powers (different from unity) of the various linear binomials entering into its factorization (8), that is,

$$(\lambda - \lambda_1)^{k_{1j}}, \quad (\lambda - \lambda_2)^{k_{2j}}, \ldots, (\lambda - \lambda_t)^{k_{tj}}$$

We call the elementary divisors of all polynomials (10) the *elementary divisors of the matrix* A and write them down in the form of array (7).

Let us now take a Jordan matrix J of order n composed of Jordan submatrices defined as follows: with each elementary divisor $(\lambda - \lambda_i)^{k_{ij}}$ of matrix A we associate a Jordan submatrix of order k_{ij} referring to the number λ_i. It is evident that only the polynomials (10) are invariant factors, different from unity, of the matrix $J - \lambda E$. Therefore, matrices $A - \lambda E$ and $J - \lambda E$ are equivalent and, hence, matrix A is similar to the Jordan matrix J.

Example. Given a matrix

$$A = \begin{pmatrix} -16 & -17 & 87 & -108 \\ 8 & 9 & -42 & 54 \\ -3 & -3 & 16 & -18 \\ -1 & -1 & 6 & -8 \end{pmatrix}$$

Reducing the matrix $A - \lambda E$ to canonical form in the usual way, we find that the invariant factors different from unity of this matrix are the polynomials

$$e_4(\lambda) = (\lambda - 1)^2 (\lambda + 2),$$
$$e_3(\lambda) = \lambda - 1$$

We see that matrix A can be reduced to the Jordan normal form even in the field of rational numbers. Its elementary divisors are the polynomials $(\lambda - 1)^2$, $\lambda - 1$ and $\lambda + 2$ and so the matrix

$$J = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix}$$

is the Jordan normal form of the matrix A.

If we wanted to find the nonsingular matrix that transforms A to J, we would have to make use of the remarks made at the end of Sec. 60.

Finally, on the basis of the foregoing results we can prove the following necessary and sufficient condition for reducing a matrix to diagonal form, a condition that immediately yields the sufficient criterion of reducibility to diagonal form that was proved in Sec. 33.

An nth-order matrix A with elements in the field P can be reduced to diagonal form if and only if all the roots of the last invariant factor $e_n(\lambda)$ of its characteristic matrix are in P (there must be no multiple roots).

Indeed, reducibility of a matrix to diagonal form is equivalent to reducibility to a Jordan form such that all Jordan submatrices have order 1. In other words, all elementary divisors of matrix A must be polynomials of degree one. However, since all invariant factors of the matrix $A - \lambda E$ are divisors of the polynomial $e_n(\lambda)$, the last condition is equivalent to all elementary divisors of the polynomial $e_n(\lambda)$ having degree one, which is what we set out to prove.

62. Minimal Polynomials

Suppose we have a square matrix A of order n with elements in the field P. If

$$f(\lambda) = \alpha_0\lambda^k + \alpha_1\lambda^{k-1} + \ldots + \alpha_{k-1}\lambda + \alpha_k$$

is an arbitrary polynomial in the ring $P[\lambda]$, then the matrix

$$f(A) = \alpha_0 A^k + \alpha_1 A^{k-1} + \ldots + \alpha_{k-1}A + \alpha_k E$$

is called the *value* of the polynomial $f(\lambda)$ for $\lambda = A$. Note, in this respect, that the constant term of the polynomial $f(\lambda)$ is multiplied by the zero power of the matrix A, that is to say, by the unit matrix E.

It can be verified readily that if

$$f(\lambda) = \varphi(\lambda) + \psi(\lambda)$$

or

$$f(\lambda) = u(\lambda)\, v(\lambda)$$

then

$$f(A) = \varphi(A) + \psi(A)$$

and, respectively,

$$f(A) = u(A)\, v(A)$$

If the polynomial $f(\lambda)$ is annihilated by the matrix A, that is,

$$f(A) = 0$$

then A will be called the *matrix root* or (where no confusion is possible) simply the *root* of the polynomial $f(\lambda)$.

Every matrix A serves as a root of some nonzero polynomial.

We know for a fact that all square matrices of order n constitute an n^2-dimensional vector space over the field P. From this it follows that the system of $n^2 + 1$ matrices

$$A^{n^2},\ A^{n^2-1},\ \ldots,\ A,\ E$$

is linearly dependent over P, that is, in P there are elements $\alpha_0,\ \alpha_1,\ \ldots,\ \alpha_{n^2},\ \alpha_{n^2+1}$, not all zero, such that

$$\alpha_0 A^{n^2} + \alpha_1 A^{n^2-1} + \ldots + \alpha_{n^2} A + \alpha_{n^2+1} E = 0$$

Thus, matrix A proved to be a root of the nonzero polynomial

$$\varphi(\lambda) = \alpha_0 \lambda^{n^2} + \alpha_1 \lambda^{n^2-1} + \ldots + \alpha_{n^2} \lambda + \alpha_{n^2+1}$$

whose degree does not exceed n^2.

The matrix A is also a root of certain polynomials whose leading coefficients are equal to unity: it suffices to take any nonzero polynomial that can be annihilated by A and divide it by its leading coefficient. The polynomial of lowest degree with leading coefficient 1 that can be annihilated by A is called the *minimal polynomial of the matrix A*. Notice that the *minimal polynomial of A is uniquely defined*, since the difference of two such polynomials would have a lower degree than each one separately, but it would also be annihilable by the matrix A.

Any polynomial $f(\lambda)$ that is annihilable by the matrix A is exactly divisible by the minimal polynomial $m(\lambda)$ of this matrix.

Actually, if

$$f(\lambda) = m(\lambda)\, q(\lambda) + r(\lambda)$$

where the degree of $r(\lambda)$ is less than the degree of $m(\lambda)$, then

$$f(A) = m(A)\, q(A) + r(A)$$

and from $f(A) = m(A) = 0$ it follows that $r(A) = 0$, which contradicts the definition of a minimal polynomial.

Let us prove the following theorem.

The minimal polynomial of a matrix A coincides with the last invariant factor $e_n(\lambda)$ of the characteristic matrix $A - \lambda E$.

Proof. Retaining notations and using the results of Sec. 59, we can write the equation

$$(-1)^n \mid A - \lambda E \mid = d_{n-1}(\lambda)\, e_n(\lambda) \tag{1}$$

whence it follows, for one thing, that the polynomials $e_n(\lambda)$ and $d_{n-1}(\lambda)$ are not zero polynomials. Next, denote by $B(\lambda)$ the adjoint of the matrix $A - \lambda E$ (see Sec. 14),

$$B(\lambda) = (A - \lambda E)^*$$

As follows from (3), Sec. 14, the equation

$$(A - \lambda E)\, B(\lambda) = \mid A - \lambda E \mid E \tag{2}$$

holds true. On the other hand, since the elements of $B(\lambda)$ are $(n-1)$th order minors (with plus or minus signs) of the matrix $A - \lambda E$, and only these minors, and the polynomial $d_{n-1}(\lambda)$ is the greatest common divisor of all these minors, it follows that

$$B(\lambda) = d_{n-1}(\lambda)\, C(\lambda) \tag{3}$$

the greatest common divisor of the elements of matrix $C(\lambda)$ being equal to 1.

From equations (2), (3) and (1) follows the equation

$$(A - \lambda E)\, d_{n-1}(\lambda)\, C(\lambda) = (-1)^n\, d_{n-1}(\lambda)\, e_n(\lambda)\, E$$

We can divide through by the nonzero factor $d_{n-1}(\lambda)$, as follows from the general remark that if $\varphi(\lambda)$ is a nonzero polynomial and $D(\lambda) = (d_{ij}(\lambda))$ is a nonzero λ-matrix [let $d_{st}(\lambda) \not\equiv 0$], then the (s, t) position in the matrix $\varphi(\lambda)\, D(\lambda)$ will be occupied by the nonzero element $\varphi(\lambda)\, d_{st}(\lambda)$. Thus,

$$(A - \lambda E)\, C(\lambda) = (-1)^n\, e_n(\lambda)\, E$$

whence

$$e_n(\lambda)\, E = (\lambda E - A)\, [(-1)^{n+1}\, C(\lambda)] \tag{4}$$

This equation shows that the remainder resulting from "left" division of the λ-matrix in the left member by the binomial $\lambda E - A$ is equal to zero. From the lemma proved at the end of Sec. 60 it follows, however, that this remainder is equal to the matrix $e_n(A)\, E = e_n(A)$. True enough, the matrix $e_n(\lambda)\, E$ may be written as a matrix λ-polynomial whose coefficients are scalar matrices, i.e., such as commute with the matrix A. Thus,

$$e_n(A) = 0$$

which is to say that the polynomial $e_n(\lambda)$ is indeed annihilated by A.

From this it follows that the polynomial $e_n(\lambda)$ is exactly divisible by the minimal polynomial $m(\lambda)$ of matrix A,

$$e_n(\lambda) = m(\lambda)\, q(\lambda) \tag{5}$$

It is clear that the leading coefficient of the polynomial $q(\lambda)$ is equal to unity.

Since $m(A) = 0$, then, by the same lemma of Sec. 60, the remainder after left-division of the λ-matrix $m(\lambda) E$ by the binomial $\lambda E - A$ is again equal to zero; that is,

$$m(\lambda) E = (\lambda E - A) Q(\lambda) \tag{6}$$

The equations (5), (4) and (6) lead to the equation

$$(\lambda E - A)[(-1)^{n+1} C(\lambda)] = (\lambda E - A)[Q(\lambda) q(\lambda)]$$

The common factor $\lambda E - A$ can be cancelled out of both sides since the leading coefficient E of this matrix λ-polynomial is a nonsingular matrix. Thus,

$$C(\lambda) = (-1)^{n+1} Q(\lambda) q(\lambda)$$

We recall, however, that the greatest common divisor of the elements of matrix $C(\lambda)$ is unity. Therefore, the polynomial $q(\lambda)$ must be of degree zero, and since its leading coefficient is unity, $q(\lambda) = 1$. Thus, by (5),

$$e_n(\lambda) = m(\lambda)$$

which completes the proof.

Since, by (1), the characteristic polynomial of matrix A is exactly divisible by the polynomial $e_n(\lambda)$, there follows from the theorem just proved the Cayley-Hamilton theorem.

Cayley-Hamilton Theorem. *Every matrix is a root of its characteristic polynomial.*

The minimal polynomial of a linear transformation. Let us first prove the following assertion.

If matrices A and B are similar and if the polynomial $f(\lambda)$ is annihilated by matrix A, then it is also annihilated by matrix B.

Indeed, let

$$B = C^{-1}AC$$

If

$$f(\lambda) = \alpha_0\lambda^k + \alpha_1\lambda^{k-1} + \ldots + \alpha_{k-1}\lambda + \alpha_k$$

then

$$\alpha_0 A^k + \alpha_1 A^{k-1} + \ldots + \alpha_{k-1}A + \alpha_k E = 0$$

Transforming both sides of this equation by matrix C, we get

$$C^{-1}(\alpha_0 A^k + \alpha_1 A^{k-1} + \ldots + \alpha_{k-1}A + \alpha_k E) C$$
$$= \alpha_0 (C^{-1}AC)^k + \alpha_1 (C^{-1}AC)^{k-1} + \ldots + \alpha_{k-1}(C^{-1}AC) + \alpha_k E$$
$$= \alpha_0 B^k + \alpha_1 B^{k-1} + \ldots + \alpha_{k-1}B + \alpha_k E = 0$$

i.e. $f(B) = 0$.

From this it follows that *similar matrices have one and the same minimal polynomial.*

Now let φ be a linear transformation of an n-dimensional linear space over the field P. The matrices that represent this transformation in different bases of space are similar. The common minimal polynomial of these matrices is termed the *minimal polynomial of the linear transformation* φ.

Using the operations (on linear transformations) introduced in Sec. 32, we can introduce the concept of the *value* of the polynomial

$$f(\lambda) = \alpha_0\lambda^k + \alpha_1\lambda^{k-1} + \ldots + \alpha_{k-1}\lambda + \alpha_k$$

from the ring $P[\lambda]$ for λ equal to the linear transformation φ; this is the linear transformation

$$f(\varphi) = \alpha_0\varphi^k + \alpha_1\varphi^{k-1} + \ldots + \alpha_{k-1}\varphi + \alpha_k\varepsilon$$

where ε is the identity transformation.

We furthermore say that the polynomial $f(\lambda)$ is *annihilated* by the linear transformation φ if

$$f(\varphi) = \omega$$

where ω is the zero transformation.

If the reader takes into account the relationship between operations on linear transformations and on matrices, it will be easy for him to prove that *the minimal polynomial of the linear transformation* φ *is that uniquely determined polynomial of minimum degree with leading coefficient* 1 *which is annihilated by the transformation* φ. After that the results obtained above, in particular the Cayley-Hamilton theorem, can be rephrased in the language of linear transformations.

CHAPTER 14

GROUPS

63. Definition of a Group. Examples

Rings and fields, which played so important a role in the previous chapters, are algebraic systems with two independent operations: addition and multiplication. However, there are many areas of mathematics and its application in which we very often encounter algebraic systems with only one algebraic operation defined. Thus, confining ourselves to examples that have already appeared in this book, we have the set of permutations of degree n (see Sec. 3) in which we defined the single operation of multiplication. On the other hand, the definition of a vector space (Sec. 8) includes the addition of vectors, whereas multiplication of vectors was not defined (notice that the multiplication of a vector by a scalar does not satisfy the definition—given in Sec. 44—of an algebraic operation).

Groups form the most important type of algebraic systems with a single operation. This concept has extensive applications and forms the subject of a whole science—the theory of groups. The present chapter may be regarded as an introduction to the theory of groups, including such elementary facts about groups as are needed by every mathematician and also, at the end, a theorem that is not so elementary.

Let us agree, as is the custom in group theory, to call the algebraic operation at hand *multiplication* and to use appropriate symbolism. It will be recalled (see Sec. 44) that an algebraic operation is always assumed to be valid and unique: for any two elements a and b of a given set the product ab exists and is a uniquely defined element of the set.

A *group* is a set G with one algebraic operation that is associative (though not necessarily commutative); the operation must have an inverse.

Because of the possible noncommutativity of the group operation, the possibility of the inverse operation signifies the following:

for any two elements a and b in G there exist in G a uniquely defined element x and a uniquely defined element y such that

$$ax = b, \qquad ya = b$$

If a group G consists of a finite number of elements, then it is called a *finite group*, and the number of elements in it is the *order* of the group. If the operation defined in G is *commutative*, then G is called a *commutative* group or an *Abelian* group.

Some simple consequences follow from the definition of a group. On the basis of reasoning already given in Sec. 44, we can assert that the associative law permits speaking in unique fashion about the *product of any finite number of elements of a group* specified (due to the possible noncommutativity of the group operation) in a definite order.

Let us examine the consequences which follow from the existence of the inverse operation.

Let an arbitrary element a be given in a group G. From the definition of a group there follows the existence in G of a uniquely defined element e_a such that $ae_a = a$; thus, this element plays the role of unity (identity) when multiplied by element a. If b is any other element of G and if y is a group element satisfying the equation $ya = b$ (its existence follows from the definition of a group), we get

$$b = ya = y\,(ae_a) = (ya)\,e_a = be_a$$

Thus, the element e_a plays the role of a right-identity with respect to all elements of the group G, and not only with respect to the initial element a; we therefore denote it by e'. From the unambiguousness implicit in the definition of the inverse operation follows the uniqueness of this element.

In similar fashion, we can prove the existence and uniqueness in the group G of an element e'' that satisfies the condition $e''a = a$ for all a in G. Indeed, the elements e' and e'' coincide since the equalities $e''e' = e''$ and $e''e' = e'$ imply $e'' = e'$. This proves that *in any group G there is a uniquely defined element e satisfying the condition*

$$ae = ea = a$$

for all a in G. This element is termed the *unit* (*identity*) *element* of G and is ordinarily denoted by the symbol 1.

From the definition of a group there also follows the existence and uniqueness, for a given element a, of elements a' and a'' such that

$$aa' = 1, \qquad a''a = 1$$

Actually, the elements a' and a'' coincide; from the equalities

$$a''aa' = a''\,(aa') = a'' \cdot 1 = a'',$$

$$a''aa' = (a''a)\,a' = 1 \cdot a' = a'$$

follows $a'' = a'$. This element is called the *inverse element* of a and is denoted by a^{-1}, that is,

$$aa^{-1} = a^{-1}a = 1$$

Thus, *every element of a group has a unique inverse element.*

From the foregoing equalities it follows that the inverse of the element a^{-1} is the element a itself. It is readily seen that the inverse of a product of several elements is the product of the inverses taken in the opposite order:

$$(a_1a_2 \ldots a_{n-1}a_n)^{-1} = a_n^{-1}a_{n-1}^{-1} \ldots a_2^{-1}a_1^{-1}$$

Finally, the unit element is its own inverse.

To check whether a given set with one operation is a group is greatly simplified by the fact that in the definition of a group the requirement that there be an inverse operation can be replaced by the assumption of the existence of a unit (identity) element and inverse elements (and only on one side, say, the right, and without any assumption about their uniqueness). This follows from the theorem which we will now prove.

A set G with a single associative operation is a group if there is at least one element e in G with the property

$$ae = a \quad \textit{for all } a \textit{ in } G$$

and if among the right-identities there is at least one element e_0 such that, relative to it, any element a in G has at least one right-inverse a^{-1}:

$$aa^{-1} = e_0$$

Proof. Let a^{-1} be one of the right-inverses of a. Then

$$aa^{-1} = e_0 = e_0e_0 = e_0aa^{-1}$$

That is, $aa^{-1} = e_0aa^{-1}$. Multiplying both sides of this equation on the right by one of the elements that are right-inverse for a^{-1}, we get $ae_0 = e_0ae_0$, whence $a = e_0a$, since e_0 is a right-identity of G. Thus, the element e_0 also turns out to be a left-identity of G. Now if e_1 is an arbitrary right-identity, e_2 an arbitrary left-identity, then from the equalities

$$e_2e_1 = e_1 \quad \text{and} \quad e_2e_1 = e_2$$

there follows $e_1 = e_2$, i.e., any right-identity is equal to any left-identity. This completes the proof of the existence and uniqueness, in the set G, of a unit element (identity) which we denote (as before) by 1.

Furthermore,

$$a^{-1} = a^{-1} \cdot 1 = a^{-1}aa^{-1}$$

That is, $a^{-1} = a^{-1}aa^{-1}$, where a^{-1} is one of the right-inverses for a. Multiplying both sides of the last equality on the right by one of

the right-inverses of a^{-1}, we get $1 = a^{-1}a$, i.e., the element a^{-1} will also serve as a left-inverse of a. Now, if a_1^{-1} is an arbitrary right-inverse of a, a_2^{-1} is an arbitrary left-inverse, then from the equalities

$$a_2^{-1}aa_1^{-1} = (a_2^{-1}a)\,a_1^{-1} = a_1^{-1},$$
$$a_2^{-1}aa_1^{-1} = a_2^{-1}\,(aa_1^{-1}) = a_2^{-1}$$

there follows $a_1^{-1} = a_2^{-1}$, which is to say, there follows the existence and uniqueness of the inverse a^{-1} of any element a in G.

It is now easy to show that the set G is a group. Indeed, the equations $ax = b$, $ya = b$ will be satisfied, as is readily seen, by the elements

$$x = a^{-1}b, \qquad y = ba^{-1}$$

The uniqueness of these solutions follows from the fact that if, say, $ax_1 = ax_2$, then, multiplying both sides of this equation on the left by a^{-1}, we get $x_1 = x_2$. The theorem is proved.

We have already encountered the concept of an isomorphism: for rings, for linear spaces and for Euclidean spaces. This concept can be defined for groups as well, and it plays just as important a role in group theory as it does in the theory of rings. Groups G and G' are termed *isomorphic* if a one-to-one correspondence can be established between them such that, under it, for any elements a, b in G and for the corresponding elements a', b' in G', to the product ab corresponds the product $a'b'$. As in Sec. 46 (for the zero element and the inverse element of a ring), it may be shown that, given an isomorphic correspondence between groups G and G', the unit element of G is associated with the unit element of G', and if a in G is associated with a' in G', then a^{-1} is associated with a'^{-1}.

Passing now to **examples of groups**, we notice that if the operation in the group G is called *addition*, then the identity (unit) element of the group is *zero* and is denoted by 0, and in place of the inverse element we speak of the *opposite element* (*additive inverse*) denoted by $-a$.

As a first instance of a group, note that, *with respect to addition, any ring* (*and, in particular, any field*) *is a group, it is an Abelian group*. This is the so-called *additive group of a ring*. This remark immediately yields a wealth of concrete examples of groups: the additive group of integers, the additive group of even numbers, additive groups of the rational numbers, the reals, the complex numbers, etc. Note that *the additive groups of integers and of even numbers are isomorphic with each other*, although the latter is only a part of the former: a mapping that associates with every integer k an even number $2k$ is one-to-one and, as can easily be verified, is even an isomorphic mapping of the former group onto the latter.

No ring is a group with respect to multiplication because the inverse operation (division) is not always possible. The situation

does not change if we pass from an arbitrary ring to a field, since division by zero does not hold in a field. However, let us consider the collection of all nonzero elements of a field. Since a field does not contain divisors of zero (that is the product of two nonzero elements is also nonzero), it follows that multiplication is an algebraic operation for this set: it will be associative and commutative. The set of all nonzero elements of a field will be closed under division. Hence, *the set of nonzero elements of any field is an Abelian group.* It is called a *multiplicative group of the field.* Instances of such groups are the multiplicative groups of the rational numbers, the real numbers, the complex numbers.

Obviously, all positive real numbers constitute a group with respect to multiplication. *This group is isomorphic to the additive group of all real numbers*: associating a real number $\ln a$ with an arbitrary positive number a, we get a one-to-one mapping of the first group onto the second group; this mapping is an isomorphism due to the equality

$$\ln (ab) = \ln a + \ln b$$

Let us now take the set of nth roots of unity in the field of complex numbers. In Sec. 19 we proved that the product of two nth roots of unity and also the inverse of an nth root of unity belong to this set of numbers. Since unity, quite naturally, belongs to this set and since multiplication of complex numbers is associative and commutative, we find that *the nth roots of unity constitute an Abelian group with respect to multiplication; it is a finite group of order n.* Thus, *for any natural number n there exist finite groups of order n.*

The group (with respect to multiplication) of the nth roots of unity is isomorphic to the additive group of the ring Z_n constructed in Sec. 45. Indeed, if ε is a primitive nth root of unity, then all elements of the first of these groups is of the form ε^k, $k = 0, 1, \ldots, n-1$. If we associate with every number ε^k an element C_k of the ring Z_n, i.e., the class of integers which yield k as remainder upon division by n, we get an isomorphic correspondence between the groups under consideration: if $0 \leqslant k \leqslant n-1$, $0 \leqslant l \leqslant n-1$ and if $k + l = nq + r$, where $0 \leqslant r \leqslant n-1$, and q is equal to 0 or 1, then $\varepsilon^k \cdot \varepsilon^l = \varepsilon^r$ and, at the same time, $C_k + C_l = C_r$.

At this point, it is worth indicating some numerical sets that are not groups. Thus, the set of all integers is not a group with respect to multiplication, the set of all positive real numbers is not a group with respect to addition, the set of all odd numbers is not a group with respect to addition, the set of all negative real numbers is not a group with respect to multiplication. All these assertions can easily be verified.

All the numerical groups considered above are of course Abelian. Instances of Abelian groups not made up of numbers are the linear

spaces: as follows from their definition (see Secs. 29, 47), *any linear space over an arbitrary field P is an Abelian group with respect to the operation of addition.*

Let us now examine examples of noncommutative groups.

The set of all nth-order matrices over the field P is not a group with respect to the operation of multiplication since the demand that there be an inverse breaks down. However, if we confine our attention to nonsingular matrices, then we get a group. Indeed, the product of two nonsingular matrices is, as we know, nonsingular, the unit matrix is nonsingular; every nonsingular matrix has an inverse which is also nonsingular and, finally, the associative law, which holds for all matrices, holds true in the particular case of nonsingular matrices. We can therefore speak of the *group of nonsingular matrices of order n* over the field P with matrix multiplication as the group operation. This group is noncommutative for $n \geqslant 2$.

The multiplication of permutations introduced in Sec. 3 is a very important example of a finite noncommutative group. We know that in the set of all permutations of degree n multiplication is an algebraic operation which is associative, although for $n \geqslant 3$ it is noncommutative, that the identity permutation E is the identity of this multiplication and that every permutation has an inverse. Thus, *the set of permutations of degree n constitutes a group with respect to multiplication; it is a finite group of order $n!$*. This group is termed a *symmetric group of degree n* and is noncommutative for $n \geqslant 3$.

In place of the set of all permutations of degree n, let us consider only the set of even permutations, which, as we know, consists of $\frac{1}{2}n!$ elements. Using the theorem, proved in Sec. 3, that the parity of a permutation coincides with the parity of the number of transpositions entering into some decomposition of this permutation into a product of transpositions, we find that *the product of two even permutations is even*. Indeed, we obtain the representation of AB as a product of transpositions by writing the appropriate decompositions of A and B one after the other. Furthermore, the associativity of multiplication of permutations is known, and the evenness of the identity permutation is obvious. Finally, the evenness of the permutation A^{-1} for the even permutation A follows at least from the fact that the notations of these permutations may be obtained one from the other by interchanging the upper and lower rows; that is to say, they contain an equal number of inversions. Thus, *the set of even permutations of degree n is a finite group of order $\frac{1}{2}n!$ with respect to multiplication*. This group is called an *alternating group of degree n*. It is easy to verify that it is noncommutative for $n \geqslant 4$, although it is commutative for $n = 3$.

Symmetric and alternating groups play a prominent role in the theory of finite groups and also in the Galois theory. Notice that it would be impossible, by analogy with alternating groups, to construct a group of odd permutations with respect to multiplication, since the product of two odd permutations is always an even permutation.

A large number of diverse examples of groups are found in the various branches of geometry. Just one simple example of this nature: the set of all rotations of a sphere about its centre is a group; it is noncommutative if we call the result of two successive rotations the product of these rotations.

64. Subgroups

A subset A of a group G is called a *subgroup* of this group if it is a group with respect to the operation defined in G.

To find out whether a subset A of group G is a subgroup of G, it is sufficient to verify that: (1) the product of any two elements of A lies in A; (2) A contains every element and the inverse of every element of A. Indeed, from the fact that the associative law holds in G it follows that it holds for elements in A; the fact that the unit element of G belongs to A follows from (2) and (1).

Many of the groups named in Sec. 63 are subgroups of other groups indicated there. For instance, the additive group of even numbers is a subgroup of the additive group of all integers, and the latter, in its turn, is a subgroup of the additive group of rational numbers. All these groups, like the additive groups of numbers in general, are subgroups of the additive group of complex numbers. The multiplicative group of positive real numbers is a subgroup of the multiplicative group of all nonzero real numbers. The alternating group of degree n is a subgroup of the symmetric group of the same degree.

There is a point to stress: the requirement contained in the definition of a subgroup that the subset A of group G be a group with respect to the group operation defined in G is essential. Thus, the multiplicative group of positive real numbers is not a subgroup of the additive group of all real numbers, although the former set is a subset of the latter.

If we take subgroups A and B in the group G, then their intersection $A \cap B$, that is, the collection of elements common to A and B, is also a subgroup of G.

Indeed, if the intersection $A \cap B$ contains elements x and y, then they lie in the subgroup A and for this reason the product xy and the inverse x^{-1} belong to A as well. By the same reasoning, the elements xy and x^{-1} belong to the subgroup B and therefore they are contained in the intersection $A \cap B$ too.

It is readily seen that this result holds true not only for two subgroups, but for any number of subgroups, whether finite or even infinite.

The subset of group G consisting of the single element 1 is obviously a subgroup of this group. This subgroup, which is contained in any other subgroup of G, is called the *unit subgroup* of group G. On the other hand, the group G itself is one of its own subgroups.

An interesting example of subgroups are the so-called *cyclic subgroups*. Let us introduce the concept of the *power* of an element a of group G. If n is any natural number, then the product of n elements equal to the element a is called the nth *power* of the element a and is denoted by a^n. *Negative powers* of element a may be defined either as elements of group G inverse to the positive powers of this element or as products of several factors equal to the element a^{-1}. These definitions actually coincide:

$$(a^n)^{-1} = (a^{-1})^n, \qquad n > 0 \tag{1}$$

To prove this, take the product of $2n$ factors, of which the first n are equal to a and the remaining ones are equal to a^{-1}, and perform the cancellations. The element equal both to the left member and the right member of (1) will be denoted by a^{-n}. Finally, let us agree to use the term *zero power* a^0 of element a for the element 1.

Note that if the operation in the group G is called addition, then in place of powers of a we should speak of *multiples* of this element and write ka.

It is easy to show that in any group G, we have for the powers of any element a for any exponents m and n (positive, negative, or zero) the following equalities:

$$a^n \cdot a^m = a^m \cdot a^n = a^{n+m}, \tag{2}$$

$$(a^n)^m = a^{nm} \tag{3}$$

We denote by $\{a\}$ the subset of G composed of all powers of the element a, including the element a itself as its first power. *The subset* $\{a\}$ *is a subgroup of the group* G: multiplication of the elements of $\{a\}$ lies in $\{a\}$ by (2); $\{a\}$ has the element 1, equal to a^0, and, finally, $\{a\}$ contains all its elements together with all the inverse elements, since from (3) follows the equality

$$(a^n)^{-1} = a^{-n}$$

The subgroup $\{a\}$ is called *a cyclic subgroup of the group* G *generated by the element* a. As is evident from (2), it is always commutative, even if the group G itself is noncommutative.

Notice that it has not been asserted above that all powers of the element a are distinct elements of the group. If this is indeed so, then a is called an *element of infinite order*. However, let there be,

among the powers of a, some which are equal, say, $a^k = a^l$ for $k \neq l$; this is always the case for finite groups, but it may also occur in an infinite group as well. If $k > l$, then

$$a^{k-l} = 1$$

which is to say that there are positive powers of the element a that are equal to unity. Let n be the least positive power of the element a equal to unity, that is,

$$(1) \quad a^n = 1, \quad n > 0,$$

$$(2) \quad \text{if } a^k = 1, \quad k > 0, \quad \text{then} \quad k \geqslant n$$

In this case we say that a is an *element of finite order*, namely, *of order* n.

If an element a is of finite order n, then all the elements

$$1, \ a, \ a^2, \ \ldots, \ a^{n-1} \tag{4}$$

will be distinct, as is clearly seen. *Any other power of the element* a, *whether positive or negative, is equal to one of the elements of* (4). Indeed, if k is any integer, then, dividing k by n, we get

$$k = nq + r, \qquad 0 \leqslant r \leqslant n$$

and so, by (2) and (3),

$$a^k = (a^n)^q \cdot a^r = a^r \tag{5}$$

Whence it follows that *if the element* a *is of finite order* n, *and* $a^k = 1$, *then* k *must be exactly divisible by* n. On the other hand, since

$$-1 = n(-1) + (n - 1)$$

it follows that *for the element* a *of finite order* n

$$a^{-1} = a^{n-1}$$

Since the system (4) contains n elements, it follows from the results obtained above that *for element* a *of finite order its order* n *coincides with the order* (*that is to say, with the number of elements*) *of the cyclic subgroup* $\{a\}$.

Finally, notice that any group has one and only one element of the first order: this is the element 1. The cyclic subgroup $\{1\}$ evidently coincides with the unit subgroup.

Cyclic groups. A group G is called a *cyclic* group if it consists of the powers of one of its elements a, that is, if it coincides with one of its cyclic subgroups $\{a\}$; here, the element a is called the *generator* of the group G. It is obvious that every cyclic group is Abelian.

An example of an infinite cyclic group is the additive group of the integers—any integer which is a multiple of the number 1;

that is to say, this number serves as the generator of the group at hand. We could also take -1 for the generator.

An example of a finite cyclic group of order n is the multiplicative group of the nth roots of unity; in Sec. 19 it is shown that all these roots are powers of one of them, namely, the primitive root.

The following theorem shows that, essentially, these examples exhaust all cyclic groups.

All infinite cyclic groups are isomorphic among themselves; all finite cyclic groups of a given order n are also isomorphic among themselves.

Indeed, an infinite cyclic group with generator a is mapped one-to-one onto the additive group of the integers if every element a^k of this group is associated with the number k; this mapping is isomorphic, since, by (2), in multiplying the powers of the element a we add the exponents. Now if we are given a finite cyclic group G of order n with generator a, then we denote by ε the primitive nth root of unity and associate with every element a^k of group G, $0 \leqslant k < n$, the number ε^k. This is a one-to-one mapping of the group G onto the multiplicative group of the nth roots of unity, the isomorphic property of which follows from (2) and (5).

This theorem enables us to speak simply about an *infinite cyclic group* or about a *cyclic group of order n*.

We now prove the following theorem.

Every subgroup of a cyclic group is itself cyclic.

Indeed, let $G = \{a\}$ be a cyclic group with generator a (infinite or finite) and let A be a subgroup of G. We assume that A is different from the unit subgroup, otherwise there would be nothing to prove. Suppose that a^k is the least positive power of a contained in A. There is such a power, since if A contains an element a^{-s}, $s > 0$, different from 1, then A also contains the inverse element a^s. Assume that A also has an element a^l, $l \neq 0$, and k does not divide l. Then if d, $d > 0$, is the greatest common divisor of the numbers k and l, there exist integers u and v such that

$$ku + lv = d$$

and therefore the subgroup A must contain the element

$$(a^k)^u \cdot (a^l)^v = a^d$$

but since under our assumptions $d < k$, we are in conflict with the choice of the element a^k. This is proof that $A = \{a^k\}$.

Decomposition of a group with respect to a subgroup. If we take subsets M and N in a group G, then the *product MN of these subsets* is to be understood as the collection of elements of G that are representable in at least one way as the product of an element of M by an element of N. From the associativity of the group opera-

tion follows the *associativity of multiplication of subsets of the group*,

$$(MN)\,P = M\,(NP)$$

One of the sets M, N may of course consist of just the one element a. In this case we get *the product aN of the element by the set or the product Ma of the set by the element.*

Suppose in G we have an arbitrary subgroup A. If x is any element of G, then the product xA is called *the left coset (of the group G with respect to the subgroup A) generated by element x. The element x naturally lies in the coset xA* since the subgroup A contains a unit element, but $x \cdot 1 = x$.

Every left coset is generated by any one of its elements, that is to say, if an element y lies in the coset xA, then

$$yA = xA \tag{6}$$

This is true because y may be represented as

$$y = xa$$

where a is an element of the subgroup A. Therefore, for any elements a' and a'' in A it will be true that

$$ya' = x\,(aa'),$$
$$xa'' = y\,(a^{-1}a'')$$

which proves (6).

From this it follows that *any two left cosets of the group G relative to the subgroup A either coincide or do not have any element in common.* Indeed, if the cosets xA and yA have a common element z, then

$$xA = zA = yA$$

Thus, the entire group G decomposes into disjoint left cosets relative to the subgroup A. This decomposition is called *the left decomposition of the group G relative to the subgroup A.*

Note that one of the left cosets of this decomposition is the subgroup A itself; this coset is generated by the element 1 or, generally, by any element a in A, since

$$aA = A$$

Naturally, taking the product Ax as *the right coset of the group G relative to the subgroup A—this coset being generated by the element x*—we obtain, in similar fashion, *a right decomposition of the group G relative to the subgroup A.* For an Abelian group, both its decompositions (left and right) relative to any subgroup will naturally coincide, so we can simply speak of the *decomposition of a group relative to a subgroup.*

For instance, the decomposition of the additive group of the integers relative to the subgroup of the multiples of the number k,

consists of k distinct cosets generated, respectively, by the numbers 0, 1, 2, . . ., $k-1$. Here, the coset generated by the number l, $0 \leqslant l \leqslant k-1$, contains all the numbers which upon division by k yield the remainder l.

In the noncommutative case, the decompositions of a group relative to a subgroup may prove to be distinct.

To illustrate, let us consider a symmetric group of degree 3, S_3; as in Sec. 3, we write its elements as cycles. For the subgroup A we take the cyclic subgroup of the element (12); it consists of the identity permutation and the permutation (12) itself. The other left cosets are: $(13)\cdot A$, consisting of the permutations (13) and (132), and $(23)\cdot A$, consisting of the permutations (23) and (123). On the other hand, the right cosets of the group S_3 relative to the subgroup A are: the subgroup A itself, the coset $A\cdot(13)$, consisting of the permutations (13) and (123), and the coset $A\cdot(23)$, consisting of the permutations (23) and (132). We see that in this case, the right decomposition differs from the left decomposition.

For the case of finite groups, the existence of decompositions of a group relative to a subgroup leads to the following important theorem.

Lagrange's theorem. *In every finite group, the order of any subgroup is a divisor of the order of the group itself.*

Indeed, in a finite group G of order n let there be given a subgroup A of order k. We consider the left decomposition of the group G relative to the subgroup A. Let it consist of j cosets; the number j is termed the *index* of the subgroup A in the group G. Every left coset xA consists of exactly k elements, since if

$$xa_1 = xa_2$$

where a_1 and a_2 are elements of A, then $a_1 = a_2$. Thus,

$$n = kj \tag{7}$$

which completes the proof.

Since the order of an element coincides with the order of its cyclic subgroup, it follows from the Lagrange theorem that *the order of any element of a finite group is a divisor of the order of the group.*

It also follows from the Lagrange theorem that *any finite group whose order is a prime number is cyclic.*

Indeed, this group must coincide with the cyclic subgroup generated by any element of it that is different from unity.

Hence, by the above-obtained description of cyclic groups, it follows that *for any prime p there is a unique, to within isomorphism, finite group of order p.*

65. Normal Divisors, Factor Groups, Homomorphisms

A subgroup A of a group G is called a *normal divisor* of this group (or an *invariant subgroup*) if the left decomposition of G with respect to A coincides with the right decomposition.

Thus, all subgroups of an Abelian group are normal divisors in it. On the other hand, in any group G both the unit subgroup and the group itself are normal divisors: both decompositions of G with respect to the unit subgroup coincide with the decomposition of the group into separate elements, and both decompositions of the group G with respect to the group itself consist of the single coset G.

Here are some of the more interesting examples of normal divisors in noncommutative groups. In the symmetric group of degree 3, S_3, the cyclic subgroup of element (123) consisting of the identity permutation and the permutations (123) and (132) is a normal divisor: in both decompositions of the group S_3 with respect to this subgroup, the second coset consists of the permutations (12), (13) and (23).

Generally, in the symmetric group S_n of degree n the alternating group A_n of degree n is a normal divisor. Indeed, the group A_n is of order $\frac{1}{2}n!$, and so any coset of the group S_n with respect to the subgroup A_n must consist of the same number of elements and, consequently, there is only one other such coset, namely, the collection of odd permutations.

In the multiplicative group of nonsingular square matrices of order n with elements in the field P, those matrices whose determinants equal 1 obviously constitute a subgroup. It will even be a normal divisor, since the class of all matrices whose determinants are equal to the determinant of the matrix M is the coset (simultaneously left and right) with respect to this subgroup, which coset is generated by the matrix M. It suffices to recall that in the multiplication of matrices the determinants are multiplied together.

The definition of a normal divisor given above may be rephrased.

A subgroup A of group G is a normal divisor of this group if for any element x in G

$$xA = Ax \tag{1}$$

That is to say, for any element x in G and an element a in A, it is possible to choose elements a' and a'' in A such that

$$xa = a'x, \qquad ax = xa'' \tag{2}$$

There are other definitions of a normal divisor equivalent to the original one. Thus, we call elements a and b of group G *conjugate* if in G there is at least one element x such that

$$b = x^{-1}ax \tag{3}$$

or we say that element b is the *transform* of element a by x. From (3) it evidently follows that

$$a = xbx^{-1} = (x^{-1})^{-1} bx^{-1}$$

A subgroup A of a group G is a normal divisor in G if and only if, together with any element of it, a, it also contains all elements conjugate to it in G.

Indeed, if A is a normal divisor in G, then, by (2), for the element a that we chose in A and for any element x in G we can find in A an element a'' such that

$$ax = xa''$$

Whence

$$x^{-1}ax = a''$$

That is, any element conjugate to a lies in A. Conversely, if a subgroup A contains, together with any element a, all elements conjugate to a, then in particular A also contains the element

$$x^{-1}ax = a''$$

whence follows the second of the equalities (2). For the same reason, A also contains the element

$$(x^{-1})^{-1} ax^{-1} = xax^{-1} = a'$$

whence follows the first of the equalities (2).

Using this result, it is easy to prove that *the intersection of any normal divisors of group G will itself be a normal divisor of this group.* Indeed, if A and B are normal divisors of G, then, as demonstrated in the preceding section, the intersection $A \cap B$ is a subgroup of G. Let c be any element of $A \cap B$ and x any element of G. Then the element $x^{-1}cx$ must lie both in A and B since both of these normal divisors contain the element c. Whence it follows that the element $x^{-1}cx$ is in the intersection $A \cap B$.

Factor group. The significance of the concept of a normal divisor is based on the fact that it is possible, in a certain very natural way, to construct a new group from the cosets with respect to a normal divisor—due to (1) there is no need in this case to distinguish between left and right cosets.

First notice that if A is an arbitrary subgroup of the group G, then

$$AA = A \tag{4}$$

since the product of any two elements of the subgroup A belongs to A and, at the same time, by multiplying all elements of A by the unit element we already get the entire subgroup A.

Let A now be a normal divisor of G. *In this case, the product of any two cosets of G with respect to A* (in the sense of multiplying sub-

sets of the group G) *will itself be a coset with respect to* A. Indeed, using the associativity of the multiplication of subsets of a group, and using equality (4) and

$$yA = Ay$$

[cf. (1)], we get

$$xA \cdot yA = xyAA = xyA \tag{5}$$

for any elements x and y of G.

Equation (5) shows that in order to find the product of two given cosets of group G with respect to the normal divisor A, we must choose in arbitrary fashion one representative in each coset (recall that every coset is generated by any one of its elements) and take the coset containing the product of these representatives.

Thus is defined the operation of multiplication in the set of all cosets of the group G with respect to the normal divisor A. We will show that *all the requirements that enter into the definition of a group are thus fulfilled.* The associativity of multiplication of cosets follows from the associativity of the multiplication of subsets of the group. The role of the unit element is played by the normal divisor A itself, which is one of the cosets of the decomposition of G with respect to A: namely, by (4) and (1), it is true that for any x in G,

$$xA \cdot A = xA, \qquad A \cdot xA = xAA = xA$$

Finally, the coset $x^{-1}A$ is the inverse of the coset xA since

$$xA \cdot x^{-1}A = 1 \cdot A = A$$

The group thus constructed is called the *factor group* of the group G with respect to the normal divisor A and is denoted G/A.

We see that every group is associated with a whole set of new groups—its factor groups with respect to different normal divisors. Here, the factor group of the group G with respect to the unit subgroup will, naturally, be isomorphic to G itself.

Every factor group G/A *of an Abelian group* G *is itself Abelian,* since from $xy = yx$ it follows that

$$xA \cdot yA = xyA = yxA = yA \cdot xA$$

Every factor group G/A *of a cyclic group* G *is cyclic,* because if G is generated by an element g, $G = \{g\}$, and if we are given an arbitrary coset xA, then there is an integer k such that

$$x = g^k$$

and so

$$xA = (gA)^k$$

The order of any factor group G/A *of a finite group* G *is a divisor of the order of the group itself.* Indeed, the order of the factor group

G/A is equal to the index of the normal divisor A in the group G, and so we can take advantage of (7) of the preceding section.

Here are some instances of factor groups. Since, in the additive group of the integers, the subgroup of multiples of the natural number k has, as shown in the preceding section, index k, the factor group of our group with respect to this subgroup is a finite group of order k; it is a cyclic group because the group under consideration is itself cyclic.

The factor group of a symmetric group S_n of degree n with respect to an alternating group A_n of degree n is a group of order 2; because 2 is prime, it is a cyclic group (see the end of the preceding section).

We have already given a description of the cosets of the multiplicative group of nonsingular matrices of order n with elements in the field P with respect to the normal divisor composed of matrices whose determinants are equal to 1. From this description it follows that the corresponding factor group is isomorphic to the multiplicative group of nonzero numbers of P.

Homomorphisms. The concepts of a normal divisor and a factor group are closely connected with the following generalization of the concept of an isomorphism.

A mapping φ of a group G onto a group G' such that to every element a of G there corresponds a unique element $a' = a\varphi$ in G' is called a *homomorphic mapping* of G onto G' (or simply a *homomorphism*) if in this mapping every element a' of G' is an image of some element a in G, $a' = a\varphi$, and if for any elements a, b of G,

$$(ab)\,\varphi = a\varphi \cdot b\varphi$$

It is quite obvious that if we also required a one-to-oneness of the mapping φ, we would obtain the already familiar definition of an isomorphism.

If φ is a homomorphism of group G onto group G' and 1 and a are, respectively, the unit element and an arbitrary element of G and, $1'$ is the unit element of G', then

$$1\varphi = 1',$$
$$(a^{-1})\,\varphi = (a\varphi)^{-1}$$

Indeed, if $1\varphi = e'$ and x' is an arbitrary element of the group G', then there is an element x in G such that $x\varphi = x'$. Whence,

$$x' = x\varphi = (x \cdot 1)\,\varphi = x\varphi \cdot 1\varphi = x' \cdot e'$$

Similarly,

$$x' = e'x'$$

and, hence, $e' = 1'$.

On the other hand, if $(a^{-1})\,\varphi = b'$, then

$$1' = 1\varphi = (aa^{-1})\,\varphi = a\varphi \cdot (a^{-1})\,\varphi = a\varphi \cdot b'$$

and, similarly,

$$1' = b' \cdot a\varphi$$

whence $b' = (a\varphi)^{-1}$.

Let us use the term *kernel* of a homomorphism φ of a group G onto a group G' for the set of elements of G which are mapped under φ into the unit element $1'$ of G'.

The kernel of any homomorphism φ of a group G is a normal divisor of G.

Indeed, if the elements a, b of G enter into the kernel of the homomorphism φ, i.e.,

$$a\varphi = b\varphi = 1'$$

then

$$(ab)\ \varphi = a\varphi \cdot b\varphi = 1' \cdot 1' = 1'$$

That is to say, the product ab is also contained in the kernel of the homomorphism φ. On the other hand, if $a\varphi = 1'$, then

$$(a^{-1})\ \varphi = (a\varphi)^{-1} = 1'^{-1} = 1'$$

which is to say that a^{-1} is also in the kernel of the homomorphism φ. Finally, if $a\varphi = 1'$, and x is an arbitrary element of the group G, then

$$(x^{-1}ax)\ \varphi = (x^{-1})\ \varphi \cdot a\varphi \cdot x\varphi = (x\varphi)^{-1} \cdot 1' \cdot x\varphi = 1'$$

The kernel of the homomorphism under consideration turned out to be a subgroup of the group G, which contains all the elements conjugate to any one of its elements; hence, it is a normal divisor.

Now let A be an arbitrary normal divisor of the group G. Associating every element x of G with that coset xA with respect to the normal divisor A in which the element lies, we obtain a mapping of the group G onto the entire factor group G/A. From the definition of multiplication in the group G/A [see (5)], it follows that this mapping is homomorphic.

The resulting homomorphism is the *canonical homomorphism* of the group G onto the factor group G/A. The normal divisor A is itself obviously the kernel of this homomorphism.

From this it follows that *only the normal divisors of the group G serve as kernels of the homomorphisms of this group*. This result can be regarded as yet another definition of a normal divisor.

It appears that all groups onto which the group G can be homomorphically mapped are actually exhausted by the factor groups of this group, and all the homomorphisms of G are exhausted by its canonical homomorphisms onto its factor groups. To be more precise, the following theorem holds.

Theorem on homomorphisms. *Suppose we have a homomorphism φ of a group G onto a group G'; let A be the kernel of this homomorphism.*

Then the group G' is isomorphic to the factor group G/A; there exists an isomorphic mapping σ of the former of these groups onto the latter such that the result of the successive mappings φ and σ coincides with the canonical homomorphism of the group G onto the factor group G/A.

Indeed, let x' be an arbitrary element of G', and let x be an element of G such that $x\varphi = x'$. Since for any element a of the kernel A of the homomorphism φ we have the equality $a\varphi = 1'$, it follows that

$$(xa)\,\varphi = x\varphi \cdot a\varphi = x' \cdot 1' = x'$$

That is, all elements of the coset xA are mapped under φ into the element x'.

On the other hand, if z is any element of the group G, such that $z\varphi = x'$, then

$$(x^{-1}z)\,\varphi = x^{-1}\varphi \cdot z\varphi = (x\varphi)^{-1} \cdot z\varphi = x'^{-1} \cdot x' = 1'$$

That is to say, $x^{-1}z$ is contained in the kernel A of the homomorphism φ. If we set $x^{-1}z = a$, then $z = xa$, or the element z is contained in the coset xA. Thus, collecting all the elements of the group G which are mapped under the homomorphism φ into the fixed element x' of the group G', we get precisely the coset xA.

The correspondence σ, which associates every element x' of G' with that coset of G by the normal divisor A which consists of all elements of G having x' as its image under φ, is a one-to-one mapping of the group G' onto the group G/A. This mapping σ is an isomorphism since if

$$x'\sigma = xA, \qquad y'\sigma = yA$$

that is,

$$x\varphi = x', \qquad y\varphi = y'$$

then

$$(xy)\,\varphi = x\varphi \cdot y\varphi = x'y'$$

and so

$$(x'y')\,\sigma = xyA = xA \cdot yA = x'\sigma \cdot y'\sigma$$

Finally, if x is an arbitrary element in G and $x\varphi = x'$ then

$$(x\varphi)\,\sigma = x'\sigma = xA$$

That is, a successive execution of the homomorphism φ and the isomorphism σ indeed maps the element x into the coset xA generated by it. The theorem is proved.

66. Direct Sums of Abelian Groups

We would like to conclude this chapter with a group-theoretic theorem that is deeper than the elementary properties of groups given above. Namely, proceeding from the description, given in Sec. 64,

of cyclic groups, we will obtain in the next section a complete description of finite Abelian groups.

As is customary in the theory of Abelian groups, we use the additive notation for the group operation: we shall speak of the sum $a + b$ of elements a and b of the group, of the zero subgroup 0, of the multiples ka of some element a, etc.

We will examine in this section a construction that will be described in detail in application to Abelian groups, though it could have been introduced at once for arbitrary (that is, not necessarily commutative) groups. This construction is suggested by the following examples. A plane regarded as a two-dimensional real linear space is an Abelian group with respect to the addition of vectors. Any straight line in this plane passing through the coordinate origin is a subgroup of the indicated group. If A_1 and A_2 are two distinct straight lines of this kind, then, as we know, any vector in the plane that issues from the origin is uniquely represented by the sum of its projections on the straight lines A_1 and A_2. Similarly, any vector of three-dimensional linear space can be uniquely written as the sum of three vectors belonging to three given straight lines A_1, A_2, and A_3, provided the lines do not lie in the same plane.

An Abelian group G is called the *direct sum* of its subgroups A_1, A_2, . . ., A_k,

$$G = A_1 + A_2 + \ldots + A_k \tag{1}$$

if every element x of G is *uniquely* written as the sum of the elements $a_1, a_2, \ldots, a_k$, taken, respectively, in the subgroups $A_1, A_2, \ldots, A_k$

$$x = a_1 + a_2 + \ldots + a_k \tag{2}$$

The notation (1) is called the *direct decomposition* of the group G. the subgroups A_i, $i = 1, 2, \ldots, k$, are *direct summands* of this decomposition, and the element a_i in (2) is a *component* of the element x in the direct summand A_i of the decomposition (1), $i = 1, 2, \ldots, k$.

If we are given a direct decomposition (1) *of a group* G *and if the direct summands* A_i *of this decomposition* (*all or some of them*), *are themselves decomposed into a direct sum,*

$$A_i = A_{i1} + A_{i2} + \ldots + A_{ik_i}, \qquad k_i \geqslant 1 \tag{3}$$

then the group G *is the direct sum of all its subgroups:*

$$A_{ij}, \quad j = 1, 2, \ldots, k_i, \quad i = 1, 2, \ldots, k$$

Indeed, for an arbitrary element x of G we have the notation (2) relative to the direct decomposition (1), and for each component a_i, $i = 1, 2, \ldots, k$, we have the notation

$$a_i = a_{i1} + a_{i2} + \ldots + a_{ik_i} \tag{4}$$

relative to the direct decomposition (3) of the group A_i. It is clear that x is the sum of all the elements a_{ij}, $j = 1, 2, \ldots, k_i$, $i = 1, 2, \ldots, k$. The uniqueness of this notation follows from the fact that we must obtain precisely equality (2) by taking any notation of the element x as a sum of elements, taken one each in the subgroups A_{ij}, and by adding the summands belonging to the same subgroup A_i, $i = 1, 2, \ldots, k$. On the other hand, each element a_i only has one notation of the type (4).

The definition of a direct sum may be restated. First let us introduce a new concept. If it is given that an Abelian group G has certain subgroups $B_1, B_2, \ldots, B_l$, then we denote by $\{B_1, B_2, \ldots, B_l\}$ the set of elements y of G which can in at least one way be written as a sum of the elements $b_1, b_2, \ldots, b_l$, taken in the subgroups $B_1, B_2, \ldots, B_l$, respectively,

$$y = b_1 + b_2 + \ldots + b_l \tag{5}$$

The set $\{B_1, B_2, \ldots, B_l\}$ *will be a subgroup of* G. We say that this subgroup *is generated* by the subgroups $B_1, B_2, \ldots, B_l$.

For the proof, let us take in $\{B_1, B_2, \ldots, B_l\}$ an element y with notation (5), and also an element y' with a similar notation,

$$y' = b_1' + b_2' + \ldots + b_l'$$

where b_i' is an element in B_i, $i = 1, 2, \ldots, l$. Then

$$y + y' = (b_1 + b_1') + (b_2 + b_2') + \ldots + (b_l + b_l'),$$
$$-y = (-b_1) + (-b_2) + \ldots + (-b_l)$$

which is to say that the elements $y + y'$ and $-y$ also have at least one notation of the type (5) and, hence, belong to the set $\{B_1, B_2, \ldots, B_l\}$, which completes the proof.

The subgroup $\{B_1, B_2, \ldots, B_l\}$ *contains each of the subgroups* B_i, $i = 1, 2, \ldots, l$. Indeed, every subgroup of the group G contains the zero element of this group and so, taking, for instance, in the subgroup B_1 any element b_1, and in the subgroups $B_2, \ldots, B_l$ the element 0, we obtain the following notation of type (5) for element b_1:

$$b_1 = b_1 + 0 + \ldots + 0$$

An Abelian group G *is the direct sum of its subgroups* $A_1, A_2, \ldots, A_k$ *if and only if it is generated by these subgroups,*

$$G = \{A_1, A_2, \ldots, A_k\} \tag{6}$$

and if the intersection of each subgroup A_i, $i = 2, \ldots, k$, *with the subgroup generated by all preceding subgroups* $A_1, A_2, \ldots, A_{i-1}$ *contains zero alone*:

$$\{A_1, A_2, \ldots, A_{i-1}\} \cap A_i = 0, \qquad i = 2, \ldots, k \tag{7}$$

Indeed, if the group G has the direct decomposition (1), then for any element x of G the notation (2) exists, and therefore we have equation (6). The validity of equations (7) follows from the uniqueness of the notation (2) for any element x: if for some i the intersection $\{A_1, A_2, \ldots, A_{i-1}\} \cap A_i$ contained a nonzero element x, then, on the one hand, x could be written as an element a_i, in A_i, i.e., $x = a_i$, and so

$$x = 0 + \ldots + 0 + a_i + 0 + \ldots + 0 \tag{8}$$

On the other hand, x, as an element of the subgroup $\{A_1, A_2, \ldots, A_{i-1}\}$, would have a notation of the form

$$x = a_1 + a_2 + \ldots + a_{i-1}$$

which is to say that

$$x = a_1 + a_2 + \ldots + a_{i-1} + 0 + \ldots + 0 \tag{9}$$

It is evident that (8) and (9) are two distinct notations of type (2) for the element x.

Conversely, let (6) and (7) hold. From (6) it follows that any element x of G has at least one notation of type (2). However, let there be two distinct notations of type (2) for some element x:

$$x = a_1 + a_2 + \ldots + a_k = a'_1 + a'_2 + \ldots + a'_k \tag{10}$$

Then we can find an i, $i \leqslant k$, such that

$$a_k = a'_k, \quad a_{k-1} = a'_{k-1}, \quad \ldots, \quad a_{i+1} = a'_{i+1} \tag{11}$$

but

$$a_i \neq a'_i$$

That is,

$$a_i - a'_i \neq 0 \tag{12}$$

From (10) and (11) follows, however, the equality

$$a_i - a'_i = (a'_1 - a_1) + (a'_2 - a_2) + \ldots + (a'_{i-1} - a_{i-1})$$

which contradicts (7) due to (12). The theorem is proved.

The concept of a direct sum may be regarded from quite a different angle. Suppose we have k arbitrary Abelian groups $A_1, A_2, \ldots, A_k$ among which there may be isomorphic groups. Denote by G the set of all possible systems of the form

$$(a_1, a_2, \ldots, a_k) \tag{13}$$

composed of elements taken one at a time in each of the groups $A_1, A_2, \ldots, A_k$. The set G will become an Abelian group if *addition* of the systems of type (13) is defined by the following rule:

$$(a_1, a_2, \ldots, a_k) + (a'_1, a'_2, \ldots, a'_k) = (a_1 + a'_1, a_2 + a'_2, \ldots, a_k + a'_k) \tag{14}$$

That is, the elements are combined separately in each of the given groups $A_1, A_2, \ldots, A_k$. Indeed, the associativity and commutativity of this addition follows from the validity of these properties in each of the specified groups; the role of zero is played by the system

$$(0_1, 0_2, \ldots, 0_k)$$

where 0_i denotes the zero element of the group A_i. $i = 1, 2, \ldots, k$. The inverse of (13) is the system

$$(-a_1, -a_2, \ldots, -a_k)$$

The Abelian group G thus constructed is called the *direct sum* of the groups $A_1, A_2, \ldots, A_k$ and is written, as above,

$$G = A_1 + A_2 + \ldots + A_k$$

This name is justified by the fact that *the group G, which is the direct sum of the groups $A_1, A_2, \ldots, A_k$ in the sense just defined, can be decomposed into the direct sum of its subgroups $A'_1, A'_2, \ldots, A'_k$, which are isomorphic, respectively, to the groups $A_1, A_2, \ldots, A_k$.*

Namely, denote by A'_i, $i = 1, 2, \ldots, k$, the set of elements of G, that is systems of type (13), with an arbitrary element a_i of group A_i in the ith position, all other positions being occupied by zeros of the corresponding groups; these will thus be systems of the form

$$(0_1, \ldots, 0_{i-1}, a_i, 0_{i+1}, \ldots, 0_k) \qquad (15)$$

The definition (14) of addition shows that the set A'_i is a subgroup of the group G. We obtain the isomorphism of this subgroup and the group A_i by associating to each system (15) an element a_i of group A_i.

It remains to prove that the group G is the direct sum of the subgroups $A'_1, A'_2, \ldots, A'_k$. Indeed, any element (13) of G may be represented as a sum of elements of the indicated subgroups:

$$\begin{aligned}(a_1, a_2, \ldots, a_k) = {} & (a_1, 0_2, \ldots, 0_k) \\ & + (0_1, a_2, 0_3, \ldots, 0_k) + \ldots + (0_1, 0_2, \ldots, 0_{k-1}, a_k)\end{aligned}$$

The uniqueness of this representation follows from the fact that distinct systems of type (13) are distinct elements of the group G.

If we have two systems of Abelian groups, $A_1, A_2, \ldots, A_k$ and $B_1, B_2, \ldots, B_k$, and the groups A_i and B_i are isomorphic, $i = 1, 2, \ldots, k$, then the groups

$$G = A_1 + A_2 + \ldots + A_k$$

and

$$H = B_1 + B_2 + \ldots + B_k$$

are also isomorphic.

Indeed, if for $i = 1, 2, \ldots, k$ there is established, between groups A_i and B_i, an isomorphism φ_i, which associates with each

element a_i of A_i an element $a_i\varphi_i$ of B_i, then the mapping φ, which associates with every element $(a_1, a_2, \ldots, a_k)$ of G an element of H defined by the equation

$$(a_1, a_2, \ldots, a_k)\,\varphi = (a_1\varphi_1, a_2\varphi_2, \ldots, a_k\varphi_k),$$

will obviously be an isomorphic mapping of the group G onto the group H.

If we have finite Abelian groups $A_1, A_2, \ldots, A_k$ of orders $n_1, n_2, \ldots, n_k$, respectively, then the direct sum G of these groups is also a finite group and its order n is equal to the product of the orders of the direct summands,

$$n = n_1 n_2 \ldots n_k \tag{16}$$

Quite true, since the number of distinct systems of type (13) whose element a_1 can assume n_1 distinct values, whose element a_2 can assume n_2 distinct values, and so on, is determined by equation (16).

Let us consider some examples.

If the order n of a finite cyclic group $\{a\}$ can be decomposed into the product of two relatively prime natural numbers,

$$n = st, \quad (s, t) = 1$$

then the group $\{a\}$ can be decomposed into the direct sum of two cyclic groups having orders s and t, respectively.

Let us use the additive notation for the group $\{a\}$. If we set $b = ta$, then

$$sb = (st)\,a = na = 0$$

but for $0 < k < s$

$$kb = (kt)\,a \neq 0$$

which is to say that the cyclic subgroup $\{b\}$ is of order s. Similarly, the cyclic subgroup $\{c\}$ of element $c = sa$ has order t. The intersection $\{b\} \cap \{c\}$ contains only zero because if $kb = lc$ for $0 < k < s$, $0 < l < t$, then

$$(kt)\,a = (ls)\,a$$

whence, since the numbers kt and ls are less than n,

$$kt = ls$$

which is impossible due to the relative primality of the numbers s and t. Finally, there are numbers u and v such that

$$su + tv = 1$$

and so

$$a = v\,(ta) + u\,(sa) = vb + uc$$

and, consequently, any element of the group $\{a\}$ may be represented as the sum of elements of the subgroups $\{b\}$ and $\{c\}$.

We call an Abelian group G *indecomposable* if it cannot be decomposed into the direct sum of two or several of its subgroups distinct from the zero subgroup. A finite cyclic group whose order is some power of the prime number p is called a *primary* cyclic group relative to the prime number p. Applying several times the assertion proved above, we find that *any finite cyclic group can be decomposed into the direct sum of primary cyclic groups relative to distinct prime numbers.* More precisely, *a cyclic group of order*

$$n = p_1^{h_1} p_2^{h_2} \dots p_s^{h_s}$$

where $p_1, p_2, \dots, p_s$ *are* ***distinct*** *prime numbers, can be decomposed into the direct sum s of cyclic groups having orders* $p_1^{h_1}, p_2^{h_2}, \dots, p_s^{h_s}$, *respectively.*

Every primary cyclic group is indecomposable.

Indeed, suppose we have a finite cyclic group $\{a\}$ of order p^k, where p is prime. If this group were decomposable, then, by (7), it would have nonzero subgroups whose intersection is zero. Actually, however, every nonzero subgroup of our group contains the nonzero element

$$b = p^{k-1}a$$

To prove this, take an arbitrary nonzero element x of our group,

$$x = sa, \quad 0 < s < p^k$$

The number s may be written as

$$s = p^l s', \quad 0 \leqslant l < k$$

where the number s' is not divisible by p and, hence, is relatively prime to it; and so there exist numbers u and v such that

$$s'u + pv = 1$$

Then

$$(p^{k-l-1}u)\, x = (p^{k-l-1}us)\, a = (p^{k-1}us')\, a$$
$$= p^{k-1}(1 - pv)\, a = (p^{k-1} - p^k v)\, a = p^{k-1}a - v\,(p^k a) = p^{k-1}a = b$$

which is to say, the element b is in the cyclic subgroup $\{x\}$.

The additive group of the integers (which is an infinite cyclic group) and also the additive group of all rational numbers are indecomposable groups.

The indecomposability of both these groups follows from the fact that in each of them there exists, for any two nonzero elements, a nonzero common multiple; that is, any two nonzero cyclic subgroups have a nonzero intersection.

Note that if the operation in an Abelian group G is termed multiplication, then instead of a direct sum we speak of a *direct product.*

The multiplicative group of nonzero real numbers can be decomposed into a direct product of the multiplicative group of positive real numbers and a group, with respect to multiplication, made up of the numbers 1 *and* -1.

Actually, the intersection of these two subgroups of our group contains only the number 1—the unit element of this group. On the other hand, every positive number is the product of the number 1 by itself, every negative number is the product of its absolute value by the number -1.

67. Finite Abelian Groups

If we take any finite set of primary cyclic groups, some of which can refer to one and the same prime number or even have the same order, i.e., be isomorphic, then the direct sum of these groups is a finite Abelian group. It turns out that this exhausts all finite Abelian groups.

Fundamental theorem of finite Abelian groups. *Every finite Abelian group G which is not a zero group can be decomposed into a direct sum of primary cyclic subgroups.*

We begin the proof of this theorem with the remark that *in the group G there will inevitably be nonzero elements of prime power orders.* Indeed, if some nonzero element x of G has order l, $lx = 0$ and if p^k, $k > 0$, is a power of the prime p such that divides the number l,

$$l = p^k m$$

then the element mx is different from zero and has order p^k.

Let

$$p_1, p_2, \ldots, p_s \tag{1}$$

be all distinct primes, some powers of which serve as the orders of certain elements of the group G. Denote any such number by p and the set of elements of G having powers of p as their orders by P.

The set P is a subgroup of the group G. Indeed, P includes the element 0 since its order is $1 = p^0$. Furthermore, if $p^k x = 0$, then $p^k(-x) = 0$ as well. Finally, if $p^k x = 0$, $p^l y = 0$ and if, say, $k \geqslant l$, then

$$p^k(x + y) = 0$$

Thus, either the number p^k or a divisor of this number, at any rate some power of p, serves as the order of the element $x + y$.

Alternately taking each of the numbers (1) for p, we obtain s nonzero subgroups

$$P_1, P_2, \ldots, P_s \tag{2}$$

The group G is the direct sum of these subgroups,

$$G = P_1 + P_2 + \ldots + P_s \tag{3}$$

True, for if x is an arbitrary element of G, then its order l can only be divisible by certain prime numbers of the system (1),

$$l = p_1^{k_1} p_2^{k_2} \ldots p_s^{k_s}$$

where $k_i \geqslant 0$, $i = 1, 2, \ldots, s$. Therefore, as was demonstrated at the end of Sec. 66, the cyclic subgroup $\{x\}$ can be decomposed into the direct sum of primary cyclic subgroups having orders $p_1^{k_1}$, $p_2^{k_2}$, ..., $p_s^{k_s}$, respectively. These primary cyclic subgroups lie in corresponding subgroups (2) and, consequently, the element x is represented in the form of a sum of elements taken one each in all or several of the subgroups (2). This proves the equality

$$G = \{P_1, P_2, \ldots, P_s\}$$

which is similar to (6) of Sec. 66.

To prove the equality similar to (7) of the same section, take any i, $2 \leqslant i \leqslant s$. Then any element y of the subgroup $\{P_1, P_2, \ldots ... ..., P_{i-1}\}$ is of the form

$$y = a_1 + a_2 + \ldots + a_{i-1}$$

where the element a_j, $j = 1, 2, \ldots, i - 1$, is in the subgroup P_j, that is, has order $p_j^{k_j}$. Then,

$$(p_1^{k_1} p_2^{k_2} \ldots p_{i-1}^{k_{i-1}})\, y = 0$$

For the order of the element y we have some divisor of the number $p_1^{k_1} p_2^{k_2} \ldots p_{i-1}^{k_{i-1}}$ and, consequently, the element y, if it is different from zero, cannot be in the subgroup P_i. This proves that

$$\{P_1, P_2, \ldots, P_{i-1}\} \cap P_i = 0$$

which is what we set out to prove.

Notice that an Abelian group, the orders of all the elements of which are powers of one and the same prime number p, is termed *primary* relative to p. Primary cyclic groups are a special case of primary groups. Thus, the subgroups (2) are primary. They are called *primary components* of the group G, and the direct decomposition (3) is called *the decomposition of this group into primary components.* Since the subgroups (2) are defined uniquely in the group G, it follows that *the decomposition of G into primary components is likewise defined uniquely.*

Quite naturally, the decomposability of any finite Abelian group into the direct sum of primary groups reduces the proof of the fundamental theorem to the case of a finite primary Abelian group P relative to some prime number p. Let us consider this case.

Let a_1 be one of the elements of the group P having the highest order in it. Furthermore, if in P there are nonzero elements, the intersection of the cyclic subgroups of which with the cyclic subgroup $\{a_1\}$ is zero only, then by a_2 we denote one of the elements of the highest order among the elements with this property; thus,

$$\{a_1\} \cap \{a_2\} = 0$$

Let the elements $a_1, a_2, \ldots, a_{i-1}$ be already chosen. Denote by $\{a_1, a_2, \ldots, a_{i-1}\}$ the subgroup of the group P generated by their cyclic subgroups:

$$\{\{a_1\}, \{a_2\}, \ldots, \{a_{i-1}\}\} = \{a_1, a_2, \ldots, a_{i-1}\} \tag{4}$$

It evidently consists of all the elements of P that can be written as the sum of multiples of the elements $a_1, a_2, \ldots, a_{i-1}$. We will say that this subgroup is generated by the elements $a_1, a_2, \ldots, a_{i-1}$. Let us now denote by a_i one of the elements of the highest order among those elements of P whose cyclic subgroups have a zero intersection with the subgroup $\{a_1, a_2, \ldots, a_{i-1}\}$. Thus

$$\{a_1, a_2, \ldots, a_{i-1}\} \cap \{a_i\} = 0 \tag{5}$$

Because of the finiteness of the group P, this process must terminate. Suppose this occurs after the elements $a_1, a_2, \ldots, a_s$ have been chosen. If by P' we denote the subgroup generated by these elements,

$$P' = \{a_1, a_2, \ldots, a_s\}$$

i.e.,

$$P' = \{\{a_1\}, \{a_2\}, \ldots, \{a_s\}\} \tag{6}$$

then, consequently, a cyclic subgroup of any nonzero element of the group P has a nonzero intersection with the subgroup P'.

The equality (6) and the equality (5), which holds true for $i = 2, 3, \ldots, s$, show that, by (4), the subgroup P' is the direct sum of the cyclic subgroups $\{a_1\}, \{a_2\}, \ldots, \{a_s\}$,

$$P' = \{a_1\} + \{a_2\} + \ldots + \{a_s\} \tag{7}$$

It remains to prove that the subgroup P' does indeed coincide with the entire group P.

Let x be any element of P having order p. Since

$$P' \cap \{x\} \neq 0$$

and the subgroup $\{x\}$ has no nonzero subgroups different from itself—recall that the order of a subgroup is a divisor of the order of the group, and the number p is prime—the subgroup $\{x\}$ is indeed contained in the subgroup P' and, hence, x belongs to P'. Thus, all elements of order p of the group P lie in the subgroup P'.

Now suppose it has been proved that all elements of P whose order does not exceed the number p^{k-1} are in the subgroup P', and let x be any element of P having order p^k. As the choice of the elements $a_1, a_2, \ldots, a_s$ shows, their orders do not increase and so we can indicate an i, $1 \leqslant i - 1 \leqslant s$, such that the orders of the elements $a_1, a_2, \ldots, a_{i-1}$ are greater than or equal to p^k, and for $i - 1 < s$ the order of the element a_i is strictly less than this number, that is to say, less than the order of the element x. Whence it follows, by the conditions to which the choice of the element a_i are subject, that if

$$Q = \{a_1, a_2, \ldots, a_{i-1}\}$$

then

$$Q \cap \{x\} \neq 0$$

However, in Sec. 66 it was proved that any nonzero subgroup of a primary cyclic group $\{x\}$ of order p^k contains the element

$$y = p^{k-1}x \tag{8}$$

Consequently, the element y lies in the intersection $Q \cap \{x\}$ and therefore in the subgroup Q as well. This enables one to write y as the sum of multiples of the elements $a_1, a_2, \ldots, a_{i-1}$,

$$y = l_1a_1 + l_2a_2 + \ldots + l_{i-1}a_{i-1} \tag{9}$$

From (8) it follows that the element y has order p. Therefore,

$$(pl_1)\, a_1 + (pl_2)\, a_2 + \ldots + (pl_{i-1})\, a_{i-1} = 0$$

That is to say, because of the existence of the direct decomposition (7),

$$(pl_j)\, a_j = 0, \quad j = 1, 2, \ldots, i - 1$$

The number pl_j must thus be divisible by the order of the element a_j, and therefore also by the number p^k, whence it follows that p^{k-1} divides l_j:

$$l_j = p^{k-1}m_j, \quad j = 1, 2, \ldots, i - 1 \tag{10}$$

Let

$$z = m_1a_1 + m_2a_2 + \ldots + m_{i-1}a_{i-1}$$

This will be an element of the subgroup Q and therefore of the subgroup P' too; by (9) and (10),

$$y = p^{k-1}z \tag{11}$$

From (8) and (11) follows the equality

$$p^{k-1}(x - z) = 0$$

That is, the order of the element

$$t = x - z$$

does not exceed p^{k-1} and, hence, by the induction hypothesis, t is contained in the subgroup P'. Therefore, element x as the sum of two elements of P', $x = z + t$, also belongs to the subgroup P'. This is proof that all elements of order p^k of the group P are contained in P'.

Consequently, our inductive proof admits of the assertion that all elements of the group P enter into the subgroup P', or $P' = P$. This concludes the proof of the fundamental theorem.

Collaterally, we have that *a finite Abelian group is primary relative to a prime number p if and only if its order is a power of p.* True enough, it was shown that any finite primary (with respect to p) Abelian group P can be decomposed into the direct sum of primary (with respect to p) cyclic groups, and for this reason the order of the group P is equal to the product of the orders of these cyclic groups, that is to say, it is a power of p. Conversely, if a finite Abelian group has order p^k, where p is prime, then the order of any one of its elements is a divisor of this number, that is, it is also some power of p, and therefore the group turns out to be primary relative to p.

The fundamental theorem does not yet exhaust the problem of a complete description of finite Abelian groups, since we have not precluded the possibility that the direct sums of two distinct sets of cyclic groups that are primary relative to certain prime numbers may prove to be isomorphic groups. Actually, this does not occur, as the following theorem shows.

If a finite Abelian group G is decomposed in two ways into a direct sum of primary cyclic subgroups,

$$G = \{a_1\} + \{a_2\} + \ldots + \{a_s\} = \{b_1\} + \{b_2\} + \ldots + \{b_t\} \quad (12)$$

then both direct decompositions have one and the same number of direct summands, $s = t$, and it is possible to establish a one-to-one correspondence between these decompositions such that the appropriate summands are cyclic groups of the same order, which is to say they are isomorphic.

Note, to begin with, that if, say, in the first of the direct decompositions (12), we collect direct summands relative to a given prime p, then their direct sum will be a primary (relative to p) subgroup of the group G and even a primary component of this group, since its order is equal to the highest power of p that divides the order of the group G. Thus combining the direct summands in each of the decompositions (12), in both cases we obtain a decomposition of G into primary components, the uniqueness of which decomposition has already been noted above.

This permits proving our theorem under the assumption that the group G is itself primary relative to the prime number p. Let the numbering of the direct summands in each of the decompositions (12) be chosen so that the orders of these summands do not increase,

that is, the elements $a_1, a_2, \ldots, a_s$ have, respectively, the orders

$$p^{k_1}, p^{k_2}, \ldots, p^{k_s}$$

for

$$k_1 \geqslant k_2 \geqslant \ldots \geqslant k_s$$

while the elements $b_1, b_2, \ldots, b_t$ have the orders

$$p^{l_1}, p^{l_2}, \ldots, p^{l_t}$$

for

$$l_1 \geqslant l_2 \geqslant \ldots \geqslant l_t$$

If the assertion of our theorem were not valid, then there would be an i, $i \geqslant 1$, such that

$$k_1 = l_1, \ldots, k_{i-1} = l_{i-1} \tag{13}$$

but

$$k_i \neq l_i$$

Naturally, $i \leqslant \min(s, t)$, since for each of the decompositions (12) the product of the orders of all direct summands is equal to the order of the group G. We will show that our assumption leads to a contradiction.

For example, let

$$k_i < l_i \tag{14}$$

Denote by H the set of elements of the group G whose orders do not exceed p^{k_i}. This is a subgroup of the group G, since if x and y are elements of H, then both $x + y$ and $-x$ have orders that do not exceed the numbers p^{k_i}.

Note that the subgroup H contains, for instance, the following elements:

$$p^{k_1-k_i}a_1, p^{k_2-k_i}a_2, \ldots, p^{k_{i-1}-k_i}a_{i-1}, a_i, a_{i+1}, \ldots, a_s$$

On the other hand, if $1 \leqslant j \leqslant i-1$, then the element $p^{k_j-k_i-1}a_j$ has order p^{k_i+1} and therefore is not in H. From this it follows that the coset $a_j + H$ (recall that we are using the additive notation!) has, as an element of the factor group G/H, the order $p^{k_j-k_i}$. Such also is the order of its cyclic subgroup $\{a_j + H\}$. We will now prove that the group G/H is the direct sum of the cyclic subgroups $\{a_j + H\}$, $j = 1, 2, \ldots, i-1$,

$$G/H = \{a_1 + H\} + \{a_2 + H\} + \ldots + \{a_{i-1} + H\} \tag{15}$$

and so its order is equal to the number

$$p^{(k_1-k_i)+(k_2-k_i)+\ldots+(k_{i-1}-k_i)} \tag{16}$$

If x is an arbitrary element of the group G, then there exists the notation

$$x = m_1a_1 + m_2a_2 + \ldots + m_sa_s$$

Suppose for $j = 1, 2, \ldots, i-1$,

$$m_j = p^{k_j - k_i} q_j + n_j$$

where

$$0 \leqslant n_j < p^{k_j - k_i} \tag{17}$$

Then

$$m_j a_j = q_j (p^{k_j - k_i} a_j) + n_j a_j$$

and since the first summand of the right member is contained in H, it follows that

$$m_j a_j + H = n_j a_j + H$$

On the other hand,

$$m_i a_i + H = H, \ldots, m_s a_s + H = H$$

And so

$$\begin{aligned} x + H &= (m_1 a_1 + H) + (m_2 a_2 + H) + \ldots + (m_s a_s + H) \\ &= (n_1 a_1 + H) + (n_2 a_2 + H) + \ldots + (n_{i-1} a_{i-1} + H) \end{aligned} \tag{18}$$

Let there also be the notation

$$x + H = (n'_1 a_1 + H) + (n'_2 a_2 + H) + \ldots + (n'_{i-1} a_{i-1} + H) \tag{19}$$

where

$$0 \leqslant n'_j < p^{k_j - k_i}, \quad j = 1, 2, \ldots, i-1 \tag{20}$$

Then the elements

$$n_1 a_1 + n_2 a_2 + \ldots + n_{i-1} a_{i-1}$$

and

$$n'_1 a_1 + n'_2 a_2 + \ldots + n'_{i-1} a_{i-1}$$

lie in one coset relative to H, i.e., their difference belongs to H and therefore

$$p^{k_i} [(n_1 - n'_1) a_1 + (n_2 - n'_2) a_2 + \ldots + (n_{i-1} - n'_{i-1}) a_{i-1}] = 0$$

From this it follows [since the first of the decompositions (12) is direct] that

$$p^{k_i} (n_j - n'_j) a_j = 0, \quad j = 1, 2, \ldots, i = 1$$

and so the number $p^{k_i} (n_j - n'_j)$ must be divisible by the order p^{k_j} of the element a_j and, hence, the difference $n_j - n'_j$ is divisible by the number $p^{k_j - k_i}$. Whence, by (17) and (20), it follows that

$$n_j = n'_j, \qquad j = 1, 2, \ldots, i-1$$

which means that the notations (18) and (19) are identical. This proves the existence of the direct decomposition (15).

Analogous arguments relative to the second of the direct decompositions (12) will show that this same factor group G/H has the direct decomposition

$$G/H = \{b_1 + H\} + \{b_2 + H\} + \ldots + \{b_{i-1} + H\} + \{b_i + H\} + \ldots$$

That is, by (13) and (14), its order must be strictly greater than the number (16). This contradiction proves the theorem.

We have thus obtained a complete survey of the finite Abelian groups. Namely, *we take all possible finite sets of the natural numbers*

$$(n_1, n_2, \ldots, n_k)$$

different from unity, but not necessarily distinct; each one of these numbers must be a power of some prime number. To each such set we associate the direct sum of cyclic groups whose orders are numbers from this set. All the finite Abelian groups thus obtained are pairwise nonisomorphic, and any other finite Abelian group is isomorphic to one of these groups.

BIBLIOGRAPHY

Higher Algebra

D. K. Faddeyev and I. S. Sominsky, *Problems in Higher Algebra* (Sbornik zadach po vysshei algebre), 9th ed., Moscow, 1968 (English translation, MIR Publishers, 1972).

E. S. Lyapin, *Course of Higher Algebra* (Kurs vysshei algebry), 2nd ed., Moscow, 1955.

L. Ya. Okunev, *Higher Algebra* (Vysshaya algebra), 2nd ed., Moscow, 1966.

G. M. Shapiro, *Higher Algebra* (Vysshaya algebra), 4th ed., Moscow, 1938.

A. K. Sushkevich, *Principles of Higher Algebra* (Osnovy vysshei algebry), 4th ed., Moscow, 1941.

S. P. Vinogradov, *Fundamentals of the Theory of Determinants* (Osnovaniya teorii determinantov), 4th ed., ONTI, Moscow-Leningrad, 1935.

Linear Algebra

M. Bôcher, *Introduction to Higher Algebra*, New York, 1938.

D. K. Faddeyev and V. N. Faddeyeva, *Computational Methods of Linear Algebra* (Vychislitelnye metody lineinoi algebry), Moscow, 1960.

R. A. Frazer, W. J. Duncan, and A. R. Collar, *Elementary Matrices and Some Applications to Dynamics and Differential Equations*, Cambridge, 1938.

F. R. Gantmacher, *The Theory of Matrices* (Teoriya matrits), 3rd ed., Moscow, 1967.

I. M. Gelfand, *Lectures in Linear Algebra* (Lektsii po lineinoi algebre), 3rd ed., Moscow, 1966.

G. B. Gurevich, *Fundamentals of the Theory of Algebraic Invariants* (Osnovy teorii algebraicheskikh invariantov). Moscow, 1948.

A. I. Maltsev, *Principles of Linear Algebra* (Osnovy lineinoi algebry), 2nd ed., Moscow, 1956.

I. V. Proskuryakov, *Problems in Linear Algebra* (Sbornik zadach po lineinoi algebre), 3rd ed., Moscow, 1967.

O. Schreier und E. Sperner, *Einführung in die analytische Geometrie und Algebra*, I, Leipzig, Berlin, 1931.

O. Schreier und E. Sperner, *Vorlesungen über Matrizen*, II, Leipzig, Berlin, 1932.

G. E. Shilov, *Introduction to the Theory of Linear Spaces* (Vvedenie v teoriyu lineinykh prostranstv), 2nd ed., Moscow, 1956.

Theory of Groups, Rings and Lattices

P. S. Aleksandrov, *Introduction to the Theory of Groups* (Vvedenie v teoriyu grupp), 2nd ed., Moscow, 1951.

Reinhold Baer, *Linear Algebra and Projective Geometry*, New York, Academy Press, 1952.

Garrett Birkhoff, *Lattice Theory*, New York, 1948.

Henri Cartan and Samuel Eilenberg, *Homological Algebra*, Princeton University Press, 1956.

N. G. Chebotarev, *Introduction to the Theory of Algebras* (Vvedenie v teoriyu algebr), Moscow, 1949.

N. Jacobson, *The Theory of Rings*, American Mathematical Society, 1943.

N. Jacobson, *Structure of Rings*, American Mathematical Society, 1956.
A. G. Kurosh, *Lectures in General Algebra* (Lektsii po obshchei algebre), Moscow, 1962.
A. G. Kurosh, *The Theory of Groups* (Teoriya grupp), 3rd ed., Moscow, 1967.
E. S. Lyapin, *Semigroups* (Polugruppy), Moscow, 1960.
L. Ya. Okunev, *Principles of Modern Algebra* (Osnovy sovremennoi algebry), Moscow, 1941.
O. Yu. Schmidt, *The Abstract Theory of Groups* (Abstraktnaya teoriya grupp), 2nd ed., Moscow, 1933 (See also O. Yu. Schmidt, *Selected Works* (Izbrannye trudy), Mathematics, Academy of Sciences, USSR, 1959).
A. K. Sushkevich, *The Theory of Generalized Groups* (Teoriya obobshchennykh grupp), GNTI, UkSSR, 1937.
B. L. van der Waerden, *Moderne Algebra*, Berlin, 1937, II, Berlin, 1940.

Theory of Fields

N. G. Chebotarev, *Principles of Galois Theory* (Osnovy teorii Galua), Part 1, ONTI, Moscow, 1934.
N. G. Chebotarev, *Theory of Algebraic Functions* (Teoriya algebraicheskikh funktsii), Moscow, 1948.
N. G. Chebotarev, *The Theory of Galois* (Teoriya Galua), ONTI, Moscow, 1936.
D. A. Grave, *Treatise on Algebraic Analysis* (Traktat po algebraicheskomu analizu), Vols. 1 and 2, Academy of Sciences, UkSSR, 1938-1939.
E. Hecke, *Vorlesungen über die Theorie der algebraischen Zahlen*, Leipzig, 1923.
W. V. D. Hodge and D. Pedoe, *Methods of Algebraic Geometry*, Cambridge University Press, Vol. 1, 1947, Vol. 2, 1952, Vol. 3, 1954.
H. Weyl, *Algebraic Theory of Numbers*, Princeton University Press, 1940.

Continuous Groups

N. G. Chebotarev *Theory of Lie Groups* (Teoriya grupp Li), Moscow, 1940,
Claude Chevalley, *Theorie des groupes de Lie*, Princeton University Press, t. 1, 1946, t. 2, 3, 1951.
F. D. Murnaghan, *The Theory of Group Representations*, Baltimore, USA, 1938.
L. S. Pontryagin, *Continuous Groups* (Nepreryvnye gruppy), 2nd ed., Moscow, 1954.
H. Weyl, *The Classical Groups—Their Invariants and Representations*, 2nd ed., Princeton, 1946.

INDEX

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: Mir Publishers, 2 Pervy Rizhsky Pereulok, Moscow, USSR.

Printed in the Union of Soviet Socialist Republics

MIR PUBLISHERS ALSO OFFER THE FOLLOWING MATH BOOKS FOR YOUR LIBRARY

DIFFERENTIAL AND INTEGRAL CALCULUS

by N. Piskunov

This text is designed as a course of mathematics for higher technical schools. It contains many worked examples that illustrate the theoretical material and serve as models for solving problems.

The first two chapters "Number. Variable. Function" and "Limit. Continuity of a Function" have been made as short as possible. Some of the questions that are usually discussed in these chapters have been put in the third and subsequent chapters without loss of continuity. This has made it possible to take up very early the basic concept of differential calculus—the derivative—which is required in the study of technical subjects. Experience has shown this arrangement of the material to be the best and most convenient for the student.

A large number of problems have been included, many of which illustrate the interrelationships of mathematics and other disciplines. The problems are specially selected (and in sufficient number) for each section of the course thus helping the student to master the theoretical material. To a large extent, this makes the use of a separate book of problems unnecessary and extends the usefulness of this text as a course of mathematics for self-instruction.

PROBLEMS IN MATHEMATICAL ANALYSIS

under the editorship
of *B.* Demidovich, D.Sc.

This collection of problems and exercises in mathematical analysis covers the maximum requirements of general courses in higher mathematics for higher technical schools. It contains over 3,000 problems sequentially arranged in Chapters I to X covering all branches of higher mathematics (with the exception of analytical geometry) given in college courses. Particular attention is given to the most important sections of the course that require established skills (the finding of limits, differentiation techniques, the graphing of functions, integration techniques, the applications of definite integrals, series, the solution of differential equations).

Each chapter begins with a brief theoretical introduction that covers the basic definitions and formulas of that section of the course. Here the most important typical problems are worked out in full. We believe that this will greatly simplify the work of the student. Answers are given to all computational problems. The problems are frequently illustrated by drawings.

THEORY OF THE FUNCTIONS OF A COMPLEX VARIABLE

by **A. Sveshnikov**, D.Sc.
and **A. Tikhonov**,
USSR Academy of Sciences

A textbook for students not without interest for theoretical physicists working in the fields of hydrodynamics and electrostatics; can be used as a reference book by post-graduate students and research workers.

Examples illustrate the application of the theory to boundary-value problems. Covers questions in saddle-point method and Wiener-Hopf equations.

PROBLEMS IN THE THEORY OF FUNCTIONS OF A COMPLEX VARIABLE

by **L. Volkovysky**,
G. Lunts,
I. Aramanovich

A book containing 1425 problems (and answers) for university students. *Contents.* Complex Numbers and Functions of a Complex Variable. Conformal Mapping Related with Elementary Functions. Integrals and Power Series. The Laurent Series. Singular Points of Single-Valued Analytic Functions. Residues and Their Applications. Functional Series. Integrals Depending on Parameter. Infinite Products. Entire and Meromorphic Functions. Cauchy's Integrals. Poisson's and Schwarz's Integral Formulas. Riemann Surfaces. Conformal Mapping. Applications to Mechanics and Physics.

HANDBOOK OF HIGHER MATHEMATICS

by **M. Vygodsky**, D.Sc.

Intended for students and engineers, teachers and sixth-form pupils as a practical reference book, or as a compact study aid giving elementary acquaintance with the subject. Contains material on the history of mathematical ideas and brief biographical notes on the mathematicians who developed them.

Contents. Analytical Plane Geometry. Analytical Solid Geometry. Basic Concepts of Mathematical Analysis. Differential Calculus. Integral Calculus. Lines in a Plane and Space. Differentiation and Integration of the Functions of Two or More Arguments. Differential Equations. Famous Curves. Tables of Logarithms.

LECTURES
IN HIGHER MATHEMATICS

by **A. Myškis**, D.Sc.

A textbook for students, revised for translation from the 3rd Russian edition.

Contents. Introduction. Quantity and Function. Analytic Geometry on a Plane. Limit. Continuity. Derivatives. Differentials. Investigating the Behaviour of Functions. Approximate Solution of Finite Functions. Interpolation. Determinants and Systems of Linear Algebraic Equations. Vectors. Complex Numbers and Functions. Functions of Several Variables. Analytic Geometry in Space. Matrices and Their Applications. Partial Derivatives. Indefinite Integrals. Definite Integrals. Differential Equations. Multiple Integrals. Series. Elements of the Probability Theory. On Modern Computers.

PROBLEMS
IN HIGHER ALGEBRA

by **D. Faddeev**,
Corr. Mem.,
USSR Academy of Sciences,
and **I. Sominsky**,
D.Sc. (Phys. and Math.)

This problem book is intended for students of universities and teachers' colleges taking the course of higher algebra.

Contents. Introduction. *Part I. Problems.* Complex Numbers. Evaluation of Determinants. Systems of Linear Equations. Matrices. Polynomials and Rational Functions of One Variable. Symmetric Functions. Linear Algebra. *Part II. Suggestions. Part III. Answers and Solutions.*

Mir Publishers' books in foreign languages are exported by V/O Mezhdunarodnaya Kniga and can be purchased or ordered through booksellers in your country dealing with V/O Mezhdunarodnaya Kniga, USSR (200, Moscow, USSR)