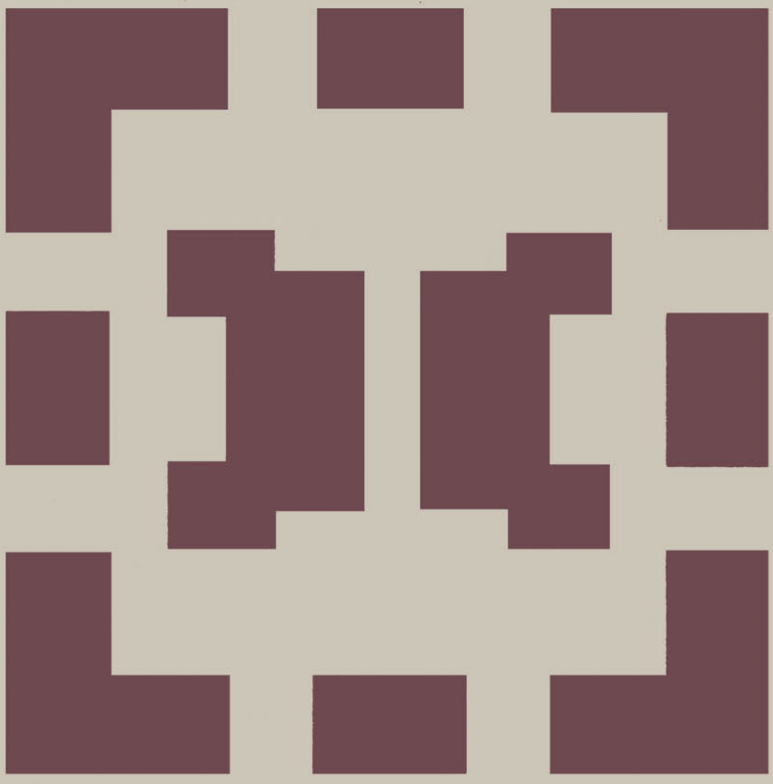
Mathematics and Its Applications

Karel Rektorys

Survey of Applicable Mathematics

Second Revised Edition



Springer Science+Business Media, LLC

Survey of Applicable Mathematics

Mathematics and Its Applications

Managing Editor:

M. HAZEWINKEL

Centre for Mathematics and Computer Science, Amsterdam, The Netherlands

Volume 280

Survey of Applicable Mathematics

Second Revised Edition

VOLUME I

Karel Rektorys
Technical University,
Prague, Slovak Republic

Springer Science+Business Media, LLC

Library of Congress Cataloging-in-Publication Data

Rektorys, Karel.
[Přehled užité matematiky. English]
Survey of applicable mathematics / Karel Rektorys. -- 2nd rev. ed.
p. cm. -- (Mathematics and its applications.

Rev. translation of: Přehled užité matematiky. 6th ed.

1. Mathematics. I. Title. II. Series: Mathematics and its applications (Kluwer Academic Publishers).

QA37.2.R44 1991
510--dc20 91-28851

ISBN 978-94-015-8310-7 ISBN 978-94-015-8308-4 (eBook)
DOI 10.1007/978-94-015-8308-4

Printed on acid-free paper

Originally published by Kluwer Academic Publishers in 1994
Softcover reprint of the hardcover 2nd edition 1994

Dedicated to the memory of

Prof. Dr. Lothar Collatz

Editor-in-Chief: Prof. RNDr. Karel Rektorys, Dr.Sc.
Associate Editor: RNDr. Emil Vitásek, C.Sc.

Authors of individual Chapters
Ass. Prof. Tomáš Cipra, C.Sc. (Chaps. 33, 34, 35, 36),
Ass. Prof. RNBr. Karel Drábek, C.Sc. (Chap.4),
Prof. RNDr. Mirorlav Fiedler, Dr.Sc. (Chap. 31),
RNDr. Jaroslav Fuka, C.Sc. (Chap. 20B, 21),
Ass. Prof. František Kejla (Chap.6, 7A),
Ass. Prof. Bořivoj Kepr (Chap. 9),
Prof. RNDr. Jindřich Nečas, Dr.Sc. (Chap.28),
Prof. RNDr. František Nožička (Chap. 23, 37),
RNDr. Milan Práger, C.Sc. (Chap. 24),
Prof. RNDr. Karel Rektorys, Dr.Sc. (Chap. 7B, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20A, 22, 26, 29),
RNDr. Jitka Segethová, C.Sc., RNDr. Karel Segeth, C.Sc. (Chap. 30),
Ass. Prof. Václav Vilhelm, C.Sc. (Chap. 1, 2, 3, 8),
RNDr. Emil Vitásek, C.Sc. (Chap. 25, 27, 32),
RNDr. Miloslav Zelenka, C.Sc. (Chap. 5).

CONTENTS

2. Trigonometric and Inverse Trigonometric Functions. Hyperbolic and Inverse Hyperbolic Functions

By VÁCLAV VILHELM

3. Some Formulae (Areas, Circumferences, Volumes, Surfaces, Centroids, Moments of Inertia)

By Václav Vilhelm

4. Plane Curves and Constructions

By Karel Drábek

5. Plane Analytic Geometry

By MILOSLAV ZELENKA

6. Solid Analytic Geometry

By FRANTIŠEK KEJLA

7. Vector Calculus

A. *Vector Algebra*

By FRANTIŠEK KEJLA

B. *Vector Analysis*

By KAREL REKTORYS

8. Tensor Calculus

By VÁCLAV VILHELM

9. Differential Geometry

By Bořivoj Kepr

A. Curves

B. *Surfaces*

10. Sequences and Series of Constant Terms. Infinite Products

By Karel Rektorys

11. Differential Calculus of Functions of a Real Variable

By KAREL REKTORYS

12. Functions of Two or More Variables

By KAREL REKTORYS

13. Integral Calculus of Functions of One Variable

By KAREL REKTORYS

14. Integral Calculus of Functions of Two and More Variables

By KAREL REKTORYS

15. Sequences and Series with Variable Terms (Sequences and Series of Functions)

by KAREL REKTORYS

16. The Space L_2. Orthogonal Systems. Fourier Series. Some Special Functions (Bessel Functions, etc.)
By KAREL REKTORYS

FOREWORD TO THE FIRST ENGLISH EDITION

By Professor F.M. ARSCOTT, M.Sc., Ph.D., F.I.M.A.,

Department of Mathematics, University of Surrey

A mathematician, pausing two-thirds of the way through the twentieth century to look back, might feel a justifiable pride in the process of his subject. Many old problems have been solved and others absorbed into wider questions, while new branches of the subject have appeared at frequent intervals and blossomed rapidly. Meanwhile, other scientific disciplines grow more mathematical; it is said that really good work in physics, chemistry or engineering requires a first degree in mathematics, and even disciplines which were never regarded as scientific are proving susceptible to mathematical analysis.

This coincidence of an explosion of mathematical activity with greatly enlarged scope for its application is, unhappily, overshadowed by a communication barrier. Between those who have mathematical knowledge and those who wish to use it there lies a great gulf. One can try to bridge this by bringing to the notice of abstract mathematicians the intriguing and challenging problems waiting for them in other fields – but mathematicians are not easily tempted from their ivory towers. This book starts, instead, from the other side, putting into the hands of the users of mathematics an array of powerful tools, of whose existence they may be unaware, with precise directions for their use. To achieve this in a reasonable compass something has to be sacrificed and the authors took the bold step of omitting virtually all proofs – an unorthodox but highly sensible procedure, since otherwise the book might have been ten times its present size. As it is, these covers contain the equivalent of a small library of standard texts on the uses of mathematics.

It is no coincidence that such a book should come from the Continent, for it is especially in Germany and eastern Europe that there flourishes the subject of "Angewandte Mathematik" – better described as "useful", "utilisable" or "applicable" mathematics rather than by the literal translation of "applied mathematics", which in Britain means something very different. For the English translation, therefore, the title "Survey of Applicable Mathematics" has been chosen.

The task of editing the translation has been interesting and congenial. We have sought to produce a text in good mathematical English while preserving all the distinctive features of the original. Notation has been left practically unchanged; only where Czech and English usages differ significantly have changes been made. Terminology has sometimes proved more difficult, such as when direct translation produced a term which, though clear and acceptable, was not generally used. In such cases we have usually retained the equivalent of the Czech original, with a note

giving the common English alternatives; thus matrices are described as "regular" rather than "non-singular", though the latter is given as an equivalent term. When, however, serious confusion might result, or a different English term has become completely standard, we have made the necessary changes.

An extensive revision of the bibliography has also been made, giving a fuller guide to current British and American literature. Translation of Russian literature have been referenced whenever they could be traced, names of Russian authors being transliterated according to the practice of the London Mathematical Society.

My colleagues and I have found the editing of this book an exciting and stimulating experience; throughout we have had the inestimable benefit of Professor Rektorys's advice and help and in commending this book to the English speaking mathematical world we would pay our own tribute to the scholarship and imagination of Professor Rektorys and his co-authors.

PREFACE TO THE FIRST CZECH EDITION

In recent years several books dealing with special fields of mathematics (for example, Angot's *Applied Mathematics for Electronic Engineering*, and others) have been published in Czechoslovakia. They have supplied readers with information, in a condensed form, about those mathematical disciplines which find employment in these particular fields.

This volume has been published as a result of the initiative of the Česká matice technická (Czech Scientific Institution for Propagation of Technical Literature). In particular, the late Professor Vyčichlo devoted much of his time and organisational powers to make clear questions concerning fundamental features and conception of this volume. The authors have attempted to produce a comprehensive work for the use of a very wide circle of readers, and the book comprises the great majority of mathematical disciplines applied in technology, yet the contributions have been prepared in such a way that a reader with only limited theoretical knowledge of mathematics can easily follow them. This volume contains, therefore, a survey of results in applicable mathematics needed by engineering graduates or other research workers, or by undergraduates and teachers of technological subjects. It is also intended to be of service to theoretical research workers in such related disciplines as physics, geodesy, etc., and to mathematicians themselves.

It was not easy to select the subject matter and to present it in a form acceptable to such a varied body of readers. During Professor Vyčichlo's lifetime an extensive survey was made in order to ascertain the views of a number of outstanding technologists; some of the opinions expressed regarding selection and presentation of subject matter showed extensive disagreement, but it was possible to formulate an outline plan for the selection of subject matter and its mode of presentation – even though some questions remained unanswered.

It was not possible to include any specialised disciplines used only in narrow fields of technology. Thus, electrical engineers may miss the theory of transmission, while readers particularly interested in solving systems of linear algebraic equations may regret the absence of reference to cracovians. On the other hand it was clearly necessary not only to include current mathematical topics but also to pay considerable attention to approximate methods. Prominent among the latter are approximate methods in algebra, including solution of systems of linear equations, of transcendental equations and algebraic equations of higher degree, and the determination of eigenvalues of matrices, while in the field of analysis we have included approximate

methods for the solution of differential equations (especially partial differential equations) and of integral equations; these are not yet adequately treated in technological literature. The book also includes comprehensive tables of integrals, of sums of series, and of solved differential equations, while in the chapter on statistics (Chapter 34) attention is devoted to the subject of quality control. The book does not, however, include the theory of computers or the technique of linear programming; these disciplines are developing so rapidly at the present time that any description would be out of date before it appeared in print.

The leading experts in our country were invited to write the contributions on individual subjects. While each author was allowed a certain degree of freedom, the editor-in-chief (of the Czech edition) has ensured the maintenance of a consistent style of treatment throughout the book. I should like to thank all the authors for their great patience and for incorporating my suggestions into their work.

The book omits proofs of the theorems and derivation of the results, but theorems and formulae are complemented with explanatory remarks and appropriate examples; in choosing these examples we have sought to include those which not only provide suitable illustration but also have practical importance. In stating the results we have borne in mind the varying standards of mathematical education and skill of the readers for whom the book is intended. In Algebra, for instance, we write: "If a, b are real or complex numbers, then ...", instead of: "If a, b are complex numbers, then ..."; a mathematician may legitimately object that the second statement is sufficient because real numbers are a special category of complex numbers, but by using the first form of statement we leave the mathematically less advanced reader in no doubt that the result is valid for real numbers as well as complex.

Some sections of this volume are not, and by their nature cannot be, truly original – for instance, the tables of integrals and of solved differential equations. These tables were abstracted from different books, namely from [26], and have been carefully checked.

Although the authors and annotators of the various chapters worked with extreme care, the possibility cannot be excluded that some errors remain undetected, and we shall be very grateful to any readers who inform us of such errors. The authors of individual chapters are responsible for their accuracy, while the editor-in-chief takes overall responsibility for the general outline of the book; he will be grateful for any criticism relating to the selection of the subject matter and its presentation.

The book is divided into chapters and sections (paragraphs), which are numbered according to the decimal system, so that 5.3 for example, means Chapter 5, Section 3. In each section the theorems, examples, etc., are numbered in order and are quoted by means of that number; if, for instance, in a certain section Example 1 is quoted, this refers to Example 1 of the current section. If, however, we refer to an example from another section, then the number of that section is given before the number of the example. Similarly a reference such as (5.3.2) relates to equation (2) of Section 5.3; thus in the running heads, we look up the number 5.3 of that section and there we find equation (2). Generally the page is also quoted for the reader's convenience.

The bibliography is to be found at the end of the book; in the text we refer to a work merely by quoting (in square brackets) its number in the list of references.

Grateful acknowledgement is due to the Česká matice technická and the Publishers of the Technological Literature (SNTL) in Prague. I am indebted also to many friends and colleagues, particularly to V. Dašek who has read the greater part of the manuscript, to I. Babuška for his great work in the organization of the project and for his many valuable suggestions, and to K. Drábek for preparing the diagrams. Thanks are also due to E. Jokl, M. Josífko, M. Pišl, Č. Vitner, J. Výborná and R. Výborný for their most careful revision of the manuscript and their many contributions to the improvement of the whole work. I have also to thank the Prometheus printing house for their extremely competent work.

Karel Rektorys

PREFACE TO THE FIRST ENGLISH EDITION

The task of translating this book into English has been pleasant but rather difficult. I must express appreciation to my colleagues Vl. Dlab, K. Komínek and R. Výborný who translated the greater part of the text, and also acknowledge the generous assistance rendered by A. Žaludová who revised the whole translation.

In the preparation and editing of this English translation I have received invaluable help from Professor F.M. Arscott and his colleagues at the University of Surrey in London. Without their co-operation it would be difficult to imagine a successful production of this English edition.

To all these individuals, and also the Iliffe Books Ltd., I want to express once more my sincere thanks.

Karel Rektorys

PREFACE TO THE SECOND REVISED ENGLISH EDITION

In the original Czech version, our Survey of Applicable Mathematics has appeared in its fifth edition this year. This fact represents a very satisfaction for the authors, because it is a testimony that they have succeeded in their primary intention to give such a book in hands of consumers of mathematics which would serve them as a sufficiently universal mathematical tool and which they could easily apply.

Nevertheless, in recent years, many fields of applicable mathematics went through considerable changes. This concerns, first of all, numerical methods, in particular those in linear algebra and differential equations, especially in the partial ones. It concerns as well mathematical statistics, new methods in economy, etc. Changes have been noticed also in so-called classical fields of mathematics. This all showed the necessity of a considerable revision of the work when the sixth Czech edition was being planned.

Simultaneously with the new Czech edition the present Second Revised English Edition was being prepared.

The revision of the book has been essential. This concerns, in particular, its second volume. Many chapters have been written quite anew. They have been Chapters 24 (on variational methods in boundary value problems), 25 (approximate solution of ordinary differential equations), 27 (the finite-difference method), 30 (numerical methods in linear algebra) and 32 (interpolation and splines). The original Chapters 33, 34 and 35 on probability and mathematical statistics were replaced by new Chapters 33 to 36. Quite new is the "economic" Chapter 37. Also Chapter 22 on functional analysis has been rewritten entirely, serving as a starting point for analytical as well as numerical methods of solution of partial differential equations. Essentially different became Chapters 28 (from the point of view of the Laplace and Fourier transforms), 18 (partial differential equations) and 19 (integral equations). The revision concerns also Chapters 23 (variational calculus), 20 (functions of a complex variable, where a new part on functions of several complex variables has been added) and 21 (with a small dictionary of conformal mappings).

The first volume of the book, containing "more classical" mathematical fields (classical algebra, geometry and calculus), was by far not undergone so many changes. Regarding the purpose of the book (as a handbook for consumers of mathematics, in the first place), the modernization had to be carried out very carefully here. For example, the chapter on differential geometry in a modern conception would be too

abstract for most of the readers. Similarly, §1.1 on some concepts of logic remained practically unchanged, being of a purely informative character, thus far from any axiomatic theory, the building up of which would have been quite inadequate from the point of view of users of this book. We also preferred a rather classical form of treating the text concerning curvilinear and surface integrals, even though the main integral theorems (as those by Gauss, Stokes, etc.) have been presented also in the symbolic of vector analysis. What has been written quite anew here are the sections on the Lebesgue and Stieltjes integrals, on the space L_2, orthogonal systems, and the Bessel functions as well as the text on approximate evaluation of definite integrals and on harmonic analysis. The new Paragraph 16.7 has been added on the possibility of treating special functions from the point of view of the theory of representation of groups.

I wish all this effort turns out useful for the readers.

Finally, I would like to express my sincere thanks to all my co-authors, especially to Dr. E. Vitásek for his great help to me when editing the book, to Ass. Prof. K. Drábek for a very careful preparation of drawings and – last but not least – to Prof. M. Hazewinkel for many good ideas and suggestions and to Kluwer Academic Publishers for their highly competent work.

Karel Rektorys

LIST OF SYMBOLS AND NOTATION

USED IN VOLUME I

Symbols and notation are arranged acording to their logical connections with various parts of mathematics.

The reader should note that it is often difficult to put a symbol or notation precisely in its apropriate place; it may happen therefore that he will have to look up a notation in a different place from that which he anticipated.

Symbol or Notation	Meaning
	Algebra
$=$	(is) equal to
$\equiv$	(is) identically equal to
$\neq$	(is) not equal to
$\not\equiv$	(is) not identically equal to
$\approx$	is approximately equal to
$\doteq$	is equal, after rounding of, to
$<$	is smaller than, is less than
$>$	greater than
$\leq$	is less than or equal to ...
$\geq$	is greater than or equal to ...
$+$	plus; positive sign
$-$	minus; negative sign
$.$, $\times$	multiplied by; this sign is often omitted, e.g. instead of $a \,.\, b$ we often write ab
:, $-$, /	divided by; over; in the text we often write, for example, $1/(2n+1)$ instead of $\frac{1}{2n+1}$; obviously, $1/2n+1$ stands for $\frac{1}{2n}+1$

Symbol or Notation	Meaning
$(\ ), [\ ], \{\ \}$	parentheses or round brackets, square brackets, curly brackets, respectively
$\sqrt[n]{a}$	the n-th root of a (the sign $\sqrt{\ }$ is called the *radicle* or the *radical* or the *radical sign*), instead of $\sqrt[2]{a}$ we write simply $\sqrt{a}$
$\lvert a \rvert$	the absolute value of the number a
a^n	$\underbrace{a \,.\, a \,.\, \ldots \,.\, a}_{n\text{-times}}$; the n-th power
a^{-n}	$\dfrac{1}{a^n}$
$n!$	$1 \,.\, 2 \,.\, 3 \,.\, \ldots \,.\, n$ (n-factorial or the factorial n); e.g. $3! = 1 \,.\, 2 \,.\, 3 = 6$
$(2n)!!$	$2 \,.\, 4 \,.\, 6 \,.\, \ldots \,.\, 2n$; e.g. $6!! = 2 \,.\, 4 \,.\, 6 = 48$
$\binom{r}{k}$	$\dfrac{r(r-1)\ldots(r-k+1)}{k!}$, r any real number
$\binom{n}{k}$, C_k^n	$\dfrac{n(n-1)\ldots(n-k+1)}{k!} = \dfrac{n!}{k!(n-k)!}$ the binomial coeficient, n a positive integer
$\sum$	the sum of, the summation sign; e.g. $\sum_{k=1}^{3} a_k = a_1 + a_2 + a_3$ $\sum_k a_k$ means: we sum over all values ok k considered
$\prod$	the product; $\prod_{k=1}^{3} a_k = a_1 a_2 a_3$
$\log_b a$	the logarithm of a to the base b
$\log a$	the common or Briggs logarithm (to the base 10)
$\ln a$	tha natural or Napierian logarithm (to the base e)
$\in$	is an element of; e.g. $x \in [a, b]$ means: x is (or lies) in the interval $[a, b]$;
$\notin$	is not an element of

Symbol or Notation	Meaning
$\subset$	the sign of inclusion; e.g. $M \subset N$ (see §1.23)
$\cup$	the union (the sum); e.g. $M \cup N$; often written $M + N$ (see §1.23)
$\cap$	the intersection (or the product); e.g. $M \cap N$ (see §1.23)
$\max(a_1, a_2, \ldots, a_n)$	the greatest of the numbers a_1, a_2, ..., a_n
$\min(a_1, a_2, \ldots, a_n)$	the least of the numbers a_1, a_2, ..., a_n
$\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$, $\boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$	n-component vector (or vector of order n) with components (coordinates) a_1, a_2, ..., a_n
$\boldsymbol{A} = \begin{bmatrix} a_{11}, & a_{12}, & a_{13} \\ a_{21}, & a_{22}, & a_{23} \end{bmatrix}$	the 2 by 3 matrix (see §1.16)
$\boldsymbol{A}'$, $\boldsymbol{A}^{\mathrm{T}}$	the transpose of a matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	the inverse of a matrix $\boldsymbol{A}$
$\boldsymbol{I}$	the identity matrix
$\boldsymbol{O}$	the zero-matrix
$\boldsymbol{A} \sim \boldsymbol{B}$	the matrices $\boldsymbol{A}$, $\boldsymbol{B}$ are equivalent
$A = \begin{vmatrix} a_{11}, & a_{12} \\ a_{21}, & a_{22} \end{vmatrix}$	the determinat of order 2 or of the second order (see §1.17)
A_{ik}	the minor belonging to the element a_{ik}
A_{ik}	the cofactor belonging to the element a_{ik}
	Geometry
$\parallel$	(is) parallel to
$\uparrow\uparrow$	(is) parallel to ... and of the same orientation
$\uparrow\downarrow$	(is) parallel to ... and of the opposite orientation
$\perp$	(is) perpendicular to

Symbol or Notation	Meaning
$\triangle$	the triangle; e.g. $\triangle ABC$ stands for a triangle with the vertices A, B, C
$^\circ$ degree $'$ minutes $''$ second	in the sexagesimal measure of angles
$\operatorname{arc}\alpha$	the arc, the radian (circular) measure of an angle α; if the magnitude of an angle α is given in degrees then $\operatorname{arc}\alpha = \frac{\pi}{180}$; e.g. for $\alpha = 90^\circ$ $\operatorname{arc}\alpha = \frac{\pi \cdot 90}{190} = \frac{\pi}{2}$
rad	the radian, the unit angle in circular measure; $1\,\text{rad} = 57^\circ 17' 44{\cdot}8''$
AB	the segment with the end-points A, B
$\overline{AB}$	the length of the segment AB
$C \prec D$	on an oriented straight line, a point C lies before a point D
(x, y)	the rectangular coordinates of a point in the plane
(x, y, z)	the rectangular coordinates of a point in space
(ρ, φ)	the polar coordinates
(ρ, φ, z)	the cylindrical coordinates of a point in space
(r, ϑ, φ)	the spherical coordinates of a point in space
Vectors in Geometry, Vector Calculus, Vector Analysis	
$\boldsymbol{a}$	a vector
$\overrightarrow{AB}$	the vector with the initial (starting) point A and the end point (terminal) point B
a, $\lvert\boldsymbol{a}\rvert$	the length (module) of a vector $\boldsymbol{a}$

Symbol or Notation	Meaning
$\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$	the principal (unit or coordinate) vectors in the axes x, y, z of a cartesian coordinate system
$\boldsymbol{r}$	the radius vector of a point (x, y, z) (a vector with the initial point $(0, 0, 0)$ and the end point (x, y, z))
$k\boldsymbol{a}$	k-multiple of the vector $\boldsymbol{a}$ (k being a scalar)
$\boldsymbol{a} \,.\, \boldsymbol{b}$, $\boldsymbol{ab}$	the scalat (inner) product of vectors $\boldsymbol{a}$, $\boldsymbol{b}$ (§7.1)
$\boldsymbol{a} \times \boldsymbol{b}$, $\boldsymbol{a} \wedge \boldsymbol{b}$, $[\boldsymbol{ab}]$	the vector (inner product of vectors $\boldsymbol{a}$, $\boldsymbol{b}$ (see §7.1)
$[\boldsymbol{abc}]$, $[\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}]$, $\boldsymbol{abc}$	the mixed product (or the trivector) of vectors $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$; $\boldsymbol{abc} = (\boldsymbol{a} \times \boldsymbol{b}) \,.\, \boldsymbol{c}$ (see §7.1)
$\boldsymbol{a}'(t)$, $\boldsymbol{a}^{(n)}(t)$	the first, the n-th derivative, respectively, of a vector $\boldsymbol{a}$ with respect to the scalar variable t, i.e. $\dfrac{\mathrm{d}\boldsymbol{a}(t)}{\mathrm{d}t}$ or $\dfrac{\mathrm{d}^{(n)}\boldsymbol{a}(t)}{\mathrm{d}t^{(n)}}$ (§7.2)
$\operatorname{grad} u$, ∇u	the gradient of u (§7.2)
$\operatorname{div} \boldsymbol{a}$, $\nabla \boldsymbol{a}$	the divergence of a vector $\boldsymbol{a}$ (§7.2)
$\operatorname{curl} \boldsymbol{a}$, $\operatorname{rot} \boldsymbol{a}$, $\nabla \times \boldsymbol{a}$	the curl of a vector $\boldsymbol{a}$
∇	the Hamilton nabla operator
a_{mn}^{ijkl}	the fourth-times contravariant and two-times covariant tensor
Analysis (Differential and Integral Calculus)	
(a, b) or $[a, b]$	an open or closed interval respectively (for details see §11.1)
$x \in [a, b]$	x belongs to the interval $[a, b]$, x lies in the interval $[a, b]$
$[a, b] \times [c, d]$	the cartesian product of the intervals $[a, b]$, $[c, d]$ (in the cartesian coordinate system in a plane the product is a rectangle with the vertices (a, c), (b, c), (b, d), (a, d))
$\{a_n\}$	a sequence with general term a_n

Symbol or Notation	Meaning
$\lim_{n\to\infty} a_n = a$	the sequence $\{a_n\}$ possesses a limit a
$\lim_{n\to\infty} a_n = +\infty$	the sequence $\{a_n\}$ diverges to $+\infty$
$\limsup_{n\to\infty} a_n$, $\overline{\lim}_{n\to\infty} a_n$	the greatest limiting point of a sequence $\{a_n\}$ (§ 10.1)
$\liminf_{n\to\infty} a_n$, $\underline{\lim}_{n\to\infty} a_n$	the least limiting point of a sequence $\{a_n\}$ (§ 10.1)
$\sum_{n=1}^{\infty} a_n$	the infinite series with general term a_n
$\prod_{n=1}^{\infty} a_n$	the infinite product with general term a_n
$f(x)$, $g(x)$, ...	function of a single variable x
$f(x,y)$, $g(x,y)$, ...	function of two variables x, y
$f(g(x))$, $f(g(x,y),h(x,y))$	composite functions
$O(f(x))$, $o(f(x))$	(see § 11.4)
$\max_{a\le x\le b} f(x)$ or $\min_{a\le x\le b} f(x)$	the maximum or minimum value of a function $f(x)$ on an interval $[a,b]$
$\sup_{a\le x\le b} f(x)$ or $\inf_{a\le x\le b} f(x)$	the least upper bound (the supremum) or the greatest lower bound (the infimum) of a function $f(x)$ on the interval $[a,b]$ (on the supremum and infimum see § 1.3)
$\lim_{x\to a} f(x) = A$	the function $f(x)$ has the limit A at the point a
$\lim_{x\to a+} f(x) = B$, $f(a+0) = B$	the function $f(x)$ possesses the right-hand limit B at the point a
$\lim_{x\to a-} f(x) = -\infty$, $f(a-0) = -\infty$	the function $f(x)$ has the infinite left-hand limit $-\infty$ at the point a
$\lim_{x\to+\infty} f(x) = C$	the function $f(x)$ has the limit C at the point $+\infty$

Symbol or Notation	Meaning
y', f', $\frac{dy}{dx}$, $\frac{df}{dx}$	the (first) derivative of the function $y = f(x)$
$y^{(n)}$, $f^{(n)}(x)$, $\frac{d^{(n)}y}{dx^{(n)}}$, $\frac{d^{(n)}f}{dx^{(n)}}$	the n-th derivative of the function $y = f(x)$; we write y'', y''', f'', $f'''(x)$ instead of $y^{(2)}$, $y^{(3)}$, $f^{(2)}(x)$, $f^{(3)}(x)$
dy, $df(x)$	the differential of the function $y = f(x)$
∂y, $\partial f(x)$	the variaton of the function $y = f(x)$
$\frac{\partial f}{\partial x}$, f'_x, f_x	the partial derivatives of the function f (of several variables) with respect to x
$\frac{\partial f}{\partial n}$, $\frac{\partial f}{\partial \nu}$	the derivative in the direction of the outward normal
$\frac{\partial^2 f}{\partial x^2}$, f''_{xx}, f_{xx}	the second partial derivative of the function f with respect to x
$\frac{\partial^2 f}{\partial x \partial y}$, f''_{xy}, f_{xy}	the second mixed derivative of the function f; $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right)$
$d_x f$	the partial differential of the function f (of several variables)
df	the total differential of the function f
$\frac{\partial(y_1, y_2, \ldots, y_n)}{\partial(x_1, x_2, \ldots, x_n)}$	the functional determinant (the Jacobian) of the system of functions y_1, y_2, ..., y_n with respect to the variables x_1, x_2, ..., x_n; cf. § 12.7
$\int$	the indefinite integral (the primitive)
$\int_a^b$	the definite integral beetween the limits a, b
$\int_a^\infty$	the improper integral (§ 13.8)

Symbol or Notation	Meaning
$\int_a^b f \, \mathrm{d}g$	the Stieltjes integral
$[f(x)]_a^b$	$f(b) - f(a)$
$\iint_\Omega$	the double integral over (or in) a region Ω
$\iiint_\Omega$	the triple integral over (or in) a region Ω
$\int_k$	the line integral over (or along) a curve k
$\iint_S$	the surface integral over a surface S
$f \in L_2(a, b)$	a function f is square integrable in the interval $[a, b]$
(f, g)	the scalar (inner) product of functions (§ 16.1)
$\|f\|$	the norm of the function (§ 16.1)
a region of type A	(see § 14.1)
a solid of type A	(see § 14.1)
a function of type B	(see § 14.1)
π	the number π; $\pi \doteq 3{\cdot}141\,592\,654$
e	the base of natural logarithmus; $\mathrm{e} \doteq 2{\cdot}718\,281\,828$
C	Euler constant; $C \doteq 0{\cdot}557\,215\,655$
M	the modulus of common logarithmus; $M = \log \mathrm{e} \doteq 0{\cdot}434\,294\,48$
m	the modulus of natural logarithms; $m = \ln \mathrm{e} \doteq 2{\cdot}302\,585\,09$
$\sin x$	the sine
$\cos x$	the cosine
$\tan x$	the tangent
$\cot x$	the cotangent

Symbol or Notation	Meaning
$\sec x$	the secant
$\operatorname{cosec} x$	the cosecant
$\arcsin x$	the arc sine
$\arccos x$	the arc cosine
$\arctan x$	the arc tangent
$\operatorname{arccot} x$	the arc cotangent
$\sinh x$	the hyperbolic sine
$\cosh x$	the hyperbolic cosine
$\tanh x$	the hyperbolic tangent
$\coth x$	the hyperbolic cotangent
$\operatorname{arsinh} x$ $\operatorname{arcosh} x$ $\operatorname{artanh} x$ $\operatorname{arcoth} x$	the inverse of the hyperbolic $\left\{\begin{matrix}\text{sine}\\ \text{cosine}\\ \text{tangent}\\ \text{cotangent}\end{matrix}\right\}$
a^x	the exponential function with the base a, or the general exponential function
e^x	the exponential function (we often write $\exp x$, particulary when the argument is rather cumbersome; e.g. $\exp(x/at) = \mathrm{e}^{x/at}$)
$\log_a x$	the logarithm of x to the base a
$\ln x$	the natural logarithm of x
$\Gamma(x)$	the Gamma function
$\mathrm{B}(x)$	the Beta function
$\mathrm{J}_\nu(x)$	the Bessel function of the first kind of order ν
$\mathrm{Y}_\nu(x)$	the Bessel function of the second kind of order ν
$\mathrm{Y}_n(x)$	the Weber (Neumann) function
$\mathrm{H}_\nu^{(1)}(x)$, $\mathrm{H}_\nu^{(2)}(x)$	the Hankel function
$\mathrm{I}_\nu(x)$, $\mathrm{K}_\nu(x)$	the modified Bessel functions

Symbol or Notation	Meaning
ber x, bei x ker x, kei x	the Kelvin functions
$P_n(x)$	the Legendre polynomials of degree n
$P_n^m(x)$, $P_{n,m}(x)$	the adjoint Legender function
$Y_{n(c)}^m(x)$, $Y_{n(s)}^m(x)$	spherical functions
$T_n(x)$	the Chebyshev polynomials of degree n
$L_n(x)$	the Laguerre polynomials of degree n
$H_n(x)$	the Hermite polynomials of degree n
$F(\alpha, \beta, \gamma, x)$	the hypergeometric series (function)
Si (x)	the sine integral (§ 13.1)
Ci (x)	the cosine integral (§ 13.1)
li (x)	the logarithmic integral (§ 13.1)
Ei (x)	(§ 13.1)
erf (x)	the error function: $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$
erfc (x)	$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2}\, dt$
$F(k, \varphi)$	the legendre elliptic integral of the first kind in the normal form
$E(k, \varphi)$	the legendre elliptic integral of the second kind in the normal form
K	the complete elliptic integral of the first kind; $K = F(k, \frac{1}{2}\pi)$
E	the complete elliptic integral of the second kind; $E = E(k, \frac{1}{2}\pi)$
sn u, cn u, dn u	jacobian functions; see § 13.12

Mathematics and Its Applications

Managing Editor:

M. HAZEWINKEL

Centre for Mathematics and Computer Science, Amsterdam, The Netherlands

Volume 281

Survey of Applicable Mathematics

Second Revised Edition

VOLUME II

Karel Rektorys
Technical University,
Prague, Slovak Republic

Springer Science+Business Media, LLC

CONTENTS

VOLUME II

18. Partial Differential Equations

By KAREL REKTORYS

19. Integral Equations

By KAREL REKTORYS

20. Functions of One and More Complex Variables

A. *Functions of One Complex Variable*

By KAREL REKTORYS

B. *Functions of Several Complex Variables*

By JAROSLAV FUKA

21. Conformal Mapping

By JAROSLAV FUKA

Numerical Methods in Conformal Mappings

22. Fundamentals of the Theory of Sets and Functional Analysis

By KAREL REKTORYS

23. Calculus of Variations

By FRANTIŠEK NOŽIČKA

A. *Problems of the First Category (Elementary Problems of the Calculus of Variations)*

B. *Problems of the Second Category (Extrema of Functionals of the Form* $\int_a^b F(x, y_1, \ldots, y_n, y_1', \ldots, y_n')\,\mathrm{d}x$)

C. *Problems of the Third Category (Extrema of Functionals of the Form* $\int_a^b F(x, y, y', .., y^{(n)})\,\mathrm{d}x$)

D. *Problems of the Fourth Category (Functionals Depending on a Function of n Variables)*

B. *Boundary Value Problems*

26. Solution of Partial Differential Equations by Infinite Series (by the Fourier Method)

By KAREL REKTORYS

27. Solution of Partial Differential Equations by the Finite-Difference Method

By EMIL VITÁSEK

28. Integral Transforms (Operational Calculus)

By JINDŘICH NEČAS

29. Approximate Solution of Fredholm's Integral Equations

By KAREL REKTORYS

30. Numerical Methods in Linear Algebra

By JITKA SEGETHOVÁ AND KAREL SEGETH

A. *Solution of Systems of Linear Algebraic Equations*

B. *Computation of Eigenvalues and Eigenvectors of Matrices*

31. Numerical Solution of Algebraic and Transcendental Equations

By Miroslav Fiedler

32. Approximation. Interpolation, Splines

By Emil Vitásek

33. Probability Theory

By Tomáš Cipra

34. Mathematical Statistics

By TOMÁŠ CIPRA

35. Topics in Statistical Inference

By TOMÁŠ CIPRA

A. *Regression Analysis. Fitting Curves to Empirical Data. Calculus of Observations*

B. *Analysis of Variance*

D. *Reliability Theory*

E. *Statistical Principles of Quality Control*

36. Stochastic Processes

By TOMÁŠ CIPRA

A. *Markov Processes*

B. *Queueing Theory*

C. *Stationary Processes*

37. Linear Programming

By FRANTIŠEK NOŽIČKA

PREFACE TO VOLUME II

This Preface has been written for those readers who do not possess the first volume of the second revised edition of our Survey and who — consequently — have not read the preface to the whole revised work.

In recent years, many fields of mathematics passed through considerable changes. In mathematics for applications, this fact concerned, first of all, numerical methods (namely in linear algebra and in ordinary and partial differential equations), mathematical statistics and related fields, etc. Changes could be noticed also in many "classical" fields of mathematics. This all gave an impulse for a deep revision of our original work.

Many modifications of the text have already been realized in the first volume, containing classical fields of algebra, geometry and analysis. However, the second volume has been revised in a much more considerable way. Most of the chapters of this volume are quite new now. This concerns, in particular, Chapters 30, 25, 24 (numerical methods in linear algebra and in ordinary as well as in partial differential equations), then 33, 34, 35, 36 (probability, statistics and related topics), 32 (interpolation, splines), 22 (functional analysis) and 37 (the "economic" chapter on linear programming). The remaining chapters of Volume II went through essential changes: In Chapter 18, sections concerning generalized and weak solutions of elliptic partial differential equations of "arbitrary order" were added, including nonlinear equations and a section on solution of evolution problems by the method of discretization in time (= by the Rothe method equipped with a new technics). Chapter 20 has been completed by sections on functions of more complex variables, Chapter 21 by a small dictionary of conformal mappings. Chapter 23 has been extended by further categories of calculus of variations, Chapter 27 by a section concerning general questions on convergence of the finite-difference method, in Chapter 28 more attention has been paid to Laplace and Fourier transforms than before. Also the text of some individual sections of these or other chapters became different, although the titles of these sections remained the same.

The whole work is divided into 37 chapters. Individual chapters are divided into sections (= paragraphs). If, for example, Theorem 2 from the same section that is being studied is to be quoted, we write "see Theorem 2" only. However, if Theorem 2 from an other section, say section 17.17, is in question, we write "see Theorem 17.17.2". So we list, in the running heads, number 17.17 of that section and there we find Theorem 2. Similarly, we write "see equation (1)" if the

first equation of the just studied section is in question, but we write "see equation (17.17.1)" if we speak about the first equation from (another) section 17.17. For the reader's convenience, the corresponding page is often quoted.

Finishing the preface, I would like to thank once more all who took part in the production of this work.

Prague, November 27th, 1991 *Karel Rektorys*

LIST OF SYMBOLS AND NOTATION USED IN VOLUME II

List of symbols and notation relating to algebra, geometry, vector calculus and analysis (including special functions) is to be found in the first volume of this work.

Symbol or Notation	Meaning
Functions of a Complex Variable	
i, j	the imaginary unit, $i^2 = -1$, $i^3 = -i$ (in electrical engineering j is used instead of i)
$\mathrm{Re}\,\alpha$, $\mathrm{R}[\alpha]$	the real part of the complex number α
$\mathrm{Im}\,\alpha$, $\mathrm{I}[\alpha]$	the imaginary part of the complex number α
$\lvert\alpha\rvert$	the absolute value (modulus) of the complex number α
$\overline{\alpha}$	the (complex) conjugate of the complex number α
$\ln z$	the natural logarithm z (a multi-valued function)
$\ln_0 z$	the principal branch of the function $\ln z$ (a single-valued function)
$\operatorname*{res}_{z=z_k} f(z)$, $\mathrm{res}[f(z)]_{z=z_k}$	the residue of a function $f(z)$ at the point $z = z_k$
$\int_{a-i\infty}^{a+i\infty} f(z)\,dz$	the integral of the function $f(z)$ along the straight line $x = a$ which is parallel to the imaginary axis
$\mathbb{R}^{2n}$, $\mathbb{C}^n$	see introduction to § 20.7
Functional Analysis	
H	the Hilbert space (§ 22.4)
B	the Banach space (§ 22.4)

Symbol or Notation	Meaning
$L_2(a, b)$, $L_2(\Omega)$	the space of functions square integrable in the interval (a, b), or in the region Ω, with the norm $\|u\|_{L_2(a,b)}$, or $\|u\|_{L_2(\Omega)}$, respectively, $\|u\|$ in brief (§ 22.2)
$W_2^{(k)}(\Omega)$, $H^k(\Omega)$, $H_k(\Omega)$	the Sobolev space (§ 22.4)
$\mathring{W}_2^{(k)}(\Omega)$, $H_0^k(\Omega)$	the Sobolev space of functions with zero traces on the boundary (§ 22.4)
$W_2^0(\Omega) = H^0(\Omega) = L_2(\Omega)$	
H_A	the energetic space (§ 22.6)
$d(u, v)$	the distance of two elements u, v in a metric space (§ 22.2)
(u, v), $\|u\|$	the scalar product, the norm
in particular:	
$\|u\|_{H^k(\Omega)}$	the norm of the function u in the Sobolev space $H^k(\Omega)$
$\operatorname{supp} u$	the support of the function u (§ 22.4)
$D^i u$	the generalized derivative of the function u (§ 22.4)
$H = L + M$, $H = L \oplus M$	the direct sum of subspaces in the Hilbert space H (§ 22.4)
$u_n \rightharpoonup u$	the weak convergence of the sequence $\{u_n\}$ (§ 22.5)
$X \subsetneq Y$	the continuous imbedding of the space X into the space Y (§ 22.4)
$D(A)$, D_A	the domain of definition of the operator A (§ 22.5)
A^{-1}	the inverse operator to the operator A (§ 22.5)
$\|A\|$	the norm of the operator A (§ 22.5)
A^*	the adjoint operator to the operator A (§ 22.5, § 22.6)
$F'(u_0, v)$, $\mathrm{D}F(u_0, v)$, $\mathrm{d}F(u_0, v)$	the Gâteaux differential of the functional F (§ 22.8)
$F''(u_0, v_0, w)$	the second Gâteaux differential (§ 22.8)

Symbol or Notation	Meaning
Probability Theory, Mathematical Statistics, Topics in Statistical Inference, Stochastic Processes	
Ω	the space of elementary events (Chap. 33) or the parameter space (Chap. 34)
$P(A)$	the probability of an event A
$P(A\|B)$	the conditional probability of an event A given an event B
$X, Y, Z, \ldots$	random variables
$P(a < X \leqq b)$	the probability that a random variable X lies in the interval (a,b]
$F(x) = P(X \leqq x)$	the distribution function of a random variable X
$p_j = P(X = x_j)$	the probability of the value x_j of a discrete random variable X
$f(x)$	the probability density of a random variable
$\mathrm{E}(X)$	the mean of a random variable X
μ	the mean of a random variable
μ_k'	the k-th moment
μ_k	the k-th central moment
$\mathrm{var}(X)$	the variance of a random variable X
σ^2	the variance of a random variable
σ	the standard deviation
γ_1	the coefficient of skewness
γ_2	the coefficient of kurtosis
x_P	the P-quantile
$\hat{x}$	the mode
$x_{0.5}$	the median
$\varphi(t)$	the characteristic function of a random variable
κ_k	the k-th cumulant
$\boldsymbol{x} = (x_1, \ldots, x_n)'$	the n-component column vector
$\boldsymbol{x}' = (x_1, \ldots, x_n)$	the n-component row vector
$F(x_1, \ldots, x_n)$	the distribution function of a random vector (the joint distribution function)

Symbol or Notation	Meaning
$f(x_1, \ldots, x_n)$	the probability density of a random vector (the joint probability density)
$F_{i_1,\ldots,i_k}(x_{i_1}, \ldots, x_{i_k})$	the marginal distribution function
$f_{i_1,\ldots,i_k}(x_{i_1}, \ldots, x_{i_k})$	the marginal probability density
$f(x_1 \mid x_2)$	the conditional probability density
$\mathrm{E}(X_1 \mid X_2 = x_2)$	the conditional mean
$\mathrm{E}(\boldsymbol{X})$	the mean of a random vector $\boldsymbol{X}$
$\mathrm{cov}(X, Y)$	the covariance of random variables X, Y
$\varrho(X, Y)$	the correlation coefficient of random variables X, Y
$\boldsymbol{\Sigma}_{\boldsymbol{X}}$	the covariance matrix of a random vector $\boldsymbol{X}$
$\varphi(t_1, \ldots, t_n)$	the characteristic function of a random vector
$N(\mu, \sigma^2)$	the normal distribution
$N(0, 1)$	the standard normal distribution
$\varphi(x)$	the probability density of $N(0, 1)$
$\Phi(x)$	the distribution function of $N(0, 1)$
u_P	the P-quantile of $N(0, 1)$
$\chi^2(n)$	the χ^2 (chi-square) distribution with n degrees of freedom
$\chi^2_P(n)$	the P-quantile of $\chi^2(n)$
$t(n)$	the t (Student) distribution with n degrees of freedom
$t_P(n)$	the P-quantile of $t(n)$
$F(n_1, n_2)$	the F (Fisher-Snedecor) distribution with n_1 and n_2 degrees of freedom
$F_P(n_1, n_2)$	the P-quantile of $F(n_1, n_2)$
$N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	the n-variate normal distribution
$X_1, \ldots, X_n$	the random sample of size n
$x_1, \ldots, x_n$	the observations of a random sample $X_1, \ldots, X_n$
n_i	the frequency of an observation x_i
$F_n(x)$	the empirical distribution function for a random sample of size n
$\overline{X}$ or $\overline{x}$	the sample mean
S^2 or s^2	the sample variance
S or s	the sample standard deviation

Symbol or Notation	Meaning
M_k' or m_k'	the sample k-th moment
M_k or m_k	the sample k-th central moment
G_1	the sample coefficient of skewness
G_2	the sample coefficient of kurtosis
r	the sample correlation coefficient
r_{ij}	the sample correlation coefficient of the i-th and j-th component of a multivariate random sample
$X_{(1)}, \ldots, X_{(n)}$	the ordered random sample
$\widetilde{X}$ or $\widetilde{x}$	the sample median
$\hat{\vartheta}$	the (point) estimator of a parameter ϑ
$b(\vartheta)$	the bias of an estimator of a parameter ϑ
$L(\boldsymbol{\vartheta})$	the likelihood function
H_0	the null hypothesis
H_1	the alternative hypothesis
W	the critical region of a statistical test
$\beta(\boldsymbol{\vartheta})$	the power function of a statistical test
α	the significance level of a statistical test or the producer's risk in an acceptance sampling
β	the consumer's risk in an acceptance sampling
n_{ij}	the frequencies in a contingency table
$n_{i.}$, $n_{.j}$	the marginal frequencies in a contingency table
e	the error variable of a regression model
$\hat{e}_1, \ldots, \hat{e}_n$	the residuals of a linear regression model
R^2	the coefficient of determination
$x_{i.}$, $x_{..}$	$x_{i.} = \sum_{p=1}^{n_i} x_{ip}$, $x_{..} = \sum_{i=1}^{I} \sum_{p=1}^{n_i} x_{ip}$ in a one-way classification with values x_{ip} $(i = 1, \ldots, I;\ p = 1, \ldots, n_i)$
$\overline{x}_{i.}$, $\overline{x}_{..}$	$\overline{x}_{i.} = x_{i.}/n_i$, $\overline{x}_{..} = x_{..}/(n_1 + \cdots + n_I)$
S_T	the total sum of squares
S_A	the A-factor sum of squares
S_e	the residual sum of squares
$R(x)$	the reliability function

Symbol or Notation	Meaning
$r(x)$	the hazard rate (the failure rate)
c	the acceptance number in an acceptance sampling
$X(t)$	the stochastic process (the random function)
X_n	the random sequence (the time series)
$p_i(t)$	the probability distribution of a Markov process at time t
$p_{ij}(t)$	the transition probability of a homogeneous Markov process
π_i	the stationary distribution of a Markov process
$p_i(n)$	the probability distribution of a Markov chain at time n
$p_{ij}(k)$	the transition probability of a homogeneous Markov chain (for $k = 1$ we write p_{ij}, in brief)
$\boldsymbol{P}(k)$	the transition matrix of a homogeneous Markov chain (for $k = 1$ we write $\boldsymbol{P}$, in brief)
q_{ij}	the transition intensity of a homogeneous Markov process
λ	the intensity of the Poisson process
$\mu(t)$	the mean of a stochastic process
$R(s, t)$	the autocovariance function of a stochastic process
$R(t)$	the autocovariance function of a stationary stochastic process
$\boldsymbol{\mu}(t)$	the mean of a multivariate stochastic process
$\boldsymbol{R}(s, t)$	the matrix autocovariance function of a multivariate stochastic process
$R_{ij}(s, t)$	the cross-covariance function of the i-th and j-th component of a multivariate stochastic process
$F(\lambda)$	the spectral distribution function of a stochastic process
$f(\lambda)$	the spectral density of a stochastic process
$I(\lambda)$	the periodogram of a stochastic process
$\psi(\lambda)$	the transfer function of a filter

Symbol or Notation	Meaning
Numerical Methods of Linear Algebra	
$\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$	the n-component vector (of order n) with components $a_1, a_2, \ldots, a_n$
$\boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$	the n-component column vector (a one-column matrix)
$\boldsymbol{u}^{\mathrm{T}} = (u_1, u_2, \ldots, u_n)$	the transpose of the vector $\boldsymbol{u}$
$\boldsymbol{A} = \begin{bmatrix} a_{11}, & a_{12}, & a_{13} \\ a_{21}, & a_{22}, & a_{23} \end{bmatrix}$	the 2 by 3 matrix (§ 1.16)
$\boldsymbol{A} = (a_{ij})$	the matrix with entries (elements) a_{ij}
$\boldsymbol{A}'$, $\boldsymbol{A}^{\mathrm{T}}$	the transpose of a matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	the inverse of a matrix $\boldsymbol{A}$
$\boldsymbol{A}^{+}$	the pseudoinverse of a matrix $\boldsymbol{A}$ (§ 30.4)
$\boldsymbol{I}$, $\boldsymbol{E}$	the identity matrix
$\boldsymbol{0}$, $\boldsymbol{o}$	the zero matrix, the zero vector
$\|\boldsymbol{a}\|$	the norm of a vector $\boldsymbol{a}$ (§ 30.3)
$\|\boldsymbol{A}\|$	the norm of a matrix $\boldsymbol{A}$ (§ 30.3)
$D = \begin{vmatrix} a_{11}, & a_{12} \\ a_{21}, & a_{22} \end{vmatrix}$	the determinant of order 2 (of the second order) (§ 1.17)
$\det \boldsymbol{A}$	the determinant of a matrix $\boldsymbol{A}$
$\mathrm{diag}(a_i)$	the diagonal matrix with diagonal entries a_i
$\varrho(\boldsymbol{A})$	the spectral radius of a matrix $\boldsymbol{A}$ (§ 30.3)
$\kappa(\boldsymbol{A})$	the condition number of a matrix $\boldsymbol{A}$ (§ 30.3)
Further Symbols and Notation	
$M = \{2, 7, 9\}$	the set M consists of the numbers 2, 7, 9
R_n, R^n	the set of ordered n-tuples of real numbers
E_n, E^n	the n-dimensional Euclidean space (the set of points of the real n-dimensional space (identified usually with R_n), equipped with the usual Euclidean metric) (§ 22.1)

Symbol or Notation	Meaning
$\{(x, y) \in E_2 \mid xy = 1)\}$	the set of such points from E_2 for which the relation $xy = 1$ holds (i.e. which lie on the hyperbola $xy = 1$)
Ω, G	the region (§ 22.1)
$\overline{\Omega} = \Omega + S = \Omega \cup S$	the closed region (= the region Ω + its boundary S)
$e^{\boldsymbol{A}x}$	the exponential function of a matrix $\boldsymbol{A}$ (§ 17.18)
$((u, v))$, $a(v, u)$	the bilinear form corresponding to a differential operator A and to the given boundary conditions (§ 18.9, § 24.6)
δy, $\delta f(x)$	the variation of a function $y = f(x)$ (§ 23.3)
δI	the variation of the functional I (§ 23.3)
F'_y or $F'_{y'}$	the partial derivative of the function F with respect to y, or y' (Chap. 23)
$\mathcal{L}(f(t))$	the Laplace transform of the function $f(t)$ (§ 28.1)
$f[x_0, \ldots, x_N]$	the n-th relative (divided) difference (§ 32.6)
$\Delta^n f(x)$	the n-th forward (ordinary) difference (§ 32.7)
$\nabla^n f(x)$	the n-th backward difference (§ 32.7)

Survey of Applicable Mathematics

Second Revised Edition

VOLUME III

Karel Rektorys
Technical University,
Prague, Slovak Republic

Springer Science+Business Media, LLC

1. ARITHMETIC AND ALGEBRA

By VÁCLAV VILHELM

References: [2], [10], [11], [12], [13], [18], [20], [21], [24], [33], [36], [38], [46], [50], [51], [53], [58], [61], [63], [69], [70], [71], [73], [77], [79], [82], [83], [88], [92], [93], [97], [98], [100], [101], [102], [103], [105], [107], [113], [120], [129], [130], [140], [141], [151], [159], [170], [171], [172], [177], [178], [179].

1.1. Some Concepts of Logic

By a *sentence* is to be understood any statement concerning which it is meaningful to say that its content is true (it holds), or false (it does not hold).

The *opposite* or *contradictory of a sentence* A (denoted by not-A or A') is a sentence defined in the following way: The sentence not-A is true if the sentence A is false, and vice versa.

Example 1. "All chairs in the room are occupied" is an example of a sentence. Its opposite is the sentence "Not all chairs in the room are occupied", i.e. "There is at least one unoccupied chair in the room".

If A, B are two sentences, then one can construct from them new sentences in various ways. First, let us introduce the concept of *implication.*

We say that "the sentence A implies the sentence B" or "B follows from A" or "if A is true, then B is true" or "B is a necessary condition for A" or "A is a sufficient condition for B" (in notation $A \Rightarrow B$), if the truth of the sentence B follows from the truth of the sentence A. (If the sentence A is false, then the sentence B can be either true or false.) In an implication $A \Rightarrow B$, A is called the *premise* (*cause*) and B the *conclusion* (*effect*) of the implication.

Example 2. The implications "If a is an integer divisible by four, then a is even" and "If the sum of the angles of a triangle is 120°, then every triangle is a right-angled triangle" are true. (The premise of the second implication is false and thus the implication is true.)

Another sentence combined from the sentences A, B is *equivalence*:

We say, that "the sentence A is equivalent to the sentence B" or "A is true if and only if B is true" or "A is a necessary and sufficient condition for B" (in notation $A \Leftrightarrow B$), if the sentences A and B are either both true or both false.

Example 3. A typical example of an equivalence is the sentence "A triangle is equilateral if and only if all its angles are equal".

REMARK 1. The equivalence $A \Leftrightarrow B$ is true if and only if both $A \Rightarrow B$ and $B \Rightarrow A$ are true.

REMARK 2. The sentence $A \Rightarrow B$ is equivalent to the sentence not-$B \Rightarrow$ not-A.

REMARK 3. Mathematical theorems usually have the form of an implication or an equivalence; e.g. "If a function $f(x)$ possesses a finite derivative at a point x_0, then it is continuous at x_0", "A quadratic equation with real coefficients has two distinct real roots if and only if its discriminant is positive."

1.2. Natural, Integral and Rational Numbers

Natural numbers are the numbers 1, 2, 3, 4, 5, ...

Natural numbers satisfy the *principle of complete* (or *mathematical*) *induction* or *finite induction*, namely:

If M is any set of natural numbers which contains the number 1 and which has the further property that if it contains the number n it also contains the number $n + 1$, then M contains *all* natural numbers.

REMARK 1. This principle is "intuitively evident": If a set M has the properties assumed in the above principle, then it contains the number 1. Hence, the property $n \in M \Rightarrow n + 1 \in M$ implies that the set M contains the numbers $1 + 1 = 2$, $2 + 1 = 3$ etc.

The principle of complete induction is the basis of "*proofs by complete induction*". To make the principle of such proofs clear, let us consider an example.

Example 1. Let $q \neq 1$. Then, for any natural number k, the formula

$$1 + q + q^2 + q^3 + \ldots + q^k = \frac{q^{k+1} - 1}{q - 1}$$

holds.

We shall prove this statement by complete induction. Let M be the set of those natural numbers k, for which the statement is valid. Evidently, the statement is true for $k = 1$ and thus $1 \in M$. Let us assume that the statement is true for $k = n$, i.e.

$n \in M$. Then

$$1 + q + \ldots + q^n + q^{n+1} = \frac{q^{n+1} - 1}{q - 1} + q^{n+1} = \frac{q^{n+1} - 1 + q^{n+2} - q^{n+1}}{q - 1} =$$

$$= \frac{q^{n+2} - 1}{q - 1};$$

hence the statement is true also for $n + 1$, i.e. $n + 1 \in M$. Since the statement holds for $k = 1$, the set M contains, in accordance with the principle of complete induction, all natural numbers and therefore the statement holds for any natural number k.

Integers are obtained by extending the set of all natural numbers by the numbers 0 (zero) and $-1, -2, -3, \ldots$.

The numbers $1, 2, 3, \ldots$ are called *positive*, the numbers $-1, -2, -3, \ldots$ *negative*.

Definition 1. The fact that an integer a is positive, or negative, is denoted by $a > 0$, or $a < 0$, respectively.

We say that a number a is *less* than or *greater* than a number b if the difference $b - a > 0$, or $b - a < 0$, and in that case we write $a < b$, or $a > b$, respectively.

REMARK 2. The notation $a \leqq b$ means that either $a < b$ or $a = b$; similarly for $a \geqq b$.

Theorem 1. *By the relation $<$ the integers are ordered. This ordering has the following properties (a, b, c, d stand for integers):*

A. *For any two integers a, b one and only one of the following relations holds:*

$$a < b, \quad a > b, \quad a = b.$$

B. $a < b, \quad b < c \Rightarrow a < c.$

C. $a < b, \quad c \leqq d \Rightarrow a + c < b + d.$

D. $a < b, \quad c > 0 \Rightarrow ac < bc.$

E. $a < b, \quad c < 0 \Rightarrow ac > bc.$

REMARK 3. The properties A–E express the *basic rules of inequalities*. D and E imply that $bc > 0 \Leftrightarrow b > 0, c > 0$ or $b < 0, c < 0$.

Rational numbers are obtained by extending the set of all integers by fractions, i.e. numbers of the form p/q with integers p and q, $q \neq 0$. The equality $p/q = p'/q'$ holds if and only if $pq' = p'q$.

Theorem 2. *Any rational number can be written in the form a/b, where a is an integer and b a natural number.*

Theorem 3. *Rational numbers can be added, subtracted, multiplied and divided; these operations satisfy the following rules* (a, b, c *stand for rational numbers*):

1. $(a + b) + c = a + (b + c)$ (*associative law for addition*).
2. $a + b = b + a$ (*commutative law for addition*).
3. *For every* a, $a + 0 = a$.
4. *For every* a, *there exists a number* $-a$ *such that* $a + (-a) = 0$.
5. $(ab)\,c = a(bc)$ (*associative law for multiplication*).
6. $ab = ba$ (*commutative law for multiplication*).
7. *For every* a, $a \cdot 1 = a$.
8. *For every* $a \neq 0$, *there exists a number* a' *such that* $aa' = 1$. (We write $a' = a^{-1}$ or $a' = 1/a$.)
9. $(a + b)\,c = ac + bc$ (*distributive law*).

REMARK 4. Addition, multiplication and division of fractions (rational numbers) are performed according to the following rules:

$$\frac{a_1}{b_1} + \frac{a_2}{b_2} = \frac{a_1 b_2 + a_2 b_1}{b_1 b_2}, \quad \frac{a_1}{b_1}\frac{a_2}{b_2} = \frac{a_1 a_2}{b_1 b_2},$$

$$\frac{\dfrac{a_1}{b_1}}{\dfrac{a_2}{b_2}} = \frac{a_1}{b_1}\frac{b_2}{a_2} = \frac{a_1 b_2}{a_2 b_1}.$$

In the last rule we assume, of course, that $a_2/b_2 \neq 0$, i.e. $a_2 \neq 0$.

Theorem 4. *The rational numbers can be ordered in the following way: If* $a = p/q$, $b = p'/q'$, *where* p, p' *are integers and* q, q' *natural numbers, then* $a \lessgtr b$ *according as* $pq' \lessgtr p'q$. *This order agrees with that of the integers and satisfies the rules* A–E *of Theorem* 1.

1.3. Real Numbers

The ordered set of the rational numbers is *dense* (i.e. between any two different rational numbers there is an infinity of rational numbers), but it has *gaps*; this means that there exist partitions of the set of the rational numbers into two non-empty classes A, B such that

1° $A \cup B$ (see Definition 1.23.2, p. 45) is the set of all rational numbers;

2° for every number $a \in A$ and every number $b \in B$, the relation $a < b$ holds;

3° the set A has no greatest number and the set B has no least number. (One can get such a partition by defining e.g. the class B to contain all positive rational numbers x satisfying $x^2 > 2$ and the class A all the other rational numbers.)

Filling up these gaps by new, so-called *irrational numbers*, we extend the set of rational numbers and so get the *real numbers* (for the detailed theory see e.g. [4]).

Theorem 1. *The rules* 1—9 *of Theorem* 1.2.3 *also hold for addition and multiplication of real numbers.*

Theorem 2. *The real numbers can be ordered in such a way that this order corresponds to that of the rational numbers and the rules* A—E *of Theorem* 1.2.1 *hold.*

Theorem 3. *Every irrational number can be expressed in the form of an infinite non-periodic decimal fraction. Rational numbers are expressed by finite or infinite periodic decimal fractions.*

Definition 1. A real number α is said to be *algebraic* if it is a root of some algebraic equation $x^n + a_1 x^{n-1} + \ldots + a_n = 0$ with rational coefficients $a_1, a_2, \ldots, a_n$. If α is not algebraic, it is called *transcendental.* For example, the numbers e, π are transcendental.

Definition 2. A set M of real numbers is said to be *bounded above* (*or bounded below*), if there exists a real number a which is greater (or less) than any number belonging to M, respectively. The set M is said to be *bounded if* it is bounded above as well as below.

Definition 3. Let M be a set of real numbers. A real number ξ is called the *least* (*exact*) *upper bound of* M (l.u.b., briefly; we shall write $\xi = \sup M$), if 1° $a \leqq \xi$ for every $a \in M$, 2° ξ is the least number having the property 1°.

Similarly: A number η is called *the greatest* (*exact*) *lower bound of* M (g.l.b.; $\eta = \inf M$) if 1° $a \geqq \eta$ for every $a \in M$, 2° η is the greatest number having the property 1°.

Example 1. Let M be the set of all numbers $0, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \ldots$ [i.e. the numbers of the form $(n-1)/n$, where $n = 1, 2, 3, \ldots$]. The set M is bounded, since every number of M is greater than, say, -1 and less than, say, 5. The least upper bound of this set is the number 1, for every $x \in M$ satisfies $x \leqq 1$ (in fact, $x < 1$) and for every (fixed) number $a < 1$ there exists a number of the form $(k-1)/k$ (k being a natural number), in the set M, such that $(k-1)/k > a$. [The choice of $k > 1/(1-a)$ is sufficient]. The greatest lower bound of the set M is evidently 0.

The following theorem states the fundamental property of the ordering of real numbers:

Theorem 4. *Every non-empty set of real numbers bounded above or bounded below possesses a least upper bound or a greatest lower bound respectively. Therefore, there are no gaps in the ordering of real numbers.*

Theorem 5. *If a set M of real numbers possesses a greatest element (maximum, denoted by* max M*), then* sup M = max M. *Similarly if there exists a least element (minimum, denoted by* min M*) in M, then* inf M = min M.

Theorem 6. *Between any two different real numbers there is an infinity of rational as well as an infinity of irrational numbers.*

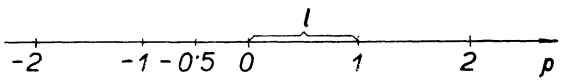


Fig. 1.1a.

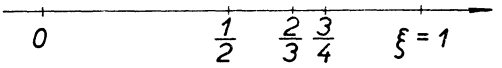


Fig. 1.1b.

REMARK 1 (The *numbered scale* or *continuum*, or *axis of real numbers*). Real numbers can be represented by points on a straight line. If we choose, on the straight line p, the origin O, a certain orientation of the straight line (Fig. 1.1a), and a unit of length l, then, to every real number a, there corresponds one and only one point A on the line p, whose coordinate is a; conversely, every point on the line p has a certain coordinate. The straight line p is then called the *numbered scale* or *continuum*. The points of the numbered scale are often identified with real numbers. A number a is less than a number b if and only if the point representing a is to the left of the point representing b on the numbered scale. Fig. 1.1b illustrates several numbers of the set M of Example 1 and the least upper bound ξ of this set.

REMARK 2. On so-called rounded off numbers and operations on them (abbreviated multiplication etc.), see Chap. 32.

1.4. Inequalities between Real Numbers. Absolute Value

Theorem 1. *Inequalities between real numbers satisfy the rules* A–E *of Theorem* 1.2.1.

Theorem 2.

$$0 < a < b \Rightarrow 0 < \frac{1}{b} < \frac{1}{a},$$

$$a < b < 0 \Rightarrow \frac{1}{b} < \frac{1}{a} < 0.$$

Theorem 3. *The inequality*

$$ax + b > 0 \tag{1}$$

with $\left\{\begin{array}{l} a > 0 \\ a < 0 \\ a = 0,\ b > 0 \end{array}\right\}$ *holds if and only if* $\left\{\begin{array}{l} x > -b/a \\ x < -b/a \\ x \text{ is arbitrary} \end{array}\right\}$.

For $a = 0$, $b \leqq 0$, *there is no* x *satisfying* (1).

Theorem 4. *The solution of the inequality*

$$ax^2 + bx + c > 0, \quad a \neq 0, \tag{2}$$

is as follows:

If the polynomial $f(x) = ax^2 + bx + c$ *has real zeros, then*

$$f(x) = a(x - \alpha_1)(x - \alpha_2), \quad \alpha_1 \leqq \alpha_2;$$

so

(a) *if* $a > 0$, *then* $f(x) > 0$ *for all* $x < \alpha_1$ *and for all* $x > \alpha_2$;

(b) *if* $a < 0$, *then* $f(x) > 0$ *for all* x *satisfying* $\alpha_1 < x < \alpha_2$. *If the polynomial* $f(x)$ *has no real roots, then it can be expressed in the form*

$$f(x) = a[(x + c)^2 + d] \quad \text{with} \quad d > 0,$$

and thus,

(a) *if* $a > 0$, *then* $f(x) > 0$ *for every real number* x;

(b) *if* $a < 0$, *then* $f(x) < 0$ *always and the inequality* $f(x) > 0$ *has no solution.*

REMARK 1. On inequalities between powers see Theorem 1.9.2, p. 14.

REMARK 2. Simple inequalities which one meets in practice are frequently reducible to inequalities of the type (1) or (2). When solving an inequality of the form $P(x)/|Q(x) > 0$, where $P(x)$ and $Q(x) \neq 0$ are polynomials, without common real zeros, the following theorem is useful: The function $P(x)/Q(x)$ changes its sign only in the neighbourhood of the zeros of odd multiplicity of the polynomials $P(x)$ and $Q(x)$. Thus, knowing the zeros of these polynomials and the sign of the function P/Q at one point where this function is non-zero, we can solve the given inequality quite easily. The procedure is illustrated in the following example.

Example 1. Let us solve the inequality

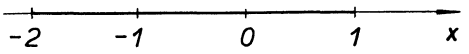
$$\frac{2x - 5}{x - 1} > 3$$

(i.e. find all real x for which this inequality holds). First, we transform the inequality to the form

$$\frac{2x - 5}{x - 1} - 3 > 0, \quad \text{i.e.} \quad \frac{-x - 2}{x - 1} > 0.$$

The polynomials $P(x) = -x - 2$ and $Q(x) = x - 1$ have the single zeros -2 and 1. Since $P(0)/Q(0) = 2 > 0$, the function P/Q is positive in the interval $(-2, 1)$ and negative in the intervals $(-\infty, -2), (1, +\infty)$. Hence the given inequality is satisfied for all x of the *open* interval $(-2, 1)$ and only for them (Fig. 1.2); these values of x represent the solution of the given inequality.

Fig. 1.2.



Definition 1. The *absolute value of a real number* a (denoted by $|a|$) is defined as follows:

$$|a| = a \quad \text{for} \quad a \geqq 0, \quad |a| = -a \quad \text{for} \quad a < 0.$$

Theorem 5. $|a| > 0$ *for* $a \neq 0$; $|0| = 0$; $|a| = \sqrt{(a^2)}$.

Theorem 6. $|a + b| \leqq |a| + |b|$ (*triangle inequality*).

Theorem 7. $||a| - |b|| \leqq |a + b|$.

Theorem 8. $|ab| = |a|\,|b|$; $\left|\frac{a}{c}\right| = \frac{|a|}{|c|}$ *for* $c \neq 0$.

Theorem 9. *Let* $k > 0$. *Then the inequality* $|a - b| < k$ *is equivalent to the inequalities* $b - k < a < b + k$. (The number $|a - b|$ is equal to the distance between the points a and b on the numbered scale.)

1.5. Further Inequalities. Means

Theorem 1 (*Hölder's Inequality*). *Let* $a_1, \ldots, a_n$, $b_1, \ldots, b_n$ *be real or complex numbers; let* $q > 1$, $q' = q/(q - 1)$. *Then*

$$\left|\sum_{k=1}^{n} a_k b_k\right| \leqq \sum_{k=1}^{n} |a_k b_k| \leqq \left(\sum_{k=1}^{n} |a_k|^q\right)^{1/q} \left(\sum_{k=1}^{n} |b_k|^{q'}\right)^{1/q'}.$$

Theorem 2 (*Cauchy's Inequality*). *Let* $a_1, \ldots, a_n$, $b_1, \ldots, b_n$ *be real or complex numbers. Then*

$$\left|\sum_{k=1}^{n} a_k b_k\right|^2 \leqq \left(\sum_{k=1}^{n} |a_k|^2\right)\left(\sum_{k=1}^{n} |b_k|^2\right)$$

(see Theorem 1 for $q = 2$).

Theorem 3 (*Minkowski's Inequality*). *Let* $a_1, \ldots, a_n, b_1, \ldots, b_n$ *be real or complex numbers,* $q \geqq 1$. *Then*

$$\left(\sum_{k=1}^{n} |a_k + b_k|^q\right)^{1/q} \leqq \left(\sum_{k=1}^{n} |a_k|^q\right)^{1/q} + \left(\sum_{k=1}^{n} |b_k|^q\right)^{1/q}.$$

Definition 1. The number $\frac{1}{n}(a_1 + \ldots + a_n)$ is called the *arithmetic mean* of the numbers $a_1, \ldots, a_n$. If these numbers are non-negative, then the number $\sqrt[n]{(a_1 a_2 \ldots a_n)}$ is said to be the geometric mean and the number $\sqrt{\left[\frac{1}{n}(a_1^2 + \ldots + a_n^2)\right]}$ the *quadratic mean* or *root-mean-square* (r.m.s.) of the numbers $a_1, \ldots, a_n$.

Theorem 4. *If* $a_1 \geqq 0, \ldots, a_n \geqq 0$, *then*

$$\sqrt[n]{(a_1 a_2 \ldots a_n)} \leqq \frac{a_1 + \ldots + a_n}{n} \leqq \sqrt{\frac{a_1^2 + \ldots + a_n^2}{n}}.$$

1.6. Complex Numbers

Complex numbers are numbers of the form $\alpha = a + \mathrm{i}b$, where a, b are real numbers and i is the so-called *imaginary unit* (in electrical engineering j is often used instead of i) which is such that

$$\mathrm{i}^2 = -1, \quad \mathrm{i}^3 = -\mathrm{i}, \quad \mathrm{i}^4 = 1.$$

Definition 1. The *equality* of two complex numbers α_1, α_2 is defined as follows: The number $\alpha_1 = a_1 + \mathrm{i}b_1$ is equal to $\alpha_2 = a_2 + \mathrm{i}b_2$ if and only if $a_1 = a_2, b_1 = b_2$.

Definition 2. *Addition* and *multiplication* of complex numbers are defined in the following way:

$$(a_1 + \mathrm{i}b_1) + (a_2 + \mathrm{i}b_2) = (a_1 + a_2) + \mathrm{i}(b_1 + b_2),$$
$$(a_1 + \mathrm{i}b_1)(a_2 + \mathrm{i}b_2) = (a_1 a_2 - b_1 b_2) + \mathrm{i}(a_1 b_2 + a_2 b_1),$$

respectively.

Theorem 1. *Addition and multiplication of complex numbers satisfy the rules* 1–9 *of Theorem* 1.2.3 (p. 4). *Complex numbers cannot be ordered in such a way that the rules* A–E *of Theorem* 1.2.1 (p. 3) *hold.*

Division of complex numbers is performed by application of the following theorem:

Theorem 2. *If*

$$\alpha = a + \mathrm{i}b \neq 0,$$

then

$$\frac{1}{\alpha} = \alpha^{-1} = \frac{a - ib}{a^2 + b^2}.$$

Definition 3. If $\alpha = a + ib$, then the real number a is called the *real part* of the number α (denoted by $\operatorname{Re}\alpha$) and the real number b the *imaginary part* of the number α (denoted by $\operatorname{Im}\alpha$). A number $\alpha = a + ib$ is said to be *pure imaginary* if $a = \operatorname{Re}\alpha = 0$ and $\operatorname{Im}\alpha \neq 0$.

REMARK 1. Some authors use the symbols $\mathrm{R}[\alpha]$ and $\mathrm{I}[\alpha]$ or script letters $\mathscr{R}(\alpha)$, $\mathscr{I}(\alpha)$, instead of $\operatorname{Re}\alpha$ and $\operatorname{Im}\alpha$, respectively.

Definition 4. The number $a - ib$ is called the *complex conjugate* of the number $\alpha = a + ib$ and is denoted by $\bar{\alpha}$.

Theorem 3. *For conjugates of complex numbers the following relations hold:*

$$\overline{\alpha + \beta} = \bar{\alpha} + \bar{\beta}, \quad \overline{\alpha\beta} = \bar{\alpha}\bar{\beta}, \quad \overline{\left(\frac{\alpha}{\gamma}\right)} = \frac{\bar{\alpha}}{\bar{\gamma}}$$

for $\gamma \neq 0$. *Further,* $\alpha = \bar{\alpha} \Leftrightarrow \alpha$ *is a real number.*

Definition 5. The *absolute value* (*modulus*) of a complex number $\alpha = a + ib$ is defined to be the real number $|\alpha| = \sqrt{(a^2 + b^2)} \geqq 0$.

Theorem 4. *The relations* $|\alpha + \beta| \leqq |\alpha| + |\beta|$, $|\alpha\beta| = |\alpha|\,|\beta|$, $|\alpha| = |\bar{\alpha}|$, $||\alpha| - |\beta|| \leqq$ $\leqq |\alpha - \beta| \leqq |\alpha| + |\beta|$ *hold.*

REMARK 2. *The Geometrical Representation of Complex Numbers* (the *Argand Diagram*) is shown in Fig. 1.3a,b. Fig. 1.3b illustrates the first of the inequalities of Theorem 4 (the so-called *triangle inequality*).

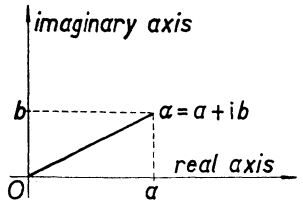


Fig. 1.3a.

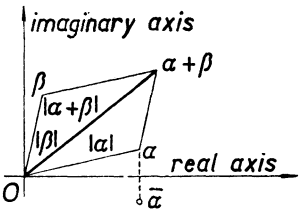


Fig. 1.3b.

Theorem 5 (*Trigonometric Form of Complex Numbers*). *Every complex number* $\alpha = a + ib \neq 0$ *can be written in the form*

$$\alpha = a + ib = r(\cos\varphi + i\sin\varphi) = re^{i\varphi},$$

where $r = |\alpha|$ *and the angle* φ (*in radian measure*) *is determined apart from an*

integral multiple of 2π *by the relations*

$$\cos\varphi = \frac{a}{\sqrt{(a^2 + b^2)}}, \quad \sin\varphi = \frac{b}{\sqrt{(a^2 + b^2)}};$$

this angle φ *is called the argument (amplitude) of the complex number* α.

The *principal value of the argument* of a complex number α (denoted by $\arg\alpha$) is the (uniquely determined) argument φ for which $-\pi < \varphi \leqq \pi$ (Fig. 1.4a).

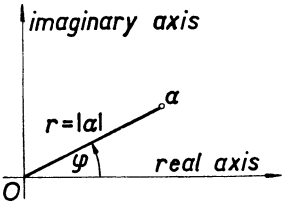


Fig. 1.4a.

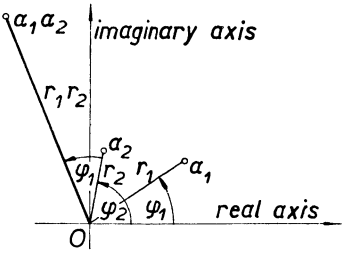


Fig. 1.4b.

Theorem 6 (*De Moivre's Formula*). *If* $\alpha = r(\cos\varphi + \mathrm{i}\sin\varphi) \neq 0$ *is a complex number, then*

$$\alpha^n = [r(\cos\varphi + \mathrm{i}\sin\varphi)]^n = r^n(\cos n\varphi + \mathrm{i}\sin n\varphi)$$

for every integer n; *in particular,*

$$(\cos\varphi + \mathrm{i}\sin\varphi)^n = \cos n\varphi + \mathrm{i}\sin n\varphi\,.$$

Theorem 7. *For* $\alpha_1 = r_1(\cos\varphi_1 + \mathrm{i}\sin\varphi_1)$ *and* $\alpha_2 = r_2(\cos\varphi_2 + \mathrm{i}\sin\varphi_2)$, *the following relation holds:*

$$\alpha_1\alpha_2 = r_1r_2[\cos(\varphi_1 + \varphi_2) + \mathrm{i}\sin(\varphi_1 + \varphi_2)]\,.$$

REMARK 3. Theorems 6 and 7 are used for multiplication, raising to powers and extracting roots of complex numbers. For example, $[\sqrt{(3)} + \mathrm{i}]^3 = [2(\cos 30° + \mathrm{i}\sin 30°)]^3 = 2^3(\cos 90° + \mathrm{i}\sin 90°) = 8\mathrm{i}$. On the use of Theorem 6 for finding roots, see § 1.21 (p. 42). Multiplication of complex numbers in the Argand diagram is performed according to Theorem 7; this can be seen in Fig. 1.4b; the number $|\alpha_1\alpha_2| = r_1r_2$ is usually determined by calculation.

1.7. Powers with Integral Exponents

(a) Powers with a Positive Integral Exponent

REMARK 1. In section (a) m, n denote natural numbers, a, b real or complex numbers.

Definition 1. The n-th *power* of a number a is the number $a^n = aa \dots a$ (n factors a); a is called the *base* and n the *exponent* of the power.

Theorem 1. $a^m a^n = a^{m+n}$, $a^n b^n = (ab)^n$, $(a^m)^n = a^{mn}$.

Theorem 2. $0^n = 0$, $a^n \neq 0$ if $a \neq 0$.

(b) Powers with any Integral Exponent

REMARK 2. In section (b) m, n denote integers, a, b real or complex numbers.

Definition 2. For $a \neq 0$, we define $a^0 = 1$.

If $a \neq 0$ and m is a negative integer, then we define $a^m = 1/a^{-m}$. The symbol a^n is thus (together with Definition 1) defined for every $a \neq 0$ and every *integral* value of n.

Theorem 3. *If $a \neq 0$, $b \neq 0$ and m, n are integers, then* $a^m a^n = a^{m+n}$, $a^n b^n = (ab)^n$, $(a^m)^n = a^{mn}$, and

$$\frac{a^n}{a^m} = a^{n-m}, \quad \frac{a^n}{b^n} = \left(\frac{a}{b}\right)^n.$$

1.8. Roots of Real Numbers

Definition 1. Let $a > 0$ be a real number, n a natural number. Then there exists exactly one *positive real* number x such that $x^n = a$. The number x is called the *n-th root* of a (denoted by $\sqrt[n]{a}$). Instead of $\sqrt[2]{a}$ we write $\sqrt{a}$.

Example 1. $\sqrt{4} = 2$; the statements $\sqrt{4} = -2$ or $\sqrt{4} = \pm 2$ are not correct.

Definition 2. For $a = 0$, we define $\sqrt[n]{0} = 0$.

Definition 3. For $a < 0$ and an *odd* n we define $\sqrt[n]{a} = -\sqrt[n]{(-a)}$ (since, $-\sqrt[n]{(-a)}$ is the only real number, whose n-th power is a). Thus, e.g. $\sqrt[3]{(-8)} = -\sqrt[3]{8} = -2$.

Theorem 1. *Let x, y be positive numbers, m, n be natural numbers. Then*

$$\sqrt[n]{(xy)} = \sqrt[n]{x}\sqrt[n]{y}, \quad \sqrt[n]{\frac{x}{y}} = \frac{\sqrt[n]{x}}{\sqrt[n]{y}}, \quad \sqrt[n]{x^k} = (\sqrt[n]{x})^k$$

(k being an integer),

$$\sqrt[m]{(\sqrt[n]{x})} = \sqrt[mn]{x}, \quad (\sqrt[n]{x})^n = x.$$

Theorem 2. *For any real number x and any even number n, we have* $\sqrt[n]{x^n} = |x|$. Thus, e.g. $\sqrt{x^2} = |x|$, but in general $\sqrt{x^2} \neq x$.

REMARK 1. On roots of complex numbers see § 1.21, p. 42.

1.9. General Powers of Real Numbers

(a) Power with a Rational Exponent

Definition 1. Let $x > 0$ be a real number, r a rational number. Then we define $x^r = \sqrt[q]{x^p}$, where p and q are integers such that $q > 0$ and $r = p/q$. Thus, if n is a natural number, then

$$x^{1/n} = \sqrt[n]{x}, \quad x^{-1/n} = \sqrt[n]{(1/x)} = 1/\sqrt[n]{x}.$$

REMARK 1. The rules for operations with powers with a rational exponent are the same as those in Theorem 1 in the next section (b).

(b) General Powers

Definition 2. For a *positive* real number x and for an arbitrary real number a, the *general power* x^a is defined as the limit of a sequence (see Definition 10.1.2, p. 336) $\{x^{a_n}\}$, where $\{a_n\}$ is an (arbitrary) sequence of rational numbers a_n such that its limit is the number a. (If, in particular, a is rational, then this definition evidently coincides with that of Definition 1 and thus x^a is the same real number according to both definitions.)

Theorem 1 (*Properties of General Powers*). *Let x, y be positive real numbers and a, b real numbers. Then the following rules hold*:

1. $1^a = 1$,
2. $x^a y^a = (xy)^a$; $\dfrac{x^a}{y^a} = \left(\dfrac{x}{y}\right)^a$; $\dfrac{1}{x^a} = \left(\dfrac{1}{x}\right)^a$;
3. $x^a x^b = x^{a+b}$; $\dfrac{x^a}{x^b} = x^{a-b}$; $x^{-a} = \dfrac{1}{x^a}$;
4. $(x^a)^b = x^{ab}$.

Theorem 2 (*Inequalities*). *Let x, y be positive real numbers and a, b real numbers. Then*

1. $x^a > 0$, $x^0 = 1$;
2. $x < y$, $a > 0 \Rightarrow x^a < y^a$;
3. $x < y$, $a < 0 \Rightarrow x^a > y^a$;
4. $x > 1$, $a < b \Rightarrow x^a < x^b$;
5. $x < 1$, $a < b \Rightarrow x^a > x^b$.

Definition 3. For $a > 0$, we define $0^a = 0$.

1.10. Logarithms

(a) The Concept and Properties of Logarithms

Definition 1. Let x, a be *positive* numbers, $a \neq 1$. Then there exists a unique real number y such that $a^y = x$; the number y is called the *logarithm of the number x to the base a* (in symbols, $\log_a x$). Thus the logarithm of a number $x > 0$ to a base a is the number $y = \log_a x$ for which $a^{\log_a x} = x$.

Theorem 1 (*Properties of Logarithms*). *Let a, b, c, x, y be real numbers, $0 < a \neq 1$, $0 < b \neq 1$, $x > 0$, $y > 0$. Then*

1. $a^{\log_a x} = x$;
2. $\log_a a = 1$, $\log_a 1 = 0$;
3. $\log_a xy = \log_a x + \log_a y$, $\log_a (x/y) = \log_a x - \log_a y$;

   $\log_a \frac{1}{x} = -\log_a x$; $\log_a x^c = c \log_a x$;
4. $\log_b x = \frac{\log_a x}{\log_a b}$; *in particular*

   $\log_{10} x \approx 0{\cdot}434\,294 \ln x$,
   $\ln x \approx 2{\cdot}302\,585 \log_{10} x$ (cf. Definition 2).
5. *For $a > 1$ and $x < y$*, $\log_a x < \log_a y$;
   for $a < 1$ and $x < y$, $\log_a x > \log_a y$.
6. *For $a > 1$ and $x > 1$*, $\log_a x > 0$;
   for $a > 1$ and $x < 1$, $\log_a x < 0$.

Definition 2. Logarithms to the base $e = 2{\cdot}718\,28\ldots$ [$e = \lim_{n\to\infty}(1 + 1/n)^n$, see Theorem 10.1.11] are called *natural* or *Napierian logarithms*. Instead of $\log_e x$ we usually write $\ln x$. Then $\log_e 10 = \ln 10 \doteq 2{\cdot}302\,585$. Logarithms to the base 10 are called *common* or *Briggs logarithms*. The value of $\log_{10} e$ is approximately 0·434 294.

REMARK 1. The use of logarithms for calculating the product and the quotient of two positive numbers, or for calculating the powers of a positive number, is apparent from the property 3. The practical procedure is described in every table of logarithms.

REMARK 2. In the following simple examples, methods of solutions of some exponential and logarithmic equations will be shown.

(b) Exponential Equations

Example 1. Solve the equation $2^{4x} \cdot 2^{x^2} = \frac{1}{16}$.

We arrange the equation in the form $2^{4x+x^2} = 2^{-4}$ and deduce (comparing the exponents) that $4x + x^2 = -4$; the problem is thus reduced to the solution of a quadratic equation (the solution is $x = -2$).

Example 2. Solve the equation $2^x = 3^{x-2} \cdot 5^x$. Taking the logarithm of each term we get $x \log_{10} 2 = (x - 2)\log_{10} 3 + x \log_{10} 5$. Hence

$$x = \frac{-2 \log_{10} 3}{\log_{10} 2 - \log_{10} 3 - \log_{10} 5}.$$

(c) Logarithmic Equations

Example 3. Solve the equation

$$[\log_{10}(x^2 + 2)]^2 - 5 \log_{10}(x^2 + 2) + 6 = 0.$$

We put $y = \log_{10}(x^2 + 2)$ and solve the equation $y^2 - 5y + 6 = 0$; this equation has two roots $y_1 = 2$, $y_2 = 3$. Thus the solution consists of those x for which either $\log_{10}(x^2 + 2) = 2$, i.e. $x^2 + 2 = 10^2$, or $\log_{10}(x^2 + 2) = 3$, i.e. $x^2 + 2 = 10^3$, that is $x = \pm\sqrt{98}$ or $x = \pm\sqrt{998}$.

Example 4. Solve the equation

$$2 \log_{10}(2x + 3) - \log_{10}(x - 2) - 1 = 0.$$

We arrange the equation in the form

$$\log_{10}(2x+3)^2 - \log_{10}(x-2) = 1\,, \quad \log_{10}\frac{(2x+3)^2}{x-2} = 1 = \log_{10} 10\,.$$

Hence $(2x+3)^2/(x-2) = 10$; the problem has thus been reduced to the solution of a quadratic equation.

In more complicated cases, numerical methods are employed (see Chap. 31).

1.11. Arithmetic and Geometric Sequences. Sums of Powers of Natural Numbers; Formulae for $a^n \pm b^n$

Definition 1. An *arithmetic sequence* is a sequence (see Definition 10.1.1, p. 336) of real or complex numbers $a_1, a_2, a_3, \ldots, a_n, \ldots$, such that $a_2 - a_1 = a_3 - a_2 = \ldots = a_{n+1} - a_n = d$ $(n = 1, 2, \ldots)$.

Theorem 1. *The relations*

$$a_n = a_1 + (n-1)\,d\,; \quad s_n = \tfrac{1}{2}n(a_1 + a_n)$$

hold, where $s_n = \sum_{i=1}^{n} a_i$ *is the sum of the first n terms.*

Definition 2. *A geometric sequence* is a sequence of real or complex numbers $b_1, b_2, \ldots, b_n, \ldots$, such that there exists a number q with the property that the relations $b_2 = b_1 q$, $b_3 = b_2 q, \ldots, b_{n+1} = b_n q$ $(n = 1, 2, \ldots)$ hold.

Theorem 2. *For* $q \neq 1$, *the relations*

$$b_n = b_1 q^{n-1}\,, \quad S_n = b_1(q^n - 1)/(q-1)$$

hold, where $S_n = \sum_{i=1}^{n} b_i$.

Theorem 3. *Sums of powers of natural numbers.*

1. $1 + 2 + \ldots + n = \dfrac{n(n+1)}{2}$;

2. $1^2 + 2^2 + \ldots + n^2 = \dfrac{n(n+1)(2n+1)}{6}$;

3. $1^3 + 2^3 + \ldots + n^3 = \dfrac{n^2(n+1)^2}{4}$;

4. $1^4 + 2^4 + \ldots + n^4 = \dfrac{n(n+1)(2n+1)(3n^2+3n-1)}{30}$;

5. $1^2 + 3^2 + 5^2 + \ldots + (2n-1)^2 = \dfrac{n(4n^2-1)}{3}$;

6. $1^3 + 3^3 + 5^3 + \ldots + (2n-1)^3 = n^2(2n^2-1)$.

Theorem 4. *Formulae for* $a^n \pm b^n$ (n, k *being natural numbers*):

1. $a^2 - b^2 = (a+b)(a-b)$, $a^2 + b^2 = (a+ib)(a-ib)$;
2. $a^3 \pm b^3 = (a \pm b)(a^2 \mp ab + b^2)$;
3. $a^n - b^n = (a-b)(a^{n-1} + a^{n-2}b + a^{n-3}b^2 + \ldots + ab^{n-2} + b^{n-1})$;
4. $a^{2k} - b^{2k} = (a+b)(a^{2k-1} - a^{2k-2}b + a^{2k-3}b^2 - \ldots - b^{2k-1})$;
5. $a^{2k+1} + b^{2k+1} = (a+b)(a^{2k} - a^{2k-1}b + a^{2k-2}b^2 - \ldots + b^{2k})$.

1.12. Permutations and Combinations

Definition 1. Every ordered n-tuple formed from n given mutually different elements is called a *permutation* of these elements.

Example 1. The permutations of three elements a, b, c are the ordered arrangements (a, b, c), (a, c, b), (b, a, c), (b, c, a), (c, a, b), (c, b, a).

Theorem 1. *The number of all (different) permutations of a collection of n elements is*

$$P_n = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot (n-1)\, n = n!$$

REMARK 1. The symbol $n!$ is read *factorial n*. For $n = 0$, $0!$ is defined as having the value 1.

Definition 2. Let $(i_1, i_2, \ldots, i_n)$ be a permutation of the numbers $1, 2, 3, \ldots, n$. We say that the numbers i_j, i_k, when $j < k$ $(1 \leqq j \leqq n,\ 1 \leqq k \leqq n)$ form an *inversion* in this permutation if $i_j > i_k$. A permutation possessing an odd, or even number of inversions is called *odd*, or *even*, respectively.

Example 2.

(a) In the permutation $(2, 4, 1, 3)$ of the numbers $1, 2, 3, 4$ each of the pairs $(2, 1)$, $(4, 1)$, $(4, 3)$ is an inversion. The permutation possesses 3 inversions; therefore it is odd.

(b) The permutation $(4, 2, 1, 3)$ possesses 4 inversions and is thus even; the permutation $(1, 2, 3, 4)$ possesses no inversion and is thus even.

Theorem 2. *If, among n elements a, b, c, ..., a occurs* α *times, b* β *times, c* γ *times, ..., then the number of all different ordered n-tuples is*

$$\frac{n!}{\alpha!\,\beta!\,\gamma!\,\ldots}.$$

Definition 3. By *combinations of n different elements taken k at a time* we mean all possible selections consisting of k different elements chosen from the n given elements, without regard to the order of selection.

Theorem 3. *The number of all combinations of n different elements taken k at a time is*

$$\binom{n}{k} = \frac{n(n-1)\dots(n-k+1)}{1 \cdot 2 \dots \dots k} = \frac{n!}{k!\,(n-k)!}.$$

REMARK 2. Besides $\binom{n}{k}$, the symbols $C(n,k)$, C_n^k, ${}_nC_k$, ${}^nC_k$ and C_k^n are also employed.

For $k = 0$, $\binom{n}{0}$ is defined as having the value 1.

Example 3. Find the number of chess matches required if there are 10 players and every player is to play once with each other.

The number of matches P is equal to the number of pairs formed out of 10 elements, i.e. it is equal to the number of combinations of 10 elements taken 2 at a time. Hence $P = \binom{10}{2} = \frac{10 \cdot 9}{2 \cdot 1} = 45$.

Theorem 4. *The symbol* $\binom{n}{k}$ *satisfies the relations*

1. $\binom{n}{k} = \binom{n}{n-k}$;
2. $\binom{n}{1} = \binom{n}{n-1} = n$, $\binom{n}{n} = \binom{n}{0} = 1$;
3. $\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$;
4. $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n-1}{k} + \dots + \binom{k}{k}$.

REMARK 3. By *combinations with repetitions of n different elements taken k at a time* we understand all possible selections consisting of k elements chosen from n given elements (without regard to the order of selection) such that each element can be repeated any number of times. The number of these combinations with repetitions is $\binom{n+k-1}{k}$. For example, all combinations with repetitions of the elements 1, 2, 3 taken 2 at a time are (1, 1), (2, 2), (3, 3), (1, 2), (1, 3), (2, 3).

Definition 4. By *permutations of n different elements taken k at a time* is meant all possible ordered arrangements consisting of k different elements chosen from the n given elements.

Theorem 5. *The number of all permutations of n different elements taken k at a time is*

$$P_{n/k} = n(n-1)\dots(n-k+1) = \frac{n!}{(n-k)!}.$$

Example 4. All permutations of the three elements 1, 2, 3 taken 2 at a time are (1, 2), (1, 3), (2, 1), (2, 3), (3, 1), (3, 2). Their number is $6 = 3!/(3-2)!$.

REMARK 4. The *permutations with repetitions of n different elements taken k at a time* are all possible ordered arrangements consisting of k elements chosen from the n given elements such that each element can be repeated any number of times. The number of these permutations with repetitions is n^k.

For example, all permutations with repetition of the elements 1, 2 taken 2 at a time are (1, 1), (2, 2), (1, 2), (2, 1).

1.13. Binomial Theorem

Theorem 1. *Let n be a natural number and let a, b be real or complex numbers. Then the following (Newton's) formula holds:*

$$(a \pm b)^n = \sum_{k=0}^{n} (\pm 1)^k \binom{n}{k} a^{n-k}b^k = a^n \pm \binom{n}{1} a^{n-1}b + \binom{n}{2} a^{n-2}b^2 + \dots + (\pm 1)^n b^n.$$

In particular

1. $(a \pm b)^2 = a^2 \pm 2ab + b^2$;
2. $(a \pm b)^3 = a^3 \pm 3a^2b + 3ab^2 \pm b^3$.

REMARK 1. The *binomial coefficients* $\binom{n}{k}$ can be readily determined by means of *Pascal's triangle*:

n	Binomial coefficients
0	1
1	1 1
2	1 2 1
3	1 3 3 1
4	1 4 6 4 1
5	1 5 10 10 5 1
	. . . . .

REMARK 2. The case where n is not a natural number is treated in Theorem 15.5.3.

1.14. Polynomials

Definition 1. Let n be a natural number and let $a_0, a_1, \ldots, a_n$ be real or complex numbers. The function $P(x)$ which may be defined for all (real or complex) numbers x by the formula

$$P(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n = \sum_{i=0}^{n} a_ix^{n-i} \tag{1}$$

is called a *polynomial* (in one variable x with real, or complex, coefficients). Besides the term polynomial the expression *rational integral function* is also used. The numbers $a_0, a_1, \ldots, a_n$ are called the *coefficients of the polynomial* $P(x)$.

Definition 2. Two polynomials $P(x)$ and $Q(x)$ are *equal* [in symbols, $P(x) = Q(x)$ or, more precisely, $P(x) \equiv Q(x)$] if, for every number a, the equality $P(a) = Q(a)$ holds.

Definition 3. The highest power of the variable x with a non-zero coefficient in the expression (1) is called the *degree of the polynomial* $P(x)$. If $a_0 \neq 0$ in (1), then $P(x)$ has degree n. (See also Theorem 1.)

Definition 4. The polynomial, all the coefficients of which are equal to zero, is called a *zero polynomial.* A zero polynomial has no degree. If $P(x)$ is a zero polynomial, we write $P(x) = 0$ or, more precisely, $P(x) \equiv 0$. Otherwise, we write $P(x) \neq 0$ or $P(x) \not\equiv 0$.

Theorem 1. *Two polynomials are equal if and only if their difference is a zero polynomial,* i.e. *if the coefficients of the corresponding powers of the variable* x *are identical.*

Theorem 2. *The sum and the difference of two polynomials of degrees* m *and* n *are polynomials of degree less than or equal to the number* $\max(m, n)$ *(or zero polynomials).*

Theorem 3. *The product of two polynomials of degrees* m *and* n *is a polynomial of degree* $m + n$.

Theorem 4. *The product of non-zero polynomials is a non-zero polynomial.*

Theorem 5. *The quotient of two polynomials need not always be a polynomial.*

How to procced in dividing a polynomial by another polynomial is shown in Example 1.

Example 1.

$$\begin{array}{l} (x^3 - 2x^2 + x - 1) : (x^2 - 3x + 2) = x + 1 \text{ (partial quotient)} \\ \underline{x^3 - 3x^2 + 2x} \\ \quad x^2 - x - 1 \\ \quad \underline{x^2 - 3x + 2} \\ \quad 2x - 3 \text{ (remainder)}. \end{array}$$

Hence, $x^3 - 2x^2 + x - 1 = (x^2 - 3x + 2)(x + 1) + 2x - 3$.

Definition 5. If the remainder on dividing a polynomial $P(x)$ by a polynomial $Q(x)$ ($Q(x) \not\equiv 0$) equals zero, then the polynomial $P(x)$ is said to be *divisible* by the polynomial $Q(x)$; $Q(x)$ is called a *divisor* of the polynomial $P(x)$.

An important result concerning the process of dividing is contained in

Theorem 6. *For every two polynomials $P(x)$ and $Q(x) \not\equiv 0$, there exist uniquely determined polynomials $S(x)$ and $R(x)$, such that*

1. $P(x) = Q(x)\,S(x) + R(x)$;
2. *$R(x)$ is either a zero polynomial or a polynomial of lower degree than the polynomial $Q(x)$.*

Definition 6. A common divisor, with highest possible degree, of the polynomials $P(x)$ and $Q(x)$ is called the *greatest common divisor* of the polynomials $P(x)$ and $Q(x)$; the polynomials $P(x)$ and $Q(x)$ are said to be *relatively prime* if their greatest common divisor has degree zero.

Theorem 7 (*Euclidean Algorithm*). *The greatest common divisor of two (non-zero) polynomials $P(x)$ and $Q(x)$ can be found in the following way:*

(i) *in accordance with Theorem 6, divide $P(x)$ by the polynomial $Q(x)$*, i.e. $P(x) = Q(x)\,S_1(x) + R_1(x)$ (*where $R_1(x)$ is the remainder*);

(ii) *divide $Q(x)$ by the polynomial $R_1(x)$*, i.e. $Q(x) = R_1(x)\,S_2(x) + R_2(x)$, *then $R_1(x)$ by the polynomial $R_2(x)$*, i.e. $R_1(x) = R_2(x)\,S_3(x) + R_3(x)$ *etc., the last remainder $R_k(x) \neq 0$ is the required greatest common divisor.*

Definition 7. The number α (in general complex) is called a *zero* of the polynomial $P(x) = \sum_{i=0}^{n} a_i x^{n-1}$ (or a *root* of this polynomial) if $P(\alpha) = \sum_{i=0}^{n} a_i \alpha^{n-i} = 0$.

Theorem 8 (*The Fundamental Theorem of Algebra*). *Every polynomial of degree $n \geqq 1$ has at least one zero.*

Theorem 9. *If a polynomial $P(x)$ has a zero α, then $P(x)$ is divisible by the linear polynomial $x - \alpha$ and vice versa.* [$x - \alpha$ is a so-called *linear factor of the polynomial $P(x)$*.]

Theorem 10 (*The Factorisation of a Polynomial into Linear Factors*). *Every polynomial* $P(x) = \sum_{i=0}^{n} a_i x^{n-i}$, $n \geqq 1$, *can be uniquely written as a product of linear factors*:

$$P(x) = a_0(x - \alpha_1)^{k_1}(x - \alpha_2)^{k_2} \ldots (x - \alpha_r)^{k_r}, \quad k_1 + k_2 + \ldots + k_r = n.$$

The numbers $\alpha_1, \ldots, \alpha_r$ *are all distinct zeros of the polynomial* $P(x)$. [α_1 is called a k_1*-fold zero*, ..., α_r *a* k_r*-fold zero* of the polynomial $P(x)$. *If* $k_j = 1$, α_j is called a *simple zero* of $P(x)$.]

Theorem 11. α *is a k-fold zero of a polynomial* $P(x)$ *if and only if it is also a zero of the first, the second, ..., the* $(k-1)$*-th derivatives of the polynomial* $P(x)$, *but is not a zero of its k-th derivative*:

$$P(\alpha) = P'(\alpha) = \ldots = P^{(k-1)}(\alpha) = 0, \quad P^{(k)}(\alpha) \neq 0.$$

(For the derivative, see § 11.5.)

Example 2. Let us consider the polynomial $P(x) = x^3 - 3x^2 + 4$. We get $P'(x) = 3x^2 - 6x$, $P''(x) = 6x - 6$, $P'''(x) = 6$. It is easy to check that $P(2) = 0$, $P'(2) = 0$, $P''(2) \neq 0$. Thus $\alpha = 2$ is a double zero of the polynomial $P(x)$. Indeed, $P(x) = (x - 2)^2(x + 1)$.

Theorem 12 (*Polynomials with Real Coefficients*).

(i) *If the polynomial* $P(x) = \sum_{i=0}^{n} a_i x^{n-i}$ *with real coefficients* a_i *has a k-fold zero* $\alpha = a + ib$, *it has also the k-fold zero* $\bar{\alpha} = a - ib$.

(ii) *A polynomial* $P(x)$ *can be uniquely factorised into linear and quadratic polynomials with real coefficients*:

$$P(x) = a_0(x - \alpha_1)^{r_1} \ldots (x - \alpha_i)^{r_i}(x^2 + p_1 x + q_1)^{s_1} \ldots (x^2 + p_j x + q_j)^{s_j},$$

where $r_1 + \ldots + r_i + 2s_1 + \ldots + 2s_j = n$ *and* $p_k^2 - 4q_k < 0$ $(k = 1, 2, \ldots, j)$, *so that* $x^2 + p_k x + q_k$ *has no real zeros* (cf. §13.3, p. 457.)

REMARK 1 (*Horner's Method*). Horner's method is used:

(i) to find the value $P(a)$ of a polynomial $P(x)$ and its derivatives at a given point a;
(ii) to divide a polynomial $P(x)$ by a linear polynomial $x - a$;
(iii) to transform a polynomial $P(x)$ by a substitution $y = x - a$.

Let $P(x) = a_0 x^n + \ldots + a_n$; let a be a real or complex number. We then construct the following (Horner's) scheme:

	a_0	a_1	a_2	...	a_n
		aa_0	ab_1	...	ab_{n-1}
a	a_0	b_1	b_2	...	$b_n = P(a)$

Write all coefficients (zero coefficients included) in the first row, leaving for a moment the second row open; in the third row under the number a_0 write a_0 again, then, under a_1, write aa_0 in the second row and $b_1 = a_1 + aa_0$ in the third row. Similarly, under a_2, write ab_1 and $b_2 = a_2 + ab_1$ in the second and third rows respectively, etc. The last number b_n is then the value $P(a)$. Moreover $P(x) = (x - a) \cdot (a_0x^{n-1} + b_1x^{n-2} + \ldots + b_{n-1}) + b_n$ so that the third row determines the quotient and the remainder on dividing the polynomial $P(x)$ by the linear polynomial $x - a$.

Applying Horner's scheme for a number a to the polynomial $a_0x^{n-1} + b_1x^{n-2} + \ldots + b_{n-1}$, we get, as the last number in the scheme, $c_{n-1} = P'(a)/1!$ Continuing in this way, we get $P''(a)/2!$ ..., $P^{(n)}(a)/n!$, successively. The following Example 3 illustrates the procedure.

Example 3. $P(x) = 5x^4 + 10x^3 + x - 1$; $a = -2$.

	5	10	0	1	−1
		−10	0	0	−2
−2	5	0	0	1	$-3 = P(-2)$
		−10	20	−40	
−2	5	−10	20	$-39 = \frac{1}{1!}P'(-2)$	
		−10	40		
−2	5	−20	$60 = \frac{1}{2!}P''(-2)$		
		−10			
−2	5	$-30 = \frac{1}{3!}P'''(-2)$			
−2	$5 = \frac{1}{4!}P^{(4)}(-2)$				

Further, by Taylor's formula for a polynomial of degree n (see § 11.10, p. 396),

$$P(x) = P(a) + \frac{1}{1!}P'(a)(x - a) + \ldots + \frac{1}{n!}P^{(n)}(a)(x - a)^n$$

so that we have in our example

$$P(x) = -3 - 39(x + 2) + 60(x + 2)^2 - 30(x + 2)^3 + 5(x + 2)^4 .$$

By the substitution $y = x + 2$, $P(x)$ is transformed into the polynomial

$$5y^4 - 30y^3 + 60y^2 - 39y - 3 .$$

From the third row of Horner's scheme we get

$$P(x) = (x + 2)(5x^3 + 1) - 3 .$$

1.15. Vectors in Algebra

Definition 1. Let n be a fixed natural number. Then, by an *n-component (n-coordinate) complex vector* (n-vector for short) $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$ we understand in algebra an ordered n-tuple of complex numbers $a_1, a_2, \ldots, a_n$. [Besides $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$ also the notation $\boldsymbol{a}(a_1, a_2, \ldots, a_n)$ is used.] All these n-component vectors (i.e. the set of all ordered n-tuples of complex numbers) form a so-called *n-dimensional vector space* V_n (over the complex numbers).

The vectors $\boldsymbol{a} = (a_1, \ldots, a_n)$, $\boldsymbol{b} = (b_1, \ldots, b_n)$ are said to be *equal* if and only if $a_1 = b_1, a_2 = b_2, \ldots, a_n = b_n$.

REMARK 1. In the same way as for complex vectors one can define *n-component real vectors* (real n-vectors); their *components*, i.e. the numbers $a_1, \ldots, a_n$ being real numbers. In the following text all concepts and theorems formulated for complex vectors are valid also for real vectors.

Addition and multiplication by a (scalar) number of n-component vectors of V_n is performed in accordance with the following definition:

Definition 2. 1. The *sum of the vectors* $\boldsymbol{a} = (a_1, \ldots, a_n)$ and $\boldsymbol{b} = (b_1, \ldots, b_n)$ is the vector $\boldsymbol{a} + \boldsymbol{b} = (a_1 + b_1, \ldots, a_n + b_n)$.

2. The *product of the vector* $\boldsymbol{a} = (a_1, \ldots, a_n)$ *and the number* c is the vector $c\boldsymbol{a} = (ca_1, \ldots, ca_n)$.

REMARK 2. We write $-\boldsymbol{a}$ instead of $(-1)\boldsymbol{a}$; thus $-\boldsymbol{a} = (-a_1, \ldots, -a_n)$.

Example 1. The sum of the vectors $\boldsymbol{a} = (1, 0, -2)$ and $\boldsymbol{b} = (3, 2, 0)$ is $\boldsymbol{a} + \boldsymbol{b} = (4, 2, -2)$; also $3\boldsymbol{a} = (3, 0, -6)$.

Theorem 1. *For the operations on vectors introduced in Definition 2, the following rules hold*:

1. $\boldsymbol{a} + \boldsymbol{b} = \boldsymbol{b} + \boldsymbol{a}$, $\boldsymbol{a} + (\boldsymbol{b} + \boldsymbol{c}) = (\boldsymbol{a} + \boldsymbol{b}) + \boldsymbol{c}$;
2. *there exists a vector* [*the so-called zero vector* $\boldsymbol{0} = (0, \ldots, 0)$] *such that* $\boldsymbol{a} + \boldsymbol{0} = \boldsymbol{a}$;
3. *for every* $\boldsymbol{a} = (a_1, \ldots, a_n)$ *and* $\boldsymbol{b} = (b_1, \ldots, b_n)$ *there exists a vector* $\boldsymbol{x}$ such *that* $\boldsymbol{a} + \boldsymbol{x} = \boldsymbol{b}$; $\boldsymbol{x} = \boldsymbol{b} - \boldsymbol{a} = (b_1 - a_1, \ldots, b_n - a_n)$;
4. $c(\boldsymbol{a} + \boldsymbol{b}) = c\boldsymbol{a} + c\boldsymbol{b}$;
5. $(c + d)\boldsymbol{a} = c\boldsymbol{a} + d\boldsymbol{a}$;
6. $c(d\boldsymbol{a}) = (cd)\boldsymbol{a}$; $0\boldsymbol{a} = \boldsymbol{0}$, $c\boldsymbol{0} = \boldsymbol{0}$;
7. *the equality* $c\boldsymbol{a} = \boldsymbol{0}$ *holds if and only if* $c = 0$ *or* $\boldsymbol{a} = \boldsymbol{0}$;
8. $-(c\boldsymbol{a}) = (-c)\boldsymbol{a} = c(-\boldsymbol{a})$.

Definition 3. We say that the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k$ of V_n are *linearly dependent* if there exist complex numbers $c_1, \ldots, c_k$, *which are not all zero*, such that $c_1\boldsymbol{a}_1 + c_2\boldsymbol{a}_2 + \ldots + c_k\boldsymbol{a}_k = \boldsymbol{0}$.

If the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k$ are not linearly dependent, we say that they are *linearly independent.*

Example 2. The vectors $\boldsymbol{a} = (1, -1, 0), \boldsymbol{b} = (0, -2, 1), \boldsymbol{c} = (2, 4, -3)$ are linearly dependent, for $2\boldsymbol{a} + (-3)\boldsymbol{b} + (-1)\boldsymbol{c} = \boldsymbol{0}$. The vectors $\mathbf{e}_1 = (1, 0, 0), \mathbf{e}_2 = (0, 1, 0), \mathbf{e}_3 = (0, 0, 1)$ are linearly independent.

Definition 4. A vector $\boldsymbol{a} \in V_n$ is said to be a *linear combination of the vectors* $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k$ of V_n if complex numbers $d_1, \ldots, d_k$ exist such that

$$\boldsymbol{a} = d_1\boldsymbol{a}_1 + \ldots + d_k\boldsymbol{a}_k .$$

Theorem 2. *Vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p$ of V_n are linearly dependent if and only if at least one of them can be expressed as a linear combination of the others.*

Example 3. The vectors $\boldsymbol{a}_1 = (3, 1, 2), \boldsymbol{a}_2 = (-1, 0, 2), \boldsymbol{a}_3 = (7, 2, 2)$ are linearly dependent, for $2\boldsymbol{a}_1 - \boldsymbol{a}_2 - \boldsymbol{a}_3 = 0$. From this equation it follows that

$$\boldsymbol{a}_1 = \tfrac{1}{2}\boldsymbol{a}_2 + \tfrac{1}{2}\boldsymbol{a}_3 , \quad \boldsymbol{a}_2 = 2\boldsymbol{a}_1 - \boldsymbol{a}_3 , \quad \boldsymbol{a}_3 = 2\boldsymbol{a}_1 - \boldsymbol{a}_2$$

so that each of them is a linear combination of the other two.

Definition 5. We say that a system $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k\}$ of vectors of V_n has the *rank* h if there are h linearly independent vectors among the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k$ but any $h+1$ vectors of $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k$ are always linearly dependent. (Then h is the maximal number of linearly independent vectors of the given system.)

Example 4. The rank of the system $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$ of the vectors of Example 2 is equal to two, for $\boldsymbol{a}, \boldsymbol{b}$ are linearly independent while $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ are linearly dependent.

Theorem 3. *Every system of n-component vectors is of rank $h \leqq n$.*

Theorem 4. *The rank of a system of n-component vectors does not change if*

1. *we change the order of the vectors in the system;*
2. *we multiply one of the vectors of the system by a non-zero number;*
3. *we add to one of the vectors a linear combination of the remaining vectors;*
4. *we drop a vector which is a linear combination of the remaining vectors of the system.*

REMARK 3. Theorem 4 is useful in determining the rank of a given system of vectors. In practice, we can find the rank also by determining the rank of the matrix whose rows are the vectors of the given system (see Remark 1.16.2 and Example 1.16.2 on p. 27).

REMARK 4. On vectors in *three-dimensional* space (scalar product, vector product, etc.) see also Chap. 7.

1.16. Matrices

Definition 1. A rectangular array $\boldsymbol{A}$ of mn real or complex numbers $a_{11}, a_{12}, \ldots, a_{mn}$ arranged in m rows and n columns is called an *m by n matrix*:

$$\boldsymbol{A} = \begin{bmatrix} a_{11}, & a_{12}, & a_{13}, & \ldots, & a_{1n} \\ a_{21}, & a_{22}, & a_{23}, & \ldots, & a_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}, & a_{m2}, & a_{m3}, & \ldots, & a_{mn} \end{bmatrix}$$

If $m = n$, we call $\boldsymbol{A}$ a *square matrix of order n*, or an *n-rowed square matrix*. The elements $a_{11}, a_{22}, a_{33}, \ldots$ of the matrix $\boldsymbol{A}$ form its *principal diagonal*, the elements $a_{1n}, a_{2,n-1}, a_{3,n-2}, \ldots$ of $\boldsymbol{A}$ form its *secondary diagonal*. The matrix, all the elements of which are equal to zero, is called a *zero matrix*.

Definition 2. The *rank of a matrix* is the rank of the system of all vectors formed by the rows of the matrix (see Definition 1.15.5, p. 25). (Cf. Theorem 2.)

Thus, a matrix $\boldsymbol{A}$ is of rank h if there are h linearly independent rows among its rows, every futher row of the matrix being a linear combination of these h rows.

Example 1. The matrix

$$\begin{bmatrix} 1, & -1, & 0 \\ 0, & -2, & 1 \\ 2, & 4, & -3 \end{bmatrix}$$

is of rank 2, for the system of the vectors $\boldsymbol{a} = (1, -1, 0)$, $\boldsymbol{b} = (0, -2, 1)$, $\boldsymbol{c} = (2, 4, -3)$ is of rank 2 (Example 1.15.4, p. 25).

Theorem 1. *For the rank h of an m by n matrix $\boldsymbol{A}$, the inequality*

$$h \leqq \min(m, n)$$

holds.

Definition 3. The matrix

$$\boldsymbol{A}' = \begin{bmatrix} a_{11}, & a_{21}, & \ldots, & a_{m1} \\ a_{12}, & a_{22}, & \ldots, & a_{m2} \\ \ldots\ldots\ldots\ldots\ldots \\ a_{1n}, & a_{2n}, & \ldots, & a_{mn} \end{bmatrix}$$

formed from the matrix $\boldsymbol{A}$ by a transposition of its elements with respect to the principal diagonal (i.e. by an interchange of its rows and columns) is called the *transpose of the matrix* $\boldsymbol{A}$ and is an n by m matrix. The notation $\boldsymbol{A}^{\mathrm{T}}$ is also used.

Theorem 2. *The rank of a matrix $\boldsymbol{A}$ and that of its transpose $\boldsymbol{A}'$ are equal.*

Theorem 3. *The rank of a matrix does not change if*

1. *we change the order of the rows of the matrix;*
2. *we multiply one of the rows by a non-zero number;*
3. *we add to one of the rows a linear combination of the remaining rows;*
4. *we drop a row of the matrix which is a linear combination of the remaining rows of the matrix.*

Thus, if we apply one of these operations, to a matrix, then the resulting matrix has the same rank as the original matrix.

REMARK 1. According to Theorem 2 we can apply the operations of Theorem 3 also to the columns without affecting the rank of the given matrix.

Theorem 4. *The matrix*

$$\boldsymbol{B} = \begin{bmatrix} b_{11}, & b_{12}, & b_{13}, & \ldots, & b_{1n} \\ 0, & b_{22}, & b_{23}, & \ldots, & b_{2n} \\ 0, & 0, & b_{33}, & \ldots, & b_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0, & 0, & \ldots, & 0, b_{kk}, & \ldots, b_{kn} \end{bmatrix}, \tag{1}$$

where $b_{11}b_{22} \ldots b_{kk} \neq 0$ *and where all elements below the principal diagonal are equal to zero, is of rank k.*

REMARK 2. Theorems 3 and 4 can be used in practice to determine the rank of a given matrix: By means of the operations 1–4 of Theorem 3 and by permutation of the columns we transform the given matrix to a matrix of the same rank and of the form (1) and then apply Theorem 4.

Example 2.

$$\boldsymbol{A} = \begin{bmatrix} 1, & 0, & 2, & 3 \\ -2, & 1, & 0, & -1 \\ -1, & 1, & 2, & 2 \\ -1, & 2, & 6, & 7 \end{bmatrix}.$$

The third row is the sum of the first and second rows; if we drop it, we get the matrix

$$\boldsymbol{A}_1 = \begin{bmatrix} 1, & 0, & 2, & 3 \\ -2, & 1, & 0, & -1 \\ -1, & 2, & 6, & 7 \end{bmatrix}.$$

Applying operations 2 and 3 we can get a matrix in which all elements of the first column of the matrix except the first are zero: First, we add twice the first row to

the second row and then we add the first to the third row. We thus obtain the matrix

$$\boldsymbol{A}_2 = \begin{bmatrix} 1, & 0, & 2, & 3 \\ 0, & 1, & 4, & 5 \\ 0, & 2, & 8, & 10 \end{bmatrix}.$$

Now we adjust the second column so as to get zero below the second element: We subtract twice the second row from the third row and thus obtain

$$\boldsymbol{A}_3 = \begin{bmatrix} 1, & 0, & 2, & 3 \\ 0, & 1, & 4, & 5 \\ 0, & 0, & 0, & 0 \end{bmatrix}.$$

In accordance with Theorem 3 we can drop the last row of this matrix. We get a matrix the rank of which is 2, according to Theorem 4. Hence the rank of $\boldsymbol{A}$ is also 2.

Definition 4. The determinant of order k (see Definition 1.17.1, p. 29) formed by the elements in the intersections of arbitrary k rows and k columns of a matrix

$$\boldsymbol{A} = \begin{bmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1}, & a_{m2}, & \ldots, & a_{mn} \end{bmatrix},$$

is called a *minor of order k of the matrix* $\boldsymbol{A}$ $[1 \leqq k \leqq \min(m, n)]$.

Theorem 5. *A matrix* $\boldsymbol{A}$ *is of rank h if and only if there exists a minor of* $\boldsymbol{A}$ *of order h different from zero, any minor of* $\boldsymbol{A}$ *of order higher than h being equal to zero.*

Example 3. Consider the matrix

$$\boldsymbol{A} = \begin{bmatrix} 1, & 0, & 2, & 3 \\ -2, & 1, & 0, & -1 \\ -1, & 2, & 6, & 7 \end{bmatrix}.$$

All its minors of order 3, namely

$$\begin{vmatrix} 1, & 0, & 2 \\ -2, & 1, & 0 \\ -1, & 2, & 6 \end{vmatrix}, \quad \begin{vmatrix} 1, & 0, & 3 \\ -2, & 1, & -1 \\ -1, & 2, & 7 \end{vmatrix}, \quad \begin{vmatrix} 1, & 2, & 3 \\ -2, & 0, & -1 \\ -1, & 6, & 7 \end{vmatrix}, \quad \begin{vmatrix} 0, & 2, & 3 \\ 1, & 0, & -1 \\ 2, & 6, & 7 \end{vmatrix}$$

are equal to zero, while the minor of order 2,

$$\begin{vmatrix} 1, & 0 \\ -2, & 1 \end{vmatrix} = 1 \neq 0.$$

Hence the rank of $\boldsymbol{A}$ is 2, in accordance with Example 2.

REMARK 3. For further results on matrices see § 1.25, p. 49.

1.17. Determinants

Definition 1. The *determinant of order n* of a square matrix

$$\boldsymbol{A} = \begin{bmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ a_{21}, & a_{22}, & \ldots, & a_{2n} \\ \ldots\ldots\ldots\ldots\ldots & & & \\ a_{n1}, & a_{n2}, & \ldots, & a_{nn} \end{bmatrix}$$

is defined as the number

$$A = \sum(-1)^r a_{1k_1} a_{2k_2} \ldots a_{nk_n},$$

where the symbol $\sum$ indicates the sum of all terms for all possible permutations $(k_1, k_2, \ldots, k_n)$ of the numbers $1, 2, \ldots, n$, the integer r being the number of inversions (Definition 1.12.2, p. 17) in the permutation $(k_1, k_2, \ldots, k_n)$; we write

$$A = \begin{vmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ a_{21}, & a_{22}, & \ldots, & a_{2n} \\ \ldots\ldots\ldots\ldots\ldots & & & \\ a_{n1}, & a_{n2}, & \ldots, & a_{nn} \end{vmatrix}. \tag{1}$$

Example 1.

$$\begin{vmatrix} a_{11}, & a_{12} \\ a_{21}, & a_{22} \end{vmatrix} = (-1)^0 a_{11}a_{22} + (-1)^1 a_{12}a_{21} = a_{11}a_{22} - a_{12}a_{21},$$

since the permutations (1, 2), and (2, 1) of the numbers 1, 2 have no inversions and 1 inversion, respectively.

Theorem 1. *The value of a determinant remains unaltered if its columns and rows are interchanged*:

$$\begin{vmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ a_{21}, & a_{22}, & \ldots, & a_{2n} \\ \ldots\ldots\ldots\ldots\ldots & & & \\ a_{n1}, & a_{n2}, & \ldots, & a_{nn} \end{vmatrix} = \begin{vmatrix} a_{11}, & a_{21}, & \ldots, & a_{n1} \\ a_{12}, & a_{22}, & \ldots, & a_{n2} \\ \ldots\ldots\ldots\ldots\ldots & & & \\ a_{1n}, & a_{2n}, & \ldots, & a_{nn} \end{vmatrix}.$$

Hence, all properties of determinants expressed in the following text for rows hold also for columns and vice versa.

Theorem 2. *The value of a determinant is unaltered if, to one of its rows, a linear combination of the remaining rows is added.*

Theorem 3. *If one of the rows is a linear combination of the remaining rows, then the value of the determinant is zero.*

Theorem 4. *The value of the determinant changes its sign if we interchange two of its rows.*

Definition 2. The determinant

$$A_{ij} = \begin{vmatrix} a_{11}, & a_{12}, & \ldots, & a_{1,j-1}, & a_{1,j+1}, & \ldots, & a_{1n} \\ a_{21}, & a_{22}, & \ldots, & a_{2,j-1}, & a_{2,j+1}, & \ldots, & a_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_{i-1,1}, & a_{i-1,2}, & \ldots, & a_{i-1,j-1}, & a_{i-1,j+1}, & \ldots, & a_{i-1,n} \\ a_{i+1,1}, & a_{i+1,2}, & \ldots, & a_{i+1,j-1}, & a_{i+1,j+1}, & \ldots, & a_{i+1,n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_{n1}, & a_{n2}, & \ldots, & a_{n,j-1}, & a_{n,j+1}, & \ldots, & a_{nn} \end{vmatrix}, \tag{2}$$

originating from the determinant A by omitting the i-th row and j-th column is called the *minor of order* $n-1$ *of the determinant* A *belonging to the element* a_{ij}.

The *cofactor* A_{ij} *of the element* a_{ij} *in the determinant* A is defined as the minor A_{ij} equipped with the sign $(-1)^{i+j}$; thus, $\mathrm{A}_{ij} = (-1)^{i+j} A_{ij}$.

Theorem 5 (*The Expansion of a Determinant According to the i-th Row*). *For the determinant* (1) *the following expansion holds*:

$$A = a_{i1}\mathrm{A}_{i1} + a_{i2}\mathrm{A}_{i2} + \ldots + a_{in}\mathrm{A}_{in} =$$
$$= (-1)^{i+1} a_{i1}A_{i1} + (-1)^{i+2} a_{i2}A_{i2} + \ldots + (-1)^{i+n} a_{in}A_{in}.$$

Theorem 6. *For* $i \neq j$,

$$a_{i1}\mathrm{A}_{j1} + a_{i2}\mathrm{A}_{j2} + \ldots + a_{in}\mathrm{A}_{jn} = 0.$$

Theorem 7 (*The Addition Rule*). *The relation*

$$\begin{vmatrix} a_1 + b_1, & a_{12}, & \ldots, & a_{1n} \\ a_2 + b_2, & a_{22}, & \ldots, & a_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_n + b_n, & a_{n2}, & \ldots, & a_{nn} \end{vmatrix} = \begin{vmatrix} a_1, & a_{12}, & \ldots, & a_{1n} \\ a_2, & a_{22}, & \ldots, & a_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_n, & a_{n2}, & \ldots, & a_{nn} \end{vmatrix} + \begin{vmatrix} b_1, & a_{12}, & \ldots, & a_{1n} \\ b_2, & a_{22}, & \ldots, & a_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ b_n, & a_{n2}, & \ldots, & a_{nn} \end{vmatrix}$$

holds, and similarly for other columns.

Theorem 8 (*The Multiplication of a Determinant by a Number*). *The relation*

$$\begin{vmatrix} ca_{11}, & a_{12}, & \ldots, & a_{1n} \\ ca_{21}, & a_{22}, & \ldots, & a_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ ca_{n1}, & a_{n2}, & \ldots, & a_{nn} \end{vmatrix} = c \begin{vmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ a_{21}, & a_{22}, & \ldots, & a_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_{n1}, & a_{n2}, & \ldots, & a_{nn} \end{vmatrix}$$

holds, and similarly for other columns.

In other words: *We multiply a determinant by a number c if we multiply by this number all the elements of a row or of a column.*

Theorem 9 (*The Multiplication of Determinants*). *The relation*

$$\begin{vmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ a_{21}, & a_{22}, & \ldots, & a_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_{n1}, & a_{n2}, & \ldots, & a_{nn} \end{vmatrix} \cdot \begin{vmatrix} b_{11}, & b_{12}, & \ldots, & b_{1n} \\ b_{21}, & b_{22}, & \ldots, & b_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ b_{n1}, & b_{n2}, & \ldots, & b_{nn} \end{vmatrix} = \begin{vmatrix} c_{11}, & c_{12}, & \ldots, & c_{1n} \\ c_{21}, & c_{22}, & \ldots, & c_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ c_{n1}, & c_{n2}, & \ldots, & c_{nn} \end{vmatrix},$$

holds, where

$$c_{ik} = a_{i1}b_{1k} + a_{i2}b_{2k} + \ldots + a_{in}b_{nk} \quad (i, k = 1, 2, \ldots, n).$$

REMARK 1 (*Evaluation of a determinant*).

1.

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

2. *Sarrus's rule* for the evaluation of a determinant of the third order:

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

3. The evaluation of a determinant of order n for $n \geqq 3$ can be reduced, according to Theorem 5, to the evaluation of a determinant of order $n - 1$. First, it is often advantageous to arrange the original determinant by means of Theorem 2 or Theorem 8 in order to get, in a certain row or column, as many zeros as possible. Then, we expand the determinant according to this row or column (Theorem 5).

Example 2.

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

$$= 2 \left(1 \begin{vmatrix} 1, & -1, & 1 \\ -1, & 2, & 0 \\ -1, & 5, & 3 \end{vmatrix} - 0 \begin{vmatrix} -1, & 2, & 2 \\ -1, & 2, & 0 \\ -1, & 5, & 3 \end{vmatrix} + 3 \begin{vmatrix} -1, & 2, & 2 \\ 1, & -1, & 1 \\ -1, & 5, & 3 \end{vmatrix} - 0 \begin{vmatrix} -1, & 2, & 2 \\ 1, & -1, & 1 \\ -1, & 2, & 0 \end{vmatrix} \right) =$$

$$= 2 \begin{vmatrix} 1, & -1, & 1 \\ -1, & 2, & 0 \\ -1, & 5, & 3 \end{vmatrix} + 6 \begin{vmatrix} -1, & 2, & 2 \\ 1, & -1, & 1 \\ -1, & 5, & 3 \end{vmatrix} = 48 .$$

We proceeded in the above evaluation as follows: First, a common factor 2 was removed from the last column; then we added the first row to the last row, finally, we expanded the determinant according to the first column.

1.18. Systems of Linear Equations

(a) Definition and Properties of Systems of Linear Equations

Definition 1. By a *system of m linear equations in n unknowns* $x_1, x_2, \ldots, x_n$ we understand the system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 , \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 , \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m \end{aligned} \tag{1}$$

($a_{11}, \ldots, a_{mn}, b_1, \ldots, b_m$ being given real or complex numbers).

By a *solution of the system* (1) we mean any ordered n-tuple of (real or complex) numbers $(\xi_1, \xi_2, \ldots, \xi_n)$, i.e. an n-component vector such that if $\xi_1, \ldots, \xi_n$ are substituted for the unknowns $x_1, \ldots, x_n$, then all the equations of the system (1) are satisfied. Two systems of linear equations (in the same number of unknowns $x_1, \ldots, x_n$) are said to be *equivalent systems of linear equations* if every solution of the first system is also a solution of the second system and vice versa. The matrix

$$\mathbf{A} = \begin{bmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ a_{21}, & a_{22}, & \ldots, & a_{2n} \\ \ldots\ldots\ldots\ldots & & & \\ a_{m1}, & a_{m2}, & \ldots, & a_{mn} \end{bmatrix}$$

is called the *matrix of the system* (1). The matrix

$$\boldsymbol{B} = \begin{bmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n}, & b_1 \\ a_{21}, & a_{22}, & \ldots, & a_{2n}, & b_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots & & & & \\ a_{m1}, & a_{m2}, & \ldots, & a_{mn}, & b_m \end{bmatrix}$$

is the so-called *augmented matrix of the system* (1).

Theorem 1 (*Theorem of Frobenius*). *The system* (1) *is solvable if and only if the rank of the matrix of the system is equal to the rank of the augmented matrix of the system.*

Theorem 2. *The system of m homogeneous equations in n unknowns*

$$\begin{aligned} a_{11}x_1 + \ldots + a_{1n}x_n &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{m1}x_1 + \ldots + a_{mn}x_n &= 0 \end{aligned} \tag{2}$$

has always the (*trivial*) *zero solution* $\mathbf{0} = (0, 0, \ldots, 0)$.

Theorem 3. *If the system* (2) *has a solution* $\xi = (\xi_1, \ldots, \xi_n)$, *then it has also the solution* $\alpha\xi = (\alpha\xi_1, \ldots, \alpha\xi_n)$, *where* α *is an arbitrary real or complex number.*

If the vectors $\xi^{(1)} = (\xi_1^{(1)}, \ldots, \xi_n^{(1)}), \ldots, \xi^{(k)} = (\xi_1^{(k)}, \ldots, \xi_n^{(k)})$ *are solutions of the system* (2), *then every linear combination of the form* $\alpha_1\xi^{(1)} + \alpha_2\xi^{(2)} + \ldots + \alpha_k\xi^{(k)}$ (see Definition 1.15.4) *is also a solution of the system* (2).

Theorem 4. *If the rank of the matrix of the system* (2) *of homogeneous equations is h, then the system* (2) *has* $n - h$ *linearly independent solutions* [in the sense of linear independence of vectors (see Definition 1.15.3)] *and every solution of the system* (2) *is a linear combination of these* $n - h$ *solutions.*

In particular, if $h = n$, then the system has only the trivial solution $\mathbf{0} = (0, \ldots, 0)$. If $m = n$ in (2), then the system has a nontrivial solution if and only if the determinant of the system is zero.

Theorem 5. *Let the rank of the matrix of the system* (1) *be h, let* $\eta = (\eta_1, \ldots, \eta_n)$ *be a solution of the system* (1) *and let* $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(n-h)}$ *be* $n - h$ *linearly independent solutions of the system* (2). *Then every solution of the system* (1) *is the sum of a soiution* $\alpha_1\xi^{(1)} + \ldots + \alpha_{n-h}\xi^{(n-h)}$ *of the homogeneous system* (2) *and the solution* η *of the system* (1); thus, the form of every solution of the system (1) is $\alpha_1\xi^{(1)} + \ldots + \alpha_{n-h}\xi^{(n-h)} + \eta$, where $\alpha_1, \ldots, \alpha_{n-h}$ are real or complex numbers.

(b) Solution of Systems of Linear Equations without the Use of Determinants

Theorem 6. *The augmented matrix* $\boldsymbol{B}$ *of the system* (1) *can be transformed by the operations* 1–4 *of Theorem* 1.16.3, p. 27 (see Example 1.16.2, p. 27), *to the*

matrix $\boldsymbol{C}$ *which has only zeros below the principal diagonal. The system of the equations in n unknowns whose augmented matrix is the matrix* $\boldsymbol{C}$ *is equivalent to the system* (1). In this way, the solution of the system (1) is transformed to the solution of a system which can be easily solved.

Example 1.

(a)
$$\begin{aligned} x - 2y + 3z &= 2, \\ 3x - y + z &= 0, \\ 3x + 4y - 7z &= -6, \\ 5y - 8z &= -6; \end{aligned} \qquad \boldsymbol{B} = \begin{bmatrix} 1, & -2, & 3, & 2 \\ 3, & -1, & 1, & 0 \\ 3, & 4, & -7, & -6 \\ 0, & 5, & -8, & -6 \end{bmatrix}.$$

The matrix $\boldsymbol{B}$ can be arranged as follows: The fourth row is a linear combination of the second and third rows and therefore can be omitted; also the third row is a linear combination of the first and second rows (namely, it is the difference of twice the second row and three times the first row) and thus can also be omitted. It is sufficient to consider the matrix

$$\begin{bmatrix} 1, & -2, & 3, & 2 \\ 3, & -1, & 1, & 0 \end{bmatrix}.$$

Here, we subtract three times the first row from the second one and get the matrix

$$\boldsymbol{C} = \begin{bmatrix} 1, & -2, & 3, & 2 \\ 0, & 5, & -8, & -6 \end{bmatrix}.$$

Thus, we solve the system

$$\begin{aligned} x - 2y + 3z &= 2, \\ 5y - 8z &= -6. \end{aligned}$$

We get $y = \frac{1}{5}(8z - 6)$, $x = \frac{1}{5}(z - 2)$. Hence, we can choose an arbitrary (complex) number for z. The system has an infinite number of solutions $x = \frac{1}{5}(\alpha - 2)$, $y = \frac{1}{5}(8\alpha - 6)$, $z = \alpha$ (α being arbitrary).

(b)
$$\begin{aligned} x - 2y + 3z &= 2, \\ 3x - y + 19z &= 0, \\ 3x + 4y - 7z &= 1, \\ 3y - 6z &= -6. \end{aligned}$$

From the matrix

$$\boldsymbol{B} = \begin{bmatrix} 1, & -2, & 3, & 2 \\ 3, & -1, & 19, & 0 \\ 3, & 4, & -7, & 1 \\ 0, & 3, & -6, & -6 \end{bmatrix}$$

we get successively

$$\begin{bmatrix} 1, & -2, & 3, & 2 \\ 0, & 5, & 10, & -6 \\ 0, & 10, & -16, & -5 \\ 0, & 3, & -6, & -6 \end{bmatrix}, \begin{bmatrix} 1, & -2, & 3, & 2 \\ 0, & 5, & 10, & -6 \\ 0, & 10, & -16, & -5 \\ 0, & 1, & -2, & -2 \end{bmatrix}, \begin{bmatrix} 1, & -2, & 3, & 2 \\ 0, & 5, & 10, & -6 \\ 0, & 0, & -36, & 7 \\ 0, & 0, & 5, & 1 \end{bmatrix},$$

$$\boldsymbol{C} = \begin{bmatrix} 1, & -2, & 3, & 2 \\ 0, & 5, & 10, & -6 \\ 0, & 0, & -36, & 7 \\ 0, & 0, & 0, & 71 \end{bmatrix}.$$

To get the solution we solve the system

$$\begin{aligned} x - 2y + 3z &= 2, \\ 5y + 10z &= -6, \\ -36z &= 7, \\ 0 &= 71; \end{aligned}$$

however, this system has no solution, for the last equation cannot be satisfied. Hence, the given system is not solvable.

REMARK 1. When rearranging the matrix $\boldsymbol{B}$ of Theorem 6 it is sometimes advantageous to interchange two columns. This can be done, provided neither of them is the last column; however, we must then interchange the unknowns in the resulting system corresponding to the interchanged columns. The procedure is obvious from Example 2.

Example 2.

$$\begin{aligned} 3x + y + 3z &= 2, \\ -x + 3z &= 3, \\ 4x - z &= 0. \end{aligned} \qquad \boldsymbol{B} = \begin{bmatrix} 3, & 1, & 3, & 2 \\ -1, & 0, & 3, & 3 \\ 4, & 0, & -1, & 0 \end{bmatrix};$$

we interchange the first and second columns:

$$\begin{bmatrix} 1, & 3, & 3, & 2 \\ 0, & -1, & 3, & 3 \\ 0, & 4, & -1, & 0 \end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} 1, & 3, & 3, & 2 \\ 0, & -1, & 3, & 3 \\ 0, & 0, & 11, & 12 \end{bmatrix}.$$

The solution to be found is then the solution of the system

$$\begin{aligned} y + 3x + 3z &= 2, \\ -x + 3z &= 3, \\ 11z &= 12. \end{aligned}$$

(c) Solution of Systems of Linear Equations by Means of Determinants

Theorem 7 (*Cramer's Rule*). *The system of n equations in n unknowns*

$$\begin{aligned} a_{11}x_1 + \ldots + a_{1n}x_n &= b_1 , \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n1}x_1 + \ldots + a_{nn}x_n &= b_n \end{aligned} \tag{3}$$

with a non-zero determinant of the system

$$D = \begin{vmatrix} a_{11}, \ldots, a_{1n} \\ \ldots\ldots\ldots \\ a_{n1}, \ldots, a_{nn} \end{vmatrix} \neq 0 ,$$

has a unique solution $(x_1, \ldots, x_n)$, *where* $x_i = D_i/D$; *here,* D_i *is the determinant obtained by replacing the i-th column of D by the column of elements forming the right-hand sides of equations* (3).

Example 3.

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

$$D = \begin{vmatrix} 3, & -2, & 1 \\ 1, & 1, & -1 \\ 2, & 0, & -3 \end{vmatrix} = -13 \neq 0 ; \quad x_1 = -\tfrac{1}{13} \begin{vmatrix} 1, & -2, & 1 \\ -2, & 1, & -1 \\ 0, & 0, & -3 \end{vmatrix} = -\tfrac{9}{13} ,$$

$$x_2 = -\tfrac{1}{13} \begin{vmatrix} 3, & 1, & 1 \\ 1, & -2, & -1 \\ 2, & 0, & -3 \end{vmatrix} = -\tfrac{23}{13} , \quad x_3 = -\tfrac{1}{13} \begin{vmatrix} 3, & -2, & 1 \\ 1, & 1, & -2 \\ 2, & 0, & 0 \end{vmatrix} = -\tfrac{6}{13} .$$

Theorem 8 (*The System of m Equations in n Unknowns*). *Let the matrix of a system and the augmented matrix of the system have the same rank h. Solution: In the matrix, we find a minor* $D_h \neq 0$ *of order h. In the h equations of the given system* (1) *containing the elements of the determinant* D_h, *we leave on the left-hand side those unknowns whose coefficients belong to* D_h. *We choose arbitrary values for the remaining unknowns, transfer them to the right-hand side and solve this system of h equations in h unknowns by Cramer's Rule. We can always proceed this way, both for homogeneous and non-homogeneous systems.*

Example 4.

(a)
$$\begin{aligned} 3x_1 - 2x_2 + x_3 - x_4 &= 2 , \\ -x_1 + 3x_3 + x_4 &= -1 , \\ x_2 + 3x_3 + 2x_4 &= 3 . \end{aligned}$$

The matrix of the system and the augmented matrix have rank 3,

$$\begin{vmatrix} 3, & -2, & 1 \\ -1, & 0, & 3 \\ 0, & 1, & 3 \end{vmatrix} = -16 .$$

Transform the system to the form

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

$$x_1 = -\tfrac{1}{16}\begin{vmatrix} 2 + x_4, & -2, & 1 \\ -1 - x_4, & 0, & 3 \\ 3 - 2x_4, & 1, & 3 \end{vmatrix} = -\tfrac{1}{16}\left[\begin{vmatrix} 2, & -2, & 1 \\ -1, & 0, & 3 \\ 3, & 1, & 3 \end{vmatrix} + x_4 \begin{vmatrix} 1, & -2, & 1 \\ -1, & 0, & 3 \\ -2, & 1, & 3 \end{vmatrix}\right] = -\tfrac{1}{16}[-31 + 2x_4] ,$$

$$x_2 = -\tfrac{1}{16}\begin{vmatrix} 3, & 2 + x_4, & 1 \\ -1, & -1 - x_4, & 3 \\ 0, & 3 - 2x_4, & 3 \end{vmatrix} = -\tfrac{1}{16}[-33 + 14x_4] ,$$

$$x_3 = -\tfrac{1}{16}\begin{vmatrix} 3, & -2, & 2 + x_4 \\ -1, & 0, & -1 - x_4 \\ 0, & 1, & 3 - 2x_4 \end{vmatrix} = -\tfrac{1}{16}[-5 + 6x_4]$$

(x_4 is arbitrary).

(b) *The system of two homogeneous equations in three unknowns*

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= 0 , \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= 0 \end{aligned}$$

is as follows.

If the rank of the matrix of the system is 2, then the solution is

$$x_1 : x_2 : x_3 = \begin{vmatrix} a_{12}, & a_{13} \\ a_{22}, & a_{23} \end{vmatrix} : \begin{vmatrix} a_{13}, & a_{11} \\ a_{23}, & a_{21} \end{vmatrix} : \begin{vmatrix} a_{11}, & a_{12} \\ a_{21}, & a_{22} \end{vmatrix} .$$

REMARK 2. On the numerical solution of systems of linear equations see Chap. 30.

1.19. Algebraic Equations of Higher Degree. General Properties

Definition 1. An equation

$$a_0x^n + a_1x^{n-1} + \ldots + a_n = 0 , \quad a_0 \neq 0 , \tag{1}$$

where $a_0, a_1, \ldots, a_n$ are real or complex numbers, is called an *algebraic equation of degree n.*

REMARK 1. On the concept of a *root*, its *multiplicity* and theorems on the number of roots see §1.14.

Theorem 1 (*Properties of Roots*). *The roots* $x_1, x_2, \ldots, x_n$ *of the equation*

$$x^n + a_1 x^{n-1} + \ldots + a_n = 0 \tag{2}$$

satisfy the relations

$$a_1 = -(x_1 + x_2 + \ldots + x_n) = -\sum_{i=1}^{n} x_i ,$$

$$a_2 = x_1x_2 + x_1x_3 + \ldots + x_{n-1}x_n = \sum_{\substack{i,j=1 \\ i<j}}^{n} x_i x_j ,$$

$$a_3 = -(x_1x_2x_3 + x_1x_2x_4 + \ldots + x_{n-2}x_{n-1}x_n) = -\sum_{\substack{i,j,k=1 \\ i<j<k}}^{n} x_i x_j x_k ,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$a_n = (-1)^n x_1x_2 \ldots x_n .$$

REMARK 2. The expressions

$$y_1 = \sum_{i=1}^{n} x_i , \quad y_2 = \sum_{\substack{i,j=1 \\ i<j}}^{n} x_i x_j , \quad \ldots , \quad y_n = x_1x_2 \ldots x_n$$

are called *elementary symmetric functions* of the variables $x_1, x_2, \ldots, x_n$.

REMARK 3. On the numerical solution of algebraic equations see Chap. 31.

Definition 2. The *resultant of two algebraic equations*

$$\begin{aligned} a_0x^m + a_1x^{m-1} + \ldots + a_m &= 0 , \quad a_0 \neq 0 , \\ b_0x^n + b_1x^{n-1} + \ldots + b_n &= 0 , \quad b_0 \neq 0 \end{aligned} \tag{3}$$

is defined as the determinant

$$\left|\begin{array}{ccccccccc} a_0, & a_1, & \ldots, & a_{m-1}, & a_m, & 0, & \ldots, & 0 & \\ 0, & a_0, & a_1, & \ldots, & a_{m-1}, & a_m, & 0, \ldots, & 0 & \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \\ 0, & \ldots, & 0, & a_0, & a_1, & \ldots, & a_{m-1}, & a_m & \\ b_0, & b_1, & \ldots, & b_{n-1}, & b_n, & 0, & \ldots, & 0 & \\ 0, & b_0, & b_1, & \ldots, & b_{n-1}, & b_n, & 0, & \ldots, & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \\ 0, & \ldots, & 0, & b_0, & b_1, & \ldots, & b_{n-1}, & b_n & \end{array}\right| \begin{array}{l} \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} n \text{ rows} \\ \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} m \text{ rows} . \end{array}$$

Theorem 2. *The equations* (3) *have a common root if and only if their resultant is equal to zero.*

1.20. Quadratic, Cubic and Biquadratic Equations

(a) A quadratic equation is of the form

(a) $ax^2 + bx + c = 0 \quad (a \neq 0)$ or
(b) $x^2 + px + q = 0 \quad (\textit{reduced form})$

Definition 1. The *discriminant* of the equation (a) is the number $D = b^2 - 4ac$ and that of the equation (b) is the number $D = p^2 - 4q$.

Theorem 1.

For $D \neq 0$, *the equation has two distinct roots;*
for $D = 0$, *the equation has one double root.*

If the coefficients of the equation are real, then
for $D > 0$, *it has two distinct real roots;*
for $D < 0$, *it has two complex conjugate roots;*
for $D = 0$, *it has only one real (double) root.*

The solution can be found:

1. by factorization into linear factors:

$$ax^2 + bx + c = a(x - x_1)(x - x_2) \quad \text{or} \quad x^2 + px + q = (x - x_1)(x - x_2),$$
$$a(x_1 + x_2) = -b, \qquad x_1 + x_2 = -p,$$
$$ax_1x_2 = c \qquad x_1x_2 = q$$

[e.g. $x^2 - 5x + 6 = 0$, $(x - 2)(x - 3) = 0$, $x_1 = 2$, $x_2 = 3$];

2. in the case of the equation $ax^2 + bx + c = 0$ by the formula

$$x_{1,2} = \frac{-b \pm \sqrt{(b^2 - 4ac)}}{2a};$$

3. in the case of the equation $x^2 + px + q = 0$ by the formula

$$x_{1,2} = -\frac{p}{2} \pm \sqrt{\left(\frac{p^2}{4} - q\right)}.$$

(b) A cubic equation is of the form

$$ax^3 + bx^2 + cx + d = 0, \quad a \neq 0. \tag{1}$$

Theorem 2. *By the substitution* $x = y - b/3a$ *and dividing by* a, *the equation* (1) *becomes*

$$y^3 + 3py + 2q = 0\,, \tag{2}$$

where

$$3p = \frac{3ac - b^2}{3a^2}\,, \quad 2q = \frac{2b^3}{27a^3} - \frac{bc}{3a^2} + \frac{d}{a}\,.$$

Definition 2. The *discriminant* of the equation (2) is the number $D = -p^3 - q^2$.

Theorem 3.

For $D \neq 0$, *the equation* (2) *has three distinct roots;*
for $D = 0$, *the equation* (2) *has either a double root* (*if* $p^3 = -q^2 \neq 0$) *or a triple zero root* (*if* $p = q = 0$).

If the coefficients of the equation (2) *are real, then*
for $D \geqq 0$, *it has three real roots which are distinct if* $D > 0$;
for $D < 0$, *it has one real and two complex conjugate roots.*

Solution (see also Chap. 31):

1. By factorization into linear factors:

$$ax^3 + bx^2 + cx + d = 0\,; \quad a(x - x_1)(x - x_2)(x - x_3) = 0$$

(x_1, x_2, x_3 are the roots);

$$x_1 + x_2 + x_3 = -\frac{b}{a}\,, \quad x_1x_2 + x_1x_3 + x_2x_3 = \frac{c}{a}\,, \quad x_1x_2x_3 = -\frac{d}{a}$$

[e.g. $x^3 + 5x^2 + 6x = 0$; $x(x^2 + 5x + 6) = x(x + 2)(x + 3)$; the roots are $x_1 = 0$, $x_2 = -2$, $x_3 = -3$].

2. *The algebraic solution* (*Tartaglia's or Cardan's Formulae*). The roots y_1, y_2, y_3 of equation (2) are

$$y_1 = u + v\,, \quad y_2 = \varepsilon_1 u + \varepsilon_2 v\,, \quad y_3 = \varepsilon_2 u + \varepsilon_1 v\,,$$

where

$$\varepsilon_{1,2} = -\frac{1}{2} \pm \mathrm{i}\frac{\sqrt{3}}{2}\,, \quad u = \sqrt[3]{[-q + \sqrt{(q^2 + p^3)}]}\,, \quad v = \sqrt[3]{[-q - \sqrt{(q^2 + p^3)}]}\,;$$

here we choose the cube roots (see § 1.21, p. 42) so that $uv = -p$. This method is not suitable if (2) has real coefficients and $D > 0$, since the real roots y_1, y_2, y_3 are expressed in terms of roots of complex numbers (the irreducible case).

3. *The trigonometric solution.* Let the coefficients p, q of the equation (2) be real and different from zero. Denote the roots by y_1, y_2, y_3. Put $r = \varepsilon\sqrt{|p|}$, where

TABLE 1.1

$p < 0$		$p > 0$	Check
$p^3 + q^2 \leqq 0$	$p^3 + q^2 > 0$		
$\cos\varphi = \frac{q}{r^3}$ *	$\cosh\varphi = \frac{q}{r^3}$	$\sinh\varphi = \frac{q}{r^3}$	
$y_1 = -2r\cos\frac{\varphi}{3}$	$y_1 = -2r\cosh\frac{\varphi}{3}$	$y_1 = -2r\sinh\frac{\varphi}{3}$	
$y_2 = 2r\cos\left(60° - \frac{\varphi}{3}\right)$	$y_2 = r\cosh\frac{\varphi}{3} + \mathrm{i}\sqrt{(3)}\, r\sinh\frac{\varphi}{3}$	$y_2 = r\sinh\frac{\varphi}{3} + \mathrm{i}\sqrt{(3)}\, r\cosh\frac{\varphi}{3}$	$y_1 + y_2 + y_3 = 0$
$y_3 = 2r\cos\left(60° + \frac{\varphi}{3}\right)$	$y_3 = r\cosh\frac{\varphi}{3} - \mathrm{i}\sqrt{(3)}\, r\sinh\frac{\varphi}{3}$	$y_3 = r\sinh\frac{\varphi}{3} - \mathrm{i}\sqrt{(3)}\, r\cosh\frac{\varphi}{3}$	

* φ is in the interval $(0°, 90°)$, $r = \varepsilon\sqrt{|p|}$ (see above).

$\varepsilon = 1$ if $q > 0$ and $\varepsilon = -1$ if $q < 0$. Then the roots can be determined by means of the trigonometric or hyperbolic functions according to Table 1.1.

If $q = 0$ in equation (2), then the equation has the common factor y and can be solved easily.

If $p = 0$ in equation (2), then (2) is a binomial equation (see § 1.21, p. 42).

(c) A biquadratic (or quartic) equation has the form:

$$ax^4 + bx^3 + cx^2 + dx + e = 0, \quad a \neq 0. \tag{3}$$

Theorem 4. *By the substitution* $x = y - b/4a$ *and dividing by* a, *the equation* (5) *becomes*

$$y^4 + py^2 + qy + r = 0 \tag{4}$$

where

$$p = -\frac{3b^2}{8a^2} + \frac{c}{a}, \quad q = \frac{b^3}{8a^3} - \frac{bc}{2a^2} + \frac{d}{a}, \quad r = -\frac{3b^4}{256a^4} + \frac{b^2c}{16a^3} - \frac{bd}{4a^2} + \frac{e}{a}.$$

Solution:

1. By factorization into linear factors:

$$ax^4 + bx^3 + cx^2 + dx + e = 0; \quad a(x - x_1)(x - x_2)(x - x_3)(x - x_4) = 0$$

(x_1, x_2, x_3, x_4 are the roots);

$$x_1 + x_2 + x_3 + x_4 = -\frac{b}{a}; \quad x_1x_2 + x_1x_3 + x_1x_4 + x_2x_3 + x_2x_4 + x_3x_4 = \frac{c}{a};$$

$$x_1x_2x_3 + x_1x_2x_4 + x_1x_3x_4 + x_2x_3x_4 = -\frac{d}{a}; \quad x_1x_2x_3x_4 = \frac{e}{a}.$$

2. *The algebraic solution.* The roots y_1, y_2, y_3, y_4 of the equation (4) are

$$y_1 = \sqrt{z_1} + \sqrt{z_2} + \sqrt{z_3}, \quad y_2 = \sqrt{z_1} - \sqrt{z_2} - \sqrt{z_3},$$
$$y_3 = -\sqrt{z_1} + \sqrt{z_2} - \sqrt{z_3}, \quad y_4 = -\sqrt{z_1} - \sqrt{z_2} + \sqrt{z_3},$$

where z_1, z_2, z_3 are the roots of the equation (*the reducing cubic*)

$$z^3 + \frac{p}{2}z^2 + \left(\frac{p^2}{16} - \frac{r}{4}\right)z - \frac{q^2}{64} = 0;$$

here, the roots $\sqrt{z_1}, \sqrt{z_2}, \sqrt{z_3}$ should be chosen (see § 1.21) such that

$$\sqrt{(z_1)}\,\sqrt{(z_2)}\,\sqrt{(z_3)} = -\frac{q}{8}.$$

REMARK 1. This method is not suitable for numerical solution (see Chap. 31).

1.21. Binomial Equations

Definition 1. An equation of the form

$$x^n - \alpha = 0, \tag{1}$$

where α is a non-zero real or complex number, is called a *binomial equation.*

Definition 2. The roots of equation (1) are said to be the *n-th roots of the number* α and are denoted in the theory of algebraic equations by the symbol $\sqrt[n]{\alpha}$; thus, in this case (in contrast to § 1.8, p. 12) $\sqrt[n]{\alpha}$ stands for any of the n roots of the equation (1).

Theorem 1. *Equation* (1) *has n simple roots* $x_1, \ldots, x_n$ *given by*

$$x_{k+1} = \sqrt[n]{(r)}\left(\cos\frac{\varphi + 2k\pi}{n} + \mathrm{i}\sin\frac{\varphi + 2k\pi}{n}\right) \quad (k = 0, 1, \ldots, n-1),$$

where $\alpha = r(\cos\varphi + \mathrm{i}\sin\varphi)$ *is the trigonometric form of the number* α, $\sqrt[n]{(r)} > 0$.

REMARK 1. By means of Theorem 1 we easily find all the n-th roots of any complex number.

Example 1. (a) $x^3 - 2 = 0$. First, $2 = 2(\cos 0 + i \sin 0)$. Hence

$$x_1 = \sqrt[3]{(2)}\,(\cos 0 + i \sin 0) = \sqrt[3]{(2)} \doteq 1{\cdot}260\,,$$

$$x_2 = \sqrt[3]{(2)}\,(\cos \tfrac{2}{3}\pi + i \sin \tfrac{2}{3}\pi) = \sqrt[3]{(2)}\left(-\frac{1}{2} + i\,\frac{\sqrt{3}}{2}\right),$$

$$x_3 = \sqrt[3]{(2)}\,(\cos \tfrac{4}{3}\pi + i \sin \tfrac{4}{3}\pi) = \sqrt[3]{(2)}\left(-\frac{1}{2} - i\,\frac{\sqrt{3}}{2}\right).$$

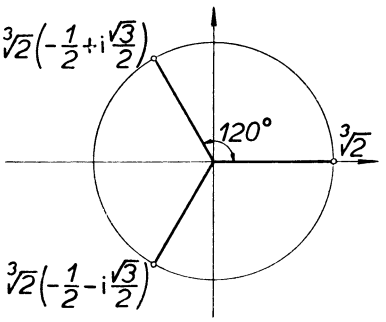


Fig. 1.5a.

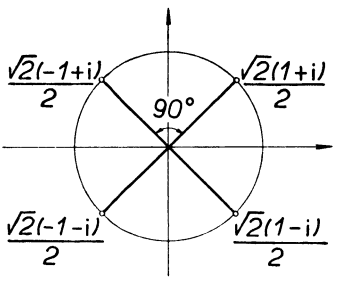


Fig. 1.5b.

(b) $x^4 + 1 = 0$. We have $-1 = \cos \pi + i \sin \pi$. Hence

$$x_1 = \cos \tfrac{1}{4}\pi + i \sin \tfrac{1}{4}\pi = \frac{\sqrt{2}}{2}(1 + i)\,,$$

$$x_2 = \cos \tfrac{3}{4}\pi + i \sin \tfrac{3}{4}\pi = \frac{\sqrt{2}}{2}(-1 + i)\,,$$

$$x_3 = \cos \tfrac{5}{4}\pi + i \sin \tfrac{5}{4}\pi = \frac{\sqrt{2}}{2}(-1 - i)\,,$$

$$x_4 = \cos \tfrac{7}{4}\pi + i \sin \tfrac{7}{4}\pi = \frac{\sqrt{2}}{2}(1 - i)\,.$$

REMARK 2. The roots of the equation $x^n - \alpha = 0$ $(\alpha \neq 0)$ form, in the Argand diagram, the vertices of an n-sided regular polygon inscribed in the circle with centre at the origin and radius $\sqrt[n]{|\alpha|} > 0$. Fig. 1.5 illustrates the roots of the equations $x^3 - 2 = 0$, $x^4 + 1 = 0$. One sees from the figure that the n-th roots can easily be constructed geometrically.

1.22. Reciprocal Equations

Definition 1. By a *reciprocal equation* we understand an equation of the form

$$a_0 x^n + a_1 x^{n-1} + \ldots + a_n = 0$$

where (a) $a_i = a_{n-i}$ $(i = 0, 1, \ldots, n)$ (*positively reciprocal equation*) or
(b) $a_i = -a_{n-i}$ $(i = 0, 1, \ldots, n)$ (*negatively reciprocal equation*).

Theorem 1. *Every positively reciprocal equation of odd degree and every negatively reciprocal equation of even degree has the root* -1.

Theorem 2. *Every negatively reciprocal equation has the root* $+1$.

Theorem 3. *Reducing a reciprocal equation by linear factors of the form* $x - 1$, $x + 1$, *we get a positively reciprocal equation of even degree*

$$a_0 x^{2m} + a_1 x^{2m-1} + \ldots + a_m x^m + \ldots + a_0 = 0$$

which, if divided by x^m, *becomes*

$$a_0\left(x^m + \frac{1}{x^m}\right) + a_1\left(x^{m-1} + \frac{1}{x^{m-1}}\right) + \ldots + a_{m-1}\left(x + \frac{1}{x}\right) + a_m = 0\,.$$

By the substitution $x + 1/x = y$, all binomials can be expressed as polynomials in y; hence we get an equation of degree m in y which in some cases can then be easily solved.

Example 1. The equation $x^6 + x^5 - 5x^4 + 5x^2 - x - 1 = 0$ is a negatively reciprocal equation of even degree; thus it has the roots $\xi_1 = -1, \xi_2 = +1$. Dividing by $(x + 1)(x - 1)$, we get a positively reciprocal equation of even degree

$$x^4 + x^3 - 4x^2 + x + 1 = 0\,, \quad \text{i.e.} \quad \left(x^2 + \frac{1}{x^2}\right) + \left(x + \frac{1}{x}\right) - 4 = 0\,.$$

By the substitution $x + 1/x = y$, we transform the equation into the form $y^2 + y - 6 = 0$ (since $x^2 + 1/x^2 = (x + 1/x)^2 - 2$) with roots $y_1 = 2$, $y_2 = -3$. Hence, the remaining four roots of the original equation are the roots of the quadratic equations

$$x + \frac{1}{x} = 2\,, \quad x + \frac{1}{x} = -3\,.$$

1.23. The Concept of a Set and the Concept of a Mapping

A *set* is a collection of certain objects, called the *elements* of the set. A set is completely determined by its elements. Thus, if the sets A, B consist of the same elements, we say that they are *equal* and write $A = B$.

Examples of sets:

(a) the set of all even numbers;
(b) the set of all points on the circumference of a given circle;
(c) the set of the numbers 1, 2, 3 [we denote it by either $\{1, 2, 3\}$ or $(1, 2, 3)$].

The *empty* (or *void*) *set* (denoted by $\emptyset$) contains no elements at all. For example, the set of all even numbers greater than 0 and less than 2 is empty.

If x is an element of the set M, we write $x \in M$; if x is not an element of this set, then we write either $x \notin M$ or x non $\in M$.

Definition 1. The set A is called a *subset* of the set B (in symbols, $A \subset B$) if every element x of the set A is also an element of the set B, i.e. if $x \in A \Rightarrow x \in B$.

REMARK 1. For the sets A, B the equality $A = B$ holds if and only if both $A \subset B$ and $B \subset A$ hold.

Definition 2. The *union* (or *sum*) *of the sets* A, B (in symbols, $A \cup B$ or $A + B$) is the set of those elements which belong to at least one of the sets. [Similarly, for a greater (even infinite) number of sets.]

Definition 3. The *intersection* (or *product*) *of the sets* A, B (in symbols, $A \cap B$ or $A \,.\, B$ or AB) is the set of elements belonging simultaneously to both A and B. (Similarly, for a greater (even infinite) number of sets.) If $A \cap B = \emptyset$, we say that A and B are *disjoint sets* (they have no common element).

Definition 4. The *difference* (or *relative complement*) *of the sets* A, B (in symbols, $A \dot{-} B$ or $A - B$ or $A \setminus B$) is the set of those elements of A which do not belong to B.

Example 1. If A is the set of real numbers x satisfying $1 \leqq x \leqq 10$ (i.e. $A = [1, 10]$) and if, similarly $B = [5, 15]$, then $A \cup B = [1, 15]$, $A \cap B = [5, 10]$, $A \dot{-} B = [1, 5)$. If $C = [1, 2]$, then $C \subset A$, $C \cap B = \emptyset$.

Definition 5. The set of all *ordered* pairs (x, y), where $x \in A$, $y \in B$, is called the *cartesian product of the sets* A, B (denoted by $A \times B$).

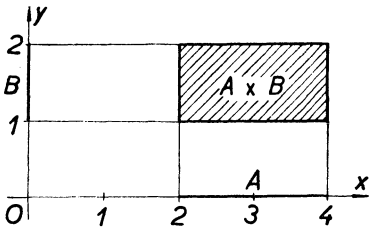


Fig. 1.6.

Example 2. If $A = [2, 4]$, $B = [1, 2]$, then $A \times B$ is the set of the ordered pairs (x, y), where $2 \leqq x \leqq 4$, $1 \leqq y \leqq 2$. If we illustrate (Fig. 1.6) the sets A and B in the plane of the coordinate axes x, y as the intervals $[2, 4]$ of the x axis and as $[1, 2]$ of the y axis, respectively, then $A \times B$ is represented by the rectangle with the vertices $(2, 1)$, $(2, 2)$, $(4, 1)$, $(4, 2)$.

REMARK 2. Let A_α be a system of sets with the index α running through a set M. Then the union and intersection of all the sets A_α are denoted by the symbols $\bigcup_{\alpha \in M} A_\alpha$ and $\bigcap_{\alpha \in M} A_\alpha$, respectively. If M is the set of the natural numbers, then the union and intersection of the sets $A_1, A_2, A_3, \ldots$ are often denoted by $\bigcup_{i=1}^{\infty} A_i$ and $\bigcap_{i=1}^{\infty} A_i$, respectively. Similarly, we write $\bigcup_{k=1}^{n} A_k$ instead of $A_1 \cup A_2 \cup \ldots \cup A_n$ and correspondingly for the intersection.

Theorem 1 (*De Morgan's Formulae*). *The relations*

$$\bigcap_{\alpha \in M} (B \cup A_\alpha) = B \cup \left(\bigcap_{\alpha \in M} A_\alpha \right); \quad \bigcup_{\alpha \in M} (B \cap A_\alpha) = B \cap \left(\bigcup_{\alpha \in M} A_\alpha \right);$$

$$B \div \bigcup_{\alpha \in M} A_\alpha = \bigcap_{\alpha \in M} (B \div A_\alpha); \quad B \div \bigcap_{\alpha \in M} A_\alpha = \bigcup_{\alpha \in M} (B \div A_\alpha)$$

hold.

Definition 6. A *mapping f of a set A into a set B* is a rule which assigns to every element $x \in A$ a definite element $y \in B$ (uniquely determined by the element x). The element y is denoted by the symbol $f(x)$ and is called the *image* of the element x. The element x is said to be the *original* or *inverse image* of the element $f(x)$. The set A is called the *domain of the mapping f*.

Definition 7. The mapping f of Definition 6 is a *mapping of the set A onto the set B* if, for every $y \in B$, there exists at least one $x \in A$ such that $y = f(x)$.

Definition 8. The mapping f is said to be *one-to-one*, if $x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2)$.

Definition 9. Let f be a one-to-one mapping of a set A onto a set B. The mapping f^{-1} which assigns to every $y \in B$ the element $f^{-1}(y) = x \in A$ such that $f(x) = y$, is called the *inverse mapping to f*.

Example 3. (a) Let A be the set of the real numbers. For $x \in A$, put $f(x) = x^2$. Then f is a mapping (not one-to-one) of the set A into the set A (not a mapping of A onto A); f is a mapping of the set A onto the set $B = [0, \infty)$ (onto the set of all real *non-negative* numbers).

(b) Let N be the set of the integers. For $x \in N$, put $f(x) = x + 5$. Then f is a one-to-one mapping of the set N onto N. For the inverse mapping f^{-1} to f, $f^{-1}(y) = y - 5$ holds.

REMARK 3. Besides the term "mapping" the terms *transformation*, *correspondence*, *operation*, *operator*, *map*, *functional*, *function* are also used, in cases where the sets A, B are in some way specialized.

REMARK 4. On the concept of a function of one real variable x see § 11.1. This function is usually denoted by $f(x)$, in contrast to mere f as in Definition 6. [In

theoretical considerations, it is often more advantageous to write only f instead of $f(x)$, for there can be no misunderstanding as to whether $f(x)$ is a function or the value of the function at the point x.]

1.24. Groups, Rings, Division Rings, Fields

Definition 1. A *group* is a non-empty set G in which multiplication is determined in some way, i.e. a rule is given which assigns to each ordered pair a, b of G a unique element $c = ab \in G$, their product. Moreover, the multiplication satisfies the following rules (laws, axioms):

1. $(ab)\,c = a(bc)$ (*associative law*).

2. For each two elements $a, b \in G$ there exist elements $x, y \in G$ such that $ax = b$ and $ya = b$.

REMARK 1. The axioms 1, 2 immediately imply that there is a unique *identity element* e in the group G such that $ea = ae = a$ for every $a \in G$. Furthermore, for each element $a \in G$, there exists a *unique inverse* a^{-1} of a such that $aa^{-1} = a^{-1}a = e$.

Definition 2. A group is called *abelian*, or *commutative*, if, for every two of its elements a, b, the relation $ab = ba$ holds.

REMARK 2. If G is an abelian group, then we frequently use additive notation, i.e. we write $a + b$ instead of ab. The identity element is denoted by 0 (zero element); the inverse of a is denoted by $-a$.

Example 1. (a) The set of all non-zero rational numbers is, with respect to multiplication, an abelian group; the number 1 is its identity element.

(b) The set of all integers is an abelian group with respect to addition; the number 0 is its identity element, the number $-a$ is the inverse of the number a.

(c) The set of all regular matrices of order n is a (non-commutative) group with respect to matrix multiplication (see Definition 1.25.3, p. 49).

Definition 3. By a *ring* (more exactly an *associative ring*) is meant a non-empty set R, in which addition and multiplication are determined in some way, i.e. rules are given which assign to each ordered pair a, $b \in R$ a unique element $a + b \in R$ (their sum) and a unique element $ab \in R$ (their product). Moreover, this addition and multiplication satisfy the following rules (laws, axioms):

1. The set R is, with respect to addition, an abelian group, i.e. for every three elements the relations $(a + b) + c = a + (b + c)$ and $a + b = b + a$ hold and there exists an element x such that $a + x = b$ (the zero element is denoted by 0).

2. Multiplication is associative and is distributive with respect to addition, i.e. for every three elements $a, b, c \in R$

$$(ab)\,c = a(bc)\,,$$
$$(a + b)\,c = ac + bc\,, \quad a(b + c) = ab + ac\,.$$

holds.

REMARK 3. In a ring R, the relation $a0 = 0a = 0$ holds for every element $a \in R$. In R, non-zero elements a, b may exist such that their product is zero: $ab = 0$. Such elements are said to be *zero-divisors*. If there is an element e in R such that $ae = ea = a$ for every $a \in R$, we say that R is a *ring with identity*. If an identity element exists in R, then it is uniquely determined.

Definition 4. A ring R is called a *commutative ring* if, for each $a, b \in R$, the relation $ab = ba$ holds.

Example 2. (a) The set of all integers is a (commutative) ring with identity with respect to addition and multiplication.

(b) The set of all even numbers is a (commutative) ring without identity with respect to addition and multiplication.

(c) The set of all square matrices of order n is a (non-commutative) ring with zero-divisors with respect to matrix addition and multiplication (see Theorem 1.25.3, p. 50).

Definition 5. A *division ring* (*skew field* or *s-field*) D is a ring with identity $e \neq 0$ such that, for every $a \in D$, $a \neq 0$, there exists an inverse $a^{-1} \in D$ such that $aa^{-1} = a^{-1}a = e$. If, moreover, the ring is commutative, then it is called a *field*.

REMARK 4. The identity element e of the division ring D is usually denoted by 1. A division ring has no zero-divisors: If the product of two elements of a division ring is equal to zero, then at least one of the elements is zero. In a division ring, to any non-zero element there corresponds a unique inverse. The set of all non-zero elements of a division ring is a group with respect to multiplication.

Example 3. (a) The set of all rational numbers* is a field (the so-called *field of rational numbers*).

(b) The set of all real numbers* is a field (the so-called *field of real numbers*).

(c) The set of all complex numbers* is a field (the so-called *field of complex numbers*).

---

* With operations of addition and multiplication defined in the usual way.

1.25. Matrices (continued). Operations on Matrices

REMARK 1. The elements of the matrices under consideration – unless otherwise stated – will always be real or complex numbers.

Definition 1. The matrix $\alpha \mathbf{A}$ obtained from a matrix $\mathbf{A}$ by multiplication of all its elements by a number α is called the (*scalar*) *product of the matrix* $\mathbf{A}$ *and the number* α:

$$\alpha \mathbf{A} = \alpha \begin{bmatrix} a_{11}, \ldots, a_{1n} \\ \ldots\ldots\ldots \\ a_{m1}, \ldots, a_{mn} \end{bmatrix} = \begin{bmatrix} \alpha a_{11}, \ldots, \alpha a_{1n} \\ \ldots\ldots\ldots\ldots \\ \alpha a_{m1}, \ldots, \alpha a_{mn} \end{bmatrix}.$$

Definition 2. The *sum* $\mathbf{A} + \mathbf{B}$ *of two* m *by* n *matrices* $\mathbf{A}$, $\mathbf{B}$ is the m by n matrix whose elements are the sums of the corresponding elements:

$$\begin{bmatrix} a_{11}, \ldots, a_{1n} \\ \ldots\ldots\ldots \\ a_{m1}, \ldots, a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11}, \ldots, b_{1n} \\ \ldots\ldots\ldots \\ b_{m1}, \ldots, b_{mn} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11}, \ldots, a_{1n} + b_{1n} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1} + b_{m1}, \ldots, a_{mn} + b_{mn} \end{bmatrix}.$$

Theorem 1. *The scalar multiplication and addition of* m *by* n *matrices satisfy the following rules*:

1. $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$.
2. $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$.
3. $\mathbf{A} + \mathbf{0} = \mathbf{A}$, *where*

$$\mathbf{0} = \begin{bmatrix} 0, \ldots, 0 \\ \ldots\ldots \\ 0, \ldots, 0 \end{bmatrix}$$

is the so-called zero-matrix.

4. *For two matrices* $\mathbf{A}$, $\mathbf{B}$ *there exists a matrix* $\mathbf{X}$ *such that* $\mathbf{A} + \mathbf{X} = \mathbf{B}$; *it is the matrix* $\mathbf{X} = \mathbf{B} + (-1)\,\mathbf{A} = \mathbf{B} - \mathbf{A}$.
5. $\alpha(\mathbf{A} + \mathbf{B}) = \alpha\mathbf{A} + \alpha\mathbf{B}$; $(\alpha + \beta)\,\mathbf{A} = \alpha\mathbf{A} + \beta\mathbf{A}$.

Definition 3. The *product* $\mathbf{AB}$ *of an* m *by* n *matrix* $\mathbf{A}$ *and an* n *by* p *matrix* $\mathbf{B}$ is the m by p matrix $\mathbf{C}$ defined as follows: If

$$\mathbf{A} = \begin{bmatrix} a_{11}, \ldots, a_{1n} \\ \ldots\ldots\ldots \\ a_{m1}, \ldots, a_{mn} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_{11}, \ldots, b_{1p} \\ \ldots\ldots\ldots \\ b_{n1}, \ldots, b_{np} \end{bmatrix}, \quad \text{then} \quad \mathbf{C} = \begin{bmatrix} c_{11}, \ldots, c_{1p} \\ \ldots\ldots\ldots \\ c_{m1}, \ldots, c_{mp} \end{bmatrix},$$

where $c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{in}b_{nj}$ $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, p)$. (In words: The *rows* of the matrix $\mathbf{A}$ are multiplied by the *columns* of the matrix $\mathbf{B}$. The number of columns of the first matrix must be equal to the number of rows of the second.)

Theorem 2. *The multiplication of matrices* $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ *satisfies the relations*

1. $(\boldsymbol{AB})\,\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$,
2. $(\boldsymbol{A} + \boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC} + \boldsymbol{BC}$, $\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$,

if the sums and products of the matrices considered are defined (i.e. if the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ *are of prescribed type).*

REMARK 2. In the following text, we restrict ourselves to *square matrices of order n*, i.e. to n by n matrices.

Theorem 3. *The set of all square matrices of order n whose elements are real or complex numbers constitutes a ring with respect to matrix addition and multiplication* (see 1.24, p. 47), *the so-called ring of square real or complex matrices. For* $n > 1$, *this ring is non-commutative (i.e., in general,* $\boldsymbol{AB} \neq \boldsymbol{BA}$); *it has an identity element – the identity matrix*

$$\boldsymbol{I} = \begin{bmatrix} 1, & 0, & \ldots, & 0 \\ 0, & 1, & \ldots, & 0 \\ \multicolumn{4}{c}{\ldots\ldots\ldots} \\ 0, & 0, & \ldots, & 1 \end{bmatrix}.$$

Furthermore, this ring has zero-divisors, i.e. there are pairs of non-zero matrices $\boldsymbol{A}$, $\boldsymbol{B}$ *such that their product is the zero matrix.*

Definition 4. A square matrix $\boldsymbol{A} = (a_{ij})$ of order n is said to be *regular* or *non-singular* if its determinant $|a_{ij}|$ is different from zero (i.e. if $\boldsymbol{A}$ is of rank n); a matrix which is not regular is called *singular.*

Theorem 4. *The determinant of the matrix* $\boldsymbol{AB}$ *– the product of square matrices* $\boldsymbol{A}$, $\boldsymbol{B}$ *of the same order – is equal to the product of the determinants of the matrices* $\boldsymbol{A}$, $\boldsymbol{B}$.

Theorem 5. *The product of regular matrices of the same order is again a regular matrix.*

Definition 5. The *inverse of a square matrix* $\boldsymbol{A}$ *of order n is a square matrix* $\boldsymbol{A}^{-1}$ of order n such that $\boldsymbol{AA}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

Theorem 6. *The inverse* $\boldsymbol{A}^{-1}$ *of a square matrix* $\boldsymbol{A}$ *exists if and only if* $\boldsymbol{A}$ *is regular. If*

$$\boldsymbol{A} = \begin{bmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ a_{21}, & a_{22}, & \ldots, & a_{2n} \\ \multicolumn{4}{c}{\ldots\ldots\ldots} \\ a_{n1}, & a_{n2}, & \ldots, & a_{nn} \end{bmatrix}, \quad \text{then} \quad \boldsymbol{A}^{-1} = \begin{bmatrix} A_{11}A^{-1}, & A_{21}A^{-1}, & \ldots, & A_{n1}A^{-1} \\ A_{12}A^{-1}, & A_{22}A^{-1}, & \ldots, & A_{n2}A^{-1} \\ \multicolumn{4}{c}{\ldots\ldots\ldots} \\ A_{1n}A^{-1}, & A_{2n}A^{-1}, & \ldots, & A_{nn}A^{-1} \end{bmatrix},$$

where A *is the determinant of the matrix* $\boldsymbol{A}$ *and* A_{ij} *is the cofactor belonging to the element* a_{ij} *in the determinant* A (see Definition 1.17.2, p. 30).

REMARK 3 (*A System of Linear Equations in Matrix Form*). Let

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 , \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{aligned} \tag{1}$$

be a system of n equations in n unknowns. Put

$$\boldsymbol{A} = \begin{bmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_{n1}, & a_{n2}, & \ldots, & a_{nn} \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Then the system of equations (1) can be rewritten in the matrix form

$$\boldsymbol{Ax} = \boldsymbol{b} . \tag{2}$$

If the determinant of the system is non-zero, then the matrix $\boldsymbol{A}$ is regular, and thus, its inverse $\boldsymbol{A}^{-1}$ exists. Multiplying (2) by this matrix $\boldsymbol{A}^{-1}$, we get

$$\boldsymbol{A}^{-1}\boldsymbol{Ax} = \boldsymbol{A}^{-1}\boldsymbol{b} , \quad \text{i.e.} \quad \boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} . \tag{3}$$

If the inverse $\boldsymbol{A}^{-1}$ of the matrix $\boldsymbol{A}$ is known, then we can, in accordance with (3), immediately write down the solution of the system (1) (see also Chap. 30).

Theorem 7. *The inverse of the product* $\boldsymbol{AB}$ *of regular matrices* $\boldsymbol{A}$, $\boldsymbol{B}$ *is equal to the product of the inverses of the matrices* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *taken in reverse order:* $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.

REMARK 4. In what follows, $\boldsymbol{A}'$ denotes the transpose of the matrix $\boldsymbol{A}$ (see Definition 1.16.3, p. 26).

Theorem 8. *The transpose of the product* $\boldsymbol{AB}$ *of two matrices* $\boldsymbol{A}$, $\boldsymbol{B}$ *is equal to the product of the transposes of the matrices* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *taken in the reverse order:* $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$. *Furthermore* $(\boldsymbol{A} + \boldsymbol{B})' = \boldsymbol{A}' + \boldsymbol{B}'$.

Definition 6. A matrix is called *symmetric* or *skew-symmetric* respectively, if $\boldsymbol{A} = \boldsymbol{A}'$, or $\boldsymbol{A} = -\boldsymbol{A}'$, i.e. if $a_{ij} = a_{ji}$, or $a_{ij} = -a_{ji}$, for $i, j = 1, 2, \ldots, n$, respectively.

REMARK 5. The diagonal elements of a skew-symmetric matrix are zero.

Theorem 9 (*A matrix expressed as a sum of a symmetric and a skew-symmetric matrix*). *A matrix* $\boldsymbol{A}$ *is a sum of the symmetric matrix* $\frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}')$ *and the skew-symmetric matrix* $\frac{1}{2}(\boldsymbol{A} - \boldsymbol{A}')$; *hence* $\boldsymbol{A} = \frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}') + \frac{1}{2}(\boldsymbol{A} - \boldsymbol{A}')$.

Theorem 10. *The product of two symmetric matrices* $\boldsymbol{A}$, $\boldsymbol{B}$ *is a symmetric matrix if and only if the matrices commute, i.e. if* $\boldsymbol{AB} = \boldsymbol{BA}$.

Theorem 11. *The rank of a skew-symmetric matrix is always an even number.*

Definition 7. A matrix $\boldsymbol{A}$ is called *orthogonal* if $\boldsymbol{AA'} = \boldsymbol{I}$, i.e. if $\boldsymbol{A'} = \boldsymbol{A}^{-1}$.

Theorem 12. *Let* $\boldsymbol{A} = (a_{ij})$ *be an orthogonal matrix of order n. Then the relations*

$$\sum_{j=1}^{n} a_{ij}^2 = 1\,, \quad \sum_{j=1}^{n} a_{ij}a_{kj} = 0 \quad (i \neq k) \tag{4}$$

hold. In words: *In an orthogonal matrix, the sum of the products of the elements of an arbitrary row and of the corresponding elements of another row is zero and the sum of the squares of the elements of an arbitrary row is unity.*

A similar statement holds for the columns:

$$\sum_{j=1}^{n} a_{ji}^2 = 1\,, \quad \sum_{j=1}^{n} a_{ji}a_{jk} = 0 \quad (i \neq k)\,. \tag{5}$$

Conversely, if the elements of a matrix $\boldsymbol{A} = (a_{ij})$ *satisfy* (4) *or* (5), *then* $\boldsymbol{A}$ *is orthogonal.*

Example 1. The rotation of the rectangular axes in a plane through an angle α is expressed by the following equations (cf. Theorem 5.13.3, p. 186):

$$\begin{aligned} x' &= x\cos\alpha + y\sin\alpha\,, \\ y' &= -x\sin\alpha + y\cos\alpha\,. \end{aligned}$$

The matrix of this transformation, i.e. the matrix

$$\begin{bmatrix} \cos\alpha, & \sin\alpha \\ -\sin\alpha, & \cos\alpha \end{bmatrix}$$

is orthogonal, as can easily be checked using Theorem 12.

Theorem 13. *The determinant of an orthogonal matrix is equal to* 1 *or to* -1.

Theorem 14. *The product of orthogonal matrices is an orthogonal matrix.*

Theorem 15. *The inverse of an orthogonal matrix is an orthogonal matrix.*

Definition 8. *The (complex) conjugate* $\overline{\boldsymbol{A}}$ *of a matrix* $\boldsymbol{A}$ (whose elements are complex numbers) is the matrix obtained from $\boldsymbol{A}$ by replacing every element a_{ij} of $\boldsymbol{A}$ by its conjugate $\overline{a_{ij}}$.

Theorem 16. *The relations*

$$\overline{\alpha\boldsymbol{A} + \beta\boldsymbol{B}} = \bar{\alpha}\overline{\boldsymbol{A}} + \bar{\beta}\overline{\boldsymbol{B}}\,; \quad \overline{\boldsymbol{AB}} = \overline{\boldsymbol{A}}\,\overline{\boldsymbol{B}}\,;$$

$$\overline{\boldsymbol{A}'} = (\overline{\boldsymbol{A}})'\,; \quad \overline{\boldsymbol{A}^{-1}} = (\overline{\boldsymbol{A}})^{-1}\,.$$

hold.

Definition 9. A matrix $\boldsymbol{A}$ is called *Hermitian*, or *skew-Hermitian*, if

$$\boldsymbol{A} = \overline{\boldsymbol{A}'}, \quad \text{or} \quad \boldsymbol{A} = -\overline{\boldsymbol{A}'}, \quad \text{respectively}.$$

Definition 10. A matrix $\boldsymbol{A}$ such that

$$\boldsymbol{A}\overline{\boldsymbol{A}'} = \boldsymbol{I}, \quad \text{i.e.} \quad \overline{\boldsymbol{A}'} = \boldsymbol{A}^{-1},$$

is called *unitary*.

Theorem 17. *A matrix* $\boldsymbol{A} = (a_{ij})$ *of order* n *is unitary if and only if the relations*

$$\sum_{j=1}^{n} a_{ij}\bar{a}_{ij} = 1, \quad \sum_{j=1}^{n} a_{ij}\bar{a}_{kj} = 0 \quad (i \neq k) \tag{6}$$

or

$$\sum_{j=1}^{n} a_{ji}\bar{a}_{jl} = 1, \quad \sum_{j=1}^{n} a_{ji}\bar{a}_{jk} = 0 \quad (i \neq k) \tag{7}$$

hold (cf. Theorem 12).

Theorem 18. *The product of unitary matrices is a unitary matrix.*

Theorem 19. *The inverse of a unitary matrix is again unitary.*

Theorem 20. *The absolute value of the determinant of a unitary matrix is* 1.

Definition 11. By the *trace of the square matrix*

$$\begin{bmatrix} a_{11}, \ldots, a_{1n} \\ \ldots\ldots\ldots \\ a_{n1}, \ldots, a_{nn} \end{bmatrix}$$

is meant the sum $a_{11} + a_{22} + \ldots + a_{nn}$ of the diagonal elements of the matrix.

1.26. Matrices Partitioned into Blocks and Operations on Them; Triangular and Diagonal Matrices

Definition 1 (*A Matrix Partitioned into Blocks*). Let $\boldsymbol{A}$ be an m by n matrix. Divide it into parts by drawing lines between certain rows and certain columns. These parts (so-called *blocks*) are again matrices and the matrix $\boldsymbol{A}$ is formed from these blocks which constitute its elements. We say that the matrix $\boldsymbol{A}$ is *partitioned into blocks*.

Example 1.

$$\boldsymbol{A} = \left[\begin{array}{cccc} 1, & 4, & 2, & 3 \\ 0, & 1, & 2, & 3 \\ 1, & -1, & 0, & 1 \end{array}\right].$$

Thus, for example,

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11}, & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21}, & \boldsymbol{A}_{22} \end{bmatrix},$$

where the individual blocks are the matrices

$$\boldsymbol{A}_{11} = \begin{bmatrix} 1, & 4 \\ 0, & 1 \end{bmatrix}, \qquad \boldsymbol{A}_{12} = \begin{bmatrix} 2, & 3 \\ 2, & 3 \end{bmatrix},$$

$$\boldsymbol{A}_{21} = \begin{bmatrix} 1, & -1 \end{bmatrix}, \quad \boldsymbol{A}_{22} = \begin{bmatrix} 0, & 1 \end{bmatrix}.$$

Theorem 1 (*Multiplication by a Scalar*). *Let a matrix* $\boldsymbol{A}$ *be partitioned into blocks* $\boldsymbol{A}_{ij}$:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11}, & \boldsymbol{A}_{12}, & \ldots, & \boldsymbol{A}_{1n} \\ \ldots\ldots\ldots\ldots\ldots & & & \\ \boldsymbol{A}_{m1}, & \boldsymbol{A}_{m2}, & \ldots, & \boldsymbol{A}_{mn} \end{bmatrix}. \tag{1}$$

Let α *be a real or complex number. Then*

$$\alpha\boldsymbol{A} = \begin{bmatrix} \alpha\boldsymbol{A}_{11}, & \alpha\boldsymbol{A}_{12}, & \ldots, & \alpha\boldsymbol{A}_{1n} \\ \ldots\ldots\ldots\ldots\ldots & & & \\ \alpha\boldsymbol{A}_{m1}, & \alpha\boldsymbol{A}_{m2}, & \ldots, & \alpha\boldsymbol{A}_{mn} \end{bmatrix}.$$

Theorem 2 (*Addition*). *Let the matrix*

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11}, & \ldots, & \boldsymbol{B}_{1n} \\ \ldots\ldots\ldots & & \\ \boldsymbol{B}_{m1}, & \ldots, & \boldsymbol{B}_{mn} \end{bmatrix}$$

be partitioned into blocks of the same type as the matrix (1). *Then*

$$\boldsymbol{A} + \boldsymbol{B} = \begin{bmatrix} \boldsymbol{A}_{11} + \boldsymbol{B}_{11}, & \ldots, & \boldsymbol{A}_{1n} + \boldsymbol{B}_{1n} \\ \ldots\ldots\ldots\ldots\ldots\ldots & & \\ \boldsymbol{A}_{m1} + \boldsymbol{B}_{m1}, & \ldots, & \boldsymbol{A}_{mn} + \boldsymbol{B}_{mn} \end{bmatrix}.$$

Theorem 3 (*Product*). *Let two matrices* $\boldsymbol{C}$, $\boldsymbol{D}$ *be partitioned into blocks*

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11}, & \ldots, & \boldsymbol{C}_{1n} \\ \ldots\ldots\ldots & & \\ \boldsymbol{C}_{m1}, & \ldots, & \boldsymbol{C}_{mn} \end{bmatrix} \quad \boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11}, & \ldots, & \boldsymbol{D}_{1p} \\ \ldots\ldots\ldots & & \\ \boldsymbol{D}_{n1}, & \ldots, & \boldsymbol{D}_{np} \end{bmatrix}$$

in such a way that the number of columns of the matrix $\boldsymbol{C}_{ij}$ *is equal to the number of rows of the matrix* $\boldsymbol{D}_{jk}$ $(i = 1, \ldots, m;\ k = 1, \ldots, p)$. *Then*

$$\boldsymbol{CD} = \begin{bmatrix} \boldsymbol{F}_{11}, & \ldots, & \boldsymbol{F}_{1p} \\ \ldots\ldots\ldots & & \\ \boldsymbol{F}_{m1}, & \ldots, & \boldsymbol{F}_{mp} \end{bmatrix},$$

where $\boldsymbol{F}_{ik} = \boldsymbol{C}_{i1}\boldsymbol{D}_{1k} + \boldsymbol{C}_{i2}\boldsymbol{D}_{2k} + \ldots + \boldsymbol{C}_{in}\boldsymbol{D}_{nk}$.

Hence: The blocks of the matrix $\boldsymbol{CD}$ are the sums of the products of the blocks forming the elements of the rows of the matrix $\boldsymbol{C}$ and the blocks forming the elements of the columns of the matrix $\boldsymbol{D}$.

REMARK 1. The products $\boldsymbol{C}_{i1}\boldsymbol{D}_{1k}, \boldsymbol{C}_{i2}\boldsymbol{D}_{2k}, \ldots$ are defined, since, according to our assumption, the number of columns of the matrix $\boldsymbol{C}_{ij}$ equals the number of rows of the matrix $\boldsymbol{D}_{jk}$.

Example 2. The relation

$$\begin{bmatrix} \boldsymbol{C}_{11}^{(2,3)}, & \boldsymbol{C}_{12}^{(2,1)} \\ \boldsymbol{C}_{21}^{(4,3)}, & \boldsymbol{C}_{22}^{(4,1)} \end{bmatrix} \begin{bmatrix} \boldsymbol{D}_{11}^{(3,4)}, & \boldsymbol{D}_{12}^{(3,2)} \\ \boldsymbol{D}_{21}^{(1,4)}, & \boldsymbol{D}_{22}^{(1,2)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{F}_{11}^{(2,4)}, & \boldsymbol{F}_{12}^{(2,2)} \\ \boldsymbol{F}_{21}^{(4,4)}, & \boldsymbol{F}_{22}^{(4,2)} \end{bmatrix},$$

holds, where the upper indices indicate the type of the corresponding matrices.

Definition 2. A matrix A partitioned into *square* blocks of the form

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11}, & \boldsymbol{0}, & \ldots, & \boldsymbol{0} \\ \boldsymbol{0}, & \boldsymbol{A}_{22}, & \boldsymbol{0}, & \ldots, \boldsymbol{0} \\ \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{0}, & \ldots, & \boldsymbol{0}, & \boldsymbol{A}_{nn} \end{bmatrix},$$

where the symbols $\boldsymbol{0}$ denote zero matrices (and which are, for the sake of brevity, omitted in the formulation of the following Theorem 4), is called the *matrix decomposed into diagonal blocks.*

Theorem 4. *The sum and the product of matrices decomposed into diagonal blocks (where corresponding blocks have the same order) is a matrix decomposed into diagonal blocks; these blocks are sums, or products, of the corresponding blocks of the given matrices, respectively:*

$$\begin{bmatrix} \boldsymbol{A}_{11} & & \\ & \ddots & \\ & & \boldsymbol{A}_{nn} \end{bmatrix} + \begin{bmatrix} \boldsymbol{B}_{11} & & \\ & \ddots & \\ & & \boldsymbol{B}_{nn} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{11} + \boldsymbol{B}_{11} & & \\ & \ddots & \\ & & \boldsymbol{A}_{nn} + \boldsymbol{B}_{nn} \end{bmatrix},$$

$$\begin{bmatrix} \boldsymbol{A}_{11} & & \\ & \ddots & \\ & & \boldsymbol{A}_{nn} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & & \\ & \ddots & \\ & & \boldsymbol{B}_{nn} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{11}\boldsymbol{B}_{11} & & \\ & \ddots & \\ & & \boldsymbol{A}_{nn}\boldsymbol{B}_{nn} \end{bmatrix}.$$

Definition 3. An upper *triangular matrix* is a square matrix of the form

$$\begin{bmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ 0, & a_{22}, & \ldots, & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 0, & 0, & \ldots, & a_{nn} \end{bmatrix},$$

where the elements below the principal diagonal are zero.

Theorem 5. *The sum and the product of upper triangular matrices of the same order is again an upper triangular matrix. The determinant of an upper triangular matrix is equal to the product of the elements in the principal diagonal.*

Definition 4. A *diagonal matrix* is a square matrix of the form

$$\begin{bmatrix} a_{11}, & 0, & \ldots, & 0 \\ 0, & a_{22}, & \ldots, & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0, & 0, & \ldots, & a_{nn} \end{bmatrix}.$$

Theorem 6. *The sum (product) of diagonal matrices of the same order is again a diagonal matrix; the elements of its principal diagonal are the sums (products) of the corresponding diagonal elements of the given matrices.*

1.27. λ-matrices, Equivalence of λ-matrices

Definition 1. A λ-matrix $\boldsymbol{A}(\lambda)$ is a square matrix whose elements are polynomials in the variable λ with real or complex coefficients.

REMARK 1. Addition and multiplication of λ-matrices and the rank of a λ-matrix are defined in the same way as in §§1.25 and 1.16, pp. 49 and 26.

Example 1. The matrix

$$\begin{bmatrix} 3 - \lambda, & 1 + \lambda \\ 1, & 5 - \lambda \end{bmatrix}$$

is a λ-matrix. Its rank is 2, for its determinant is a non-zero polynomial $(3 - \lambda) \cdot (5 - \lambda) - (\lambda + 1) \not\equiv 0$. However, if we substitute for λ a particular numerical value, then we obtain another matrix (no longer the original λ-matrix) whose rank can be smaller. For example, if $\lambda = 2$, we get a matrix of rank 1.

REMARK 2. λ-matrices include also ordinary matrices as a particular case where the elements are polynomials of zero degree or zero polynomials.

Definition 2. By an *elementary transformation of a given λ-matrix* $\boldsymbol{A}(\lambda)$ we understand one of the following rearrangements of the matrix:

1. an interchange of two rows or two columns of the matrix;
2. multiplication of a row or a column by a non-zero number;
3. addition of a row, or a column, multiplied by a polynomial $\varphi(\lambda)$ to another row, or column, respectively.

Definition 3 (*The Equivalence of λ-matrices*). A λ-matrix $\boldsymbol{A}(\lambda)$ is said to be *equivalent* to a λ-matrix $\boldsymbol{B}(\lambda)$ if the matrix $\boldsymbol{B}(\lambda)$ can be obtained from $\boldsymbol{A}(\lambda)$ by a finite number of elementary transformations. In this case, we write $\boldsymbol{A}(\lambda) \sim \boldsymbol{B}(\lambda)$.

Theorem 1. *λ-matrices $\mathbf{A}(\lambda)$, $\mathbf{B}(\lambda)$ of order n are equivalent if and only if there exist λ-matrices $\mathbf{C}(\lambda)$, $\mathbf{D}(\lambda)$ of order n such that their determinants are non-zero (real or complex) numbers and*

$$\mathbf{B}(\lambda) = \mathbf{C}(\lambda)\,\mathbf{A}(\lambda)\,\mathbf{D}(\lambda)\,.$$

Theorem 2. *Equivalent λ-matrices have the same rank* (the converse does not hold – see Theorem 6).

Theorem 3. *Two matrices $\mathbf{A}$, $\mathbf{B}$ of the same order whose elements are real or complex numbers are equivalent if and only if they have the same rank.*

Theorem 4. *A λ-matrix $\mathbf{A}(\lambda)$ of order n is equivalent to one and only one of the λ-matrices of the form*

$$\begin{bmatrix} E_1(\lambda), & 0, & 0, & \ldots, & 0 \\ 0, & E_2(\lambda), & 0, & \ldots, & 0 \\ 0, & 0, & E_3(\lambda), & \ldots, & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0, & 0, & 0, & \ldots, & E_n(\lambda) \end{bmatrix}, \tag{1}$$

where the polynomials $E_i(\lambda)$ are either zero polynomials or have the coefficient of the highest power of λ equal to 1, and the polynomial $E_{j+1}(\lambda)$ is divisible by the polynomial $E_j(\lambda)$ $(j = 1, 2, \ldots, n-1)$. If the rank of $\mathbf{A}(\lambda)$ is h, then $E_1(\lambda)\,E_2(\lambda)\ldots$ $\ldots E_h(\lambda) \not\equiv 0$, $E_{h+1}(\lambda) = E_{h+2}(\lambda) = \ldots = E_n(\lambda) \equiv 0$.

Definition 4. The polynomials $E_1(\lambda), \ldots, E_n(\lambda)$ are called *invariant factors of the matrix $\mathbf{A}(\lambda)$*; the form (1) is called the *rational canonical form of the matrix $\mathbf{A}(\lambda)$*.

Example 2. Rearrangements transforming the matrix of Example 1 to the rational canonical form reduce the given matrix successively to the following matrices:

$$\begin{bmatrix} 3-\lambda, & 1+\lambda \\ 1, & 5-\lambda \end{bmatrix} \sim \begin{bmatrix} 1, & 5-\lambda \\ 3-\lambda, & 1+\lambda \end{bmatrix} \sim \begin{bmatrix} 1, & 5-\lambda \\ 0, & (1+\lambda)+(5-\lambda)(-3+\lambda) \end{bmatrix} =$$

$$= \begin{bmatrix} 1, & 5-\lambda \\ 0, & -\lambda^2+9\lambda-14 \end{bmatrix} \sim \begin{bmatrix} 1, & 0 \\ 0, & -\lambda^2+9\lambda-14 \end{bmatrix} \sim \begin{bmatrix} 1, & 0 \\ 0, & \lambda^2-9\lambda+14 \end{bmatrix}.$$

First, we interchanged the rows; then we added the first row multiplied by $-(3-\lambda)$ to the second one; then we added to the second column the first one multiplied by $\lambda - 5$ and, finally, we multiplied the second row by the number -1. Thus, the invariant factors of the matrix are the polynomials $E_1(\lambda) = 1$, $E_2(\lambda) = \lambda^2 - 9\lambda + 14$.

Theorem 5. *The greatest common divisor $D_i(\lambda)$ (the so-called i-th determinant divisor) of all the i-rowed minors of a matrix $\mathbf{A}(\lambda)$ satisfies the relation*

$$D_i(\lambda) = cE_1(\lambda)\,E_2(\lambda)\ldots E_i(\lambda)\,,$$

where $E_j(\lambda)$ are the invariant factors of $\mathbf{A}(\lambda)$ and c is a non-zero real or complex number.

Theorem 6. *Two λ-matrices $\mathbf{A}(\lambda)$, $\mathbf{B}(\lambda)$ of the same order are equivalent if and only if they have the same invariant factors.*

Definition 5 (*Elementary Divisors of a λ-matrix*). Let a matrix $\mathbf{A}(\lambda)$ have the invariant factors $E_1(\lambda), \ldots, E_n(\lambda)$. We can factorize each of these polynomials in the variable λ into the product of powers of distinct linear factors $(\lambda - \alpha)^k$. Then every such power of a linear factor $(\lambda - \alpha)^k$ is called an *elementary divisor of the matrix* $\mathbf{A}(\lambda)$. The elementary divisors of the matrix $\mathbf{A}(\lambda)$ form the so-called *system of elementary divisors of the matrix* $\mathbf{A}(\lambda)$ (see Examples 3–5).

Example 3. Let a matrix $\mathbf{A}(\lambda)$ of order 5 have the invariant factors $E_1(\lambda) = 1$, $E_2(\lambda) = \lambda$, $E_3(\lambda) = \lambda(\lambda + 1)^2$, $E_4(\lambda) = \lambda^2(\lambda + 1)^2$, $E_5(\lambda) = 0$. Then the system of its elementary divisors is $\lambda, \lambda, \lambda^2, (\lambda + 1)^2, (\lambda + 1)^2$.

Example 4. The matrix of Example 2 has the elementary divisors $\lambda - 2$, $\lambda - 7$.

Theorem 7. *The invariant factors of a given matrix are uniquely determined by the order, rank and system of the elementary divisors.*

Example 5. Let us determine invariant factors of a matrix $\mathbf{A}(\lambda)$ of order 5 and rank 4 if the system of its elementary divisors is $\lambda - 1$, $\lambda - 1$, $(\lambda - 1)^2$, $(\lambda + 1)^2$.

In order to find the invariant factors, let us use Theorem 4. Since $h = 4$, we have $E_5(\lambda) = 0$. Now $E_1(\lambda)\, E_2(\lambda)\, E_3(\lambda)\, E_4(\lambda) = (\lambda - 1)\,(\lambda - 1)\,(\lambda - 1)^2\,(\lambda + 1)^2$. Since $E_1(\lambda)$, $E_2(\lambda)$, $E_3(\lambda)$ are divisors of $E_4(\lambda)$, we get immediately (using Definition 5) $E_4(\lambda) = (\lambda - 1)^2\,(\lambda + 1)^2$. Now $E_1(\lambda)\, E_2(\lambda)\, E_3(\lambda) = (\lambda - 1)\,(\lambda - 1)$, $E_1(\lambda)$, $E_2(\lambda)$ are divisors of $E_3(\lambda)$. Hence, $E_3(\lambda) = \lambda - 1$. Similarly, we find that $E_2(\lambda) = \lambda - 1$, and, finally, $E_1(\lambda) = 1$.

Theorem 8. *Two λ-matrices $\mathbf{A}(\lambda)$, $\mathbf{B}(\lambda)$ of order n are equivalent if and only if they have the same rank and the same system of elementary divisors.*

Theorem 9. *Let a λ-matrix $\mathbf{A}(\lambda)$ be partitioned into diagonal blocks:*

$$\mathbf{A}(\lambda) = \begin{bmatrix} \mathbf{A}_{11}(\lambda), & \mathbf{0}, & \ldots, & \mathbf{0} \\ \mathbf{0}, & \mathbf{A}_{22}(\lambda), & \ldots, & \mathbf{0} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{0}, & \mathbf{0}, & \ldots, & \mathbf{A}_{nn}(\lambda) \end{bmatrix}.$$

Then the system of elementary divisors of the matrix $\mathbf{A}(\lambda)$ is the collection of the systems of all elementary divisors of the diagonal blocks, i.e. of the λ-matrices $\mathbf{A}_{11}(\lambda), \ldots, \mathbf{A}_{nn}(\lambda)$.

1.28. Similar Matrices; the Characteristic Matrix and Characteristic Polynomial of a Matrix

Definition 1. We define two square matrices $\boldsymbol{A}$, $\boldsymbol{B}$ of order n, the elements of which are real or complex numbers, to be *similar* if a regular matrix $\boldsymbol{P}$ of order n exists, the elements of which are real or complex numbers respectively, for which the relation

$$\boldsymbol{B} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$$

holds.

Definition 2. By the *characteristic matrix* of a square matrix $\boldsymbol{A}$, we understand the λ-matrix $\lambda\boldsymbol{I} - \boldsymbol{A}$, where $\boldsymbol{I}$ is the identity matrix (p. 50). Thus, if

$$\boldsymbol{A} = \begin{bmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ a_{21}, & a_{22}, & \ldots, & a_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_{n1}, & a_{n2}, & \ldots, & a_{nn} \end{bmatrix},$$

then

$$\lambda\boldsymbol{I} - \boldsymbol{A} = \begin{bmatrix} \lambda - a_{11}, & -a_{12}, & \ldots, & -a_{1n} \\ -a_{21}, & \lambda - a_{22}, & \ldots, & -a_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ -a_{n1}, & -a_{n2}, & \ldots, & \lambda - a_{nn} \end{bmatrix}.$$

Definition 3. The determinant of the matrix $\lambda\boldsymbol{I} - \boldsymbol{A}$ is said to be the *characteristic polynomial of the matrix* $\boldsymbol{A}$. Its zeros are called the *eigenvalues* (or *characteristic values* or *characteristic numbers*) *of the matrix* $\boldsymbol{A}$.

Example 1. The characteristic polynomial of the matrix

$$\begin{bmatrix} 1, & 2 \\ -1, & 1 \end{bmatrix}$$

is

$$f(\lambda) = \begin{vmatrix} \lambda - 1, & -2 \\ 1, & \lambda - 1 \end{vmatrix} = (\lambda - 1)(\lambda - 1) + 2 = \lambda^2 - 2\lambda + 3\,;$$

its zeros $1 \pm \mathrm{i}\sqrt{2}$ are the eigenvalues of the given matrix. (On numerical methods for evaluation of eigenvalues see Chap. 30.)

Theorem 1. *The eigenvalues of an upper triangular matrix*

$$\boldsymbol{A} = \begin{bmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ 0, & a_{22}, & \ldots, & a_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ 0, & 0, & \ldots, & a_{nn} \end{bmatrix}$$

are equal to the elements in the principal diagonal, i.e. to the numbers $a_{11}, a_{22}, \ldots, a_{nn}$.

Theorem 2. *If the characteristic polynomial of a matrix* $\mathbf{A}$ *of order* n *has* n *simple zeros* $\alpha_1, \ldots, \alpha_n$, *then the system of elementary divisors of the characteristic matrix* $\lambda \mathbf{I} - \mathbf{A}$ *is* $\lambda - \alpha_1, \ldots, \lambda - \alpha_n$.

Theorem 3. *The product of all elementary divisors of the characteristic matrix* $\lambda \mathbf{I} - \mathbf{A}$ *of a given matrix* $\mathbf{A}$ *is equal to the characteristic polynomial of the matrix* $\mathbf{A}$.

Theorem 4. *Two matrices* $\mathbf{A}$, $\mathbf{B}$ *of the same order are similar if and only if their characteristic matrices* $\lambda \mathbf{I} - \mathbf{A}$, $\lambda \mathbf{I} - \mathbf{B}$ *are equivalent, i.e. if they have the same elementary divisors.*

Theorem 5. *Similar matrices have the same characteristic polynomial, and thus also the same eigenvalues.*

Theorem 6. *Similar matrices have the same traces* (see Definition 1.25.11, p. 53).

Definition 4. By a *Jordan block* (*of order* k) is meant a matrix of order k of the form

$$\begin{bmatrix} \varrho, & 1, & 0, & \ldots, & 0, & 0 \\ 0, & \varrho, & 1, & \ldots, & 0, & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0, & 0, & 0, & \ldots, & \varrho, & 1 \\ 0, & 0, & 0, & \ldots, & 0, & \varrho \end{bmatrix}, \tag{1}$$

where ϱ is a real or complex number.

A matrix decomposed into diagonal blocks (see Definition 1.26.2) which are Jordan blocks is called a *Jordan matrix.*

Theorem 7. *The characteristic matrix of the Jordan block* (1) *has the single elementary divisor* $(\lambda - \varrho)^k$. *Hence, using Theorem* 1.27.9, *one can determine the system of elementary divisors of the characteristic matrix of a given Jordan matrix.*

Example 2. The matrix

$$\mathbf{A} = \begin{bmatrix} 1, & 0, & 0 \\ 0, & 2, & 1 \\ 0, & 0, & 2 \end{bmatrix}$$

is a Jordan matrix; its Jordan blocks are the matrices $[1]$ and $\begin{bmatrix} 2, & 1 \\ 0, & 2 \end{bmatrix}$. Hence, the elementary divisors of the characteristic matrix $\lambda \mathbf{I} - \mathbf{A}$ are $\lambda - 1$, $(\lambda - 2)^2$.

Theorem 8. *Every square matrix* $\boldsymbol{A}$ *of order* n *is similar to a Jordan matrix of order* n; *if* $\boldsymbol{A}$ *is similar to two Jordan matrices, then these matrices differ only in the order of arrangement of their diagonal blocks.*

REMARK 1. The following two examples indicate the method if determining (at least theoretically) a Jordan matrix which is similar to a given matrix $\boldsymbol{A}$.

Example 3. Let

$$\boldsymbol{A} = \begin{bmatrix} 1, & 2, & 0 \\ 0, & 2, & 0 \\ -2, & -2, & -1 \end{bmatrix}.$$

The characteristic polynomial is

$$\begin{vmatrix} \lambda - 1, & -2, & 0 \\ 0, & \lambda - 2, & 0 \\ 2, & 2, & \lambda + 1 \end{vmatrix} = (\lambda - 1)(\lambda + 1)(\lambda - 2);$$

this polynomial has simple zeros $1, -1, 2$, and therefore, by Theorem 2, the system of elementary divisors of the characteristic matrix is $\lambda - 1$, $\lambda + 1$, $\lambda - 2$. According to Theorem 4, the characteristic matrix of the required Jordan matrix $\boldsymbol{B}$ has the same system of elementary divisors; hence we can find this Jordan matrix $\boldsymbol{B}$ by Theorem 7. The matrix $\boldsymbol{B}$ is decomposed into three Jordan blocks of order 1 corresponding to the elementary divisors $(\lambda - 1)$, $(\lambda + 1)$, $(\lambda - 2)$. Hence

$$\boldsymbol{B} = \begin{bmatrix} 1, & 0, & 0 \\ 0, & -1, & 0 \\ 0, & 0, & 2 \end{bmatrix}.$$

Example 4. Let

$$\boldsymbol{A} = \begin{bmatrix} 3, & 1, & -3 \\ -7, & -2, & 9 \\ -2, & -1, & 4 \end{bmatrix}.$$

The characteristic polynomial is

$$\begin{vmatrix} \lambda - 3, & -1, & 3 \\ 7, & \lambda + 2, & -9 \\ 2, & 1, & \lambda - 4 \end{vmatrix} = (\lambda - 1)(\lambda - 2)^2.$$

Since it does not have simple zeros, Theorem 2 cannot be applied. We therefore first determine invariant factors $E_1(\lambda)$, $E_2(\lambda)$, $E_3(\lambda)$ of the matrix $\lambda\boldsymbol{I} - \boldsymbol{A}$. Since their product is equal to the product of all the elementary divisors, Theorem 3 shows that it is equal to $(\lambda - 1)(\lambda - 2)^2$. Hence either $E_3(\lambda) = (\lambda - 1)(\lambda - 2)^2$, $E_2(\lambda) = E_1(\lambda) = 1$, or $E_3(\lambda) = (\lambda - 1)(\lambda - 2)$, $E_2(\lambda) = \lambda - 2, E_1(\lambda) = 1$ [see Example 1.27.5, p. 58]. Now, $E_1(\lambda)E_2(\lambda)$ is the greatest common divisor of all minors of

order 2 of the matrix

$$\lambda I - A = \begin{bmatrix} \lambda - 3, & -1, & 3 \\ 7, & \lambda + 2, & -9 \\ 2, & 1, & \lambda - 4 \end{bmatrix}.$$

One of these minors is $\begin{vmatrix} -1, & 3 \\ 1, & \lambda - 4 \end{vmatrix} = -\lambda + 1$. Thus, $E_2(\lambda)$ cannot be equal to $\lambda - 2$, for $\lambda - 2$ is not a factor of the binomial $\lambda - 1$. So $E_2(\lambda) = 1$ and the system of elementary divisors of the matrix $\lambda I - A$ consists of the polynomials $\lambda - 1$, $(\lambda - 2)^2$. The corresponding Jordan matrix is therefore

$$\begin{bmatrix} 1, & 0, & 0 \\ 0, & 2, & 1 \\ 0, & 0, & 2 \end{bmatrix}.$$

Alternative method: By elementary transformations (Definition 1.27.2, p. 56), we bring the matrix $\lambda I - A$ to the rational canonical form and then, by Theorem 1.27.9, we determine its elementary divisors.

Theorem 9. *The eigenvalues of a Hermitian matrix are real numbers.*

Theorem 10. *Let* A *be a symmetric matrix whose elements are real numbers. Then its eigenvalues are real numbers.*

Theorem 11. *Let* A *be a Hermitian matrix. Then there exists a unitary matrix* U *such that the matrix* $U^{-1}AU$ *is diagonal (and real).* $U^{-1}AU$ *is the Jordan form of the matrix* A.

Theorem 12. *Let* A *be a symmetric matrix the elements of which are real numbers. Then there exists a real orthogonal matrix* P *such that the matrix* $P^{-1}AP$ *is diagonal.* $P^{-1}AP$ *is the Jordan form of the matrix* A.

REMARK 2. A method for finding the matrix P of Theorem 12 is given in Example 1.29.3, p. 65.

Theorem 13. *Let* U *be a unitary matrix. Then there exists a unitary matrix* V *such that* $V^{-1}UV$ *is diagonal and the absolute value of each of its elements in the principal diagonal is* 1. $V^{-1}UV$ *is the Jordan form of the matrix* U.

1.29. Quadratic and Hermitian Forms

Definition 1. A *quadratic form in n variables* $x_1, x_2, \ldots, x_n$ is a polynomial of the form

$$\begin{aligned} f(x_1, \ldots, x_n) = a_{11}x_1^2 + 2a_{12}x_1x_2 + & \quad \ldots + 2a_{1n}x_1x_n + \\ + \; a_{22}x_2^2 \; + 2a_{23}x_2x_3 + \ldots + & \; 2a_{2n}x_2x_n + \\ + \; \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \; + a_{nn}x_n^2 \, , \end{aligned}$$

briefly

$$f(x_1, \ldots, x_n) = \sum_{i,j=1}^{n} a_{ij} x_i x_j \quad (a_{ij} = a_{ji}), \tag{1}$$

where a_{ij} are real or complex numbers. In the case, where the a_{ij} are real, we say that the form $f(x_1, \ldots, x_n)$ is *real.*

Definition 2. The symmetric matrix of order n

$$\boldsymbol{A} = \begin{bmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ a_{21}, & a_{22}, & \ldots, & a_{2n} \\ \ldots\ldots\ldots\ldots\ldots & & & \\ a_{n1}, & a_{n2}, & \ldots, & a_{nn} \end{bmatrix}$$

is called the *matrix of the quadratic form* (1), its rank being the *rank of the quadratic form* (1).

Definition 3 (*Linear Mapping*). The mapping

$$\begin{aligned} y_1 &= q_{11}x_1 + q_{12}x_2 + \ldots + q_{1n}x_n, \\ y_2 &= q_{21}x_1 + q_{22}x_2 + \ldots + q_{2n}x_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= q_{n1}x_1 + q_{n2}x_2 + \ldots + q_{nn}x_n, \end{aligned}$$

briefly

$$y_i = \sum_{j=1}^{n} q_{ij} x_j \quad (i = 1, 2, \ldots, n) \tag{2}$$

where q_{ij} are fixed real or complex numbers, assigns to every ordered n-tuple (i.e. n-component vector) $\boldsymbol{x} = (x_1, \ldots, x_n)$ an ordered n-tuple (i.e. n-component vector) $\boldsymbol{y} = (y_1, \ldots, y_n)$ and is called the *linear mapping* (*of the n-dimensional vector space* V_n *into itself*).

The matrix

$$\boldsymbol{Q} = \begin{bmatrix} q_{11}, & q_{12}, & \ldots, & q_{1n} \\ \ldots\ldots\ldots\ldots\ldots & & & \\ q_{n1}, & q_{n2}, & \ldots, & q_{nn} \end{bmatrix}$$

is called the *matrix of the linear mapping* (2). The mapping (2) is said to be *regular* if the matrix $\boldsymbol{Q}$ is regular.

Theorem 1. *If the mapping* (2) *is regular, then there exists an inverse linear mapping* $x_i = \sum_{j=1}^{n} p_{ij} y_j$ $(i = 1, \ldots, n)$, *whose matrix* $\boldsymbol{P}$ *is inverse to the matrix* $\boldsymbol{Q}$.

Theorem 2. *The composition of two linear mappings* $z_i = \sum_{j=1}^{n} r_{ij} y_j$, $y_i = \sum_{j=1}^{n} s_{ij} x_j$

with matrices $\mathbf{R}$, $\mathbf{S}$ *is a linear mapping* $z_i = \sum_{j=1}^{n} t_{ij}x_j$ *with matrix* $\mathbf{T} = \mathbf{RS}$. *If both mappings are regular, then the composite mapping is also regular.*

Definition 4. If in a quadratic form $f(x_1, \ldots, x_n)$ we substitute for the variables $x_1, \ldots, x_n$ the variables $y_1, \ldots, y_n$ by means of a linear mapping

$$x_i = \sum_{j=1}^{n} p_{ij}y_j \quad (i = 1, 2, \ldots, n), \tag{3}$$

we say that we apply the *linear substitution* (3) to $f(x_1, \ldots, x_n)$. If the mapping (3) is regular, then the corresponding linear substitution is said to be *regular*. If the numbers p_{ij} are real, then the substitution (3) is *real*.

REMARK 1 (*Matrix Notation for Linear Mappings and Quadratic Forms*). If we denote by $\mathbf{x}$ the n by 1 matrix $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ and by $\mathbf{y}$ the matrix $\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, then the linear mapping (3) can be written in the matrix form $\mathbf{x} = \mathbf{Py}$, where $\mathbf{P}$ is the matrix (of order n) of the mapping (3). Similarly, the quadratic form (1) can be written in the matrix form, $f(x_1, \ldots, x_n) = \mathbf{x'Ax}$ where $\mathbf{x'} = \begin{bmatrix} x_1, x_2, \ldots, x_n \end{bmatrix}$.

Example 1. In matrix notation, the form $x_1^2 - 4x_1x_2 + 2x_2^2$ is written

$$\mathbf{x'}\begin{bmatrix} 1, & -2 \\ -2, & 2 \end{bmatrix}\mathbf{x}, \quad \text{where} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

since

$$\begin{bmatrix} x_1, x_2 \end{bmatrix}\begin{bmatrix} 1, & -2 \\ -2, & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 - 2x_2, & -2x_1 + 2x_2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} =$$

$$= x_1^2 - 2x_1x_2 - 2x_1x_2 + 2x_2^2 .$$

Theorem 3. *A quadratic form* $f(x_1, \ldots, x_n) = \sum_{i,j=1}^{n} a_{ij}x_ix_j$ *with matrix* $\mathbf{A}$ *is transformed by a linear substitution* (3) (*i.e. by a substitution* $\mathbf{x} = \mathbf{Py}$) *into the form*

$$g(y_1, \ldots, y_n) = (\mathbf{Py})'\,\mathbf{A}(\mathbf{Py}) = \mathbf{y}'(\mathbf{P'AP})\,\mathbf{y} = \sum_{i,j=1}^{n} b_{ij}y_iy_j$$

with matrix $\mathbf{B} = \mathbf{P'AP}$, *where* $\mathbf{P}$ *is the matrix of the linear substitution* (3).

REMARK 2. Square matrices $\mathbf{A}$, $\mathbf{B}$ of order n are said to be *congruent* if there exists a regular matrix $\mathbf{P}$ such that $\mathbf{B} = \mathbf{P'AP}$.

Theorem 4. *The quadratic form* $g(y_1, \ldots, y_n)$ *obtained from a form* $f(x_1, \ldots, x_n)$ *by a regular linear substitution* $x_i = \sum_{j=1}^{n} p_{ij} y_j$ $(i = 1, \ldots, n)$ *has the same rank as the form* $f(x_1, \ldots, x_n)$.

Theorem 5. *For every (for every real) quadratic form* $f(x_1, \ldots, x_n)$ *of rank* h *there exists a regular linear substitution* (3) [*a real regular linear substitution* (3)] *which transforms the form* $f(x_1, \ldots, x_n)$ *into the form*

$$g(y_1, \ldots, y_n) = c_1 y_1^2 + c_2 y_2^2 + \ldots + c_n y_n^2 ,$$

where $c_1, \ldots, c_n$ *are complex (real) numbers, precisely* h *of which are non-zero.*

Example 2. A method of finding such a substitution will be shown in the following example: Let $f(x_1, x_2, x_3) = x_1 x_2 + 4x_2 x_3 - 2x_1 x_3$. Since the form does not contain the square of any variable, we transform it, first, by the regular linear substitution $x_1 = z_1 + z_2$, $x_2 = z_1 - z_2$, $x_3 = z_3$ into the form $h(z_1, z_2, z_3) = z_1^2 - z_2^2 + 2z_1 z_3 - 6z_2 z_3$. This can be rewritten in the form $h(z_1, z_2, z_3) = (z_1 + z_3)^2 - z_3^2 - z_2^2 - 6z_2 z_3$. We then apply the regular linear transformation $t_1 = z_1 + z_3, t_2 = z_2, t_3 = z_3$ thus obtaining the form $k(t_1, t_2, t_3) = t_1^2 - t_2^2 - t_3^2 - 6t_2 t_3 = t_1^2 - (t_2 + 3t_3)^2 + 9t_3^2 - t_3^2$ which, by means of the regular linear substitution $y_1 = t_1, y_2 = t_2 + 3t_3, y_3 = t_3$ is transformed into the form $g(y_1, y_2, y_3) = y_1^2 - y_2^2 + 8y_3^2$ as required. By combining the applied substitutions, we find that $f(x_1, x_2, x_3)$ is transformed into this final form by the regular linear substitution $x_1 = y_1 + y_2 - 4y_3$, $x_2 = y_1 - y_2 + 2y_3$, $x_3 = y_3$.

Theorem 6. *For every real quadratic form* $f(x_1, \ldots, x_n)$ *with a matrix* $\boldsymbol{A}$, *there exists a (real) regular linear substitution* (3) *with an orthogonal matrix* $\boldsymbol{P}$ *which transforms the form f into the form*

$$g(y) = \alpha_1 y_1^2 + \alpha_2 y_2^2 + \ldots + \alpha_n y_n^2 , \tag{4}$$

where $\alpha_1, \ldots, \alpha_n$ *are the eigenvalues of the matrix* $\boldsymbol{A}$ *(and are real* – see Theorem 1.28.10, p. 62).

Example 3. The problem is to find a transformation $x_i = \sum_{j=1}^{3} p_{ij} y_j$ $(i = 1, 2, 3)$ with an orthogonal matrix $\boldsymbol{P} = (p_{ij})$ which brings the given quadratic form $f(x_1, x_2, x_3) = 2x_1^2 + x_2^2 - 4x_1 x_2 - 4x_2 x_3$ into the form $g(y_1, y_2, y_3)$ given by (4). The matrix of the form $f(x_1, x_2, x_3)$ is

$$\boldsymbol{A} = (a_{ij}) = \begin{bmatrix} 2, & -2, & 0 \\ -2, & 1, & -2 \\ 0, & -2, & 0 \end{bmatrix};$$

its eigenvalues, i.e. the roots of the equation

$$\begin{vmatrix} \lambda - 2, & 2, & 0 \\ 2, & \lambda - 1, & 2 \\ 0, & 2, & \lambda \end{vmatrix} = 0$$

are $1, -2, 4$. According to Theorem 6, there exists a transformation $x_i = \sum_{j=1}^{3} p_{ij} y_j$ with an orthogonal matrix $\boldsymbol{P} = (p_{ij})$ which brings the form $f(x_1, x_2, x_3)$ into the form $g(y_1, y_2, y_3) = y_1^2 - 2y_2^2 + 4y_3^2$. The matrix $\boldsymbol{P}$ (and, thus, the required transformation) can be found as follows: If we denote by $\boldsymbol{B} = (b_{ij})$ the matrix of the form $g(y_1, y_2, y_3)$, i.e.

$$\boldsymbol{B} = \begin{bmatrix} 1, & 0, & 0 \\ 0, & -2, & 0 \\ 0, & 0, & 4 \end{bmatrix},$$

then, by Theorem 3, $\boldsymbol{B} = \boldsymbol{P}'\boldsymbol{A}\boldsymbol{P}$. Since $\boldsymbol{P}$ is orthogonal, i.e. by Definition 1.25.7, $\boldsymbol{P}' = \boldsymbol{P}^{-1}$, the equality $\boldsymbol{B} = \boldsymbol{P}'\boldsymbol{A}\boldsymbol{P}$ can be written in the form $\boldsymbol{P}\boldsymbol{B} = \boldsymbol{A}\boldsymbol{P}$. This means that $\sum_{l=1}^{3} a_{il} p_{lk} = \sum_{j=1}^{3} p_{ij} b_{jk}$ $(i, k = 1, 2, 3)$. In our problem, we thus get the following equations:

$$\text{(a)} \quad \begin{aligned} 2p_{11} - 2p_{21} &= p_{11}, \\ -2p_{11} + p_{21} - 2p_{31} &= p_{21}, \\ -2p_{21} &= p_{31}; \end{aligned} \qquad \text{(b)} \quad \begin{aligned} 2p_{12} - 2p_{22} &= -2p_{12}, \\ -2p_{12} + p_{22} - 2p_{23} &= -2p_{22}, \\ -2p_{22} &= -2p_{32}; \end{aligned}$$

$$\text{(c)} \quad \begin{aligned} 2p_{13} - 2p_{23} &= 4p_{13}, \\ -2p_{13} + p_{23} - 2p_{33} &= 4p_{23}, \\ -2p_{23} &= 4p_{33}. \end{aligned}$$

These are three systems of homogeneous equations, each of them being of rank 2. Solving, for example, the system a) we can confine ourselves to the first and third equations from which there follows

$$p_{11} : p_{21} : p_{31} = \begin{vmatrix} -2, & 0 \\ -2, & -1 \end{vmatrix} : \begin{vmatrix} 0, & 1 \\ -1, & 0 \end{vmatrix} : \begin{vmatrix} 1, & -2 \\ 0, & -2 \end{vmatrix} = 2 : 1 : -2.$$

Since $p_{11}^2 + p_{21}^2 + p_{31}^2 = 1$ (see Theorem 1.25.12, p. 52), we get

$$p_{11} = \frac{\pm 2}{\sqrt{(2^2 + 1^2 + 2^2)}} = \pm\tfrac{2}{3}, \quad p_{21} = \pm\tfrac{1}{3}, \quad p_{31} = \mp\tfrac{2}{3}.$$

Similarly, we find that $p_{12} = \pm\frac{1}{3}$, $p_{22} = \pm\frac{2}{3}$, $p_{32} = \pm\frac{2}{3}$, $p_{13} = \pm\frac{2}{3}$, $p_{23} = \mp\frac{2}{3}$, $p_{33} = \pm\frac{1}{3}$. Thus, the matrix $\boldsymbol{P}$ can be chosen as follows:

$$\boldsymbol{P} = \begin{bmatrix} \frac{2}{3}, & \frac{1}{3}, & \frac{2}{3} \\ \frac{1}{3}, & \frac{2}{3}, & -\frac{2}{3} \\ -\frac{2}{3}, & \frac{2}{3}, & \frac{1}{3} \end{bmatrix}.$$

Theorem 7 (*Sylvester's Law of Inertia*). *Any real quadratic form* $f(x_1, \ldots, x_n)$ *of rank* h *can be transformed by a real regular linear substitution into the form*

$$g(y_1, \ldots, y_n) = y_1^2 + y_2^2 + \ldots + y_{s_1}^2 - y_{s_1+1}^2 - \ldots - y_{s_1+s_2}^2 \quad (s_1 + s_2 = h). \quad (5)$$

The substitution transforming $f(x_1, \ldots, x_n)$ *into the form* (5) *is not unique; however, the number* s_1 *of positive signs as well as the number* $s_2 = h - s_1$ *of the negative sings in the resulting form is always the same.*

Definition 5. The number $s_1 - s_2$ in Theorem 7 is called the *signature of the form* $f(x_1, \ldots, x_n)$.

Theorem 8. *A real quadratic form* $f(x_1, \ldots, x_n)$ *can be transformed by a real regular linear substitution into a form* $g(y_1, \ldots, y_n)$ *if and only if both forms have the same rank and signature.*

Definition 6. Let $f(x_1, \ldots, x_n)$ be a real quadratic form.

(a) The form $f(x_1, \ldots, x_n)$ is called *positive* (or *negative*) *definite* if, for every non-zero n-tuple $(\alpha_1, \ldots, \alpha_n)$ of real numbers $\alpha_1, \ldots, \alpha_n$ (briefly: for any real non-zero n-tuple), the number $f(\alpha_1, \ldots, \alpha_n)$ is positive (or negative).

(b) The form $f(x_1, \ldots, x_n)$ is called *positive* (or *negative*) *semidefinite* if, for every non-zero real n-tuple $(\alpha_1, \ldots, \alpha_n)$, the inequality $f(\alpha_1, \ldots, \alpha_n) \geqq 0$ [or $f(\alpha_1, \ldots, \alpha_n) \leqq 0$] holds and at the same time there exist non-zero real n-tuples $(\beta_1, \ldots, \beta_n)$ such that $f(\beta_1, \ldots, \beta_n) = 0$.

(c) The form $f(x_1, \ldots, x_n)$ is said to be *indefinite* if there are non-zero real n-tuples $(\alpha_1, \ldots, \alpha_n)$ and $(\beta_1, \ldots, \beta_n)$ such that $f(\alpha_1, \ldots, \alpha_n) > 0$ and $f(\beta_1, \ldots, \beta_n) < 0$.

REMARK 3. The matrix of a positive, or negative, definite quadratic form is called *positive* (or *negative*) *definite*. In the following theorems, some conditions for a symmetric matrix to be positive definite are introduced.

Theorem 9. *Let* $f(x_1, \ldots, x_n)$ *be a real quadratic form of rank* h *and signature* s.

1. $f(x_1, \ldots, x_n)$ *is positive* (*or negative*) *definite if and only if* $h = n$ *and* $s = n$ (*or* $s = -n$); *the form can be transformed by a real regular linear substitution into a sum of positive* (*or negative*) *squares of all* n *variables.*

2. $f(x_1, \ldots, x_n)$ *is positive* (*or negative*) *semidefinite if and only if* $h < n$ *and* $s = h$ (*or* $s = -h$).

3. $f(x_1, \ldots, x_n)$ *is indefinite if* $-h < s < h$.

REMARK 4. If a form $f(x_1, \ldots, x_n)$ is positive definite, or semidefinite, then the form $-f(x_1, \ldots, x_n)$ is obviously negative definite, or semidefinite, respectively. Therefore, we can restrict our consideration to positive definite or positive semidefinite forms.

Theorem 10. *Let $f(x_1, \ldots, x_n)$ be a real quadratic form and let $\mathbf{A}$ be its matrix. The form $f(x_1, \ldots, x_n)$ is positive definite, or semidefinite, if and only if all the eigenvalues of the matrix $\mathbf{A}$ are positive, or non-negative, respectively.*

Theorem 11. *A real quadratic form $f(x_1, \ldots, x_n)$ with a matrix*

$$\begin{bmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ \ldots\ldots\ldots\ldots\ldots \\ a_{n1}, & a_{n2}, & \ldots, & a_{nn} \end{bmatrix}$$

is positive definite if and only if all the principal minors

$$|a_{11}|, \begin{vmatrix} a_{11}, & a_{12} \\ a_{21}, & a_{22} \end{vmatrix}, \begin{vmatrix} a_{11}, & a_{12}, & a_{13} \\ a_{21}, & a_{22}, & a_{23} \\ a_{31}, & a_{32}, & a_{33} \end{vmatrix}, \ldots, \begin{vmatrix} a_{11}, & \ldots, & a_{1n} \\ \ldots\ldots\ldots \\ a_{n1}, & \ldots, & a_{nn} \end{vmatrix}$$

of the matrix $\mathbf{A}$ are positive.

Example 4. *The form* $a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2$ is positive definite if and only if $a_{11} > 0$, $a_{11}a_{22} - a_{12}^2 > 0$; it is semidefinite if $a_{11}a_{22} - a_{12}^2 = 0$; it is indefinite, if $a_{11}a_{22} - a_{12}^2 < 0$.

Definition 7. A *Hermitian (quadratic) form in n variables* $x_1, \ldots, x_n$ is a polynomial of the form

$$f(x_1, x_2, \ldots, x_n) = \sum_{i,j=1}^{n} a_{ij}x_i\bar{x}_j, \quad a_{ij} = \bar{a}_{ji} \quad (i, j = 1, \ldots, n), \tag{6}$$

where the bar indicates a conjugate complex number. The matrix

$$\begin{bmatrix} a_{11}, & \ldots, & a_{1n} \\ \ldots\ldots\ldots \\ a_{n1}, & \ldots, & a_{nn} \end{bmatrix}$$

of a Hermitian form is a Hermitian matrix, i.e. $\mathbf{A} = \bar{\mathbf{A}}'$ holds.

Theorem 12. *A Hermitian form with a matrix $\mathbf{A}$ is transformed by a linear substitution $x_i = \sum p_{ij}y_j$ into the Hermitian form with the matrix $\mathbf{B} = \mathbf{P}'\mathbf{A}\bar{\mathbf{P}}$.*

REMARK 5. Matrices $\mathbf{A}$, $\mathbf{B}$ are said to be *conjunctive (Hermitian congruent)* if there exists a regular matrix $\mathbf{P}$ such that $\mathbf{B} = \mathbf{P}'\mathbf{A}\bar{\mathbf{P}}$.

Theorem 13. *If $(\alpha_1, \ldots, \alpha_n)$ is an arbitrary n-tuple of real or complex numbers and if $f(x_1, \ldots, x_n)$ is a Hermitian form, then the number $f(\alpha_1, \ldots, \alpha_n)$ is real.*

REMARK 6. In the same way as for real quadratic forms, we define the rank and signature of a Hermitian form, and also positive (negative) definite, semidefinite and indefinite Hermitian forms (see Definitions 2, 5 and 6). Theorems formulated for real quadratic forms hold also for Hermitian forms; in such formulation, instead of real regular linear substitutions we have complex regular linear substitutions and in Theorem 6 we must replace "orthogonal matrix $\mathbf{P}$" by "unitary matrix $\mathbf{P}$".

2. TRIGONOMETRIC AND INVERSE TRIGONOMETRIC FUNCTIONS. HYPERBOLIC AND INVERSE HYPERBOLIC FUNCTIONS

By VÁCLAV VILHELM

References: [26], [43], [56], [59], [68], [103], [125], [126], [135], [185].

2.1. Measurement of Angles (Measurement by Degrees and Circular Measure)

If theoretical problems are under consideration, angles are not measured in degrees, but in radians (circular measure): The magnitude of an angle α is given by the length l of the arc, intercepted by the arms of the angle α on the unit circle with centre at the vertex of the angle (Fig. 2.1). We shall denote the magnitude of the angle α in circular measure again by α; sometimes, instead of α, the notation arc α° is employed, α° denoting the magnitude of the angle α expressed in degrees (in the sexagesimal system).

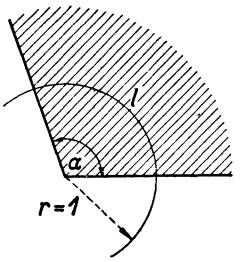


Fig. 2.1.

Theorem 1. *The relationship between circular measure and degrees is*

$$\alpha = \operatorname{arc} \alpha^\circ = \frac{\pi}{180^\circ} \alpha^\circ .$$

Example 1. The angle of 90° is in circular measure

$$\alpha = \frac{\pi}{180^\circ} . 90^\circ = \tfrac{1}{2}\pi .$$

Definition 1. The angle ϱ, whose circular measure is 1, is called the *radian*; its magnitude measured in degrees (in the sexagesimal system) is

$$\varrho^\circ = 180^\circ/\pi \doteq 57{\cdot}295\,779\,5^\circ = 57^\circ 17'44{\cdot}806'' .$$

In centesimal measure,

$$\varrho^g = 400^g/2\pi \doteq 63{\cdot}661\,977^g \text{ (grades)} .$$

In particular,

$$360^\circ = 2\pi\,,\ 180^\circ = \pi\,,\ 90^\circ = \tfrac{1}{2}\pi\,,\ 60^\circ = \tfrac{1}{3}\pi\,,\ 45^\circ = \tfrac{1}{4}\pi\,,\ 30^\circ = \tfrac{1}{6}\pi\,.$$

2.2. Definition of Trigonometric Functions

Definition 1. The trigonometric functions of an angle α in the interval $[0, 2\pi)$ are defined by means of a unit circle or (for acute angles) by means of a right angled triangle (Fig. 2.2) as follows:

$$\begin{aligned}
&\sin\alpha = \widetilde{AB} && (\text{for } 0 < \alpha < \tfrac{1}{2}\pi,\ \sin\alpha = a/c)\,,\\
&\cos\alpha = \widetilde{OA} && (\text{for } 0 < \alpha < \tfrac{1}{2}\pi,\ \cos\alpha = b/c)\,,\\
&\tan\alpha = \sin\alpha/\cos\alpha,\ \alpha \neq \tfrac{1}{2}\pi, \tfrac{3}{2}\pi && (\text{for } 0 < \alpha < \tfrac{1}{2}\pi,\ \tan\alpha = \widetilde{CD} = a/b)\,,\\
&\cot\alpha = \cos\alpha/\sin\alpha,\ \alpha \neq 0,\ \pi && (\text{for } 0 < \alpha < \tfrac{1}{2}\pi,\ \cot\alpha = \widetilde{EF} = b/a)\,,\\
&\sec\alpha = 1/\cos\alpha,\ \alpha \neq \tfrac{1}{2}\pi, \tfrac{3}{2}\pi && (\text{for } 0 < \alpha < \tfrac{1}{2}\pi,\ \sec\alpha = c/b)\,,\\
&\operatorname{cosec}\alpha = 1/\sin\alpha,\ \alpha \neq 0,\ \pi && (\text{for } 0 < \alpha < \tfrac{1}{2}\pi,\ \operatorname{cosec}\alpha = c/a)\,.
\end{aligned}$$

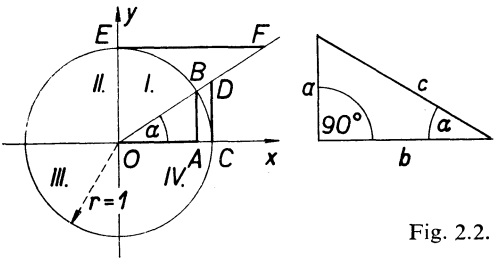


Fig. 2.2.

Here, $\widetilde{AB}$ is the directed length of the segment AB, i.e. $\widetilde{AB} > 0$ if AB is in the same direction and $\widetilde{AB} < 0$ if AB is in the opposite direction to the positive direction of the y-axis. The other lengths are used with a similar meaning (for example $\widetilde{OA} > 0$ if OA is in the same direction as the positive direction of the x-axis).

Further we define:

Definition 2.

$$\sin(2k\pi + \alpha) = \sin\alpha\,,\quad \cos(2k\pi + \alpha) = \cos\alpha\,, \tag{1}$$

$$\tan(k\pi + \alpha) = \tan\alpha\,,\quad \cot(k\pi + \alpha) = \cot\alpha \tag{2}$$

for an arbitrary integer k. In this way, the functions $\sin \alpha$ and $\cos \alpha$ are defined for all real α, the function $\tan \alpha$ for all real α different from $\frac{1}{2}\pi + k\pi$ and the function $\cot \alpha$ for all real α different from $k\pi$.

REMARK 1. The functions $\sin \alpha$ and $\cos \alpha$ are periodic functions with period 2π; the functions $\tan \alpha$ and $\cot \alpha$ are periodic functions with period π.

REMARK 2. In the case where an angle is measured in radians, instead of α we often write the letter x as is usual in the case of functions, where the letter x stands for the independent variable; we thus speak about the functions $\sin x$, $\cos x$, $\tan x$, $\cot x$.

2.3. Behaviour of Trigonometric Functions. Their Fundamental Properties

REMARK 1. In Fig. 2.3, x denotes the angle measured in radians; the figure represents the graph of the functions $\sin x$, $\cos x$, $\tan x$, $\cot x$ for x in the interval $[-\pi, 2\pi]$.

Fundamental properties:

1.

$$
\begin{aligned}
&-1 \leqq \sin \alpha \leqq 1, \\
&-1 \leqq \cos \alpha \leqq 1, \\
&-\infty < \tan \alpha < +\infty, \\
&-\infty < \cot \alpha < +\infty.
\end{aligned}
$$

2.

$$
\begin{aligned}
\sin(-\alpha) &= -\sin \alpha; \\
\cos(-\alpha) &= \cos \alpha; \\
\tan(-\alpha) &= -\tan \alpha; \\
\cot(-\alpha) &= -\cot \alpha.
\end{aligned}
$$

2.4. Relations Among Trigonometric Functions of the Same Angle

1.

$$\sin^2 \alpha + \cos^2 \alpha = 1; \quad \tan \alpha = \frac{\sin \alpha}{\cos \alpha}, \quad \cot \alpha = \frac{\cos \alpha}{\sin \alpha},$$

$$\sec \alpha = \frac{1}{\cos \alpha}, \quad \operatorname{cosec} \alpha = \frac{1}{\sin \alpha},$$

$$1 + \tan^2 \alpha = \frac{1}{\cos^2 \alpha}, \quad 1 + \cot^2 \alpha = \frac{1}{\sin^2 \alpha}.$$

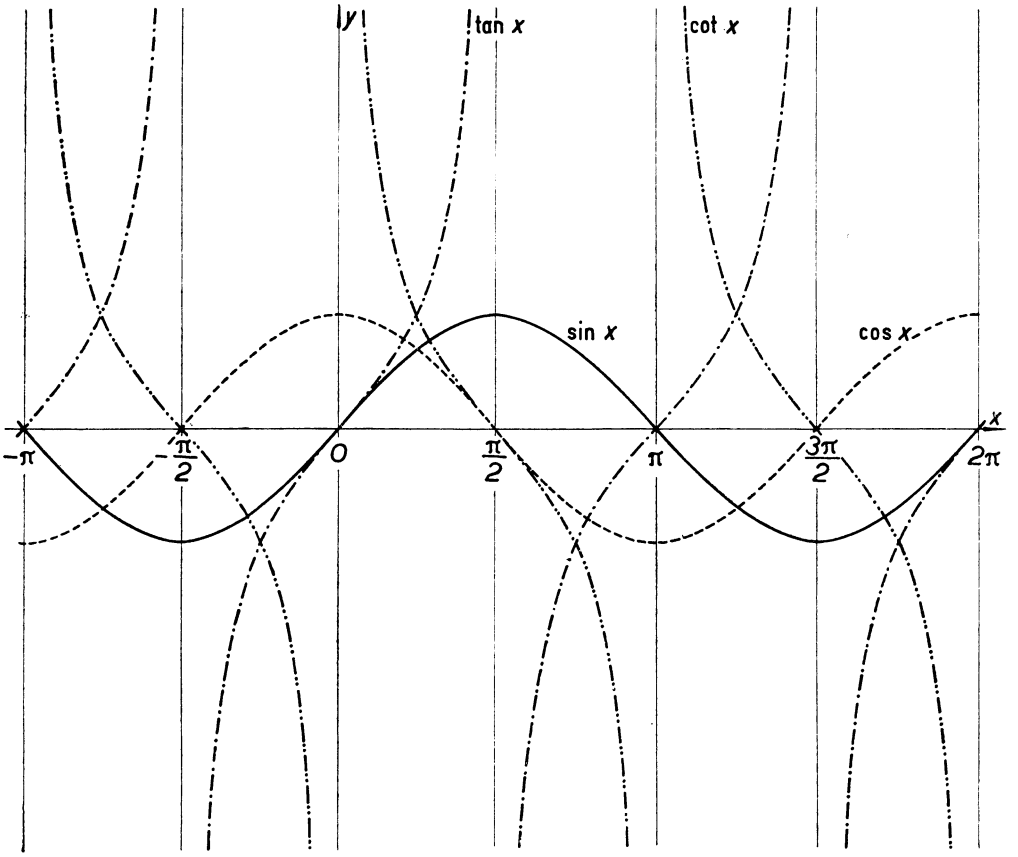


Fig. 2.3.

TABLE 2.1

Signs of trigonometric functions in individual quadrants

Function	Quadrant			
	I	II	III	IV
sin α	+	+	−	−
cos α	+	−	−	+
tan α	+	−	+	−
cot α	+	−	+	−

TABLE 2.2 *Values of trigonometric functions for some special angles*

Degrees Radians	0° 0	30° $\frac{1}{6}\pi$	45° $\frac{1}{4}\pi$	60° $\frac{1}{3}\pi$	90° $\frac{1}{2}\pi$	120° $\frac{2}{3}\pi$	135° $\frac{3}{4}\pi$	150° $\frac{5}{6}\pi$
$\sin\alpha$	0	$\frac{1}{2}$	$\frac{1}{2}\sqrt{2}$	$\frac{1}{2}\sqrt{3}$	1	$\frac{1}{2}\sqrt{3}$	$\frac{1}{2}\sqrt{2}$	$\frac{1}{2}$
$\cos\alpha$	1	$\frac{1}{2}\sqrt{3}$	$\frac{1}{2}\sqrt{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\frac{1}{2}\sqrt{2}$	$-\frac{1}{2}\sqrt{3}$
$\tan\alpha$	0	$\frac{1}{3}\sqrt{3}$	1	$\sqrt{3}$		$-\sqrt{3}$	-1	$-\frac{1}{3}\sqrt{3}$
$\cot\alpha$		$\sqrt{3}$	1	$\frac{1}{3}\sqrt{3}$	0	$-\frac{1}{3}\sqrt{3}$	-1	$-\sqrt{3}$

Degrees Radians	180° π	210° $\frac{7}{6}\pi$	225° $\frac{5}{4}\pi$	240° $\frac{4}{3}\pi$	270° $\frac{3}{2}\pi$	300° $\frac{5}{3}\pi$	315° $\frac{7}{4}\pi$	330° $\frac{11}{6}\pi$
$\sin\alpha$	0	$-\frac{1}{2}$	$-\frac{1}{2}\sqrt{2}$	$-\frac{1}{2}\sqrt{3}$	-1	$-\frac{1}{2}\sqrt{3}$	$-\frac{1}{2}\sqrt{2}$	$-\frac{1}{2}$
$\cos\alpha$	-1	$-\frac{1}{2}\sqrt{3}$	$-\frac{1}{2}\sqrt{2}$	$-\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}\sqrt{2}$	$\frac{1}{2}\sqrt{3}$
$\tan\alpha$	0	$\frac{1}{3}\sqrt{3}$	1	$\sqrt{3}$		$-\sqrt{3}$	-1	$-\frac{1}{3}\sqrt{3}$
$\cot\alpha$		$\sqrt{3}$	1	$\frac{1}{3}\sqrt{3}$	0	$-\frac{1}{3}\sqrt{3}$	-1	$-\sqrt{3}$

TABLE 2.3 *Reduction of trigonometric functions to the first quadrant*

Function	$\beta = 90° \pm \alpha$	$\beta = 180° \pm \alpha$	$\beta = 270° \pm \alpha$	$\beta = 360° \pm \alpha$
$\sin\beta$	$+\cos\alpha$	$\mp\sin\alpha$	$-\cos\alpha$	$\pm\sin\alpha$
$\cos\beta$	$\mp\sin\alpha$	$-\cos\alpha$	$\pm\sin\alpha$	$+\cos\alpha$
$\tan\beta$	$\mp\cot\alpha$	$\pm\tan\alpha$	$\mp\cot\alpha$	$\pm\tan\alpha$
$\cot\beta$	$\mp\tan\alpha$	$\pm\cot\alpha$	$\mp\tan\alpha$	$\pm\cot\alpha$

2.

$$|\sin\alpha| = \sqrt{(1-\cos^2\alpha)} = \frac{|\tan\alpha|}{\sqrt{(1+\tan^2\alpha)}} = \frac{1}{\sqrt{(1+\cot^2\alpha)}};$$

$$|\cos\alpha| = \sqrt{(1-\sin^2\alpha)} = \frac{1}{\sqrt{(1+\tan^2\alpha)}} = \frac{|\cot\alpha|}{\sqrt{(1+\cot^2\alpha)}};$$

$$|\tan\alpha| = \frac{|\sin\alpha|}{\sqrt{(1-\sin^2\alpha)}} = \frac{\sqrt{(1-\cos^2\alpha)}}{|\cos\alpha|} = \frac{1}{|\cot\alpha|};$$

$$\tan\alpha = \frac{1}{\cot\alpha}; \quad \cot\alpha = \frac{1}{\tan\alpha}.$$

REMARK 1. The absolute value must be used in relations 2, since, for example, $\sin 30° = \sqrt{(1-\cos^2 30°)}$, but $\sin 270° = -\sqrt{(1-\cos^2 270°)}$. For a definite α we have $\sin\alpha = \sqrt{(1-\cos^2\alpha)}$ or $\sin\alpha = -\sqrt{(1-\cos^2\alpha)}$ according to the sign of $\sin\alpha$ in the corresponding quadrant (Table 2.1).

Similarly for the other formulae in which the absolute values occur.

2.5. The Addition Formulae, the Multiple-angle and Half-angle Formulae

1.

$$\sin(\alpha \pm \beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta;$$

$$\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta;$$

$$\tan(\alpha \pm \beta) = \frac{\tan\alpha \pm \tan\beta}{1 \mp \tan\alpha\tan\beta};$$

$$\cot(\alpha \pm \beta) = \frac{\cot\alpha\cot\beta \mp 1}{\cot\beta \pm \cot\alpha}.$$

2. $\sin n\alpha$, $\cos n\alpha$, for n a natural number can be determined by De Moivre's theorem (Theorem 1.6.6, p. 11),

$$\cos n\alpha + \mathrm{i}\sin n\alpha = (\cos\alpha + \mathrm{i}\sin\alpha)^n = \sum_{k=0}^{n}\binom{n}{k}\cos^k\alpha\,(\mathrm{i}\sin\alpha)^{n-k}.$$

3.

$$\sin 2\alpha = 2\sin\alpha\cos\alpha; \quad \sin 3\alpha = 3\sin\alpha - 4\sin^3\alpha;$$

$$\sin n\alpha = n\sin\alpha\cos^{n-1}\alpha - \binom{n}{3}\sin^3\alpha\cos^{n-3}\alpha + \binom{n}{5}\sin^5\alpha\cos^{n-5}\alpha - \ldots.$$

4.

$$\cos 2\alpha = \cos^2\alpha - \sin^2\alpha\,; \quad \cos 3\alpha = 4\cos^3\alpha - 3\cos\alpha\,;$$

$$\cos n\alpha = \cos^n\alpha - \binom{n}{2}\sin^2\alpha\cos^{n-2}\alpha + \binom{n}{4}\sin^4\alpha\cos^{n-4}\alpha - \ldots.$$

5.

$$\tan 2\alpha = \frac{2\tan\alpha}{1-\tan^2\alpha}\,; \quad \tan 3\alpha = \frac{3\tan\alpha - \tan^3\alpha}{1-3\tan^2\alpha}\,;$$

$$\tan n\alpha = \frac{n\tan\alpha - \binom{n}{3}\tan^3\alpha + \binom{n}{5}\tan^5\alpha - \ldots}{1 - \binom{n}{2}\tan^2\alpha + \binom{n}{4}\tan^4\alpha - \binom{n}{6}\tan^6\alpha + \ldots}.$$

6.

$$\cot 2\alpha = \frac{\cot^2\alpha - 1}{2\cot\alpha}\,; \quad \cot 3\alpha = \frac{\cot^3\alpha - 3\cot\alpha}{3\cot^2\alpha - 1}\,;$$

$$\cot n\alpha = \frac{\cot^n\alpha - \binom{n}{2}\cot^{n-2}\alpha + \binom{n}{4}\cot^{n-4}\alpha - \ldots}{n\cot^{n-1}\alpha - \binom{n}{3}\cot^{n-3}\alpha + \binom{n}{5}\cot^{n-5}\alpha - \ldots}.$$

7.

$$\left|\sin\frac{\alpha}{2}\right| = \sqrt{\left[\tfrac{1}{2}(1-\cos\alpha)\right]}\,; \quad \left|\tan\frac{\alpha}{2}\right| = \sqrt{\frac{1-\cos\alpha}{1+\cos\alpha}}\,;$$

$$\tan\frac{\alpha}{2} = \frac{1-\cos\alpha}{\sin\alpha} = \frac{\sin\alpha}{1+\cos\alpha}\,;$$

$$\left|\cos\frac{\alpha}{2}\right| = \sqrt{\left[\tfrac{1}{2}(1+\cos\alpha)\right]}\,; \quad \left|\cot\frac{\alpha}{2}\right| = \sqrt{\frac{1+\cos\alpha}{1-\cos\alpha}}\,;$$

$$\cot\frac{\alpha}{2} = \frac{1+\cos\alpha}{\sin\alpha} = \frac{\sin\alpha}{1-\cos\alpha}.$$

8.

$$\sin\alpha = \frac{2\tan\frac{1}{2}\alpha}{1+\tan^2\frac{1}{2}\alpha}\,; \quad \cos\alpha = \frac{1-\tan^2\frac{1}{2}\alpha}{1+\tan^2\frac{1}{2}\alpha}.$$

2.6. Sum, Difference, Product of Trigonometric Functions, Powers of Trigonometric Functions

1.

$$\sin\alpha + \sin\beta = 2\sin\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2};$$

$$\sin\alpha - \sin\beta = 2\cos\frac{\alpha+\beta}{2}\sin\frac{\alpha-\beta}{2};$$

$$\cos\alpha + \cos\beta = 2\cos\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2};$$

$$\cos\alpha - \cos\beta = -2\sin\frac{\alpha+\beta}{2}\sin\frac{\alpha-\beta}{2};$$

$$\tan\alpha \pm \tan\beta = \frac{\sin(\alpha \pm \beta)}{\cos\alpha\cos\beta};$$

$$\cot\alpha \pm \cot\beta = \frac{\sin(\beta \pm \alpha)}{\sin\alpha\sin\beta};$$

$$\tan\alpha \pm \cot\beta = \pm\frac{\cos(\alpha \mp \beta)}{\cos\alpha\sin\beta}.$$

2.

$$\sin\alpha\sin\beta = \tfrac{1}{2}[\cos(\alpha-\beta) - \cos(\alpha+\beta)];$$

$$\cos\alpha\cos\beta = \tfrac{1}{2}[\cos(\alpha-\beta) + \cos(\alpha+\beta)];$$

$$\sin\alpha\cos\beta = \tfrac{1}{2}[\sin(\alpha-\beta) + \sin(\alpha+\beta)];$$

$$\tan\alpha\tan\beta = \frac{\tan\alpha+\tan\beta}{\cot\alpha+\cot\beta}; \quad \cot\alpha\cot\beta = \frac{\cot\alpha+\cot\beta}{\tan\alpha+\tan\beta};$$

$$\tan\alpha\cot\beta = \frac{\tan\alpha+\cot\beta}{\tan\beta+\cot\alpha}.$$

3.

$$\sin^2\alpha = \tfrac{1}{2}(1-\cos 2\alpha); \quad \sin^3\alpha = \tfrac{1}{4}(3\sin\alpha - \sin 3\alpha);$$

$$\cos^2\alpha = \tfrac{1}{2}(1+\cos 2\alpha); \quad \cos^3\alpha = \tfrac{1}{4}(\cos 3\alpha + 3\cos\alpha);$$

$$\sin^4\alpha = \tfrac{1}{8}(\cos 4\alpha - 4\cos 2\alpha + 3); \quad \cos^4\alpha = \tfrac{1}{8}(\cos 4\alpha + 4\cos 2\alpha + 3).$$

REMARK 1. Higher powers can be found by De Moivre's theorem (see relations 2, 3 and 4 of the previous § 2.5).

2.7. Trigonometric Sums

Theorem 1. For an arbitrary real α, an arbitrary real $x \neq 2k\pi$ (k being an integer) and for n a natural number we have

$$\sin x + \sin 2x + \ldots + \sin nx = \frac{\sin \frac{1}{2}nx}{\sin \frac{1}{2}x} \sin \tfrac{1}{2}(n+1)x ;$$

$$\cos x + \cos 2x + \ldots + \cos nx = \frac{\sin \frac{1}{2}nx}{\sin \frac{1}{2}x} \cos \tfrac{1}{2}(n+1)x ;$$

$$\sum_{j=1}^{n} \sin(\alpha + jx) = \frac{\sin \frac{1}{2}nx}{\sin \frac{1}{2}x} \sin\left[\alpha + \tfrac{1}{2}(n+1)x\right] ;$$

$$\sum_{j=1}^{n} \cos(\alpha + jx) = \frac{\sin \frac{1}{2}nx}{\sin \frac{1}{2}x} \cos\left[\alpha + \tfrac{1}{2}(n+1)x\right] .$$

2.8. Trigonometric Equations

Trigonometric equations are equations in the unknown x of the form

$$f(\cos x, \sin x, \tan x, \cot x, x) = 0 . \tag{1}$$

A trigonometric equation can be solved either by employing numerical methods (see Chap. 31), or, in some simple cases, by rearranging the equation using suitable formulae, to contain only one trigonometric function; then we solve the equation for this function.

Example 1. $\sin x - \cos^2 x + \frac{1}{4} = 0$; we rearrange the equation by means of the relation $\cos^2 x = 1 - \sin^2 x$ and put $y = \sin x$. We thus obtain the equation $y^2 + y - \frac{3}{4} = 0$ with the roots $y_1 = \frac{1}{2}$, $y_2 = -\frac{3}{2}$. There is no real solution corresponding to the root y_2 (since $|\sin x| \leqq 1$); the root $y_1 = \frac{1}{2}$ gives the solutions $x = \frac{1}{6}\pi + 2k\pi$, $x = \frac{5}{6}\pi + 2k\pi$ (k being any integer).

Example 2. $a \cos x + b \sin x = c$ $(ab \neq 0)$. We put $a = r \cos \lambda$, $b = r \sin \lambda$, $r > 0$. Then $\tan \lambda = b/a$, $r = a/\cos \lambda = b/\sin \lambda$. The angle λ is determined to within an integral multiple of 2π. The equation is transformed into the form $r \cos x \cos \lambda + r \sin x \sin \lambda = c$, i.e. $\cos(x - \lambda) = c/r$. We get, in general, two values for $x - \lambda$ which are determined to within an integral multiple of 2π (provided, of course, that $|c/r| \leqq 1$).

REMARK 1. If relation (1) is satisfied for all real x for which the expression* $f(\cos x, \sin x, \tan x, \cot x, x)$ has a meaning, then it is called a *trigonometric identity*.

* Other angles $y, z, \ldots$ may also be contained in this expression.

Example 3. Let us decide whether the relation

$$\frac{\sin x + \sin y}{\cos x + \cos y} = \tan \frac{x + y}{2} \tag{2}$$

is a trigonometric identity.

We try to arrange the left-hand side in the form $\tan \frac{1}{2}(x + y)$. Applying formulae 1 of § 2.6 (p. 76) we get the left-hand side in the form

$$\frac{2 \sin \frac{1}{2}(x + y) \cos \frac{1}{2}(x - y)}{2 \cos \frac{1}{2}(x + y) \cos \frac{1}{2}(x - y)} = \tan \tfrac{1}{2}(x + y) ;$$

this means that the relation (2) is a trigonometric identity.

2.9. Plane Trigonometry

(a) Right-angled Triangle (Fig. 2.4a)

REMARK 1. In this section the following symbols for the elements of a right-angled triangle will be used: a, b enclose the right angle, c is the hypotenuse; A, B, C are the vertices opposite to the sides a, b, c, respectively; α, β, $90°$ are the interior angles corresponding to the vertices A, B, C, respectively; h is the altitude; P-the area.

TABLE 2.4

Formulae for determining the remaining elements of a right-angled triangle if two elements are given

The given elements	The other elements of the triangle				
a, α	$\beta = 90° - \alpha$	$b = a \cot \alpha$	$c = \frac{a}{\sin \alpha}$	$h = a \cos \alpha$	$P = \frac{1}{2}a^2 \cot \alpha$
c, α	$\beta = 90° - \alpha$	$a = c \sin \alpha$	$b = c \cos \alpha$	$h = \frac{1}{2}c \sin 2\alpha$	$P = \frac{1}{4}c^2 \sin 2\alpha$
a, b	$\tan \alpha = \frac{a}{b}$	$\tan \beta = \frac{b}{a}$	$c = \frac{a}{\sin \alpha} = \sqrt{(a^2 + b^2)}$	$h = a \cos \alpha = b \sin \alpha$	$P = \frac{1}{2}ab$
a, c	$\sin \alpha = \frac{a}{c}$	$\cos \beta = \frac{a}{c}$	$b = c \cos \alpha = c \sin \beta = \sqrt{(c^2 - a^2)}$	$h = a \cos \alpha = a \sin \beta$	$P = \frac{1}{4}c^2 \sin 2\alpha = \frac{1}{2}a^2 \tan \beta$

(b) General (Scalene) Triangle (Fig. 2.4b)

REMARK 2. In this section the following symbols for the elements of a triangle will be used: a, b, c are the sides; A, B, C the vertices opposite to the sides a, b, c, respectively; α, β, γ the interior angles corresponding to the vertices A, B, C, respectively; r is the radius of the inscribed circle; R the radius of the circumscribed circle; h_a, h_b, h_c are the altitudes corresponding to the vertices A, B, C or to the sides a, b, c, respectively, P the area and $s = \frac{1}{2}(a + b + c)$.

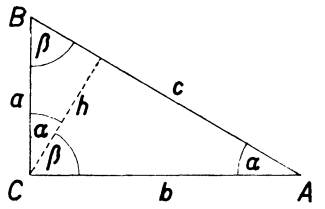


Fig. 2.4a.

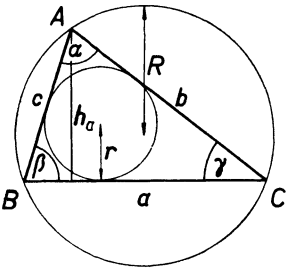


Fig. 2.4b.

Theorem 1. *Fundamental relations*:

1. $\dfrac{a}{\sin\alpha} = \dfrac{b}{\sin\beta} = \dfrac{c}{\sin\gamma} (= 2R)$ (*the Sine Theorem*).

2. $a^2 = b^2 + c^2 - 2bc\cos\alpha$ (*the Cosine Theorem*).

3. $\dfrac{a+b}{a-b} = \dfrac{\tan\frac{1}{2}(\alpha+\beta)}{\tan\frac{1}{2}(\alpha-\beta)}$ (*the Tangent Theorem*).

Theorem 2. *Further relations*:

4. $a = b\cos\gamma + c\cos\beta$.

5. $\dfrac{a+b}{c} = \dfrac{\cos\frac{1}{2}(\alpha-\beta)}{\cos\frac{1}{2}(\alpha+\beta)}$; $\dfrac{a-b}{c} = \dfrac{\sin\frac{1}{2}(\alpha-\beta)}{\sin\frac{1}{2}(\alpha+\beta)}$.

6. $\sin\dfrac{\alpha}{2} = \sqrt{\dfrac{(s-b)(s-c)}{bc}}$; $\cos\dfrac{\alpha}{2} = \sqrt{\dfrac{s(s-a)}{bc}}$.

7. $\tan\dfrac{\alpha}{2} = \dfrac{P}{s(s-a)}$; $\tan\dfrac{\alpha}{2} = \dfrac{r}{s-a}$.

8. $\tan\alpha = \dfrac{a\sin\gamma}{b - a\cos\gamma}$.

9. $r = s \tan \frac{\alpha}{2} \tan \frac{\beta}{2} \tan \frac{\gamma}{2}$.

10. $a = 2R \sin \alpha$; $R = \frac{abc}{4P}$.

11. $h_a = b \sin \gamma = c \sin \beta$; $\frac{1}{h_a} : \frac{1}{h_b} : \frac{1}{h_c} = a : b : c$; $\frac{1}{h_a} + \frac{1}{h_b} + \frac{1}{h_c} = \frac{1}{r}$.

12. $s = 4R \cos \frac{\alpha}{2} \cos \frac{\beta}{2} \cos \frac{\gamma}{2}$.

13. *The length of the median corresponding to the side c*:

$$t_c = \tfrac{1}{2}\sqrt{[2(a^2 + b^2) - c^2]} = \tfrac{1}{2}\sqrt{[a^2 + b^2 + 2ab \cos \gamma]};$$

$$t_a^2 + t_b^2 + t_c^2 = \tfrac{3}{4}(a^2 + b^2 + c^2).$$

14. *The length of the bisector of the angle* γ:

$$u_\gamma = \frac{2\sqrt{[abs(s - c)]}}{a + b} = \frac{\sqrt{[[ab[(a + b)^2 - c^2]]]}}{a + b} = \frac{2ab \cos \frac{1}{2}\gamma}{a + b}.$$

15. *The radius of the circumscribed circle*:

$$R = \frac{a}{2 \sin \alpha}.$$

16. *The radius of the inscribed circle*:

$$r = 4R \sin \frac{\alpha}{2} \sin \frac{\beta}{2} \sin \frac{\gamma}{2} = \frac{abc}{4Rs} = \sqrt{\frac{(s - a)(s - b)(s - c)}{s}}.$$

17. *The area of the triangle*:

$$P = \tfrac{1}{2}ab \sin \gamma = a^2 \frac{\sin \beta \sin \gamma}{2 \sin \alpha} = r^2 \cot \frac{\alpha}{2} \cot \frac{\beta}{2} \cot \frac{\gamma}{2} = 2R^2 \sin \alpha \sin \beta \sin \gamma,$$

$$P = \sqrt{[s(s - a)(s - b)(s - c)]} \quad (\textit{Heron's Formula}).$$

Theorem 3. *Solution of a general triangle*:

1. *Given the elements* a, β, γ $(\beta + \gamma < 180°)$:

$$\alpha = 180° - (\beta + \gamma); \quad b = \frac{a \sin \beta}{\sin \alpha}; \quad c = \frac{a \sin \gamma}{\sin \alpha}.$$

2. *Given the elements* a, b, γ:

$$\tfrac{1}{2}(\alpha + \beta) = 90° - \tfrac{1}{2}\gamma; \quad \tan \tfrac{1}{2}(\alpha - \beta) = \frac{a - b}{a + b} \cot \tfrac{1}{2}\gamma;$$

hence we determine the angles α, β:

$$\alpha = \tfrac{1}{2}(\alpha + \beta) + \tfrac{1}{2}(\alpha - \beta)\,; \quad \beta = \tfrac{1}{2}(\alpha + \beta) - \tfrac{1}{2}(\alpha - \beta)\,;$$
$$c = \sqrt{(a^2 + b^2 - 2ab\cos\gamma)}\,.$$

An alternative method:

$$\tan\alpha = \frac{a\sin\gamma}{b - a\cos\gamma}\,; \quad \tan\beta = \frac{b\sin\gamma}{a - b\cos\gamma}\,;$$
$$c = \frac{a\sin\gamma}{\sin\alpha} = \frac{a - b\cos\gamma}{\cos\beta}\,.$$

3. *Given the elements* a, b, α:

$$\sin\beta = \frac{b\sin\alpha}{a}\,; \quad \gamma = 180^\circ - (\alpha + \beta)\,;$$
$$c = \frac{a\sin\gamma}{\sin\alpha}\,; \quad P = \tfrac{1}{2}ab\sin\gamma\,.$$

If $a > b$, *then* $\beta < 90^\circ$ *and there exists a single solution.*
If $a = b$, *there exists a single solution for* $\alpha < 90^\circ$.
If $a < b$ (*and*, $\alpha < 90^\circ$, *of course*), *then*

1° *for* $b\sin\alpha < a$ *there exist two solutions* (*there are two angles* β, *satisfying the relation* $\beta_2 = 180^\circ - \beta_1$);
2° *for* $b\sin\alpha = a$ *there exists a single solution* ($\beta = 90^\circ$);
3° *there is no solution for* $b\sin\alpha > a$.

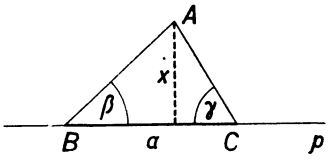


Fig. 2.5.

4. *Given the elements* a, b, c:

If the sum of any two sides is greater than the third side then a single solution exists and is given by

$$\cos\alpha = \frac{b^2 + c^2 - a^2}{2bc}\,, \quad \tan\frac{\alpha}{2} = \frac{P}{s(s-a)}\,, \quad P = \sqrt{[s(s-a)(s-b)(s-c)]}$$

and similarly for the angles β, γ.

Example 1. The problem is to find the distance x of an inaccessible point A from a straight road p (see Fig. 2.5).

On the road, the points B, C have been chosen, the distance a between them determined and the angles $ABC = \beta$, $ACB = \gamma$ measured. By Theorem 2 (formula 17), the area P of the triangle $\triangle ABC$ is

$$P = \frac{a^2 \sin \beta \sin \gamma}{2 \sin \alpha}, \quad \alpha = 180^\circ - (\beta + \gamma).$$

In addition, $P = \frac{1}{2}ax$. Hence $x = a \sin \beta \sin \gamma / \sin \alpha$.

2.10. Spherical Trigonometry

(a) Great Circle on a Sphere; Spherical (Euler's) Triangle

Definition 1. By a *great circle* on a given sphere we mean any circle lying on this sphere, whose centre coincides with the centre of the sphere. Through two points A, B on a sphere, which do not lie on the same diameter, one and only one great circle can be drawn; the smaller of the two arcs cut off by the points A, B on this circle has the shortest length d of all the curves on the given sphere joining the points A, B. This number d is called the *spherical distance* of the points A, B. The spherical distance of opposite points on a sphere equals the semi-circumference of a great circle.

Definition 2 (*Spherical Triangle*). Let A, B, C be three points on a sphere which do not lie on the same great circle. If we draw the three arcs $\overset{\frown}{AB}$, $\overset{\frown}{AC}$, $\overset{\frown}{BC}$ of the great circles which do not intersect except at the points A, B, C, the spherical surface splits into two spherical triangles with vertices A, B, C. If we choose, in particular, the arcs $\overset{\frown}{AB}$, $\overset{\frown}{AC}$, $\overset{\frown}{BC}$ to be of lengths equal to the spherical distances of their end points A, B, C, then the smaller of the two spherical triangles obtained (i.e. the one lying inside the trihedral angle formed by the half-lines OA, OB, OC emanating from the centre O of the sphere, see Fig. 2.6) is called an *Euler triangle.*

REMARK 1. In what follows we deal only with Euler triangles; moreover, we choose (except in Theorem 3) the radius of the sphere $r = 1$.

Definition 3. The lengths a, b, c of the corresponding arcs $\overset{\frown}{BC}$, $\overset{\frown}{AC}$, $\overset{\frown}{AB}$ of the great circles are called the *sides* of the spherical triangle $\triangle ABC$. Thus, they are determined by the angles BOC, AOC, AOB of the half-lines OA, OB, OC and are measured in radians or in degrees (Fig. 2.6).

Definition 4. The interior angles of the faces of the trihedral $OABC$ are called the angles α, β, γ of the spherical triangle $\triangle ABC$. They are measured in radians or in degrees (Fig. 2.6).

REMARK 2. The half-lines joining the centre of the sphere with the vertices of the spherical triangle $\triangle ABC$ form the basic trihedral $OABC$. The so-called *polar trihedral* $OA'B'C'$ has its edges normal to the faces of the basic trihedral and defines on the sphere an Euler spherical triangle $\triangle A'B'C'$, which is polar to $\triangle ABC$. The sides of the polar triangle are $a = 180° - \alpha$, $b = 180° - \beta$, $c = 180° - \gamma$; its angles are $\alpha = 180° - a$, $\beta = 180° - b$, $\gamma = 180° - c$. Thus, substituting the supplements of the angles for the sides and the supplements of the sides for the angles in any formula, we get a new formula.

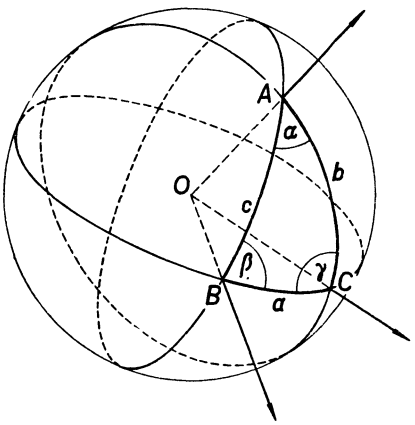


Fig. 2.6.

Fundamental properties of spherical triangles:

Theorem 1. *The sides and the angles of an Euler triangle are less than* 180° (*less than* π).

Theorem 2. *The sum of the angles* α, β, γ *of a spherical triangle is always greater than* 180°.

Definition 5. The number

$$\varepsilon^\circ = \alpha^\circ + \beta^\circ + \gamma^\circ - 180^\circ$$

is called the *spherical excess of a spherical triangle.*

Theorem 3. *The area of a spherical triangle is*

$$P = \frac{\varepsilon^\circ}{180^\circ} \pi r^2 ,$$

where r *is the radius of the sphere and* ε° *the excess of the triangle expressed in degrees.*

(b) Right-angled Spherical Triangle

REMARK 3. In this section, c denotes the hypotenuse, a, b the sides "enclosing" the right angle and α, β the angles opposite to the sides a, b, respectively (Fig. 2.7).

Theorem 4 (*Napier's Rule*). *We ascribe the hypotenuse c, the angles* α, β *and the complements of the sides* $90° - a$, $90° - b$ *to the vertices of a pentagon in the order indicated in Fig.* 2.8. *Then, the cosine of an arbitrary element equals the product*

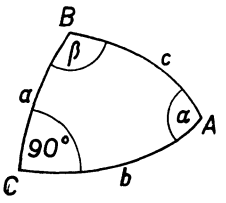


Fig. 2.7.

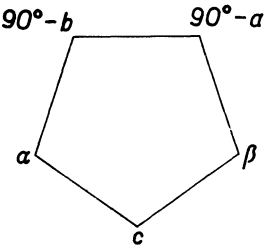


Fig. 2.8.

of the sines of the two opposite elements or the product of the cotangents of the two adjacent elements. In this way, we obtain the formulae

1. $\cos c = \cos a \cos b$,
2. $\cos c = \cot \alpha \cot \beta$,
3. $\cos \alpha = \cos a \sin \beta$,
4. $\cos \beta = \sin \alpha \cos b$,
5. $\sin a = \sin \alpha \sin c$,
6. $\sin b = \sin \beta \sin c$,
7. $\cos \alpha = \tan b \cot c$,
8. $\cos \beta = \tan a \cot c$,
9. $\sin a = \tan b \cot \beta$,
10. $\sin b = \tan a \cot \alpha$.

Theorem 5. *Spherical excess*:

$$\tan \frac{\varepsilon}{2} = \tan \frac{a}{2} \tan \frac{b}{2}.$$

Theorem 6. *The solution of a right-angled spherical triangle* (Table 2.5):

TABLE 2.5

Given elements	Number in parentheses denotes the corresponding formula of Theorem 4
a, b	c (1), α (10), β (9)
a, c	b (1), α (5), β (8)
a, α	b (10), c (5), β (3)
a, β	b (9), c (8), α (3)
c, α	a (5), b (7), β (2)
α, β	a (3), b (4), c (2)

(c) General (Oblique) Spherical Triangle

REMARK 4. Let us denote the sides of the triangle $\triangle ABC$ by a, b, c and the angles corresponding to the vertices A, B, C by α, β, γ, respectively (Fig. 2.9).

Theorem 7. *Fundamental formulae for an Euler triangle*:

1. $\dfrac{\sin a}{\sin \alpha} = \dfrac{\sin b}{\sin \beta} = \dfrac{\sin c}{\sin \gamma}$ (*the Sine Theorem*).

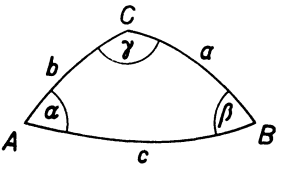


Fig. 2.9.

2. $\cos a = \cos b \cos c + \sin b \sin c \cos \alpha$ (*the Cosine Theorem for the sides*).

3. $\cos \alpha = -\cos \beta \cos \gamma + \sin \beta \sin \gamma \cos a$ (*the Cosine Theorem for the angles*).

4. (a) $\cos a \sin b = \sin a \cos b \cos \gamma + \sin c \cos \alpha$;
   (b) $\cot a \sin b = \sin \gamma \cot \alpha + \cos \gamma \cos b$.

5. (a) $\cos \alpha \sin \beta = \sin \gamma \cos a - \sin \alpha \cos \beta \cos c$;
   (b) $\cot \alpha \sin \beta = \sin c \cot a - \cos c \cos \beta$.

Theorem 8. *Further formulae for an Euler triangle*:

6. $\tan \frac{1}{2}(a + b) = \dfrac{\cos \frac{1}{2}(\alpha - \beta)}{\cos \frac{1}{2}(\alpha + \beta)} \tan \frac{1}{2}c$.

7. $\tan \frac{1}{2}(a - b) = \dfrac{\sin \frac{1}{2}(\alpha - \beta)}{\sin \frac{1}{2}(\alpha + \beta)} \tan \frac{1}{2}c$.

8. $\tan \frac{1}{2}(\alpha + \beta) = \dfrac{\cos \frac{1}{2}(a - b)}{\cos \frac{1}{2}(a + b)} \cot \frac{1}{2}\gamma$.

9. $\tan \frac{1}{2}(\alpha - \beta) = \dfrac{\sin \frac{1}{2}(a - b)}{\sin \frac{1}{2}(a + b)} \cot \frac{1}{2}\gamma$.

10. $\cos \frac{1}{2}(\alpha + \beta) \cos \frac{1}{2}c = \cos \frac{1}{2}(a + b) \sin \frac{1}{2}\gamma$.

11. $\sin \frac{1}{2}(\alpha + \beta) \cos \frac{1}{2}c = \cos \frac{1}{2}(a - b) \cos \frac{1}{2}\gamma$.

12. $\cos \frac{1}{2}(\alpha - \beta) \sin \frac{1}{2}c = \sin \frac{1}{2}(a + b) \sin \frac{1}{2}\gamma$.

13. $\sin \frac{1}{2}(\alpha - \beta) \sin \frac{1}{2}c = \sin \frac{1}{2}(a - b) \cos \frac{1}{2}\gamma$.

In formulae 14 *and* 15 *the notation*

$$s = \tfrac{1}{2}(a + b + c), \quad s_1 = s - a, \quad s_2 = s - b, \quad s_3 = s - c,$$
$$\sigma = \tfrac{1}{2}(\alpha + \beta + \gamma), \quad \sigma_1 = \sigma - \alpha, \quad \sigma_2 = \sigma - \beta, \quad \sigma_3 = \sigma - \gamma$$

is used.

14. $$\cot\frac{a}{2} = \frac{1}{\cos\sigma_1}\sqrt{\frac{\cos\sigma_1\cos\sigma_2\cos\sigma_3}{-\cos\sigma}}.$$

15. $$\tan\frac{\alpha}{2} = \frac{1}{\sin s_1}\sqrt{\frac{\sin s_1 \sin s_2 \sin s_3}{\sin s}}.$$

Theorem 9. *The solution of a general spherical triangle is shown in Table* 2.6.

TABLE 2.6

Given elements	Number in parentheses denotes the corresponding formula of Theorems 7 and 8
a, b, γ	$\frac{\alpha+\beta}{2}$ (8), $\frac{\alpha-\beta}{2}$ (9), c (e. g. 10 or 12)
α, β, c	$\frac{a+b}{2}$ (6), $\frac{a-b}{2}$ (7), γ (e.g. 11 or 12)
a, b, c	α (15), similarly β and γ
α, β, γ	a (14), similarly b and c
a, b, α*	β (1), γ (9), c (7)
α, β, a**	b (1), c (7), γ (9)

2.11. Inverse Trigonometric Functions

Inverse trigonometric functions are the functions arcsin x (or $\sin^{-1} x$), arccos x ($\cos^{-1} x$), arctan x ($\tan^{-1} x$), arccot x ($\cot^{-1} x$), which are inverse (see § 11.1, p. 362) to the trigonometric functions.

REMARK 1. In this section, the angles are expressed in circular measure.

* If $\sin b \sin\alpha > \sin a$, then no solution exists. If $\sin b \sin\alpha = \sin a$ there is a single solution (the triangle is right-angled). If $\sin b \sin\alpha < \sin a$, it is necessary to distinguish two cases: 1° if a is nearer to 90° than b, then there exists one solution (β and b are of the same kind, i.e. both either acute or obtuse); 2° if b is nearer to 90° than a, then there are two solutions or no solution according to whether a and α are of the same or of different kinds, acute or obtuse, respectively.

** The discussion of this case can be obtained from the discussion of the case* by substituting throughout the sides a, b, c for the angles α, β, γ and vice versa.

Definition 1. The function $y = \arcsin x$ is inverse to the function $x = \sin y$ $(-\frac{1}{2}\pi \leqq y \leqq \frac{1}{2}\pi)$; it is defined for x in the interval $[-1, 1]$. Thus: If $-1 \leqq \leqq x \leqq 1$, then $\arcsin x$ is the unique angle y in the range $[-\frac{1}{2}\pi, \frac{1}{2}\pi]$ such that $\sin y = x$.*

Definition 2. The function $y = \arccos x$ is inverse to the function $x = \cos y$ $(0 \leqq y \leqq \pi)$; it is defined for x in the interval $[-1, 1]$. Thus: If $-1 \leqq x \leqq 1$, then $\arccos x$ is the unique angle y in the range $[0, \pi]$ such that $\cos y = x$.

Definition 3. The function $y = \arctan x$ is inverse to the function $x = \tan y$ $(-\frac{1}{2}\pi < y < \frac{1}{2}\pi)$; it is defined for all real x. Thus, if x is a real number, then $\arctan x$ is the unique angle y in the range $(-\frac{1}{2}\pi, \frac{1}{2}\pi)$ such that $\tan y = x$.

Definition 4. The function $y = \operatorname{arccot} x$ is inverse to the function $x = \cot y$ $(0 < y < \pi)$; it is defined for all real x. Thus, if x is a real number, then $\operatorname{arccot} x$ is the unique angle y in the range $(0, \pi)$ such that $\cot y = x$. (The range $(-\frac{1}{2}\pi, \frac{1}{2}\pi)$ is sometimes used.)

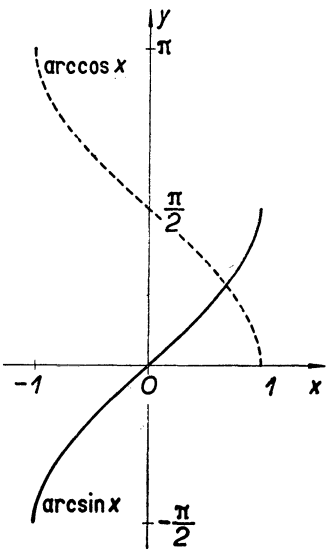


Fig. 2.10.

REMARK 2. The graphs of the functions $\arcsin x$, $\arccos x$, $\arctan x$, $\operatorname{arccot} x$ are illustrated in Fig. 2.10 and 2.11.

Theorem 1. *The values of the inverse trigonometric functions at some special points*:

$$\arcsin 0 = 0, \quad \arcsin \tfrac{1}{2} = \tfrac{1}{6}\pi, \quad \arcsin 1 = \tfrac{1}{2}\pi, \quad \arcsin(-1) = -\tfrac{1}{2}\pi;$$

*) In English literature this function is more usually called *the principal value of* $\arcsin x$, the general function $\arcsin x$ being the (multi-valued) function inverse to $x = \sin y$, and similarly for the other inverse functions.

$\arccos 0 = \frac{1}{2}\pi$, $\arccos \frac{1}{2} = \frac{1}{3}\pi$, $\arccos 1 = 0$, $\arccos(-1) = \pi$;

$\arctan 0 = 0$, $\arctan 1 = \frac{1}{4}\pi$, $\lim_{x \to +\infty} \arctan x = \frac{1}{2}\pi$, $\lim_{x \to -\infty} \arctan x = -\frac{1}{2}\pi$;

$\operatorname{arccot} 0 = \frac{1}{2}\pi$, $\operatorname{arccot} 1 = \frac{1}{4}\pi$, $\lim_{x \to +\infty} \operatorname{arccot} x = 0$, $\lim_{x \to -\infty} \operatorname{arccot} x = \pi$.

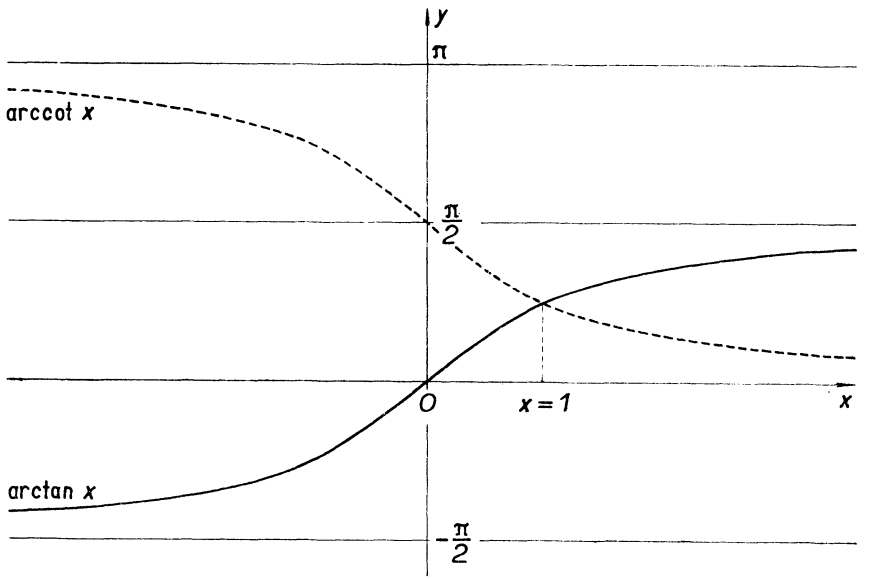


Fig. 2.11.

Theorem 2. *Fundamental formulae and relations among inverse trigonometric functions (if the domain of validity is not mentioned, then the formula holds for all x):*

1. $\arcsin(\sin x) = x \quad (|x| \leqq \frac{1}{2}\pi)$, $\arccos(\cos x) = x \quad (0 \leqq x \leqq \pi)$.
2. $\sin(\arcsin x) = x$, $\cos(\arccos x) = x \quad (|x| \leqq 1)$.
3. $\arctan(\tan x) = x \quad (|x| < \frac{1}{2}\pi)$, $\operatorname{arccot}(\cot x) = x \quad (0 < x < \pi)$.
4. $\tan(\arctan x) = x$.
5. $\cot(\operatorname{arccot} x) = x$.
6. $\arcsin x + \arccos x = \frac{1}{2}\pi \quad (|x| \leqq 1)$.
7. $\arctan x + \operatorname{arccot} x = \frac{1}{2}\pi$.
8. $\arcsin(-x) = -\arcsin x \quad (|x| \leqq 1)$.
9. $\arccos(-x) = \pi - \arccos x \quad (|x| \leqq 1)$.
10. $\arctan(-x) = -\arctan x$.

11. $\operatorname{arccot}(-x) = \pi - \operatorname{arccot} x$.

12. $\arcsin x = \arctan \dfrac{x}{\sqrt{(1 - x^2)}} \quad (|x| < 1)$.

13. $\arccos x = \operatorname{arccot} \dfrac{x}{\sqrt{(1 - x^2)}} \quad (|x| < 1)$.

14. $\arctan x = \arcsin \dfrac{x}{\sqrt{(1 + x^2)}}$.

15. $\operatorname{arccot} x = \arccos \dfrac{x}{\sqrt{(1 + x^2)}}$.

16. $\arctan x = \operatorname{arccot} \dfrac{1}{x} \quad (x > 0)$.

17. $\arcsin x = \arccos \sqrt{(1 - x^2)}$, $\arccos x = \arcsin \sqrt{(1 - x^2)} \quad (0 \leqq x \leqq 1)$.

18. $\arcsin x + \arcsin y =$
$= \arcsin [x \sqrt{(1 - y^2)} + y \sqrt{(1 - x^2)}] \quad (xy \leqq 0 \text{ or } x^2 + y^2 \leqq 1)$,
$= \pi - \arcsin [x \sqrt{(1 - y^2)} + y \sqrt{(1 - x^2)}] \quad (x > 0, y > 0 \text{ and } x^2 + y^2 > 1)$,
$= -\pi - \arcsin [x \sqrt{(1 - y^2)} + y \sqrt{(1 - x^2)}] \quad (x < 0, y < 0 \text{ and } x^2 + y^2 > 1)$.

19. $\arcsin x - \arcsin y =$
$= \arcsin [x \sqrt{(1 - y^2)} - y \sqrt{(1 - x^2)}] \quad (xy \geqq 0 \text{ or } x^2 + y^2 \leqq 1)$,
$= \pi - \arcsin [x \sqrt{(1 - y^2)} - y \sqrt{(1 - x^2)}] \quad (x > 0, y < 0 \text{ and } x^2 + y^2 > 1)$,
$= -\pi - \arcsin [x \sqrt{(1 - y^2)} - y \sqrt{(1 - x^2)}] \quad (x < 0, y > 0 \text{ and } x^2 + y^2 > 1)$.

20. $\arccos x + \arccos y = \arccos [xy - \sqrt{(1 - x^2)} \sqrt{(1 - y^2)}] \quad (x + y \geqq 0)$,
$= 2\pi - \arccos [xy - \sqrt{(1 - x^2)} \sqrt{(1 - y^2)}] \quad (x + y < 0)$.

21. $\arccos x - \arccos y = -\arccos [xy + \sqrt{(1 - x^2)} \sqrt{(1 - y^2)}] \quad (x \geqq y)$,
$= \arccos [xy + \sqrt{(1 - x^2)} \sqrt{(1 - y^2)}] \quad (x < y)$.

22. $\arctan x + \arctan y = \arctan \dfrac{x + y}{1 - xy} \quad (xy < 1)$,

$= \pi + \arctan \dfrac{x + y}{1 - xy} \quad (xy > 1, x > 0)$,

$= -\pi + \arctan \dfrac{x + y}{1 - xy} \quad (xy > 1, x < 0)$.

23. $\arctan x - \arctan y = \arctan \dfrac{x - y}{1 + xy} \quad (xy > -1)$,

$$= \pi + \arctan\frac{x - y}{1 + xy} \quad (xy < -1, x > 0),$$

$$= -\pi + \arctan\frac{x - y}{1 + xy} \quad (xy < -1, x < 0).$$

2.12. Hyperbolic Functions

Definition 1. The functions $\sinh x$ (*hyperbolic sine*), $\cosh x$ (*hyperbolic cosine*) and $\tanh x$ (*hyperbolic tangent*) are defined for all real x as follows:

$$\sinh x = \tfrac{1}{2}(e^x - e^{-x}), \quad \cosh x = \tfrac{1}{2}(e^x + e^{-x}),$$

$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{\sinh x}{\cosh x}.$$

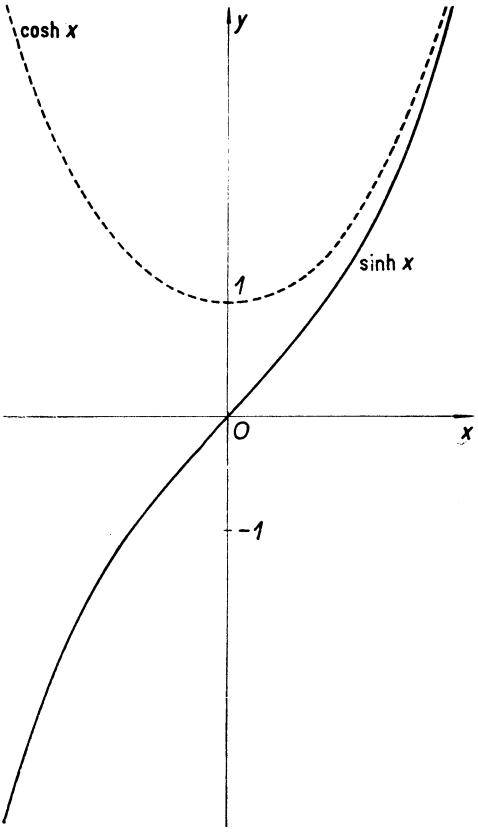


Fig. 2.12a.

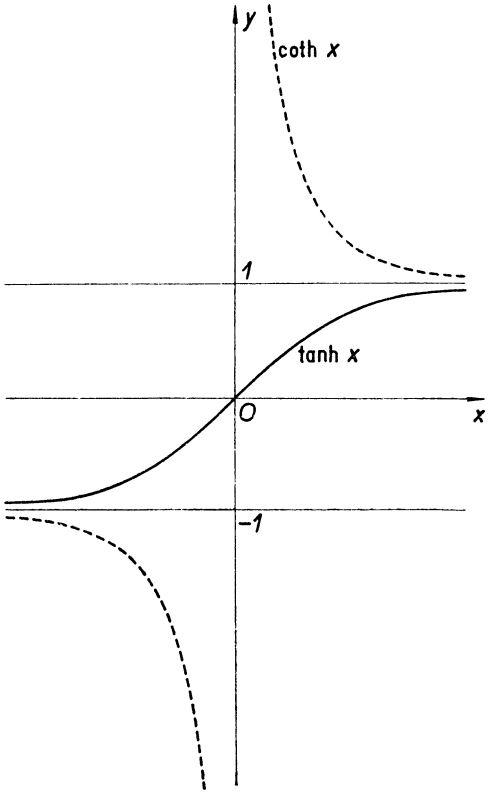


Fig. 2.12b.

For $x \neq 0$, the function $\coth x$ (*hyperbolic cotangent*) is defined by the relation

$$\coth x = \frac{e^x + e^{-x}}{e^x - e^{-x}} = \frac{1}{\tanh x}.$$

Further, the following functions are defined

$$\operatorname{sech} x = \frac{1}{\cosh x} \quad (\textit{hyperbolic secant})$$

$$\operatorname{cosech} x = \frac{1}{\sinh x} \quad \text{for} \quad x \neq 0 \quad (\textit{hyperbolic cosecant}).$$

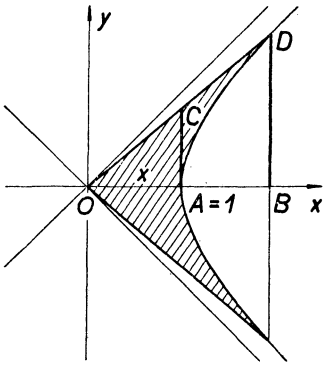


Fig. 2.13.

REMARK 1. The behaviour of the hyperbolic functions can be seen in Fig. 2.12a,b.

REMARK 2. The hyperbolic functions stand in a similar relation to an equiangular hyperbola with semi-axis of length 1 as do the trigonometric functions to a unit circle; the independent variable (argument) $x \geqq 0$ denotes the area of the hyperbolic sector (the shaded area in Fig. 2.13). Here,

$$\sinh x = \widetilde{BD}, \quad \cosh x = \widetilde{OB}, \quad \tanh x = \widetilde{AC}.$$

Theorem 1. *Relations between hyperbolic functions*:

1. $\cosh^2 x - \sinh^2 x = 1$,
2. $\cosh x + \sinh x = e^x$,
   $\cosh x - \sinh x = e^{-x}$.
3. $\sinh(-x) = -\sinh x$,
   $\cosh(-x) = \cosh x$,
   $\tanh(-x) = -\tanh x$,
   $\coth(-x) = -\coth x$.

4. $|\sinh x| = \sqrt{(\cosh^2 x - 1)} = \dfrac{|\tanh x|}{\sqrt{(1 - \tanh^2 x)}} = \dfrac{1}{\sqrt{(\coth^2 x - 1)}}$.

5. $\cosh x = \sqrt{(\sinh^2 x + 1)} = \dfrac{|\coth x|}{\sqrt{(\coth^2 x - 1)}} = \dfrac{1}{\sqrt{(1 - \tanh^2 x)}}$.

6. $\tanh x = \dfrac{\sinh x}{\sqrt{(\sinh^2 x + 1)}}$.

7. $\sinh(x \pm y) = \sinh x \cosh y \pm \cosh x \sinh y$,

$\cosh(x \pm y) = \cosh x \cosh y \pm \sinh x \sinh y$,

$\tanh(x \pm y) = \dfrac{\tanh x \pm \tanh y}{1 \pm \tanh x \tanh y}$, $\coth(x \pm y) = \dfrac{1 \pm \coth x \coth y}{\coth x \pm \coth y}$.

8. $\sinh 2x = 2 \sinh x \cosh x$, $\cosh 2x = \sinh^2 x + \cosh^2 x$,

$\tanh 2x = \dfrac{2 \tanh x}{1 + \tanh^2 x}$, $\coth 2x = \dfrac{1 + \coth^2 x}{2 \coth x}$.

9. *De Moivre's Theorem*: $(\cosh x \pm \sinh x)^n = \cosh nx \pm \sinh nx$.

10. $\sinh x \pm \sinh y = 2 \sinh \dfrac{x \pm y}{2} \cosh \dfrac{x \mp y}{2}$,

$\cosh x + \cosh y = 2 \cosh \dfrac{x + y}{2} \cosh \dfrac{x - y}{2}$,

$\cosh x - \cosh y = 2 \sinh \dfrac{x + y}{2} \sinh \dfrac{x - y}{2}$,

$\tanh x \pm \tanh y = \dfrac{\sinh(x \pm y)}{\cosh x \cosh y}$.

11. *Relations between hyperbolic and trigonometric functions* (see Remark 20.4.4):

$$\sin ix = i \sinh x, \quad \cos ix = \cosh x,$$
$$\tan ix = i \tanh x, \quad \cot ix = -i \coth x.$$

2.13. Inverse Hyperbolic Functions

Inverse hyperbolic functions are the functions arsinh x ($\sinh^{-1} x$), arcosh x ($\cosh^{-1} x$), artanh x ($\tanh^{-1} x$), arcoth x ($\coth^{-1} x$) which are inverse (see 11.1, p. 400) to the hyperbolic functions.

Definition 1. The function $y = \operatorname{arsinh} x$ is inverse to the function $x = \sinh y$; it is defined for all real x. Thus: If x is a real number, then $\operatorname{arsinh} x$ is the unique number y such that $\sinh y = x$.

Definition 2. The function $y = \operatorname{arcosh} x$ is inverse to the function $x = \cosh y$ considered only in the interval $[0, \infty)$; it is defined for every x in the interval $[1, \infty)$. Thus: If $1 \leqq x < +\infty$, then $\operatorname{arcosh} x$ is the unique number y in the interval $[0, \infty)$ such that $\cosh y = x$.*)

Fig. 2.14a.

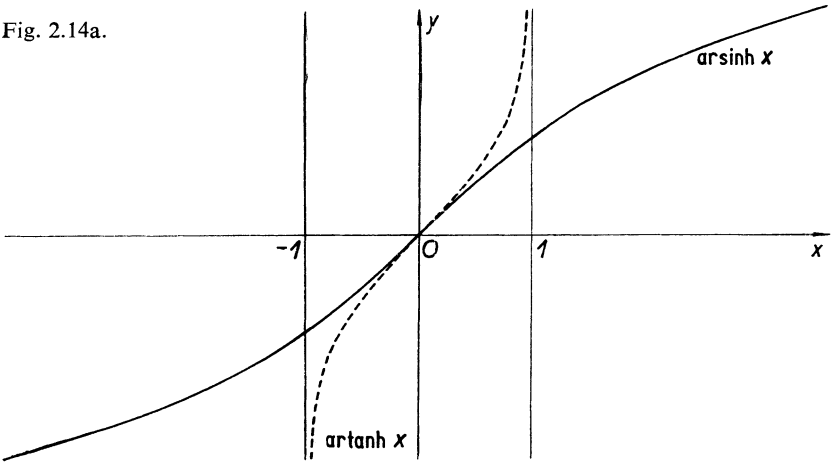
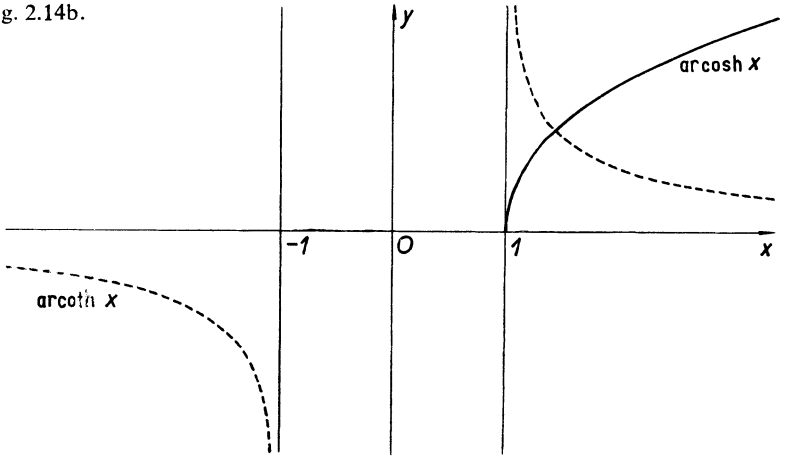


Fig. 2.14b.



Definition 3. The function $y = \operatorname{artanh} x$ is inverse to the function $x = \tanh y$; it is defined for all x in the interval $(-1, 1)$. Thus: If $-1 < x < 1$, then $\operatorname{artanh} x$ is the unique number y such that $\tanh y = x$.

*) In English literature the function $\operatorname{arcosh} x$ is more usually defined as the two-valued function inverse to $x = \cosh y$.

Definition 4. The function $y = \operatorname{arcoth} x$ is inverse to the function $x = \coth y$; it is defined for all x satisfying $|x| > 1$. Thus: If $|x| > 1$, then arcoth x is the unique number y such that $\coth y = x$.

REMARK 1. The graphs of the inverse hyperbolic functions are illustrated in Fig. 2.14a,b.

Theorem 1. *Inverse hyperbolic functions expressed by means of logarithms*:

$$\operatorname{arsinh} x = \ln\left[x + \sqrt{(x^2 + 1)}\right], \quad \operatorname{arcosh} x = \ln\left[x + \sqrt{(x^2 - 1)}\right] \quad (x \geqq 1),$$

$$\operatorname{artanh} x = \tfrac{1}{2} \ln \frac{1 + x}{1 - x} \quad (|x| < 1), \quad \operatorname{arcoth} x = \tfrac{1}{2} \ln \frac{x + 1}{x - 1} \quad (|x| > 1).$$

Theorem 2. *Relations between the inverse hyperbolic functions*:

1. $$\operatorname{arsinh} x = \operatorname{artanh} \frac{x}{\sqrt{(x^2 + 1)}}, \quad |\operatorname{arsinh} x| = \operatorname{arcosh} \sqrt{(x^2 + 1)}.$$

2. $$\operatorname{artanh} x = \operatorname{arsinh} \frac{x}{\sqrt{(1 - x^2)}} \quad (|x| < 1),$$
$$= \operatorname{arcoth} \frac{1}{x} \quad (|x| < 1, \ x \neq 0).$$

3. $$\operatorname{arsinh} x \pm \operatorname{arsinh} y = \operatorname{arsinh}\left[x \sqrt{(1 + y^2)} \pm y \sqrt{(1 + x^2)}\right],$$
$$|\operatorname{arcosh} x \pm \operatorname{arcosh} y| = \operatorname{arcosh}\left[\left[xy \pm \sqrt{[(x^2 - 1)(y^2 - 1)]}\right]\right]$$
$$(x \geqq 1, \ y \geqq 1),$$
$$\operatorname{artanh} x \pm \operatorname{artanh} y = \operatorname{artanh} \frac{x \pm y}{1 \pm xy} \quad (|x| < 1, \ |y| < 1).$$

3. SOME FORMULAE (AREAS CIRCUMFERENCES, VOLUMES, SURFACES, CENTROIDS, MOMENTS OF INERTIA)

By VÁCLAV VILHELM

References: [26], [28], [68].

3.1. Area, Circumference, Centroid and Moments of Inertia of Plane Figures

REMARK 1. For the calculation of areas and circumferences of plane figures by means of integrals see § 14.9.

(a) The Triangle (Fig. 3.1). Consider a triangle ABC, denoting its sides by a, b, c, interior angles by α, β, γ, altitudes by h_a, h_b, h_c, radius of the inscribed circle by r, radius of the circumscribed circle by R, area by P, semi-perimeter by $s = \frac{1}{2}(a + b + c)$, medians by t_a, t_b, t_c, centroid by T. The following relations hold:

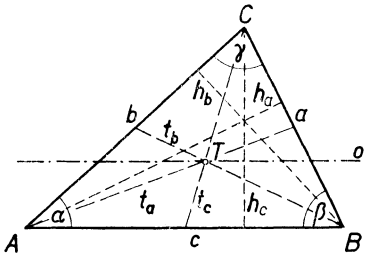


Fig. 3.1.

$$P = \tfrac{1}{2}ah_a = \tfrac{1}{2}bh_b = \tfrac{1}{2}ch_c\,, \tag{1}$$

$$= \sqrt{[s(s-a)(s-b)(s-c)]} \quad (\textit{Heron's formula})\,, \tag{2}$$

$$= \frac{abc}{4R} = 2R^2 \sin\alpha \sin\beta \sin\gamma\,, \tag{3}$$

$$= rs = r^2 \cot\frac{\alpha}{2} \cot\frac{\beta}{2} \cot\frac{\gamma}{2}\,, \tag{4}$$

$$= \tfrac{1}{2}ab \sin\gamma\,. \tag{5}$$

If $x_1, y_1; x_2, y_2; x_3, y_3$ are the coordinates of the vertices A, B, C of a triangle in a cartesian coordinate system, then

$$\pm P = \tfrac{1}{2}\begin{vmatrix} x_1, & y_1, & 1 \\ x_2, & y_2, & 1 \\ x_3, & y_3, & 1 \end{vmatrix} = \tfrac{1}{2}\begin{vmatrix} x_2 - x_1, & y_2 - y_1 \\ x_3 - x_1, & y_3 - y_1 \end{vmatrix}; \tag{6}$$

the minus sign relates to the case where the determinant is negative.

The coordinates of the centroid T (the point of intersection of the medians t_a, t_b, t_c) are

$$x_T = \tfrac{1}{3}(x_1 + x_2 + x_3), \quad y_T = \tfrac{1}{3}(y_1 + y_2 + y_3). \tag{7}$$

The moment of inertia about a median axis o, i.e. an axis through the centroid parallel to the side c, or about the side c is

$$I_0 = \tfrac{1}{36}ch_c^3, \quad \text{or} \quad I_c = \tfrac{1}{12}ch_c^3, \quad \text{respectively}. \tag{8}$$

The area of a *right-angled* triangle ABC with hypotenuse c (hence, $\gamma = 90°$) is

$$P = \tfrac{1}{2}ab = \tfrac{1}{2}a^2 \tan\beta = \tfrac{1}{4}c^2 \sin 2\alpha. \tag{9}$$

REMARK 2. For trigonometric formulae concerning a triangle see § 2.9, p. 78.

(b) The Quadrilateral (Fig. 3.2). Consider a quadrilateral with sides a, b, c, d and with vertices A, B, C, D (the sides intersecting only at the vertices). Let u_1, u_2 be its diagonals, φ the angle between them, and h_1, h_2, the altitudes of the triangles

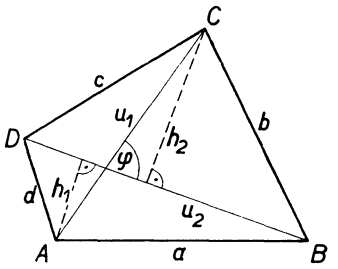


Fig. 3.2.

ABD, BDC, dropped from the points A, C, respectively. Then the area P of the quadrilateral is

$$P = \tfrac{1}{2}u_1u_2 \sin\varphi = \tfrac{1}{2}(h_1 + h_2)u_2 \tag{10}$$

(if the quadrilateral is not convex, then u_2 in (10) is the inner diagonal).

If the vertices of a convex quadrilateral lie on a circle then the area of the quadrilateral is

$$P = \sqrt{[(s-a)(s-b)(s-c)(s-d)]} \tag{11}$$

where $s = \tfrac{1}{2}(a + b + c + d)$.

A *trapezium* (Fig. 3.3a) is a (convex) quadrilateral two opposite sides of which are parallel. The area P is given by the formulae (10), (11) and

$$P = \tfrac{1}{2}(a + c)\, h\,. \tag{12}$$

The centroid T lies on the segment MN, where M, N, are the mid-points of the sides a, c, respectively, at distance $h_a = h(a + 2c)/[3(a + c)]$ from the side a.

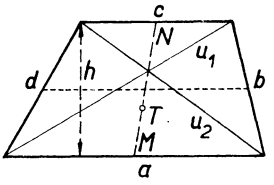


Fig. 3.3a.

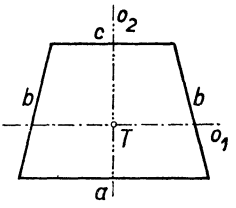


Fig. 3.3b.

The moments of inertia of an isosceles trapezium of altitude h about the median axes o_1, o_2 (Fig. 3.3b) are

$$I_{o_1} = \frac{h^3(a^2 + 4ac + c^2)}{36(a + c)}\,; \quad I_{o_2} = \frac{h(a^4 - c^4)}{48(a - c)}\,. \tag{13}$$

A *parallelogram* (Fig. 3.4) is a quadrilateral the opposite sides of which are parallel and, consequently, of the same length. If $\gamma = 90°$, we get a rectangle or a square. The area P of a parallelogram is given by formulae (10), (11), (12) (where $a = c$, $b = d$) and

$$P = ab \sin \gamma\,. \tag{14}$$

A *rhombus* is a parallelogram with $a = b$. Then, $\varphi = 90°$ and

$$P = a^2 \sin \gamma = \tfrac{1}{2}u_1 u_2\,. \tag{15}$$

A square is a rhombus with $\gamma = 90°$.

The centroid T of a parallelogram lies at the point of intersection of the diagonals.

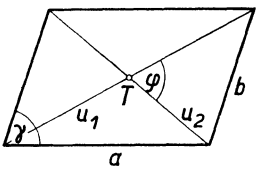


Fig. 3.4.

The moment of inertia of a parallelogram about the diagonal u_1 is

$$I_{u_1} = \tfrac{1}{48}u_1 u_2^3 \sin^3 \varphi = \tfrac{1}{24}P u_2^2 \sin^2 \varphi\,. \tag{16}$$

The moment of inertia of a rectangle with sides a, b about a median axis o parallel to the side a is

$$I_0 = \tfrac{1}{12}ab^3\,. \tag{17}$$

(c) The Polygon. The area can be determined by dividing the polygon into simple figures, for example into triangles (see Fig. 3.5a).

A regular polygon (Fig. 3.5b) has all its sides and all its angles equal. Let n be the number of sides, a their common length, $\alpha = 360°/n$ the central angle, r the radius of the inscribed circle, R the radius of the circumscribed circle, P the area, C the

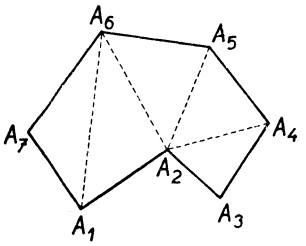


Fig. 3.5a.

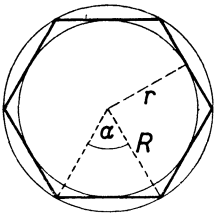


Fig. 3.5b.

circumference of the regular polygon. Then

$$P = \tfrac{1}{2}nar = \tfrac{1}{4}na^2 \cot\frac{\alpha}{2} = nr^2 \tan\frac{\alpha}{2} = \tfrac{1}{2}nR^2 \sin\alpha\,, \tag{18}$$

$$C = na = 2nR \sin\frac{\alpha}{2} = 2nr \tan\frac{\alpha}{2}\,. \tag{19}$$

TABLE 3.1

Calculation of the elements of regular polygons

n	$\frac{P}{a^2}$	$\frac{P}{R^2}$	$\frac{P}{r^2}$	$\frac{R}{a}$	$\frac{a}{R}$	$\frac{a}{r}$	$\frac{R}{r}$	$\frac{r}{R}$
3	0·433 0	1·299 0	5·196 2	0·577 4	1·732 1	3·464 1	2·000 0	0·500 0
4	1·000 0	2·000 0	4·000 0	0·707 1	1·414 2	2·000 0	1·414 2	0·707 1
5	1·720 5	2·377 6	3·632 7	0·850 7	1·175 6	1·453 1	1·236 1	0·809 0
6	2·598 1	2·598 1	3·464 1	1·000 0	1·000 0	1·154 7	1·154 7	0·866 0
7	3·633 9	2·736 4	3·371 0	1·152 4	0·867 8	0·963 1	1·109 9	0·901 0
8	4·828 4	2·828 4	3·313 7	1·306 6	0·765 4	0·828 4	1·082 4	0·923 9
9	6·181 8	2·892 5	3·275 7	1·461 9	0·684 0	0·727 9	1·064 2	0·939 7
10	7·694 2	2·938 9	3·249 2	1·618 0	0·618 0	0·649 8	1·051 5	0·951 1
12	11·196 2	3·000 0	3·215 4	1·931 9	0·517 6	0·535 9	1·035 3	0·965 9
15	17·642 4	3·050 5	3·188 3	2·404 9	0·415 8	0·425 1	1·022 3	0·978 1
16	20·109 4	3·061 5	3·182 6	2·562 9	0·390 2	0·397 8	1·019 6	0·980 8
20	31·568 8	3·090 2	3·167 7	3·196 2	0·312 9	0·316 8	1·012 5	0·987 7
24	45·574 5	3·105 8	3·159 7	3·830 6	0·261 1	0·263 3	1·008 6	0·991 4
32	81·225 4	3·121 4	3·151 7	5·101 1	0·196 0	0·197 0	1·004 8	0·995 2
48	183·084 6	3·132 6	3·146 1	7·644 9	0·130 8	0·131 1	1·002 1	0·997 9
64	325·687 5	3·136 5	3·144 1	10·190 0	0·098 1	0·098 3	1·001 2	0·998 8

The centroid of a regular polygon of n vertices is at its centre, the moment of inertia about an arbitrary axis o passing through the centre is

$$I_0 = \tfrac{1}{96}nar(12r^2 + a^2) = \tfrac{1}{24}P(6R^2 - a^2). \tag{20}$$

(d) The Circle. Let r denote the radius, d the diameter, P the area and C the circumference of the circle. Then

$$P = \pi r^2 = \tfrac{1}{4}\pi d^2 = \tfrac{1}{4}Cd \approx 0{\cdot}785\,4d^2\ ; \tag{21}$$

$$C = 2\pi r = \pi d \approx 3{\cdot}141\,59d\,. \tag{22}$$

The centroid of a circle is at its centre.

The moment of inertia about an axis o passing through the centre is

$$I_0 = \tfrac{1}{4}\pi r^4\,. \tag{23}$$

REMARK 3. For the measurement of angles and for conversion of angles measured in degrees into radians and vice versa, see § 2.1, p. 69.

The length l of a circular arc of radius r, corresponding to the central angle α (Fig. 3.6):

$$l = r \operatorname{arc}\alpha \text{ (arc } \alpha \text{ denotes the magnitude of the angle } \alpha \text{ in radians)}, \tag{24}$$

$$l = \frac{\pi r\alpha}{180} \approx 0{\cdot}017\,453r\alpha \text{ (the angle in degrees)}, \tag{25}$$

$$l \approx \sqrt{(t^2 + \tfrac{16}{3}h^2)}\,. \tag{26}$$

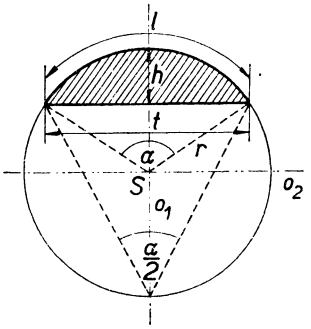


Fig. 3.6.

A *segment of a circle* (Fig. 3.6, the shaded area). Let r be the radius, l the length of the arc, t the length of the chord, α the central angle (in degrees), h the altitude of the segment, P the area of the segment. Then

$$t = 2\sqrt{(2hr - h^2)} = 2r\sin\frac{\alpha}{2}, \quad h = \frac{t}{2}\tan\frac{\alpha}{4}, \tag{27}$$

$$P = \tfrac{1}{2}r^2\left(\frac{\pi\alpha}{180} - \sin\alpha\right) = \tfrac{1}{2}[lr - t(r - h)]\,. \tag{28}$$

The centroid T lies on the bisector o_1 of the central angle (Fig. 3.6); its distance from the centre S is

$$\overline{TS} = \frac{4r\sin^3\tfrac{1}{2}\alpha}{3(\frac{1}{180}\pi\alpha - \sin\alpha)}\,. \tag{29}$$

The moments of inertia about the axes o_1, o_2 (Fig. 3.6) are

$$I_{o_1} = \tfrac{1}{48}r^4\left(\frac{\pi\alpha}{30} - 8\sin\alpha + \sin 2\alpha\right), \quad I_{o_2} = \tfrac{1}{16}r^4\left(\frac{\pi\alpha}{90} - \sin 2\alpha\right). \tag{30}$$

A *sector of a circle* (Fig. 3.7). The area

$$P = \frac{\pi r^2\alpha}{360} = \tfrac{1}{2}rl\,, \tag{31}$$

where α stands for the magnitude of the central angle in degrees.

The centroid T lies on the bisector o_1 of the central angle; its distance from the centre S (Fig. 3.7) is

$$\overline{TS} = \frac{240r\sin\tfrac{1}{2}\alpha}{\pi\alpha}\,. \tag{32}$$

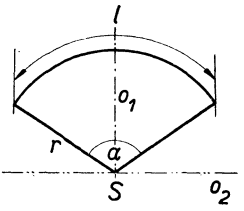


Fig. 3.7.

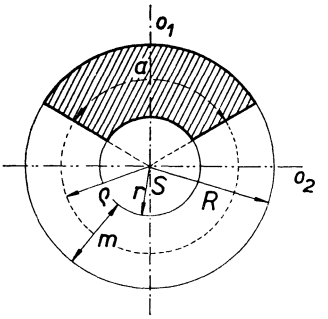


Fig. 3.8.

The moments of inertia about the axes o_1, o_2 (Fig. 3.7) are

$$I_{o_1} = \tfrac{1}{8}r^4\left(\frac{\pi\alpha}{180} - \sin\alpha\right),$$

$$I_{o_2} = \tfrac{1}{8}r^4\left(\frac{\pi\alpha}{180} + \sin\alpha\right). \tag{33}$$

An *annulus* (Fig. 3.8). Let r be the radius of the inner circle, R the radius of the outer circle, $\varrho = \frac{1}{2}(r + R)$, $m = R - r$, and P the area. Then

$$P = \pi(R^2 - r^2) = 2\pi\varrho m\,. \tag{34}$$

The centroid lies at the centre S.

The moment of inertia about the median axis o_1 is

$$I_{o_1} = \tfrac{1}{4}\pi(R^4 - r^4)\,. \tag{35}$$

A *sector of an annulus* (with central angle α in degrees, see Fig. 3.8, the shaded area). The area P is given by

$$P = \frac{\pi\alpha}{360}(R^2 - r^2) = am\,. \tag{36}$$

The centroid T lies on the bisector o_1 of the central angle; its distance from the centre S is

$$\overline{TS} = \frac{4}{3}\,\frac{R^3 - r^3}{R^2 - r^2}\,\frac{\sin\frac{1}{2}\alpha}{\frac{1}{180}\pi\alpha}\,. \tag{37}$$

The moments of inertia about the axes o_1, o_2 (Fig. 3.8) are

$$I_{o_1} = \tfrac{1}{8}(R^4 - r^4)\left(\frac{\pi\alpha}{180} - \sin\alpha\right),\quad I_{o_2} = \tfrac{1}{8}(R^4 - r^4)\left(\frac{\pi\alpha}{180} + \sin\alpha\right). \tag{38}$$

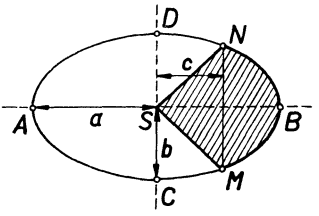


Fig. 3.9.

(e) The Ellipse (Fig. 3.9). Let $\overline{SA} = a$ be the semi-major axis, $\overline{SC} = b$ the semi-minor axis, $e = \sqrt{(a^2 - b^2)}/a$ the eccentricity of the ellipse, C its circumference, P its area. The following relations hold:

$$P = \pi ab\,, \tag{39}$$

$$C = 4aE(e, \tfrac{1}{2}\pi) \tag{40}$$

where $E(e, \frac{1}{2}\pi) = \int_0^{\pi/2} \sqrt{(1 - e^2\sin^2\varphi)}\,\mathrm{d}\varphi$ is the so-called complete elliptic integral of the second kind (see § 13.12). The following approximate formulae hold for the

circumference of an ellipse:

$$C \approx \pi[1{\cdot}5(a+b) - \sqrt{(ab)}], \quad C \approx \pi(a+b)\frac{64-3l^4}{64-16l^2}, \quad \text{where } l = \frac{a-b}{a+b}. \tag{41}$$

The circumference of an ellipse with semi-axes a, b can be calculated by using Table 3.2. The circumference C is given by the formula $C = ak$.

The centroid of an ellipse lies at the centre S.

The moments of inertia about the axes a, b are as follows:

$$I_a = \tfrac{1}{4}\pi ab^3, \quad I_b = \tfrac{1}{4}\pi a^3 b. \tag{42}$$

An *elliptic sector* (Fig. 3.9, the shaded area) has the area

$$P = ab \arccos\frac{c}{a}. \tag{43}$$

REMARK 4. For further properties of the ellipse see §§ 4.2 and 5.10.

TABLE 3.2

$\frac{b}{a}$	k	$\frac{b}{a}$	k	$\frac{b}{a}$	k	$\frac{b}{a}$	k	$\frac{b}{a}$	k
0·00	4·000 0	0·20	4·202 0	0·40	4·602 6	0·60	5·105 4	0·80	5·672 3
01	C01 1	21	218 6	41	625 8	61	132 4	81	702 0
02	003 7	22	235 6	42	649 2	62	159 6	82	731 7
03	007 8	23	253 1	43	672 8	63	187 0	83	761 5
04	013 1	24	271 0	44	696 6	64	214 5	84	791 5
05	019 4	25	289 2	45	720 7	65	242 1	85	821 5
06	026 7	26	307 8	46	745 0	66	269 9	86	851 6
07	034 8	27	326 8	47	769 5	67	297 8	87	881 9
08	043 8	28	346 2	48	794 2	68	325 9	88	912 2
09	053 5	29	365 9	49	819 1	69	354 1	89	942 6
0·10	064 0	0·30	385 9	0·50	844 2	0·70	382 4	0·90	973 2
11	075 2	31	406 2	51	869 5	71	410 8	91	6·003 8
12	087 0	32	426 9	52	895 0	72	439 4	92	034 5
13	099 4	33	447 9	53	920 7	73	468 1	93	065 3
14	112 5	34	469 2	54	946 6	74	496 9	94	096 2
15	126 1	35	490 8	55	972 6	75	525 8	95	127 1
16	140 3	36	512 6	56	998 8	76	554 9	96	158 2
17	155 0	37	534 7	57	5·025 2	77	584 1	97	189 3
18	170 2	38	557 1	58	051 8	78	613 4	98	220 5
19	185 9	39	579 7	59	078 5	79	642 8	99	251 8

(f) The Hyperbola (Fig. 3.10). Let $OA = a$ be the semi-major axis, $OC = b$ be the semi-minor axis, $e = \sqrt{(a^2 + b^2)}/a$ the eccentricity.

A *segment MBN of the hyperbola* (Fig. 3.10, the shaded area) has the area

$$P = xy - ab \ln\left(\frac{x}{a} + \frac{y}{b}\right) = xy - ab \operatorname{arcosh} \frac{x}{a}, \tag{44}$$

where $y = (b/a)\sqrt{(x^2 - a^2)}$.

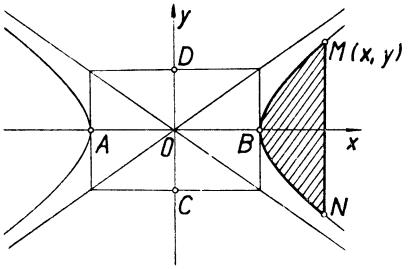


Fig. 3.10.

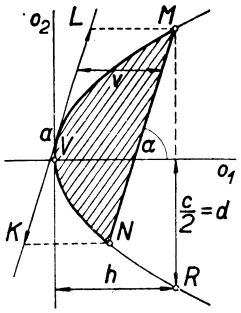


Fig. 3.11.

REMARK 5. For further properties of the hyperbola see § 4.3, 5.11.

(g) The Parabola (Fig. 3.11). The area of a *segment MVN of the parabola* (the shaded area) is

$$P = \tfrac{2}{3}av \sin \alpha\,; \tag{45}$$

it is thus equal to two-thirds of the area of the parallelogram $KLMN$.

The length l of an arc MVR of a parabola is

$$l = \tfrac{1}{2}\sqrt{[c^2 + (4h)^2]} + \frac{c^2}{8h} \ln\left[\frac{4h}{c} + \frac{1}{c}\sqrt{[c^2 + (4h)^2]}\right]; \tag{46}$$

the following relation holds approximately (for small h/c):

$$l \approx c\left[1 + \frac{8}{3}\left(\frac{h}{c}\right)^2 - \frac{32}{5}\left(\frac{h}{c}\right)^4\right]. \tag{47}$$

The centroid T of a parabolic segment MVR (Fig. 3.11) lies on the axis o_1 of the parabola; its distance from the vertex V is

$$\overline{TV} = \tfrac{3}{5}h\,. \tag{48}$$

The moments of inertia of a parabolic segment MVR about the axes o_1, o_2 (Fig. 3.11) are

$$I_{o_1} = \tfrac{4}{15}hd^3\,, \quad I_{o_2} = \tfrac{4}{7}h^3d\,. \tag{49}$$

REMARK 6. For further properties of the parabola see § 4.4, 5.12.

3.2. Volume, Surface, Centroid and Moments of Inertia of Solids

REMARK 1. For the calculation of volumes and surfaces of solids by means of integrals see § 14.9.

REMARK 2. In the following text, V always denotes the volume, S the total and Q the lateral area* of the surface of the respective solids.

(a) The Prism (Fig. 3.12). Let a be the length of the lateral edge or the slant height, h the height of the prism (i.e. the distance between the planes of the upper and lower bases), P the area of the base, N the area of the normal section (the plane section which is perpendicular to the lateral edges). Then

$$V = Ph = Na\,, \tag{1}$$

$$Q = C_N a\,, \quad S = 2P + C_N a \tag{2}$$

where C_N is the circumference of the normal section.

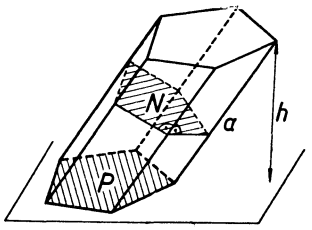


Fig. 3.12.

The centroid lies at the mid-point of the segment connecting the centroids of the two bases of the prism.

A *truncated triangular prism* (i.e. cut off by a plane non-parallel to the plane of the base; Fig. 3.13), whose lateral edges are of lengths a, b, c, has the volume

$$V = \tfrac{1}{3}N(a + b + c)\,. \tag{3}$$

A *parallelepiped* is a prism, the base of which is a parallelogram.

* i.e. area of the slant faces or of the curved surface.

A *right parallelepiped* (i.e. a right prism, the base of which is a rectangle or a square), whose edges are of lengths a, b, c, has the volume

$$V = abc \tag{4}$$

and the surface area

$$S = 2(ab + ac + bc) . \tag{5}$$

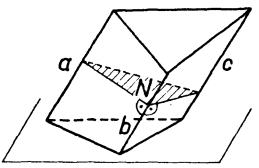


Fig. 3.13.

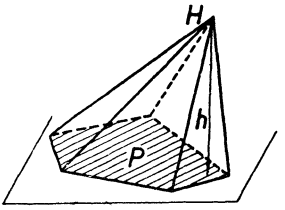


Fig. 3.14.

The moment of inertia about the median axis o which is parallel to the edge c is

$$I_o = \tfrac{1}{12}abc(a^2 + b^2) .$$

A *cube* is a right parallelepiped whose edges are of the same length a; the volume and surface are given by

$$V = a^3 , \quad S = 6a^2 . \tag{6}$$

(b) The Pyramid (Fig. 3.14). Let h be the height of the pyramid (the distance of the apex H from the plane of the base), P the area of the base. Then

$$V = \tfrac{1}{3}Ph . \tag{7}$$

The centroid lies on the segment connecting the apex and the centroid of the base; its distance from the base is $\frac{1}{4}h$.

A *triangular pyramid* with one vertex at the origin of the cartesian coordinate system, the other three vertices being (x_i, y_i, z_i) $(i = 1, 2, 3)$, has the volume equal to one-sixth of the absolute value of the determinant

$$D = \begin{vmatrix} x_1, & y_1, & z_1 \\ x_2, & y_2, & z_2 \\ x_3, & y_3, & z_3 \end{vmatrix} , \quad \text{i.e.} \quad V = \tfrac{1}{6}|D| .$$

A *regular pyramid* (i.e. a pyramid whose base is a regular polygon, and whose altitude passes through the centre of the base). The lateral area

$$Q = \tfrac{1}{2}Cl \tag{8}$$

where C is the circumference of the base and l the length of the perpendicular from the apex to (any) of the edges of the base.

A *frustum of a pyramid* (Fig. 3.15) (the bases lie in parallel planes). Let P_1, P_2 be the areas of the bases and h the height (the distance between the two bases). Then

$$V = \tfrac{1}{3}h[P_1 + P_2 + \sqrt{(P_1P_2)}]\,. \tag{9}$$

If a frustum of a pyramid is *regular*, then the lateral area is

$$Q = \tfrac{1}{2}(C_1 + C_2)\,l \tag{10}$$

where C_1, C_2 are the circumferences of the bases and l is the altitude of the trapezoid formed by an (arbitrary) lateral face.

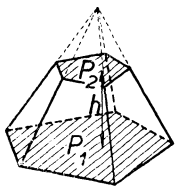


Fig. 3.15.

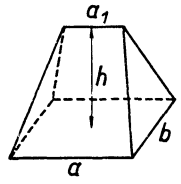


Fig. 3.16a.

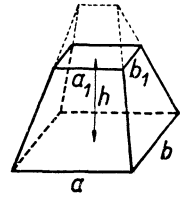


Fig. 3.16b.

REMARK 3. (a) A *dihedral angle* (Fig. 3.16a) (the base is a rectangle with sides a, b, one pair of opposite slant faces is formed by two congruent isosceles triangles, the other pair by two congruent isosceles trapezia). The volume:

$$V = \tfrac{1}{6}(2a + a_1)\,bh\,. \tag{11}$$

The centroid lies on the segment connecting the centre of the upper edge a_1 and the centre of the base; its distance from the base is

$$z = \frac{h(a + a_1)}{2(2a + a_1)}\,. \tag{12}$$

(b) An *obelisk* (Fig. 3.16b) (the bases are rectangles with sides a, b and a_1, b_1, the opposite slant faces make the same angle with the base, but they do not intersect at one point). The volume

$$V = \tfrac{1}{6}h[(2a + a_1)\,b + (2a_1 + a)\,b_1]\,. \tag{13}$$

The centroid lies on the segment connecting the centres of the two bases; its distance from the lower base is

$$z = \frac{h}{2}\,\frac{a(b + b_1) + a_1(b + 3b_1)}{a(2b + b_1) + a_1(b + 2b_1)}\,. \tag{14}$$

(c) The Cylinder (Fig. 3.17). Let h be the height, l the length of the side, P the area of the base, N the area of the normal section (plane section per-

pendicular to the sides) of the cylinder. Then, the volume V and the lateral area Q of the cylinder are given by:

$$V = Ph = Nl\,, \quad Q = C_P h = C_N l\,, \tag{15}$$

where C_P, and C_N, are the circumferences of the base, and of the normal section, respectively.

The centroid lies at the mid-point of the segment connecting the centroids of upper and lower bases of the cylinder.

A *right circular cylinder*. The base is a circle of radius r, lying in a plane which is perpendicular to the side of the cylinder, h is the height. Then

$$V = \pi r^2 h\,, \quad Q = 2\pi r h\,,$$
$$S = 2\pi r(r + h)\,. \tag{16}$$

The moment of inertia about the axis of revolution o is

$$I_o = \tfrac{1}{2}\pi r^4 h\,. \tag{17}$$

A *truncated right circular cylinder* (Fig. 3.18). Let h_1 be the shortest and h_2 the longest side of the cylinder. Then

$$V = \pi r^2 \frac{h_1 + h_2}{2}\,, \quad Q = \pi r(h_1 + h_2)\,,$$
$$S = \pi r\left[\left[h_1 + h_2 + r + \sqrt{\left[r^2 + \left(\frac{h_2 - h_1}{2}\right)^2\right]}\right]\right]. \tag{18}$$

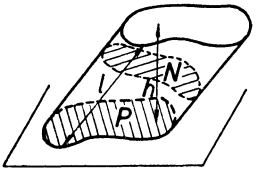


Fig. 3.17.

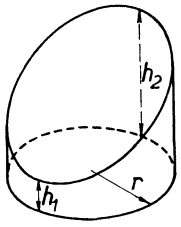


Fig. 3.18.

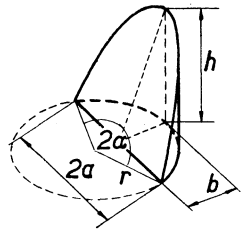


Fig. 3.19.

A *segment of a right circular cylinder* – a *cylindrical angle* (Fig. 3.19). Using the notation of Fig. 3.19, we have

$$V = \frac{h}{3b}\left[a(3r^2 - a^2) + 3r^2(b - r)\,\alpha\right] = \frac{hr^3}{b}\left(\sin\alpha - \tfrac{1}{3}\sin^3\alpha - \alpha\cos\alpha\right),$$
$$Q = \frac{2rh}{b}\left[(b - r)\,\alpha + a\right], \tag{19}$$

with the angle α measured in radians ($0 < \alpha \leqq \pi$). For $\alpha = \frac{1}{2}\pi$, we have $a = b = r$ and

$$V = \tfrac{2}{3}r^2h\,, \quad Q = 2rh\,. \tag{20}$$

A *hollow right circular cylinder* – a *tube* (Fig. 3.20). Let r be the inner radius, R the outer radius, $a = R - r$ the thickness, $\varrho = \frac{1}{2}(r + R)$ the mean radius, h the height. Then

$$V = \pi(R^2 - r^2)\,h = \pi ah(2R - a) = \pi ah(2r + a) = 2\pi\varrho ah\,. \tag{21}$$

The moment of inertia about the axis of revolution o is

$$I_o = \tfrac{1}{2}\pi h(R^4 - r^4)\,. \tag{22}$$

(d) The Cone (Fig. 3.21). Let h be the height, P the area of the base. Then

$$V = \tfrac{1}{3}Ph\,. \tag{23}$$

The centroid lies on the segment connecting the apex and the centroid of the base; its distance from the base is $\frac{1}{4}h$.

A *right circular cone.* Its base is a circle of radius r, and the line passing through the apex and through the centre of the base (the axis of the cone) is perpendicular to the plane of the base; let h be the height. Then

$$V = \tfrac{1}{3}\pi r^2h\,, \quad Q = \pi rl\,, \quad S = \pi r(r + l) \tag{24}$$

where $l = \sqrt{(r^2 + h^2)}$ is the length of the side of the cone.

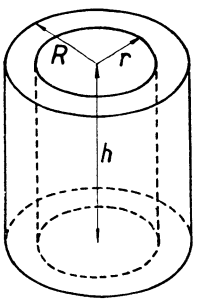


Fig. 3.20.

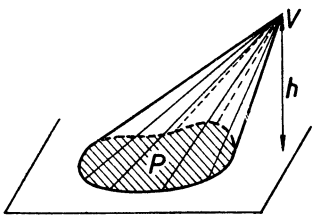


Fig. 3.21.

The moment of inertia about the axis of revolution o is

$$I_o = \tfrac{1}{10}\pi r^4h\,. \tag{25}$$

A *frustum of a right circular cone* (Fig. 3.22). Using the notation of Fig. 3.22,

we have

$$V = \tfrac{1}{3}\pi h(R^2 + Rr + r^2), \quad Q = \pi(R + r)\, a \tag{26}$$

where $a = \sqrt{[h^2 + (R - r)^2]}$ is the length of the side of the frustum.

The centroid lies on the axis of revolution o; its distance from the lower base (of radius R) is

$$z = \frac{h(R^2 + 2Rr + 3r^2)}{4(R^2 + Rr + r^2)}. \tag{27}$$

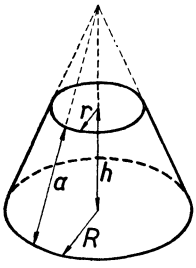


Fig. 3.22.

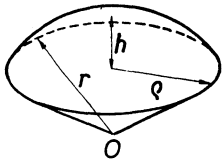


Fig. 3.23.

The moment of inertia about the axis of revolution o is

$$I_o = \frac{\pi h(R^5 - r^5)}{10(R - r)}. \tag{28}$$

(e) The Sphere. If r is the radius of the sphere, then

$$V = \tfrac{4}{3}\pi r^3 \approx 4{\cdot}188\,8r^3, \quad S = 4\pi r^2 \approx 12{\cdot}566r^2. \tag{29}$$

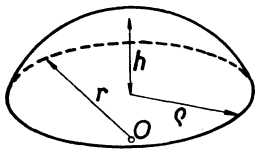


Fig. 3.24.

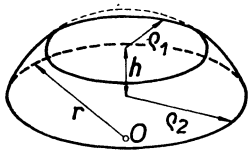


Fig. 3.25.

The moment of inertia about the axis o passing through the centre of the sphere

$$I_o = \tfrac{8}{15}\pi r^5. \tag{30}$$

A *sector of a sphere* (Fig. 3.23). Using the notation of Fig. 3.23, we have

$$V = \tfrac{2}{3}\pi r^2 h, \quad S = \pi r(2h + \varrho). \tag{31}$$

A *segment of a sphere* (Fig. 3.24). Using the notation of Fig. 3.24, we have

$$V = \tfrac{1}{6}\pi h(3\varrho^2 + h^2) = \tfrac{1}{3}\pi h^2(3r - h)\,,$$
$$S = 2\pi rh + \pi\varrho^2\,, \quad Q = 2\pi rh\,. \tag{32}$$

A *spherical layer* (Fig. 3.25). Using the notation of Fig. 3.25, we have

$$V = \tfrac{1}{6}\pi h(3\varrho_2^2 + 3\varrho_1^2 + h^2)\,,$$
$$S = \pi(2rh + \varrho_1^2 + \varrho_2^2)\,, \quad Q = 2\pi rh\,. \tag{33}$$

A *spherical ring* is the part of a spherical layer, obtained by removing from it an inscribed frustum of a cone (or a cylinder). If a is the length of the side of the inscribed frustum of cone (or of the cylinder), then the volume V of the spherical ring is

$$V = \tfrac{1}{6}\pi ha^2\,. \tag{34}$$

(f) The Ellipsoid with semi-axes a, b, c has the lateral area

$$S = 2\pi c^2 + \frac{2\pi b}{\sqrt{(a^2 - c^2)}}\left[c^2F(k, \varphi) + (a^2 - c^2)\,E(k, \varphi)\right], \tag{35}$$

where

$$k = \frac{a}{b}\sqrt{\frac{b^2 - c^2}{a^2 - c^2}}\,, \quad \varphi = \arccos\frac{c}{a}$$

and $F(k, \varphi)$, $E(k, \varphi)$ are the elliptic integrals of the first and second kinds (see § 13.12, p. 552).

The volume of an ellipsoid

$$V = \tfrac{4}{3}\pi abc\,. \tag{36}$$

A *prolate spheroid* is formed when an ellipse with semi-axes a, b $(a > b)$ is rotated around its major axis; its surface is

$$S = 2\pi\left(b^2 + ab\frac{\arcsin e}{e}\right), \quad e = \frac{\sqrt{(a^2 - b^2)}}{a}\,. \tag{37}$$

An *oblate spheroid* is formed when an ellipse with semi-axes a, b $(a > b)$ is rotated around the minor axis; its surface is

$$S = 2\pi\left(a^2 + \frac{b^2}{2e}\ln\frac{1 + e}{1 - e}\right), \quad e = \frac{\sqrt{(a^2 - b^2)}}{a}\,. \tag{38}$$

The moment of inertia of a spheroid about the semi-axis a is

$$I_a = \tfrac{8}{15}\pi ab^4\,. \tag{39}$$

(g) The Paraboloid of Revolution (Fig. 3.26). The volume bounded by a paraboloid of revolution and by a plane perpendicular to its axis at distance h from the vertex O (the radius of the base being r) is

$$V = \tfrac{1}{2}\pi r^2 h \tag{40}$$

the lateral area is

$$Q = \frac{\pi r}{6h^2}\left[(r^2 + 4h^2)^{3/2} - r^3\right]. \tag{41}$$

The centroid lies on the axis of revolution o; its distance from the vertex O of the paraboloid is $\frac{2}{3}h$.

The moment of inertia about the axis o is

$$I_o = \tfrac{1}{6}\pi r^4 h\,. \tag{42}$$

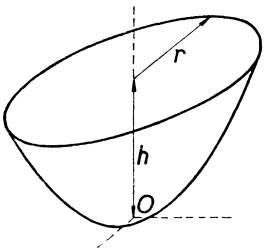


Fig. 3.26.

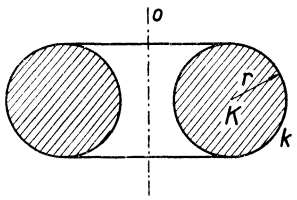


Fig. 3.27.

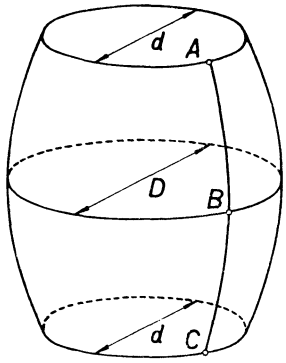


Fig. 3.28.

(h) The Torus (annuloid, ring) (Fig. 3.27) is formed by rotation of a circle k of radius r, with centre K around the axis o, lying in the plane of the circle at distance R $(R > r)$ from the centre K.

$$V = 2\pi^2 R r^2 \approx 19{\cdot}739 R r^2\,, \tag{43}$$

$$S = 4\pi^2 R r \approx 39{\cdot}478 R r\,. \tag{44}$$

The moment of inertia of a torus about the axis of revolution o is

$$I_o = \tfrac{1}{2}\pi^2 R r^2 (4R^2 + 3r^2)\,. \tag{45}$$

(i) The Cask (Fig. 3.28). The diameter of the upper and lower bases is d, the diameter of the central section is D, the height is h.

For a *circular* shape (ABC being an arc of a circle)

$$V \approx 0{\cdot}262 h(2D^2 + d^2)\,. \tag{46}$$

For a *parabolic* shape (ABC being an arc of a parabola)

$$V \approx \frac{\pi}{60}\, h(8D^2 + 4Dd + 3d^2)\,. \tag{47}$$

4. PLANE CURVES AND CONSTRUCTIONS

By KAREL DRÁBEK

References: [6], [118], [143], [155], [162], [187].

4.1. The Circle

A circle (for the definition see § 5.9) with centre S and radius r will be denoted by $k(S, r)$.

By the *construction of a circle* we mean the determination of its centre and radius from certain given conditions (with the help of fundamental theorems of plane geometry).

Theorem 1. *The circle is axially symmetrical about any line passing through its centre S (and called a diameter) and, hence, it is radially symmetrical about its centre S* (Fig. 4.1).

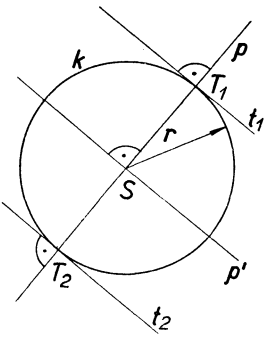


Fig. 4.1.

Theorem 2. *The tangent at a point of a circle is perpendicular to the line connecting this point and the centre of the given circle; consequently, all the normals of a circle pass through the centre of the circle.*

Theorem 3. *The tangents at the points of intersection of a circle and a diameter are parallel.*

In what follows the term diameter will normally be used in the sense of a so-called *bounded diameter*, i.e. the segment determined by the points of intersection of the diameter and the circle (ellipse, hyperbola, etc.), or its length.

Definition 1. The diameter of a circle parallel to the tangents at the end points of a given diameter is called the *conjugate diameter* to the original diameter.

Hence, conjugate diameters of a circle are perpendicular.

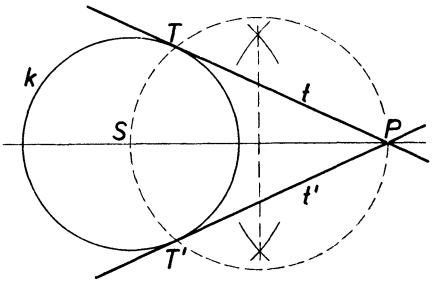


Fig. 4.2.

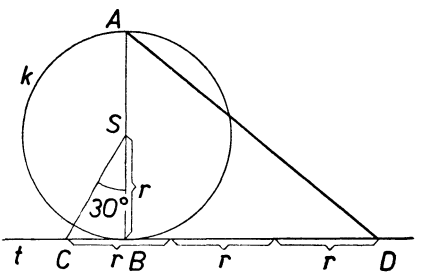


Fig. 4.3.

Construction 1 of the tangents to a circle $k(S, r)$ from an external point P (i.e. from a point whose distance from the centre S of the circle k is $d > r$): the circle constructed on the diameter PS, the so-called *Thalet's circle*, meets the given circle k at two points T, T' (Fig. 4.2) which are the points of contact of the tangents $t \equiv PT$, $t' \equiv PT'$ from the given point P.

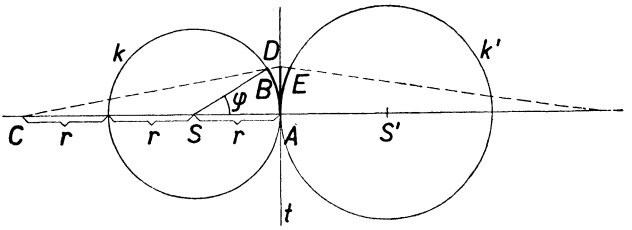


Fig. 4.4.

Definition 2. The construction of a segment equal in length to the circumference of a circle, or of a circular arc, is called the *rectification of the circle*, or of the circular arc, respectively.

In practice, i.e. using a ruler and a pair of compasses, these constructions for a circle are only approximate.

Construction 2 (*Kochański's rectification of a circle* (Fig. 4.3)). At the point B of a diameter AB we construct the tangent t and determine the point C of intersection

of t and of the other arm of the angle $BSC = 30°$. On CB produced we find the point D such that $\overline{CD} = 3r$. Then $\overline{AD} \approx \pi r$.

Since the error reaches only 1 mm for $r \approx 17$ m, it need not be taken into account in our constructions.

Construction 3 (*Sobotka's rectification of a circular arc* $\overset{\frown}{AB}$ (Fig. 4.4)). We determine the point C on the half-line AS such that $\overline{AC} = 3r$. The line CB meets the tangent t constructed at the point A of the circle k at a point D. Then $\overline{AD} \approx \overset{\frown}{AB}$.

This construction is very accurate for arcs corresponding to angles $\varphi \leqq 30°$. For example, for $\varphi = 30°$ we get an error of 1 mm for $r \approx 2{\cdot}5$ m. Therefore, greater arcs are divided into parts in order to rectify arcs corresponding to angles $\varphi \leqq 30°$ with sufficient accuracy.

By an inverse construction we can wind a given segment onto a circle or transfer an arc of a circle onto another circle (Fig. 4.4).

4.2. The Ellipse

For the definition of the ellipse see § 5.10 (p. 183). We denote the foci by F_1, F_2 (Fig. 4.5); the line connecting a point of the ellipse and a focus is called a *focal radius*.

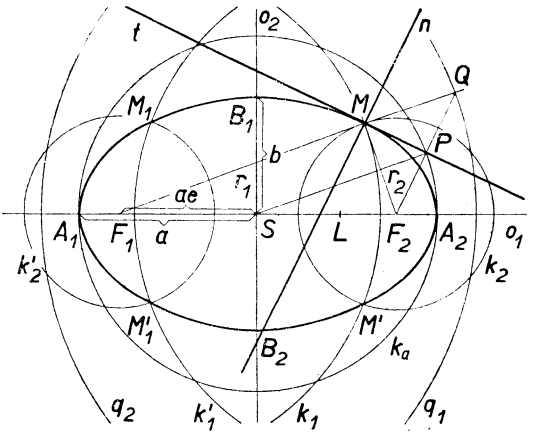


Fig. 4.5.

Theorem 1. *The ellipse is a curve symmetrical about the axis connecting both foci (the major axis) and about the perpendicular bisector of the segment* F_1F_2 *(the minor axis) and hence it is radially symmetrical about the point S of intersection of the axes of the ellipse (the centre of the ellipse).*

The points A_1, A_2 of the ellipse on the major axis are called the *major vertices*, the points B_1, B_2 on the minor axis the *minor vertices*. The length $\overline{A_1S} = \overline{A_2S} = a$ is called the *semi-major axis*, and the length $\overline{B_1S} = \overline{B_2S} = b$ the *semi-minor axis*. Let the length $F_1S = F_2S = ae$, so that the ratio $F_1S/A_1S = e$. Then e is called the *eccentricity* of the ellipse.

Theorem 2. *Between the lengths a, b and the eccentricity e, the relation $a^2e^2 = a^2 - b^2$ holds.*

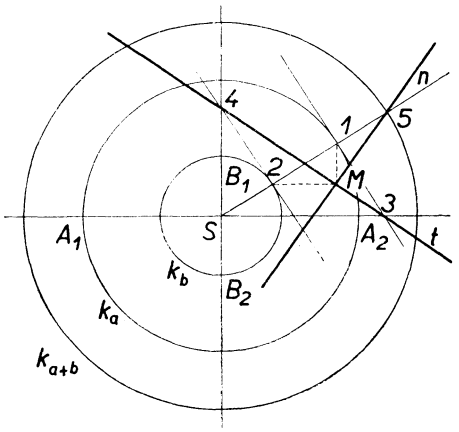


Fig. 4.6.

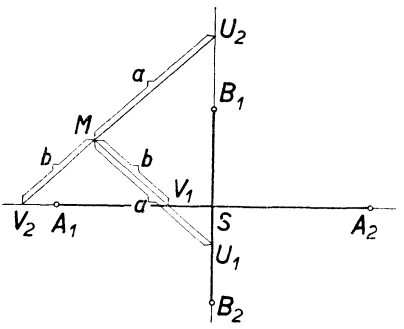


Fig. 4.7.

Construction 1 of points of an ellipse with semi-axes a, b by means of its definition (Fig. 4.5): We determine the major and minor vertices of the ellipse and, by Theorem 2, we determine the foci F_1, F_2 on the major axis. We choose an arbitrary point L between the points F_1, F_2 and describe circles of radii $\overline{A_1L}$, about one focus and $\overline{A_2L}$ about the other. The points of intersection M, M' of these circles k_1, k_2 are points of the ellipse. By an interchange of the foci as centres of the constructed circles, we get two further points M_1, M_1' of the ellipse. This construction is not accurate in the vicinity of the major vertices A_1, A_2.

Construction 2 of points of an ellipse with given semi-axes using affinity with a circle (Fig. 4.6): Let the (*vertex*) circles k_a, k_b with centres at the point S and radii a, b be cut by a radius from the point S at the points *1*, *2*. The line through the point *1* parallel to the minor axis and the line through the point *2* parallel to the major axis intersect at a point M of the ellipse. The construction is always accurate, for the auxiliary lines intersect at right angles.

Construction 3 of an ellipse with given semi-axes a, b (Fig. 4.7).

(a) By means of the difference of the semi-axes: If the segment $\overline{U_1V_1} = a - b$ is moved along two perpendicular lines, then the point M (exterior to the segment $\overline{U_1V_1}$) describes the ellipse with semi-axes $\overline{MU_1} = a$, $\overline{MV_1} = b$.

(b) By means of the sum of the semi-axes: If the segment $\overline{U_2V_2} = a + b$ is moved along two perpendicular lines, then the point M (interior to the segment $\overline{U_2V_2}$) describes the ellipse with semi-axes $\overline{MU_2} = a$, $\overline{MV_2} = b$.

This construction is often used to determine the length of one of the semi-axes, given the other semi-axis, the position of the axes and a point of the ellipse.

Theorem 3. *The tangent, or the normal, at a given point of an ellipse bisects the angle between the focal radii which contains, or does not contain, a major vertex of the ellipse, respectively.*

Construction 4 of the tangent and normal at a point M of an ellipse using the ciclres k_a, k_b, k_{a+b} (Fig. 4.6): The required tangent is the line connecting the point M of the ellipse and the point of intersection *3* (or *4*) of the tangent constructed at the point *1* (or 2) of the circle k_a (or k_b) and the major (or the minor) axis of the ellipse. The normal to the ellipse at the point M joins the point M and the point *5* which is the point of intersection of the half line *S1* and the circle k_{a+b} (of centre S and radius $a + b$).

The following theorems are important for the construction of tangents from an external point of an ellipse and for some constructions of the ellipse (Fig. 4.5).

Theorem 4. *The locus of points Q which are reflections of one focus of an ellipse in its tangents is the circle q having its centre at the other focus and radius 2a.*

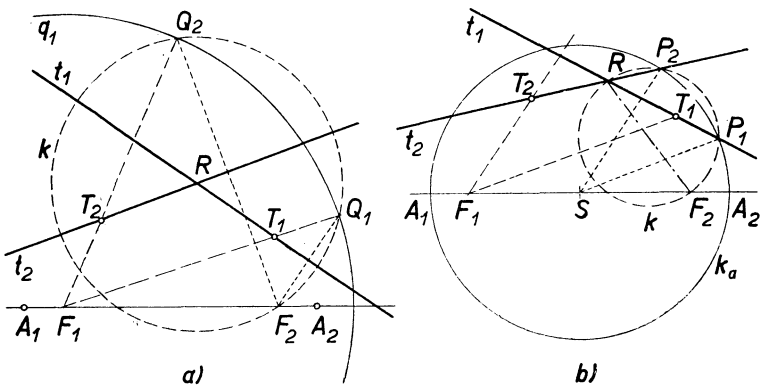


Fig. 4.8.

Theorem 5. *The locus of the feet P of perpendicular lines dropped from the foci of an ellipse to its tangents is the vertex circle $k_a(S, a)$.*

Theorem 6. *The locus of the centres of circles touching the circle $q_2(F_2, 2a)$ and passing through its internal point F_1 is the ellipse with foci F_1, F_2 and with its major axis of length 2a.*

Theorem 7. *Let the vertex of a right angle move along the circle $k_a(S, a)$ so that one of its arms passes through an internal point F_1 of the circle k_a; then the other arm is a tangent to the ellipse with focus F_1, centre S and semi-major axis of length a.*

Construction 5 of tangents to an ellipse from an external point R:

(a) By means of the circle q_1 (Fig. 4.8a): We determine the points of intersection Q_1, Q_2 of the circles $k(R, \overline{RF_2})$ and $q_1(F_1, 2a)$. The perpendicular bisectors of the segments $\overline{Q_1F_2}$, $\overline{Q_2F_2}$ are the tangents t_1, t_2 from the point R to the ellipse. The

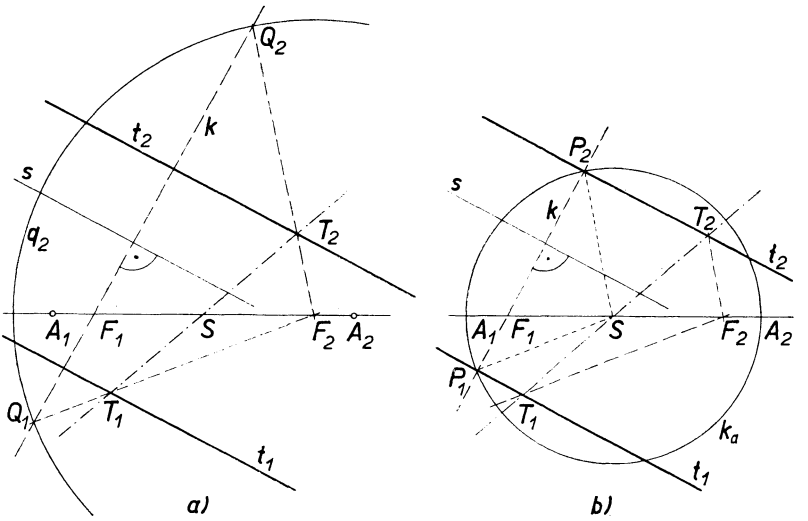


Fig. 4.9.

points of contact T_1, T_2 are the points of intersection of the tangents t_1, t_2 and the lines connecting the points Q_1, Q_2 and the focus F_1 (i.e. the focus about which the circle q_1 is described).

(b) By means of the vertex circle $k_a(S, a)$ (Fig. 4.8b): We determine the points of intersection P_1, P_2 of the circle k_a and the Thalet circle drawn on the diameter RF_2. The lines connecting P_1 and P_2 and the point R are tangents t_1, t_2 of the ellipse. The points of contact T_1, T_2 are the points of intersection of the tangents t_1, t_2 and the lines through the focus F_1 parallel to SP_1, SP_2, respectively.

Construction 6 of tangents to an ellipse, which are parallel to a given direction s:

(a) By means of the circle q_2 (Fig. 4.9a): The line k through the point F_1 perpendicular to the direction s intersects the circle $q_2(F_2, 2a)$ at points Q_1, Q_2; the perpendicular bisectors of the segments Q_1F_1, Q_2F_1 are the required tangents t_1, t_2.

(b) By means of the vertex circle k_a (Fig. 4.9b): The line k through the point F_1 perpendicular to the direction s intersects the circle $k_a(S, a)$ at points P_1, P_2: then the required tangents t_1, t_2 pass through P_1, P_2 and are parallel to the direction s.

The line connecting the points of contact T_1, T_2 of the parallel tangents t_1, t_2 passes through the centre S of the ellipse and is called the *conjugate diameter to the direction s*. Tangents to an ellipse parallel to a given direction a always exist.

Construction 7 (*the Rytz construction*) of the axes of an ellipse given by conjugate diameters M_1M_2, N_1N_2: On the perpendicular erected to one of the diameters, say M_1M_2, at the centre S (Fig. 4.10), we draw the segment $\overline{SR} = \overline{M_1S}$, join the points R and N_1 and describe a circle through the centre of the ellipse about the

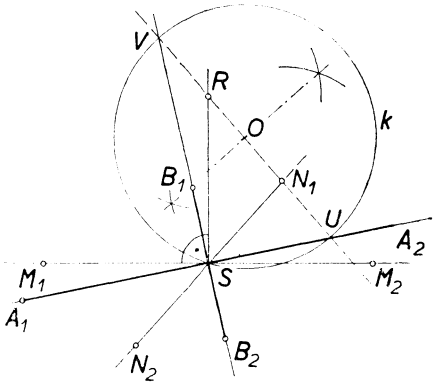


Fig. 4.10.

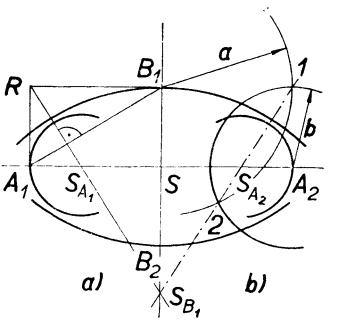


Fig. 4.11.

point O as centre, where O is the mid-point of RN_1. This circle intersects the line RN_1 in two points U, V through which the required axes pass (the major axis lies always within the acute angle made by the given conjugate diameters). Furthermore, $a = \overline{RU} = \overline{N_1V}$, $b = \overline{RV} = \overline{N_1U}$.

Construction 8 of the centres of curvature at the vertices of an ellipse:

(a) A perpendicular dropped from the vertex R of the rectangle SA_1RB_1 (Fig. 4.11a) to its diagonal A_1B_1 intersects the major, or minor axis at the centre of curvature corresponding to the major, or minor vertex of the ellipse, respectively.

(b) The line connecting the points of intersection *1* and *2* of the circles $k_1(A_2, b)$ $k_2(B_1, a)$ intersects the major, or minor axis at the required centres of curvature (Fig. 4.11b).

The circle with its centre at a centre of curvature constructed as above, which passes through the corresponding vertex (the *osculating circle* of the vertex) approximates to the given ellipse in the neighbourhood of the vertex.

4.3. The Hyperbola

For the definition of the hyperbola see § 5.11 (p. 184). We denote the foci by F_1, F_2 (Fig. 4.12); by a *focal radius*, denoted by r_1, r_2, we shall again mean a line connecting a point of the hyperbola and a focus.

Theorem 1. *The hyperbola is a curve symmetrical about the axis connecting both foci (the major axis) and about their perpendicular bisector (the minor axis) and*

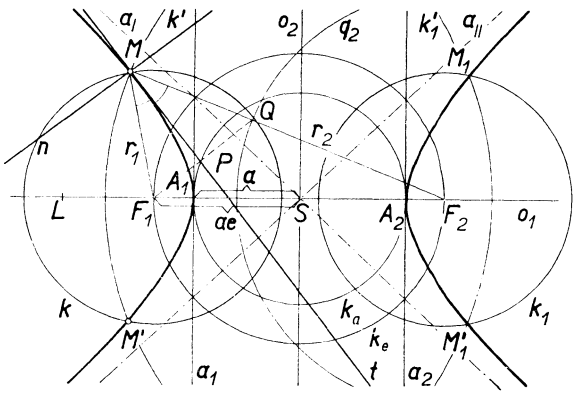


Fig. 4.12.

hence it is radially symmetrical about the point S of intersection of the axes of the hyperbola (the centre of the hyperbola).

The points A_1, A_2 of the hyperbola on the major axis are called the *major vertices*. The length $\overline{A_1S} = \overline{A_2S} = a$ is called the *semi-major axis*. Let the length $F_1S = F_2S = ae$, so that the ratio $F_1S/A_1S = e$. Then e is called the *eccentricity of the hyperbola*.

Construction 1 of points of a hyperbola given by the semi-major axis and focal distance ae (Fig. 4.12): We chose an arbitrary point L outside the segment F_1F_2 and describe circles of radii A_1L about one focus and A_2L about the other. The points of intersection M, M' of these circles are points of the hyperbola. Interchanging the foci as centres of the constructed circles, we get two further points M_1, M_1' of the hyperbola.

From Construction 1, it is evident that the points of a hyperbola lie on two branches. The points of one branch satisfy the relation $r_1 - r_2 = 2a$ while the points of the other branch satisfy $r_2 - r_1 = 2a$. All points of a hyperbola (excepting the major vertices) lie outside the strip bounded by the lines a_1, a_2 parallel to the minor axis and passing through the points A_1, A_2.

Theorem 2. *The tangent, or the normal, at a point of a hyperbola bisects the angle between the focal radii which contains, or does not contain, the major vertices respectively.*

The following theorems are important for the construction of tangents from an external point of a hyperbola (i.e. from a point for which the absolute value of the difference of the focal radii is less than $2a$) and for some constructions of the hyperbola (Fig. 4.12):

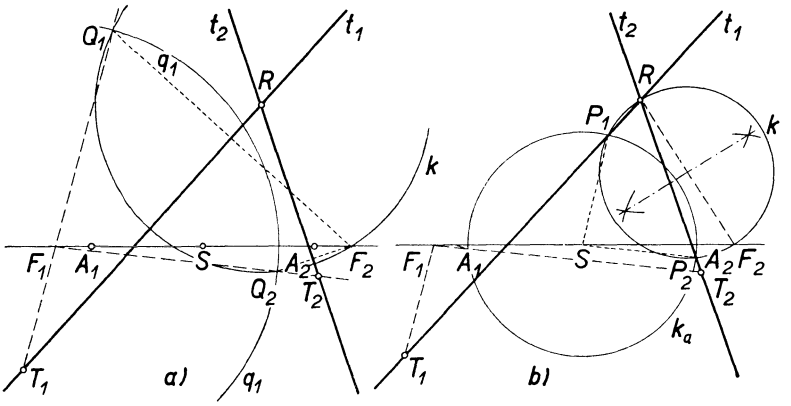


Fig. 4.13.

Theorem 3. *The locus of points Q which are reflections of one focus of a hyperbola in its tangents is the circle q having its centre at the other focus and radius equal to the length of the major axis $2a$.*

Theorem 4. *The locus of the feet P of perpendicular lines dropped from the foci of a hyperbola to its tangents is the (vertex) circle $k_a(S, a)$.*

Theorem 5. *The locus of the centres of circles touching the circle $q_2(F_2, 2a)$ and passing through the external point F_1 of the circle is the hyperbola with foci F_1, F_2 and with its major axis of length $2a$.*

Theorem 6. *Let the vertex of a right angle move along the circle $k_a(S, a)$ so that one of its arms passes through an external point F_1 of the circle k_a; then the other arm is a tangent to the hyperbola with focus F_1 and vertex circle $k_a(S, a)$.*

Construction 2 of tangents to a hyperbola from an external point R (Fig. 4.13a,b):

(a) By means of the circle q: We determine the points of intersection Q_1, Q_2 of the circles $k(R, \overline{RF_2})$ and $q_1(F_1, 2a)$. The perpendicular bisectors of the segments $\overline{Q_1F_2}$, $\overline{Q_2F_2}$ are the tangents t_1, t_2 from the point R to the hyperbola. The points of contact T_1, T_2 are the points of intersection of the tangents t_1, t_2 and the lines connecting the points Q_1, Q_2 and the focus F_1 (about which the circle q_1 is described).

(b) By means of the vertex circle k_a: We determine the points of intersection P_1, P_2 of the Thalet circle k drawn on the diameter RF_2 and the vertex circle k_a. The lines connecting P_1 and P_2 and the point R are tangents t_1, t_2 of the hyperbola. The points of contact T_1, T_2 are the points of intersection of the tangents t_1, t_2 and the lines through the focus F_1 parallel to SP_1, SP_2.

When constructing the tangents from the centre S of a hyperbola, we obtain the points of contact on these tangents a_{I}, a_{II} as points at infinity. We usually extend the locus defining the hyperbola to include these points.

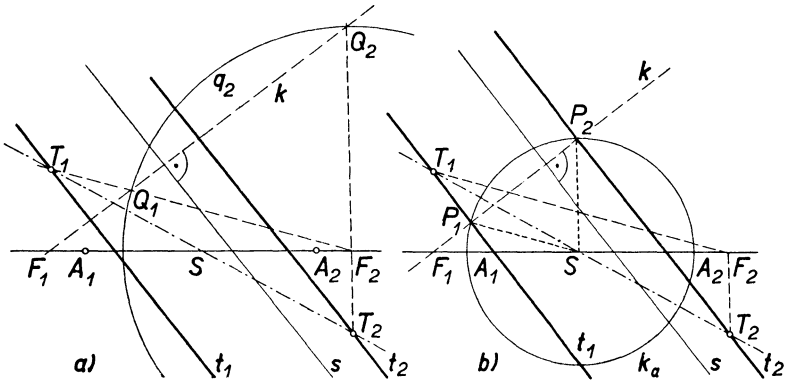


Fig. 4.14.

Definition 1. The tangents a_{I}, a_{II} from the centre S to a hyperbola are called the *asymptotes*; their directions ${}^{s}a_{\mathrm{I}}$, ${}^{s}a_{\mathrm{II}}$ which determine the points of the hyperbola at infinity are called the *directions of the asymptotes.*

Construction 3 of tangents parallel to a given direction s:

(a) By means of the circle q_2 (Fig. 4.14a): The line k through the point F_1 perpendicular to the direction s intersects the circle $q_2(F_2, 2a)$ at points Q_1, Q_2; the perpendicular bisectors of the segments Q_1F_1, Q_2F_1 are the required tangents t_1, t_2 which are parallel to s.

(b) By means of the vertex circle k_a (Fig. 4.14b): The line k through the point F_1 perpendicular to the direction s intersects the circle $k_a(S, a)$ at points P_1, P_2 through which pass the required tangents t_1, t_2 which are parallel to s.

The line connecting the points of contact T_1, T_2 of parallel tangents t_1, t_2 passes through the centre S of a hyperbola and is called the *conjugate diameter to the direction* s.

If φ is the acute angle between an asymptote and the major axis of a hyperbola, and if ψ is the acute angle between the direction s and the major axis, then tangents parallel to the given direction s exist only for $\psi > \varphi$.

The following theorems can be advantageously used when constructing a hyperbola:

Theorem 7. *The segments on an arbitrary secant of a hyperbola (intersecting the asymptotes) between the points of the hyperbola and the asymptotes are equal. In particular: The point of contact of a tangent to a hyperbola bisects its segment between the asymptotes.*

Theorem 8. *The parallelograms formed by the asymptotes of a hyperbola and by the lines constructed through points of the hyperbola parallel to the asymptotes are of a constant area. In particular: The triangles formed by the asymptotes and the tangents of a hyperbola are of a constant area.*

By means of variable parallelograms of constant area we can construct points of a hyperbola with given asymptotes. Further, by means of variable triangles of constant area we can construct tangents to a hyperbola with given asymptotes; in particular the vertex tangent and the vertex of the hyperbola can be determined.

Theorem 9. *The perpendicular drawn to an asymptote of a hyperbola at the point of intersection of the asymptote and a vertex tangent intersects the major axis of the hyperbola at the centre of curvature of the vertex.*

4.4. The Parabola

For the definition of the parabola see § 5.12 (p. 185). The focus will be denoted by F, the directrix by f (Fig. 4.15). The point F does not lie on the line f.

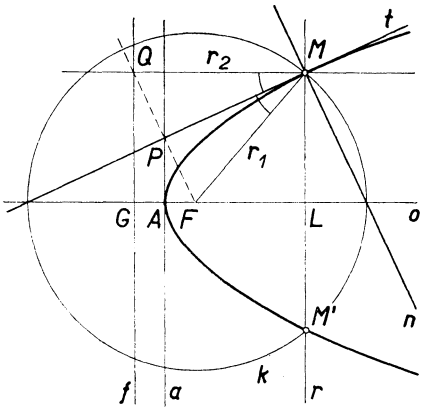


Fig. 4.15.

If M is a point of the parabola, then $MF = r_1$ is one of the focal radii of the point M; as the second (focal) radius the line through the point M perpendicular to the directrix f is to be understood.

The point A of the parabola bisecting the distance of the focus F from the directrix f (this distance is called the *parameter*) is said to be the *vertex of the parabola*; the tangent a at the point A is called the *vertex tangent*.

Theorem 1. *The parabola is a curve symmetrical about the axis, i.e. the line perpendicular to the directrix through the focus F.*

Construction 1 of points of the parabola given by a focus F and a directrix f: Through an arbitrary point L on AF produced we construct the line r parallel to the directrix f. If G is the point of intersection of the axis o and the directrix f, then the points of intersection of the circle $k(F, \overline{GL})$ and the line r are points M, M' of the parabola.

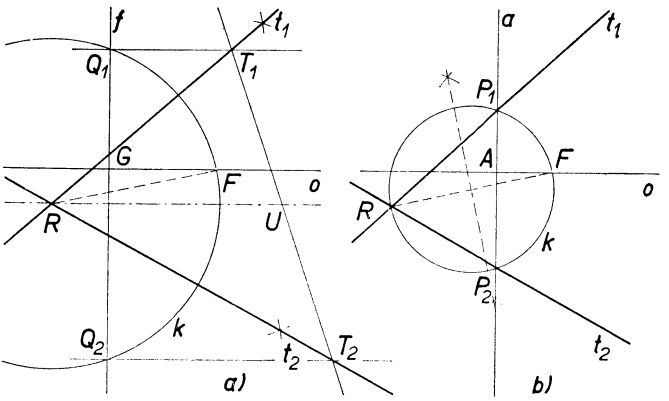


Fig. 4.16.

Theorem 2. *The tangent, or the normal, at a point of the parabola bisects the angle between the focal radii in which the vertex of the parabola lies, or does not lie, respectively.*

Theorem 3. *The locus of points Q which are reflections of the focus F of the parabola in its tangents is the directrix f.*

Theorem 4. *The locus of the feet P of perpendicular lines dropped from the focus F of the parabola to its tangents is the vertex tangent a.*

Theorem 5. *The locus of centres of the circles, touching a line f and passing through a point F (which does not lie on f) is a parabola with focus F and directrix f.*

Theorem 6. *Let the vertex of a right angle move along a straight line a so that one of its arms passes through a point F (which does not lie on a); then the other arm is a tangent to the parabola with focus F and vertex tangent a.*

Construction 2 of tangents to a parabola from an external point R whose distance from the focus F is greater than the distance from the directrix f:

(a) By means of the directrix f (Fig. 4.16a): The circle $k(R, \overline{RF})$ intersects the directrix f at the points Q_1, Q_2; the perpendicular bisectors of the segments Q_1F, Q_2F

are the required tangents t_1, t_2. The points of contact T_1, T_2 are the points of intersection of the tangents t_1, t_2 and the lines through Q_1, Q_2 parallel to the axis of the parabola.

(b) By means of the vertex tangent a (Fig. 4.16b): The circle on the diameter RF intersects the vertex tangent a at the points P_1, P_2 through which pass the required tangents $t_1 \equiv RP_1$, $t_2 \equiv RP_2$. The points of contact should be determined as in (a); consequently, the construction (b) is not convenient in this case.

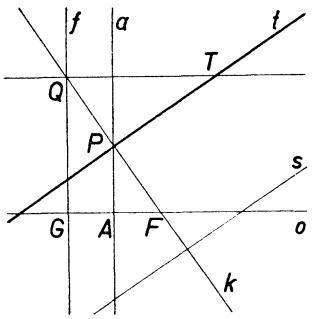


Fig. 4.17.

Construction 3 of the tangent parallel to a given direction s (Fig. 4.17): A perpendicular k from the focus F to the given direction s intersects the vertex tangent a at the point P and the directrix f at the point Q. The tangent $t \parallel s$ passes through the point P (perpendicularly to FQ) and its point of contact T is the point of intersection of the tangent t and the line through Q parallel to the axis of the parabola.

Definition 1. The distance between the point of contact T and the point of intersection of a tangent, or a normal, of the parabola and its axis is called the *length of the tangent*, or the *length of the normal*, briefly the *tangent*, or the *normal*, respectively. The rectangular projection of the tangent, or the normal, onto the axis of the parabola is called the *sub-tangent*, or the *sub-normal*, respectively.

Theorem 7. *A sub-tangent is bisected by the vertex. A sub-normal is of constant length equal to the parameter. The segment which is the sum of the sub-tangent and the sub-normal is bisected by the focus.*

Theorem 8. *The line connecting the point of intersection of two tangents to a parabola and the midpoint of the corresponding chord of contact is parallel to the axis of the parabola (and is called a diameter of the parabola).*

From this theorem, it follows that all diameters of a parabola are parallel.

Theorem 9. *The circle circumscribed about a triangle formed by three tangents to a parabola passes through its focus.*

Theorem 10. *The radius of curvature at the vertex of a parabola is equal to the parameter.*

Construction 4 of a parabola given by two tangents t_1, t_2 with points of contact T_1, T_2: We determine (Fig. 4.18) the diameter p of the parabola by means of the point of intersection R of the tangents t_1, t_2 and the mid-point U of the chord T_1T_2. Denote by *1*, *2* the points of intersection of the lines through the points T_1, T_2 parallel to the diameter p and a line r (arbitrarily chosen) through the point R, respectively. The diagonals $1T_1$, $2T_2$ of the constructed trapezium $12T_1T_2$ meet at a point T of the parabola; the tangent t at T is parallel to r.

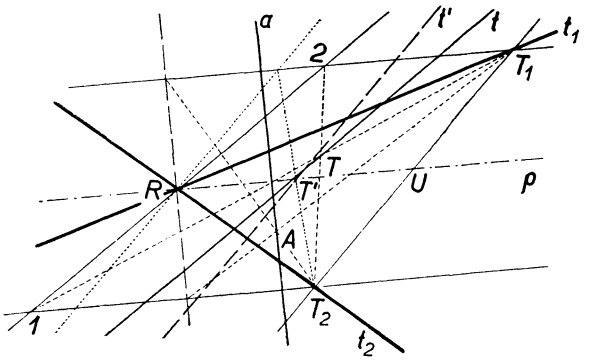


Fig. 4.18.

In particular: If $r \perp p$, we get the vertex A and the vertex tangent a. If $r \parallel T_1T_2$ (then the trapezium becomes a parallelogram) we get a tangent t' parallel to the chord T_1T_2 whose point of contact T' bisects the segment RU.

4.5. Parabolas and Hyperbolas of Higher Degree (Power Curves)

Definition 1. A curve given by the equation

$$y = ax^n \tag{1}$$

is called a *power curve* (a being constant, n rational, x positive, in general). For $n > 1$, we get the so-called *parabolas of higher degree*, for $n < -1$ we get the *hyperbolas of higher degree.*

If $|n| \in (0, 1)$ and $a > 0$, in general, we can write (interchanging the role of the coordinates)

$$x = by^{1/n}, \quad \text{i.e.} \quad x = by^m, \quad \text{where} \quad b = a^{-1/n}, \quad m = 1/n > 1 \quad \text{or} \quad < -1. \tag{2}$$

Theorem 1. *A tangent t (at a given point $P(x_0, y_0)$) cuts off on the y-axis an intercept equal to $(1 - n)\, y_0$. The length of a sub-tangent s_t^x on the x-axis, or s_t^y on the y-axis, is $|x_0/n|$, or $|ny_0|$, respectively.*

Theorem 2a. *The tangent to the parabola* (1), *or* (2), *at the origin O is the x-axis, or y-axis, respectively.*

Theorem 2b. *The asymptotes of the hyperbola* (1) *are the x-axis and y-axis.*

Theorem 3. *The length of an arc of the parabola* (1) *from the point O to the point* $P(x_0, y_0)$ *is given by the integral*

$$s = \int_0^{|x_0|} \sqrt{(1 + a^2 n^2 x^{2n-2})}\, dx ,$$

which can be expressed in an elementary way if $1/(2n - 2)$ *or* $1/(2n - 2) + \frac{1}{2}$ *is an integer.*

Theorem 4. *The area bounded by a parabola of a higher degree, by the x-axis and by the ordinate of a point with abscissa* x_0 *is given by* $P = |x_0 y_0|/(n + 1)$.

Construction 1 of points of the cubical parabola $y = ax^3$ (Fig. 4.19a), or of points of the semicubical parabola $y^2 = ax^3$ (Fig. 4.19b) passing through a given point $P(x_0, y_0)$: We divide the coordinates x_0, y_0 of the point $P(x_0, y_0)$ into an equal number of parts of the same length. If M is the foot of a perpendicular dropped from the point P to the x-axis, we describe a semicircle on MP and erect perpendiculars to the x-axis at the points of subdivision of the segment OM.

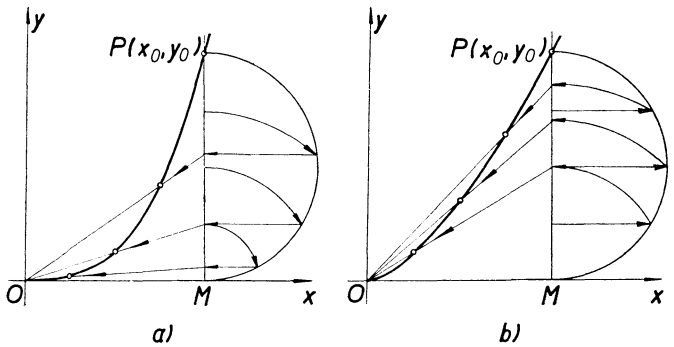


Fig. 4.19.

(a) Points of the *cubical* parabola: The circles with the centre M passing through the points of subdivision of the ordinate MP meet the semicircle on MP at points which we project rectangularly back onto MP. The lines connecting these projections and the origin O intersect the perpendiculars constructed at the points of subdivision of the segment OM at points of a cubical parabola.

(b) Points of the *semicubical* (Neil's) parabola: We project the points of subdivision of the ordinate y_0 parallel to the x-axis onto the semicircle on MP; we turn the points of intersection obtained in this way back onto the segment MP by circles

with centre M. The lines connecting the points so obtained and the origin O intersect the perpendiculars at the points of subdivision of the segment OM at points of a semicubical parabola.

4.6. The Cyclic Curves

Definition 1. By rolling a curve h (the *generating curve* or *moving polhode*), without slipping, along a fixed curve p (the *basic curve* or *fixed polhode*) each point of a plane moving with the curve h describes a curve called a *trochoid.*

Theorem 1. *The fixed polhode is the locus of points which are instantaneous centres of rotation for the respective stages of the motion. The moving polhode is the locus of points which become instantaneous centres of rotation during the motion. Both polhodes always touch, at a point which is an instantaneous centre of rotation.*

Theorem 2. *The normal at a point of a trochoid passes through the instantaneous centre of rotation.*

In what follows we consider only the cases in which both polhodes are circles, or one of them is a circle and the other a straight line.

(a) The cycloids

Definition 2. By rolling a circle h along a straight line p without slipping, each point of the circle describes a *simple* (*general, normal*) *cycloid.* If the original position of the generating point coinciding with the point of contact of the circle h and the straight line p is the origin O and the straight line p is the x-axis, then

$$x = r(t - \sin t), \quad y = r(1 - \cos t) \tag{1}$$

are parametric equations of this simple cycloid; here, r is the radius of the generating circle h and t is the angle through which the rolling circle has turned at any instant.

Construction 1 of points of a simple cycloid (Fig. 4.20). We divide the circumference of the circle h and its rectified length on the tangent p at the point A into the same number of equal parts (there are 12 in Fig. 4.20). Consequently, $\widehat{A1} = \overline{AI}$, $\widehat{12} = \overline{III}$, ... Perpendiculars constructed through the points on the straight line p determine on the line through the point H parallel to the straight line p (i.e. on the path of the point H) the centres $H_1, H_2, \ldots$ of circles $h_1, h_2, \ldots$. Lines through the points of subdivision $1, 2, 3, \ldots$ of the circle h, parallel to the straight line p, meet the circles $h_1, h_2, \ldots$ at the points $A_1, A_2, \ldots$ of the cycloid.

Theorem 3. *The normal to a simple cycloid at a given point passes through the corresponding point of contact of the generating circle h on the given straight line p (i.e. through the instantaneous centre of rotation). The tangent to a simple cycloid at a given point passes through the point of the circle h which is diametrically opposite to the instantaneous centre of rotation.*

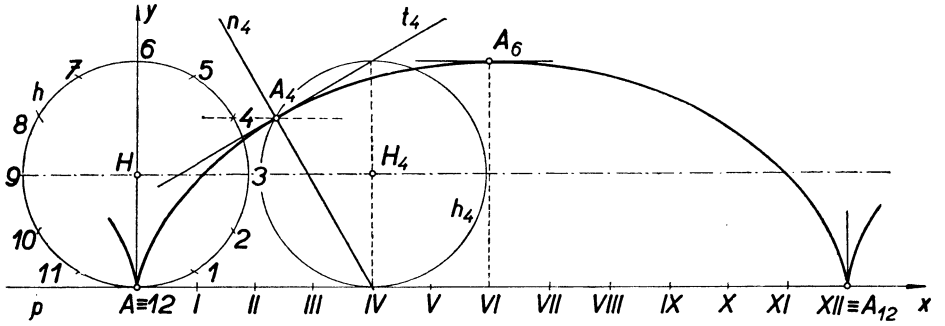


Fig. 4.20.

Theorem 4. *The length of a normal is*

$$n = \left|2r \sin \frac{t}{2}\right| = \sqrt{(2ry)} .$$

Theorem 5. *The radius of curvature at a point other than the cuspidal point of a simple cycloid is*

$$R = \left|4r \sin \frac{t}{2}\right| = 2\sqrt{(2ry)} = 2n ;$$

thus, at the vertex

$$R = 4r .$$

Theorem 6. *The length of arc (on a single branch) of a simple cycloid measured from the cuspidal point to the point $P(x, y)$ is*

$$s = 4r\left(1 - \cos \frac{t}{2}\right);$$

thus, the length of the entire branch is

$$s = 8r .$$

Theorem 7. *The area bounded by the x-axis and by a branch of a simple cycloid is*

$$P = 3\pi r^2 .$$

Definition 3. If a circle h rolls along a fixed straight line p, without slipping, an internal or external point moving with the circle h describes a *curtate*, or a *prolate, cycloid*, respectively.

Theorem 8. *If t is the angle through which the generating circle h has rolled and if $d \leqq r$ is the distance between the moving point P and the centre H of the circle h, the parametric equations of the curtate, or prolate, cycloid traced out by P are given by the equations*

$$x = rt - d \sin t\,, \quad y = r - d \cos t\,. \tag{2}$$

Construction 2 of points of a curtate, or a prolate, cycloid (Fig. 4.21): We attach to the generating circle h a concentric circle h' of radius $r' = \overline{HB} < r$ or h'' of radius $r'' = \overline{HC} > r$, respectively. Then, on the appropriate radius of an instantaneous position of the generating circle h_k with centre H_k $(k = 1, 2, \ldots)$ we determine the position of the circle h'_k, or h''_k, and, consequently, get the point B_k, or C_k of a curtate or prolate cycloid respectively.

Theorem 9. *The normal at a point of a curtate (prolate) cycloid passes through the point of contact of an instantaneous position of the generating circle h and the straight line p.*

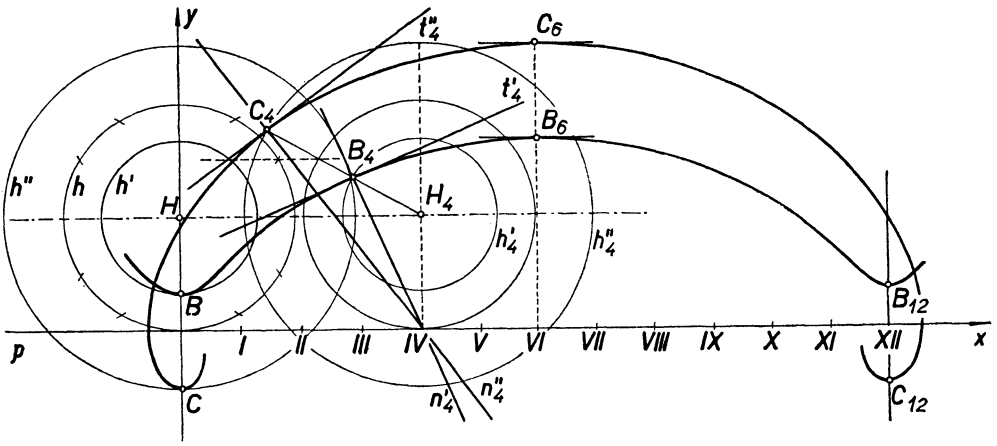


Fig. 4.21.

Theorem 10. *The radius of curvature at the points of a minimum or a maximum of a curtate (prolate) cycloid is*

$$R = \frac{(r - d)^2}{d}\,, \quad \text{or} \quad R = \frac{(r + d)^2}{d}\,, \quad \text{respectively}\,.$$

A simple cycloid has an infinite number of *cuspidal points*, a curtate cycloid has

an infinite number of *points of inflexion* and a prolate cycloid an infinite number of *double points* (so-called *nodes*).

(b) The epicycloids and hypocycloids

Definition 4. If a generating circle h of radius r rolls along the exterior, or the interior circumference of a fixed circle p of radius $\bar{r}$, then each point of the circle h describes a *simple* (*general, normal*) *epicycloid*, or *hypocycloid*, respectively.

Theorem 11. *The equations*

$$x = (\bar{r} \pm r)\cos t \mp r\cos\frac{\bar{r} \pm r}{r}t, \quad y = (\bar{r} \pm r)\sin t - r\sin\frac{\bar{r} \pm r}{r}t \qquad (3)$$

are parametric equations of a simple epicycloid (*the upper sign*) *or hypocycloid* (*the lower sign*).

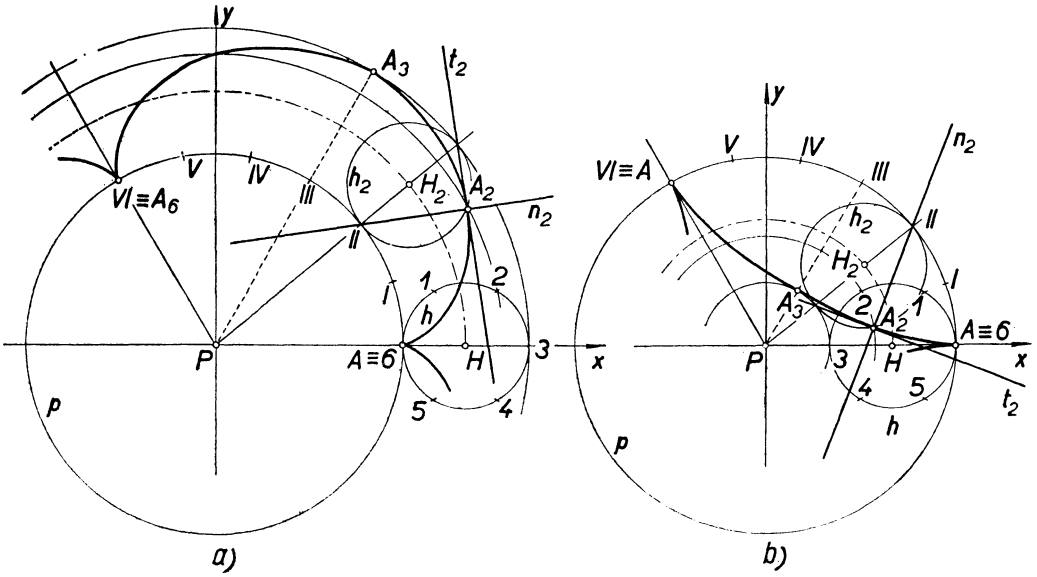


Fig. 4.22.

If $\lambda = \bar{r}/r$ is an integer, then λ denotes the number of branches of the curve formed by a single rotation of the circle h around the circle p. If $\lambda = p/q$ is a rational number, then the curve consists of p branches which are formed by q rotations of the circle h around the circle p. For an irrational λ, the curve contains an infinite number of branches.

Construction 3 of points of a simple epicycloid, or hypocycloid: We divide the circumference of the circle h into a certain number of equal parts (Fig. 4.22a,b), to give the points $A, 1, 2, \ldots$.

On the circle p we determine an arc whose length is equal to the circumference of the circle h, and divide it by points $A, I, II, \ldots$ into an equal number of parts of the same length as those on h. We describe concentric circles through the points $1, 2, \ldots$ about the centre P of the circle p and find the points $H_1, H_2, \ldots$ of intersection of the radii $PI, PII, \ldots$ with the circle described about P through the centre H of the circle h. Then the circle $h_1(H_1, r)$ meets the circle $k_1(P, \overline{PI})$ at a point A_1 of a simple epicycloid, or hypocycloid, respectively, etc.

Theorem 12. *The radius of curvature at a point (other than a cuspidal point) of a simple epicycloid, or hypocycloid, is*

$$R = \left| \frac{4r(\bar{r} \pm r)}{\bar{r} \pm 2r} \sin \frac{\bar{r}t}{2r} \right| ;$$

the length of arc (on the same branch) from the point $t = 0$ to a point t is

$$s = \frac{8r(\bar{r} \pm r)}{\bar{r}} \sin^2 \frac{\bar{r}t}{4r}$$

($r < \bar{r}$ is assumed for the hypocycloid). In particular: The radius of curvature at a vertex is

$$R = \left| \frac{4r(\bar{r} \pm r)}{\bar{r} \pm 2r} \right|$$

and the length of one branch is

$$s = \frac{8r(\bar{r} \pm r)}{\bar{r}} .$$

Here the positive sign holds for an epicycloid, the negative sign for a hypocycloid.

Definition 5. The curve described by an internal, or an external, point rotating with the generating circle h is called a *curtate*, or a *prolate, epicycloid (hypocycloid)*, respectively.

Theorem 13. *If $d \lessgtr r$ is the distance between the generating point and the centre H of the circle h, then*

$$x = (\bar{r} \pm r) \cos t \mp d \cos \frac{\bar{r} \pm r}{r} t , \quad y = (\bar{r} \pm r) \sin t - d \sin \frac{\bar{r} \pm r}{r} t$$

are parametric equations of the curves of Definition 5 (the upper signs refer to an epicycloid, the lower ones to a hypocycloid).

For the construction of points of a curtate or a prolate epicycloid (hypocycloid) we employ again the concentric circle h', or h'', attached to the circle h, as in Construction 2 above.

Example 1. If $\bar{r} = r$, $d > r$, we get a prolate epicycloid (the *limaçon of Pascal*) of parametric equations (Fig. 4.23)

$$x = 2r \cos t - d \cos 2t ,$$

$$y = 2r \sin t - d \sin 2t ,$$

whose equation in rectangular cartesian coordinates (by elimination of the parameter t and translation of the origin to the point $(d, 0)$) is

$$(x^2 + y^2 + 2dx)^2 = 4r^2(x^2 + y^2) .$$

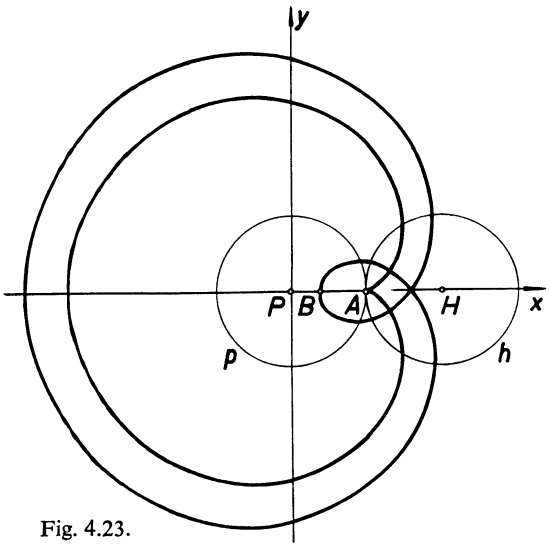


Fig. 4.23.

Example 2. For $r = \frac{1}{2}\bar{r}$, the equations of a simple hypocycloid are

$$x = \bar{r} \cos t , \quad y = 0 ;$$

hence, it is a segment of length $2\bar{r}$ on the x-axis.

The equations of a curtate hypocycloid are

$$x = \left(\frac{\bar{r}}{2} + d\right) \cos t , \quad y = \left(\frac{\bar{r}}{2} - d\right) \sin t ;$$

hence it is an ellipse with a semi-major axis of length $r + d$ on the x-axis and a semi-minor axis of length $r - d$ on the y-axis.

Example 3. For $r = \bar{r}$ (Fig. 4.23) the parametric equations of a (simple) epicycloid (the *cardioid*) are

$$x = r(2 \cos t - \cos 2t) , \quad y = r(2 \sin t - \sin 2t) .$$

If the origin of cartesian coordinates is at the centre of the fixed curve and the cuspidal point lies on the x-axis, then we get the equation of the curve in the form

$$(x^2 + y^2)^2 - 6r^2(x^2 + y^2) + 8r^3x - 3r^4 = 0;$$

if the origin is at the double point (and the x-axis is the axis of symmetry), then we get the equation

$$(x^2 + y^2 + 2rx)^2 - 4r^2(x^2 + y^2) = 0.$$

The equation of the cardioid in polar coordinates is

$$\rho = 2r(1 - \cos\varphi).$$

A cardioid can also be obtained as an orthogonal pedal curve (see Definition 9.10.1, p. 303) of a circle for a pole on the circle.

Example 4. For $r = \frac{1}{2}\bar{r}$ (Fig. 4.24), the parametric equations of a simple epicycloid (the *nephroid*) are

$$x = r(3\cos t - \cos 3t), \quad y = r(3\sin t - \sin 3t);$$

the equation of the curve in cartesian coordinates is

$$(x^2 + y^2 - 4r^2)^3 - 108r^4y^2 = 0.$$

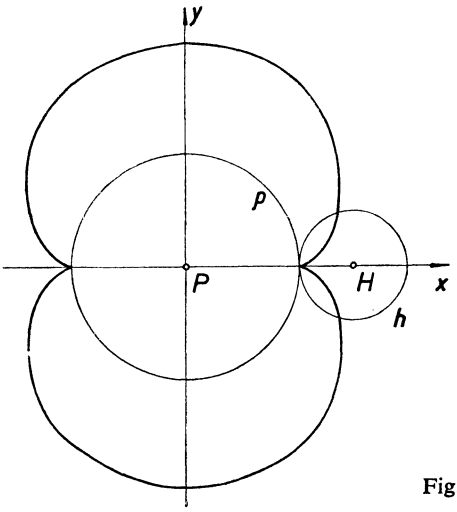


Fig. 4.24.

Example 5. For $r = \frac{1}{3}\bar{r}$ (Fig. 4.25), the parametric equations of a (simple) hypocycloid (*Steiner's hypocycloid*) are

$$x = r(2\cos t + \cos 2t),$$
$$y = r(2\sin t - \sin 2t);$$

the equation of the curve in cartesian coordinates is

$$(x^2 + y^2)^2 + 8rx(3y^2 - x^2) + 18r^2(x^2 + y^2) - 27r^4 = 0 .$$

Example 6. For $r = \frac{1}{4}\bar{r}$, the parametric equations of a (simple) hypocycloid (the *astroid*, Fig. 4.26) are

$$x = r(3\cos t + \cos 3t), \quad y = r(3\sin t - \sin 3t);$$

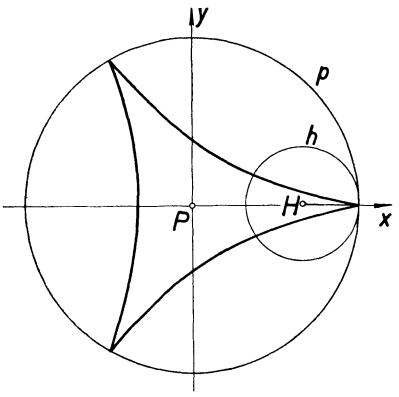


Fig. 4.25.

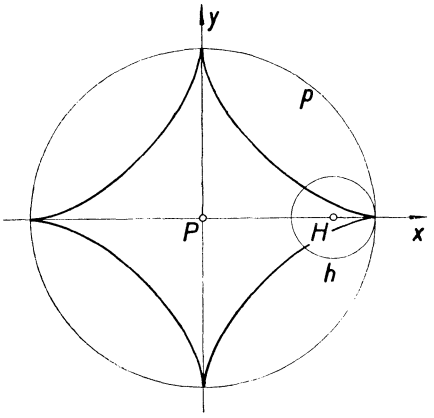


Fig. 4.26.

the equation of the curve in cartesian coordinates is

$$x^{2/3} + y^{2/3} = (4r)^{2/3} .$$

For the curves given in the above examples we can use Theorem 12 to determine the radius of curvature at any point or the length of an arc, in particular the radius of curvature at a vertex or the length of a branch.

(c) The involute of a circle

Definition 6. Any point of a plane rotating with a straight line h, rolling on a fixed circle p, describes an *involute of a circle.*

Theorem 14. *For a fixed circle* $p(0, r)$ *and a generating point* $A(r + d, 0)$ *the parametric equations of the involute are*

$$x = (r + d)\cos t + rt\sin t, \quad y = (r + d)\sin t - rt\cos t,$$

where t *is the angle between the x-axis and the radius of the circle* p *perpendicular to the position of the straight line* h.

Definition 7. For $d = 0$, the (simple, general, normal) *circular involute* is generated, for $d > 0$ (the generating point and the circle p are on opposite sides of the straight

line h) a *curtate involute* is generated, for $d < 0$ (the generating point and the circle p lie on the same side of the straight line h) we get a *prolate involute.*

Construction 4 of points of a (simple) circular involute (Fig. 4.27): We divide the circumference of the given circle p into a certain number of equal parts (for example, into 12) by the points $A, 1, 2, \ldots$; we rectify the arc corresponding to one part and then we determine on the tangent to the circle p at every point of subdivision the

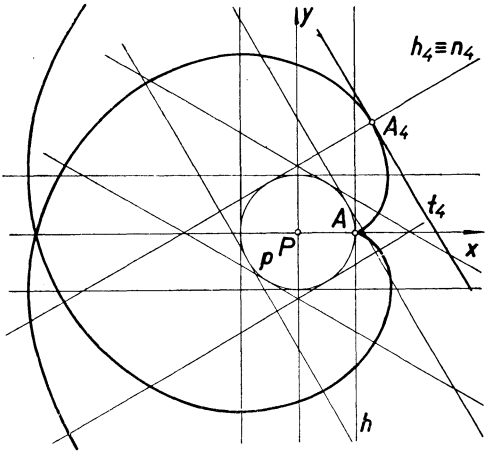


Fig. 4.27.

point at a distance equal to the length of the corresponding number of arcs: at the point *1* at a distance of one arc, at the point *2* of two arcs, etc.

Theorem 15. *The normals to a circular involute are tangents to the fixed circle p which is therefore an involute of the given curve* (see Definition 9.8.3, p. 297) (*and consequently, the locus of centres of curvature*).

Theorem 16. *For the radius of curvature of a circular involute we have*

$$R = rt\,,$$

and for its length of arc

$$s = \tfrac{1}{2}rt^2\,.$$

Example 7. For $d = -r$, a prolate involute, the *spiral of Archimedes* (see § 4.7, Fig. 4.29) is generated; its parametric equations are

$$x = rt \sin t\,, \quad y = -rt \cos t\,;$$

the equation in polar coordinates is

$$\varrho = r\varphi\,.$$

(d) Construction of centres of curvature of cyclic curves

Construction 5 at a point M which is not a vertex of the curve (Fig. 4.28): The centre of curvature S_M lies on the normal $n \equiv SM$. We construct a perpendicular k through the point S to the normal n and find the point of intersection *1* of the lines k, MH. Then S_M is the intersection of n and $1P$.

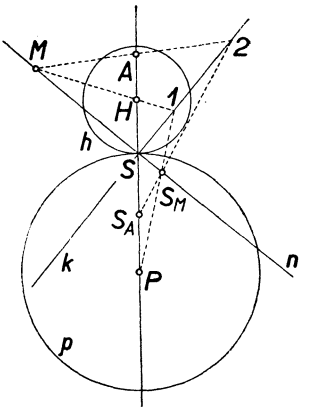


Fig. 4.28.

Construction 6 of the centre of curvature at a vertex A of the curve (i.e. at a point lying on PH) (Fig. 4.28): Using Construction 5, we construct the centre of curvature S_M for an arbitrary point M of the curve. We determine the point of intersection *2* of k and MA, and then S_A is the intersection of $2S_M$ and HP.

4.7. Spirals

Definition 1. A curve generated by a point moving uniformly along a polar radius rotating uniformly around its pole is called a *spiral of Archimedes* (Fig. 4.29).

Construction 1 of points of a spiral of Archimedes (Fig. 4.29): After one revolution, the distance of the moving point M from the origin O is equal to r_0. We divide the angle 2π and the segment r_0 (in the figure, $r_0 = \overline{OM}_{12}$) into n (say, 12) equal parts. Starting at the origin O, we successively mark off segments of lengths r_0/n, $2r_0/n$, ... on the corresponding polar radii. The end points of the segments are points of a spiral of Archimedes.

Theorem 1. *The equation of a spiral of Archimedes in polar coordinates is*

$$\varrho = \frac{r_0}{2\pi}\varphi = a\varphi$$

(*where* r_0, *and hence* $a = r_0/2\pi$, *is a given constant*).

The acute angle between a tangent to the curve and the polar radius at the point of contact increases with the polar radius and converges to the value $\frac{1}{2}\pi$.

Theorem 2. *The length of a polar sub-normal* s_n *is constant and equal to* a. (*The polar sub-normal is the segment between the pole and the point* N *of intersection of the normal* n *at the point* M *under consideration with the perpendicular constructed at the pole* O *to the polar radius.*)

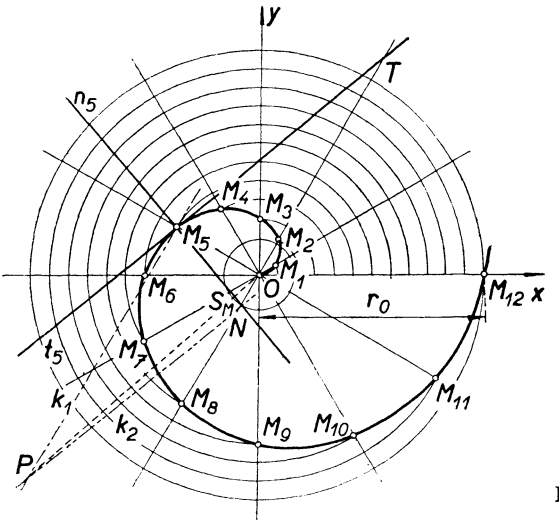


Fig. 4.29.

At a given point of a spiral of Archimedes we construct its normal and tangent by using Theorem 2.

The segment $\overline{OT} = s_t$, where T is the point of intersection of the tangent t at a point M of the curve and a perpendicular constructed at the pole to the polar radius of the point, is called a *polar sub-tangent.* Thus, for a spiral of Archimedes the equation

$$s_t = \frac{\varrho^2}{s_n} = \frac{\varrho^2}{a} = a\varphi^2$$

holds.

Construction 2 of the centre of curvature at a point of a spiral of Archimedes (Fig. 4.29): The perpendicular k_1 constructed at the point M to its polar radius, meets the perpendicular k_2 constructed at the point N to the normal n, at a point P. The centre of curvature S_M is the intersection of n and PO.

Definition 2. The arc of a spiral of Archimedes for which $2(n-1)\pi \leqq \varphi < 2n\pi$, is called the *n-th coil of the curve.*

Theorem 3. *The individual coils of a spiral of Archimedes are equidistant curves.*

Since the motion of a generating point on the polar radius can be accomplished in two directions, a spiral of Archimedes has two branches symmetrical about the x-axis.

Definition 3. A curve for which (in polar coordinates) the product of the length of the polar radius and the argument is constant, is called a *hyperbolic spiral or reciprocal spiral* (Fig. 4.30).

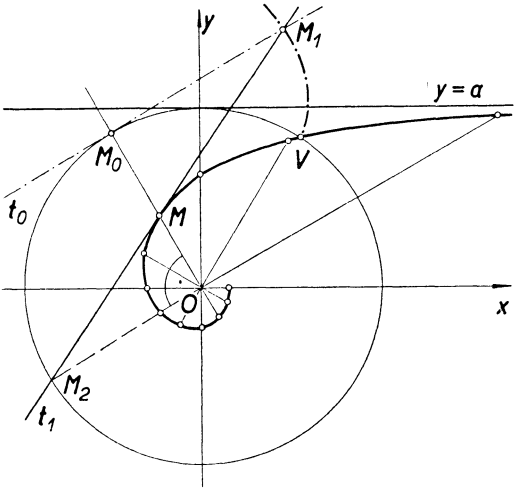


Fig. 4.30.

Theorem 4. *The equation of a hyperbolic spiral in polar coordinates is*

$$\varrho = \frac{a}{\varphi} \quad (\textit{where } a \textit{ is constant}).$$

If $a < 0$, then also $\varphi < 0$ and we get the second branch (not illustrated in the figure) which is symmetrical to the first one about the x-axis.

Theorem 5. *The straight line $y = a$ is an asymptote of the hyperbolic spiral; the pole O is an asymptotic point.*

Theorem 6. *For a hyperbolic spiral, the length of the polar sub-tangent s_t is constant and equal to a.*

Construction 3 of points and of tangents to a hyperbolic spiral (Fig. 4.30): By Theorem 6, the end points of polar sub-tangents lie on the circle $k(O, a)$. The point V of the circle k, where the polar radius makes with the axis an angle $\varphi = 1$ (in circular measure) is also a point of the hyperbolic spiral. On an (arbitrary) radius OM_0 we find a point M of the hyperbolic spiral and the tangent in the following way: We connect the point of intersection M_1 of the tangent t_0 to the circle k at the point M_0 and the evolute e of the point V of the circle k, with the point M_2 of the polar sub-tangent and thus get a tangent t to the hyperbolic spiral, which meets the polar radius OM_0 at the point of contact M.

Definition 4. A curve making a constant angle with the polar radii at its points is called a *logarithmic* (*equiangular, logistic*) *spiral* (Fig. 4.31). It can also be characterized as the curve whose length of arc between a fixed and a variable point is proportional to the polar radius of the latter point.

Theorem 7. *The equation of a logarithmic spiral in polar coordinates is*

$$\varrho = ae^{b\varphi},$$

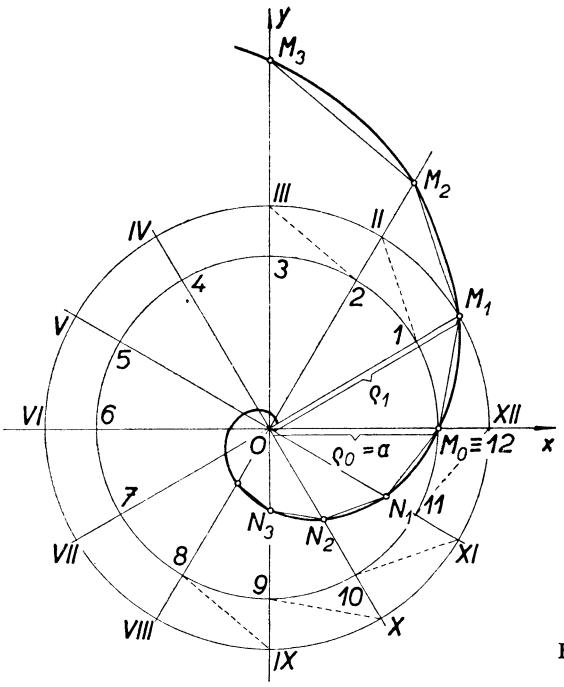


Fig. 4.31.

where $a, b > 0$ *are constant,* φ *is the angle* (*in radians*) *between the polar radius and the polar axis and* e *is the base of natural logarithms.*

Theorem 8. *If the angles* φ *form an arithmetic progression, then the corresponding polar radii* ϱ *form a geometric progression.*

Theorem 9. *The pole* O *is an asymptotic point of a logarithmic spiral. For* $\varphi = 0$, $\varrho_0 = a$.

Construction 4 of points of a logarithmic spiral (Fig. 4.31): We divide the angle 2π into n (say, 12) equal angles (Fig. 4.31) and calculate two adjacent polar radii $\varrho_0 = a$, $\varrho_1 = ae^{b(\pi/6)}$. The triangles OM_0M_1, OM_1M_2, ... are similar. We describe circles k_0, k_1 with radii ϱ_0, ϱ_1 about the pole O and mark on them the points M_0, *1*, *2*, ..., and *XII*, M_1, *II*, ..., determined by the polar radii. Then, the line through the point M_1 parallel to the line joining 1, *II* meets the polar radius ϱ_2 at the point M_2 of the

logarithmic spiral, the line through the point M_2 parallel to the line joining *2, III* meets the polar radius ϱ_3 at the point M_3, etc. Similarly, the point N_1 of the polar radius corresponding to the angle $\varphi = -\frac{1}{6}\pi$ can be obtained as the point of intersection of the polar radius and the straight line through the point M_0 drawn parallel to the line joining *XII, 11*, etc.

Theorem 10. *The tangent at a point of a logarithmic spiral makes with its polar radius an angle* ϑ *satisfying* $\tan \vartheta = 1/b$. *For a polar sub-tangent, or sub-normal, the relations*

$$s_t = \frac{\varrho}{b}, \quad s_n = b\varrho$$

hold, respectively.

Construction 5 of the tangent at a point of a logarithmic spiral (Fig. 4.32): On the polar radius OM we determine the point Q such that $\overline{OQ} = 1$. On the perpendicular through the point O to the polar radius OM we determine the point R such that $\overline{OR} = 1/b$ (the sense of OR being such that a rotation from OR to OM is positive). The angle OQR is equal to ϑ and hence the line parallel to QR through the point M is the required tangent t at the point M of the logarithmic spiral. Other tangents at points of a logarithmic spiral can be constructed by translation of the constant angle ϑ so obtained (see Definition 4 and Theorem 10).

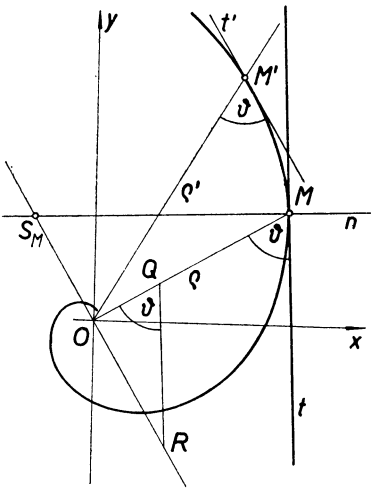


Fig. 4.32.

Theorem 11. *The radius of curvature at a point of a logarithmic spiral is*

$$R = \varrho \sqrt{(1 + b^2)}$$

and it is equal to the length of the polar normal. The centre of curvature lies at the point of intersection of the normal and the perpendicular through the point O to the polar radius of the point.

Definition 5. Curves satisfying the polar equation

$$\varrho^m = a^m \sin m\varphi$$

are called *sinusoidal spirals.*

Theorem 12. *By a rotation of the coordinate system through an angle* $\varphi = (\frac{1}{2}\pi/m) - \varphi$ *the equation of a sinusoidal spiral becomes*

$$\varrho^m = a^m \cos m\psi .$$

Theorem 13. *For rational m, sinusoidal spirals are algebraic curves; for irrational m, they are transcendental curves.*

Example 1. For special values of m, we get the following sinusoidal spirals:

(a) $m = 1$; $\varrho = a \cos \varphi$, the sinusoidal spiral is a circle given by the equation $x^2 + y^2 = ax$;

(b) $m = 2$; $\varrho^2 = a^2 \cos 2\varphi$, the sinusoidal spiral is a *lemniscate of Bernoulli* satisfying the equation $(x^2 + y^2)^2 = a^2(x^2 - y^2)$ (see § 4.11);

(c) $m = -1$; $\varrho = a/\cos \varphi$, the sinusoidal spiral is a straight line given by the equation $x = a$;

(d) $m = -2$; $\varrho^2 = a^2/\cos 2\varphi$, the sinusoidal spiral is a rectangular hyperbola satisfying the equation $x^2 - y^2 = a^2$;

(e) $m = \frac{1}{2}$; $\varrho = a \cos^2 \frac{1}{2}\varphi$, the sinusoidal spiral is a cardioid given by the equation $\varrho = 2r(1 + \cos \varphi)$, which can be obtained by use of the relation $2 \cos^2 \frac{1}{2}\varphi = 1 + \cos \varphi$ on putting $a = 4r$.

(f) $m = -\frac{1}{2}$; $\varrho = a/\cos^2 \frac{1}{2}\varphi$, the sinusoidal spiral is a parabola satisfying the equation $y^2 = 4a(a - x)$.

4.8. The Clothoid (Cornu Spiral)

Definition 1. A curve whose radius of curvature R at a point M is inversely proportional to the length s of the arc between this point and a fixed point O is called the *clothoid or Cornu Spiral* (Fig. 4.33).

Theorem 1. *The intrinsic equation of a clothoid* (see Definition 9.4.3 and Remark 9.4.10, pp. 280, 281) *is*

$$R = \frac{a^2}{s} .$$

Theorem 2. *Parametric equations of a clothoid with the arc s as a parameter are given by the Fresnel integrals* (§ 13.12)

$$x = \int_0^s \cos \frac{s^2}{2a^2}\, ds , \quad y = \int_0^s \sin \frac{s^2}{2a^2}\, ds .$$

If the angle $\varphi = \frac{1}{2}s^2/a^2$ *of the tangent at the point under consideration is taken as a parameter, then the equations are of the form*

$$x = \frac{a}{\sqrt{2}}\int_0^{\varphi} \frac{\cos\varphi}{\sqrt{\varphi}}\,\mathrm{d}\varphi\,, \quad y = \frac{a}{\sqrt{2}}\int_0^{\varphi} \frac{\sin\varphi}{\sqrt{\varphi}}\,\mathrm{d}\varphi\,.$$

If $\varphi = \frac{1}{2}\pi t^2$, *then the parametric equations have the form*

$$x = a\sqrt{\pi}\int_0^{t} \cos\frac{\pi t^2}{2}\,\mathrm{d}t\,, \quad y = a\sqrt{\pi}\int_0^{t} \sin\frac{\pi t^2}{2}\,\mathrm{d}t\,.$$

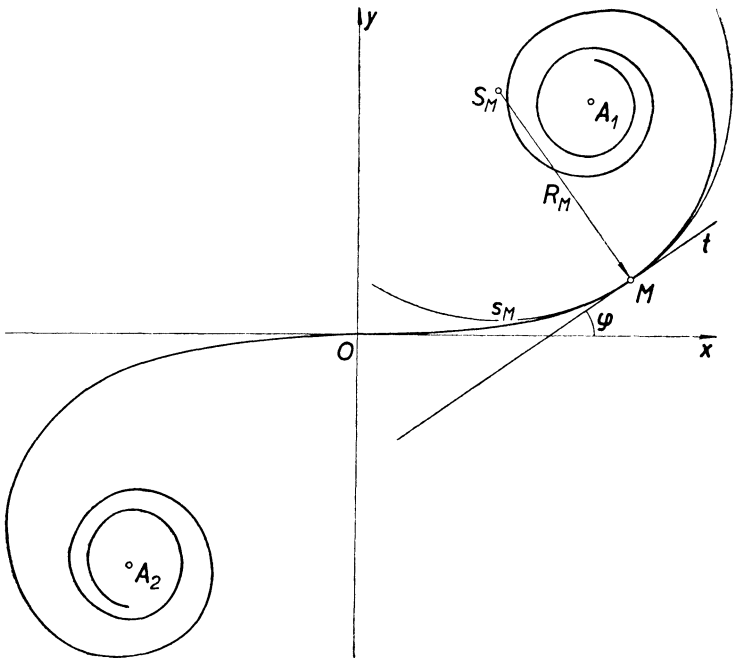


Fig. 4.33.

Theorem 3. *A clothoid is symmetrical about the point O which is a point of inflexion and it touches the x-axis at this point.*

Theorem 4. *The points* $(\frac{1}{2}a\sqrt{\pi}, \frac{1}{2}a\sqrt{\pi})$, $(-\frac{1}{2}a\sqrt{\pi}, -\frac{1}{2}a\sqrt{\pi})$ *are asymptotic points for a clothoid.*

Theorem 5. *The tangents of a clothoid at points for which*

$$\frac{s^2}{2a^2} = k\pi \quad (k = 0, 1, 2, \ldots)$$

are parallel to the x-axis, and tangents at the points for which

$$\frac{s^2}{2a^2} = \frac{2k+1}{2}\pi \quad (k = 0, 1, 2, \ldots)$$

are parallel to the y-axis.

Theorem 6. *For the angle φ made by a tangent to the curve and the x-axis, Theorem 2 and Definition 1 yield*

$$\varphi = \frac{s}{2R}.$$

Theorem 7. *The following relations hold between the quantities a, s, R, φ (in circular measure)*:

$$\text{(a)} \quad a = \sqrt{(sR)} = \frac{s}{\sqrt{(2\varphi)}} = R\sqrt{(2\varphi)};$$

$$\text{(b)} \quad s = \frac{a^2}{R} = 2\varphi R = a\sqrt{(2\varphi)};$$

$$\text{(c)} \quad R = \frac{a^2}{s} = \frac{s}{2\varphi} = \frac{a}{\sqrt{(2\varphi)}};$$

$$\text{(d)} \quad \varphi = \frac{s}{2R} = \frac{s^2}{2a^2} = \frac{a^2}{2R^2}.$$

For practical use, the Fresnel integrals are tabulated. The constant a is the *parameter* determining the relative magnitude of the curve. If, for example, $a = 200$, then all the longitudinal values of the corresponding clothoid are double the values for the parameter $a = 100$.

4.9. The Exponential Curve

Definition 1. The curve whose equation in cartesian coordinates is

$$y = ab^{cx} \left(\text{or} \quad x = \frac{1}{c}\log_b \frac{y}{a} \right),$$

where $a > 0$, $b > 0$, c are constants, is called the *exponential curve.*

For $b = e$, we get

$$y = ae^{cx} \left(\text{or} \quad x = \frac{1}{c}\ln \frac{y}{a} \right).$$

Theorem 1. *The curves* $y = ab^{cx}$, $y = ab^{-cx}$ *are symmetrical about the y-axis. Both curves have only positive ordinates y and pass through the point* $A(0, a)$. *The x-axis is their common asymptote.*

Theorem 2. *Three points* $P(x, y)$, $P_1(x_1, y_1)$, $P_2(x_2, y_2)$ *of the curve satisfy the following relation*:

$$\left(\frac{y}{y_1}\right)^{x_1 - x_2} = \left(\frac{y_1}{y_2}\right)^{x - x_1}.$$

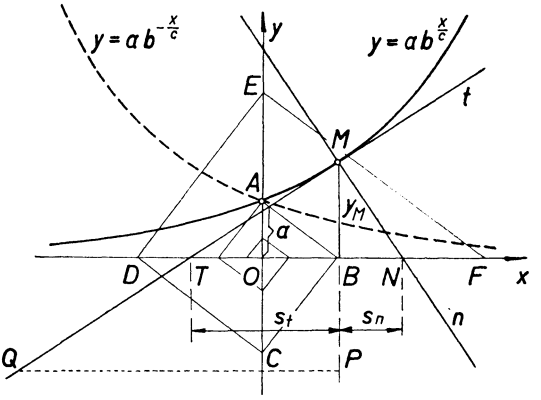


Fig. 4.34.

Construction 1 of points of the curve (Fig. 4.34): The ordinates y form a geometric progression if the abscissae x form an arithmetic progression. Hence we construct the points $A(0, a)$ and $B(ab^c, 0)$, and then draw successive perpendiculars $BC \perp AB$, $CD \perp BC, \ldots$; the segments $\overline{OA}$, $\overline{OB}$, $\overline{OC}, \ldots$ are now the ordinates of the points whose abscissae are $x = 0, 1, 2, \ldots$. By a reverse procedure we obtain the points of the curve for $x = -1, -2, \ldots$.

Theorem 3. *The sub-tangent of an exponential curve* $y = ae^{cx}$ *(with respect to the x-axis) has the constant value*

$$s_t = -\frac{1}{c}.$$

For the sub-normal we have $s_n = cy^2$. *The length of the tangent is* $t = \sqrt{(y^2 + 1/c^2)}$, *and of the normal is* $n = y\sqrt{(c^2y^2 + 1)}$.

For a curve $y = ab^{cx}$, *the sub-tangent is* $s_t = -1/(c \ln b)$.

Theorem 4. *The radius of curvature R of an exponential curve is given by the expression*

$$R = \frac{\sqrt{(y^2 + (1/c^2))^3}}{y} \quad c = \frac{n^3}{c^2y^4} = \frac{n^3}{s_n^2}.$$

For the point for which $y = \sqrt{(2)}/2c$, *the radius R is minimal and equal to* $3\sqrt{(3)}/2c$.

Construction 2 of the radius of curvature of the curve $y = ab^{cx}$ at a point M (Fig. 4.34): Since $y/t = (s_n + s_t)/R$, we construct on the perpendicular through the point M to the x-axis, the point P such that $\overline{MP} = (s_n + s_t)$. The tangent t and the perpendicular to the y-axis at the point P meet at the point Q; then $R = \overline{MQ} = \overline{MS}_M$.

4.10. The Catenaries (Chainettes)

(a) The general catenary

Definition 1. A curve satisfying the equation

$$y = \tfrac{1}{2}a(e^{x/a} + e^{-x/a}), \quad \text{i.e.} \quad y = a \cosh \frac{x}{a}$$

is called a *general catenary* (Fig. 4.35).

A heavy homogeneous perfectly flexible cable suspended by two points assumes the form of a general catenary.

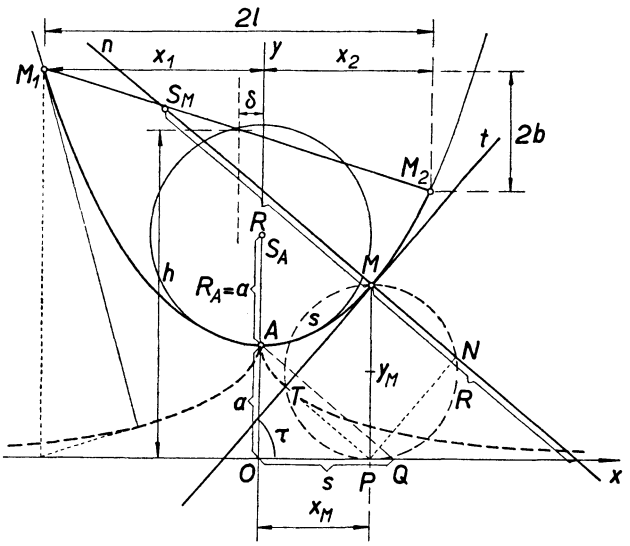


Fig. 4.35.

Theorem 1. *A general catenary is symmetrical about the y-axis, on which it has its vertex A at a distance a from the origin O; the value a is called the parameter of the catenary.*

Theorem 2. *A general catenary and the parabola* $y = a + x^2/2a$ *have three-point contact at the vertex* $A(0, a)$. *Also an ellipse with centre* $S(0, 4a)$, *major axis of length* $6a$ *on the y-axis and semi-minor axis* $a\sqrt{3}$ *has three-point contact with the general catenary at the common vertex A.*

Theorem 3. *By a translation of the origin of the coordinate system to the point of suspension M_1 (with the abscissa $-m$ in the original system) the equation of a general catenary becomes*

$$y = a\left(\cosh\frac{x-m}{a} - \cosh\frac{m}{a}\right).$$

Theorem 4. *The angle τ between a tangent and the x-axis satisfies (in the original system)*

$$\tan\tau = \sinh\frac{x}{a} = \frac{1}{a}\sqrt{(y^2 - a^2)}, \quad \cos\tau = \frac{a}{y}.$$

Construction 1 of the tangent and the normal at a point M of a general catenary (Fig. 4.35): The circle on the ordinate MP of the point M and the circular arc about the point M, or P, of radius $r = a$ intersect at the points N, or T, respectively. The point T is a point of the tangent, the point N is a point of the normal at the given point M.

Theorem 5. *The arc s of a general catenary (measured from its vertex A) is*

$$s = a\sinh\frac{x}{a} = \sqrt{(y^2 - a^2)} = a\tan\tau\,;$$

thus it is proportional to the tangent of the angle made by the tangent at the end point of the arc with the x-axis.

According to Construction 1, the arc of Theorem 5 is equal to the segment $\overline{MT} = \overline{OQ}$ (where Q is the point of the x-axis for which $AQ = y_M$); thus the arc is equal to the (rectangular) projection of the ordinate y_M of the point M onto the tangent.

Theorem 6. *The radius of curvature R and the length n of the corresponding normal are equal:*

$$R = n = a\cosh^2\frac{x}{a} = \frac{y^2}{a} = \frac{a}{\cos^2\tau}.$$

For the vertex, $R = a$.

Theorem 7. *The area enclosed by the x-axis, the y-axis, the arc of a general catenary and the ordinate of a given point is given by*

$$P = a^2\sinh\frac{x}{a} = as\,.$$

Example 1. The determination of the parameter a and the position of the axes, given the length of the cable $2s$, the horizontal distance between the points of suspension $2l$ and the difference of the heights $2b$ (Fig. 4.35).

To solve the problem we use the relations

$$\frac{\sqrt{(s^2 - b^2)}}{l} = \frac{\sinh u}{u}; \quad a = \frac{l}{u}.$$

Putting $c = (1/l)\sqrt{(s^2 - b^2)}$, we determine u from the equation $\sinh u = cu$ (for example, by using the tables of the function $\Theta(u) = \sinh u/u$) and hence the parameter a.

Then, the distance of the x-axis from the centre of the segment joining the points of suspension is

$$h = s \coth u .$$

The displacement of the y-axis in the direction of the lower point of suspension is $\delta = av$, where $\tanh v = b/s$.

The angles between the tangents at the points of suspension M_1, M_2 and the x-axis are $\tan \alpha_i = \sinh x_i/a$ $(i = 1, 2)$, where x_1, x_2 are the abscissae of the given points of suspension.

Theorem 8. *The involute of a catenary, called the tractrix, has the equation*

$$x = a \ln \frac{a - \sqrt{(a^2 - y^2)}}{y} + \sqrt{(a^2 - y^2)} .$$

Its tangent is of a constant length a.

The points of a general catenary can be constructed using tables of the hyperbolic cosine.

(b) The catenary of constant strength

Definition 2. The curve satisfying the equation

$$e^{y/a} \cos \frac{x}{a} = 1 , \quad \text{i.e.} \quad y = -a \ln \cos \frac{x}{a} ,$$

where $a > 0$ and where x satisfies the inequalities

$$a(4k - 1)\frac{\pi}{2} < x < a(4k + 1)\frac{\pi}{2} \quad (k \text{ an integer}) ,$$

is called a *catenary of constant strength.*

A heavy perfectly flexible and inelastic cable whose cross-section varies in such a manner that its resistance to breakage is constant, assumes, after being suspended, the form of a catenary of constant strength (Fig. 4.36).

Theorem 9. *A catenary of constant strength consists of an infinite number of congruent branches, touching the x-axis at the points* $x = 2k\pi a$ *and having the straight lines* $x = a(4k \pm 1)\,\pi/2$ (k *an integer*) *as asymptotes.*

Theorem 10. *The angle* τ *between a tangent at a point of a catenary of constant strength and the x-axis is proportional to the abscissa of the point of contact:*

$$\tau = \frac{x}{a}.$$

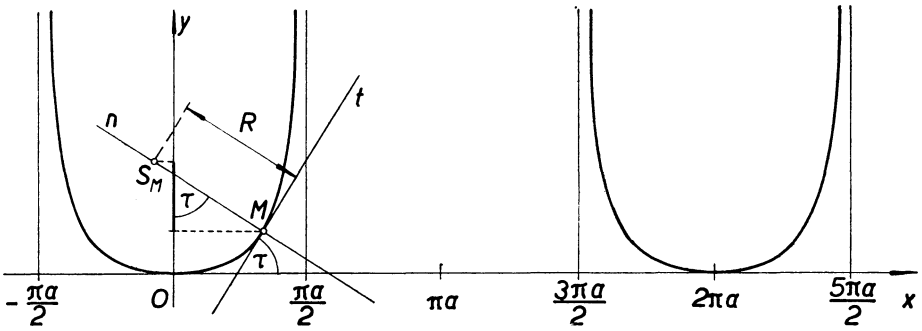


Fig. 4.36.

Theorem 11. *The radius of curvature at a point of the curve under consideration is*

$$R = \frac{a}{\cos\dfrac{x}{a}} = \frac{a}{\cos\tau};$$

thus, since $R\cos\tau = a$, *the rectangular projection of the radius of curvature on the y-axis is constant.*

Theorem 12. *For an arc s of a catenary of constant strength the relation*

$$s = a\ln\tan\left(\frac{x}{2a} + \frac{\pi}{4}\right).$$

holds.

Theorem 13. *The area enclosed by a branch of a catenary of constant strength, by both asymptotes and by the x-axis is*

$$P = \pi a^2 \ln 2\,.$$

4.11. Examples of Some Algebraic Curves

Example 1 (The *cissoid of Diocles*; Fig. 4.37). We construct the tangent $t \parallel y$ to the circle k of diameter a, with the centre on the x-axis and passing through the origin O. We draw lines through O intersecting k at the points $1, 2, \ldots$, and t at the points $1', 2', \ldots$. On every such line, we mark the point whose distance from the origin is

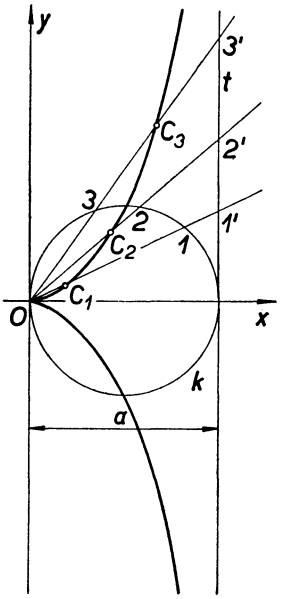


Fig. 4.37.

equal to the length of the segment determined by the points of intersection of the line with the circle and the tangent, i.e. $OC_1 = 11'$, $OC_2 = 22', \ldots$. Then $C_1, C_2, \ldots$ are the points of a cissoid of Diocles.

The equation of the curve in polar coordinates is

$$\varrho = a \sin \varphi \tan \varphi = a \frac{\sin^2 \varphi}{\cos \varphi},$$

and in cartesian coordinates

$$x(x^2 + y^2) - ay^2 = 0 \quad \left(\text{or} \quad y^2 = \frac{x^3}{a - x}\right).$$

The parametric equations are

$$x = \frac{at^2}{1 + t^2}, \quad y = \frac{at^3}{1 + t^2}.$$

A cissoid of Diocles is an algebraic curve of the third degree. The tangent t (of equation $x = a$) to the circle k is an asymptote of the curve.

A cissoid of Diocles is the orthogonal pedal curve of a parabola for a pole at the vertex of the parabola.

Example 2 (The *folium of Descartes*, Fig. 4.38). A cissoid of the ellipse $x^2 - xy + y^2 + a(x + y) = 0$, $a > 0$, with regard to the straight line $x + y + a = 0$ for

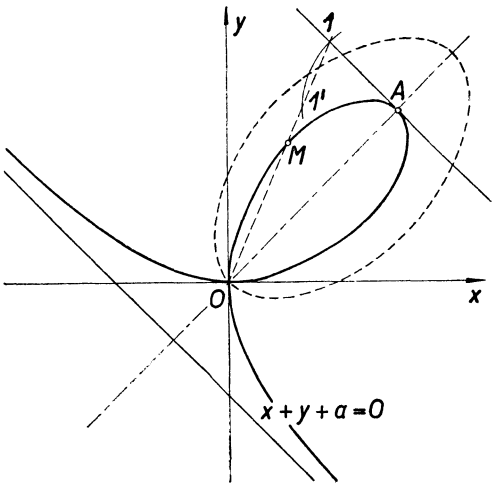


Fig. 4.38.

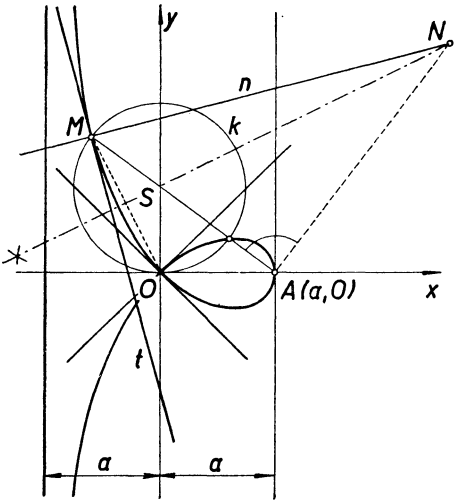


Fig. 4.39.

the pole O is called the *folium of Descartes*. The equation of this curve in cartesian coordinates is

$$x^3 + y^3 - 3axy = 0\,,$$

and in polar coordinates

$$\varrho = \frac{3a \sin\varphi \cos\varphi}{\sin^3\varphi + \cos^3\varphi}\,;$$

the parametric equations are

$$x = \frac{3at}{1 + t^3}\,, \quad y = \frac{3at^2}{1 + t^3}\,.$$

A folium of Descartes is a curve of the third degree symmetrical about the straight line $y = x$; at the point O, it has a node with the x-axis and y-axis as tangents; at the point $A(\frac{3}{2}a, \frac{3}{2}a)$, it has a vertex; the straight line $x + y + a = 0$ is its asymptote.

Construction 1 of points of a folium of Descartes: We draw a line through the pole O to meet the tangent constructed at the vertex A at the point 1. On this line we de-

termine the point I' such that $AI = AI'$; we then construct the harmonic point M to the point I with regard to O and I' (i.e. the cross-ratio $(O, I', I, M) = -1$). The point M is a point of the folium of Descartes.

Example 3 (The *strophoid*; Fig. 4.39). We intersect the pencil of circles having the x-axis as a common tangent, with the common point of contact at the origin, by the diameters drawn through the point $A(a, 0)$. The end points of the diameters lie on a (straight) strophoid, whose equation in polar coordinates is

$$\varrho = a \frac{\cos 2\varphi}{\cos \varphi},$$

and in cartesian coordinates is

$$x(x^2 + y^2) - a(x^2 - y^2) = 0 \quad \left(\text{or} \quad y^2 = x^2 \frac{a - x}{a + x}\right);$$

the parametric equations are

$$x = \frac{a(1 - t^2)}{1 + t^2}, \quad y = \frac{at(1 - t^2)}{1 + t^2}.$$

The curve is symmetrical about the x-axis, it has a node with the tangents $y = \pm x$ at the origin O and the straight line $x + a = 0$ is its asymptote.

Construction 2 of the normal and tangent at a point M of a strophoid: The perpendicular bisector of the segment $\overline{OM}$ intersects the line through the point A perpendicular to AM at the point N of the normal n; having obtained the normal, the tangent at M can be determined.

Example 4 (The *lemniscate of Bernoulli*; Fig. 4.40). This curve is a rectangular pedal curve of the rectangular hyperbola $x^2 - y^2 = a^2$. Its equation is

$$(x^2 + y^2)^2 = a^2(x^2 - y^2),$$

or in the polar form

$$\varrho^2 = a^2 \cos 2\varphi.$$

Its parametric equations are

$$x = \frac{at(1 + t^2)}{1 + t^4}, \quad y = \frac{at(1 - t^2)}{1 + t^4}.$$

The vertices A_1, A_2 of the rectangular hyperbola are the vertices of the lemniscate, at the point O there is a double point (of inflexion) with the tangents $y = \pm x$ (which are the asymptotes of the hyperbola).

The lemniscate of Bernoulli is one of the *Cassinian ovals*, i.e. its points have a constant product (equal to $\frac{1}{2}a^2$) of their distances from two fixed points $(\pm\frac{1}{2}a\sqrt{2}, 0)$.

Construction 3 of points of a lemniscate of Bernoulli: We intersect the circle with centre O and radius $\frac{1}{2}a\sqrt{2}$ by a line, for example, from the vertex A_1, at the points C_1, C_2. Then $r_1 = \overline{A_1C_1}$, $r_2 = \overline{A_1C_2}$ are the focal radii of a point of the lemniscate of Bernoulli.

The polar form shows that φ is restricted to the intervals $(-\frac{1}{4}\pi, \frac{1}{4}\pi)$, $(\frac{3}{4}\pi, \frac{5}{4}\pi)$, and thus the curve lies within the right angles made by the tangents at the point O, containing the x-axis.

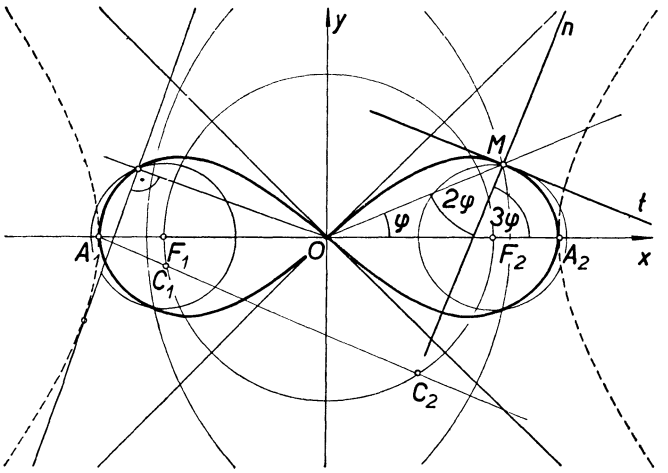


Fig. 4.40.

A lemniscate of Bernoulli has two axes of symmetry and, hence, it is radially symmetrical about their point of intersection. At the points whose coordinates are $(\pm\frac{1}{4}a\sqrt{6}, \pm\frac{1}{4}a\sqrt{2})$, i.e. for which $\varphi = \pm\pi, \frac{5}{6}, \frac{7}{6}\pi$ and $\varrho = \frac{1}{2}\sqrt{2}$, the tangents are parallel to the x-axis.

The angle between a tangent and the polar axis is equal to $\pm\frac{1}{2}\pi + 3\varphi$, the angle between a normal and the polar axis is equal to 3φ, and the angle between a normal and the polar radius is equal to 2φ.

Example 5 (The *conchoid of Nicomedes*; Fig. 4.41). We intersect a fixed straight line $x = a$ by a pencil of straight lines with vertex (pole) at O. On each line of the pencil we mark off segments of a constant length b on both sides of the point of intersection with the fixed straight line. The end points of the segments lie on a conchoid of Nicomedes whose equation in polar coordinates is

$$\varrho = \frac{a}{\cos\varphi} \pm b\,,$$

and in rectangular coordinates is

$$(x^2 + y^2)(x - a)^2 - b^2x^2 = 0\,.$$

A conchoid of Nicomedes consists of two branches; it is symmetrical about the x-axis, and the straight line $x = a$ is its asymptote. If $b > a$, then a branch has a node at O; if $b = a$, then it has a cusp at O; if $b < a$, then O is an isolated point.

The normals of all conchoids corresponding to the points of a given polar radius pass through a point N which is the point of intersection of the perpendicular to the polar radius at the pole O with the line parallel to the x-axis through the point of intersection P of the polar radius and the straight line $x = a$.

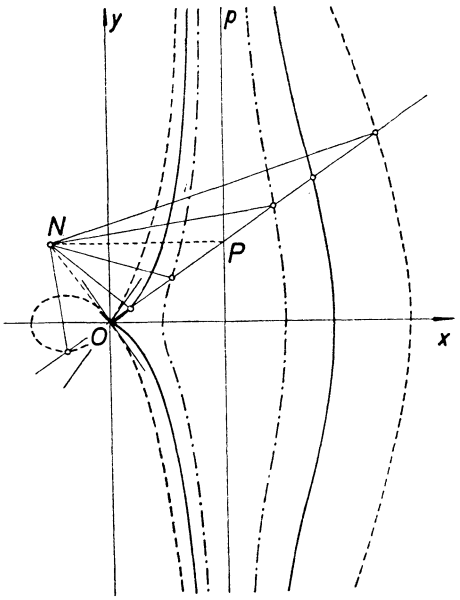


Fig. 4.41.

Example 6 (The *conchoid of a circle*; Fig. 4.42). We intersect the circle $\varrho = a \cos \varphi$ by a pencil of straight lines with the vertex (pole) at O and mark off segments of a constant length b on both sides from the point of intersection of a straight line of the pencil and the circle. The end points of the segments lie on a conchoid of the circle whose equation in polar coordinates is

$$\varrho = a \cos \varphi \pm b ,$$

and in rectangular coordinates is

$$(x^2 + y^2 - ax)^2 - b^2(x^2 + y^2) = 0 .$$

The conchoid of a circle (the *limaçon of Pascal*) is symmetrical about the x-axis, and has a double point at the pole O. (For $b < a$ the double point is a node, for $b = a$ it is a cuspidal point (the curve is a cardioid) and for $b > a$ it is an isolated

point.) The equations of the tangents at the pole are

$$x\sqrt{(a^2 - b^2)} \pm by = 0\,;$$

the equation of the double tangent is

$$x + \frac{b^2}{4a} = 0\,;$$

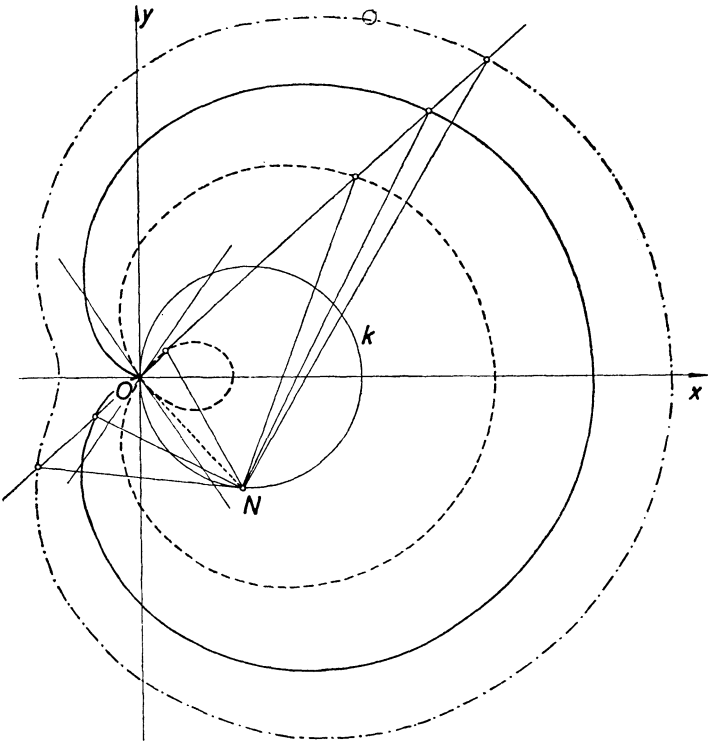


Fig. 4.42.

the points of contact have the ordinates

$$y = \frac{\pm b\sqrt{(4a^2 - b^2)}}{4a}\,.$$

The normals of all conchoids corresponding to the points of a given polar radius pass through a point N which is the point of intersection of the perpendicular erected to the polar radius at the pole O and the fixed circle.

4.12. The Sine Curves

To express periodical phenomena that repeat, without change, after a certain time, we use the *sine functions*. The least value of the constant, which, being added to the argument, does not change the value of the function, is called a *primitive period*.

Example 1. $y = \sin x$ (Fig. 4.43). The primitive period is 2π, the zero points $0, \pm\pi, \pm 2\pi, \ldots$ are points of inflexion of the curve and the tangents at these points make an angle of $\pm\frac{1}{4}\pi$ with the x-axis.

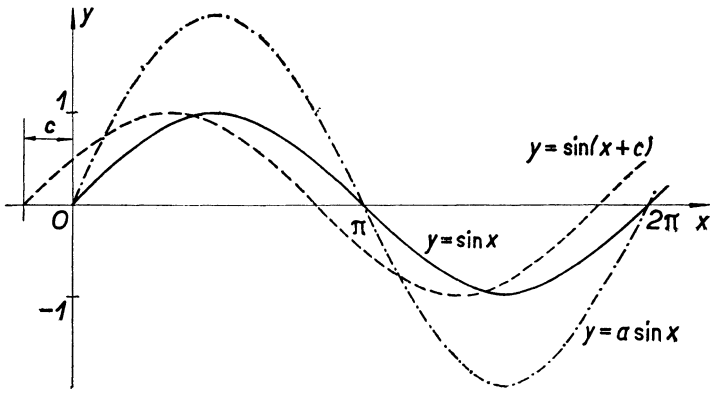


Fig. 4.43.

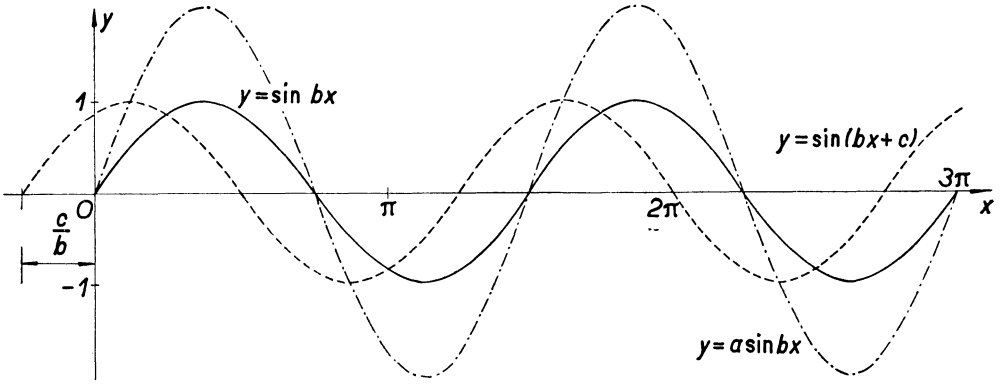


Fig. 4.44.

Example 2. $y = \sin(x + c)$. The graph can be obtained from the graph of Example 1 (Fig. 4.43) by a translation through a distance $-c$ in the direction of the x-axis.

Example 3. $y = a \sin x$ (Fig. 4.43). The graph can be obtained from the graph of Example 1 by multiplying the ordinates y by a.

Example 4. $y = \sin bx$ (Fig. 4.44). The primitive period is $2\pi/b$ and the zero points (of inflexion) are $x = \pm k\pi/b$ ($k = 0, 1, 2, \ldots$). The tangents at these points have the direction of the hypotenuse of the right-angled triangle one side of which is on the x-axis and is of unit length, the other side being of length b. The graph can be obtained from the graph of Example 1 by a change of the x-coordinates in the ratio $1/b$. The coefficient b indicates the number of waves coming into the length 2π and is called the *circular (angular) frequency*.

Example 5. $y = a \sin bx$ (Fig. 4.44). The graph can be obtained from the graph of Example 4 by multiplying the ordinates y by a.

Example 6. $y = a \sin(bx + c)$. The graph can be obtained from the graph of Example 5 (Fig. 4.44) by a translation through a distance $-c/b$ in the positive direction of the x-axis. By this function, so-called simple *harmonic motion* is given. The notation

$$y = a \sin(\omega t + \varphi) \tag{1}$$

is often used, with *amplitude* $a > 0$, *circular frequency* ω and *phase displacement* φ ($|\varphi| < \pi$). The period is $T = 2\pi/\omega$; the frequency is $n = \frac{1}{2}\omega/\pi = 1/T$.

Equation (1) can be put in the form

$$y = a_1 \sin \omega t + a_2 \cos \omega t, \tag{2}$$

where $a_1 = a \cos \varphi$, $a_2 = a \sin \varphi$. Conversely, if the form (2) is given, we get (1) by putting

$$a = \sqrt{(a_1^2 + a_2^2)}, \quad \varphi = \arctan \frac{a_2}{a_1} + k\pi,$$

where

$$\begin{aligned} k &= 0 \text{ for } a_1 > 0, \\ k &= 1 \text{ for } a_1 < 0, \quad a_2 > 0, \\ k &= -1 \text{ for } a_1 < 0, \quad a_2 < 0. \end{aligned}$$

4.13. The Curves of Oscillations

(a) Undamped (continuous) oscillations

(α) *Free undamped oscillations* are accomplished by a particle of mass m, on which a force Cy proportional to the displacement y from an equilibrium position is exerted; in dynamics, C is called the *spring constant*. The motion is given by the following differential equation:

$$\ddot{y} + \frac{C}{m} y = 0.$$

The solution is (see §17.13 and equation (4.12.1))

$$y = a \sin(\omega_0 t + \varphi), \tag{1}$$

where $\omega_0 = \sqrt{(C/m)}$ and a, φ are constants given by the initial conditions of the motion.

The composition of two harmonic motions

I. IDENTICAL CIRCULAR FREQUENCIES ω

$$a \sin(\omega t + \varphi_1) + b \sin(\omega t + \varphi_2) = A_1 \sin \omega t + A_2 \cos \omega t = A \sin(\omega t + \varphi), \tag{2}$$

where

$$A_1 = a \cos \varphi_1 + b \cos \varphi_2, \quad A_2 = a \sin \varphi_1 + b \sin \varphi_2, \quad A = \sqrt{(A_1^2 + A_2^2)},$$

$$\tan \varphi = \frac{A_2}{A_1}.$$

In particular, for equal amplitudes $b = a$ we get

$$a \sin(\omega t + \varphi_1) + a \sin(\omega t + \varphi_2) = 2a \cos \frac{\varphi_1 - \varphi_2}{2} \sin(\omega t + \varphi),$$

where

$$\varphi = \frac{\varphi_1 + \varphi_2}{2}. \tag{3}$$

Thus, in the case of equal frequencies the sum is a harmonic motion of the same frequency.

II. IDENTICAL AMPLITUDES, DIFFERENT FREQUENCIES

$$a \sin \omega_1 t + a \sin \omega_2 t = 2a \cos \frac{\omega_1 - \omega_2}{2} t \sin \frac{\omega_1 + \omega_2}{2} t. \tag{4}$$

(β) *Forced undamped (continuous) oscillation* is the motion of a particle of mass m under a periodically varying force $P \sin \omega t$ in addition to the force Cy. This motion satisfies the differential equation

$$\ddot{y} + \frac{C}{m} y = \frac{P}{m} \sin \omega t.$$

The solution is (see § 17.14):

$$y = Y \sin \omega t + a \sin (\omega_0 t + \varphi) \quad (\omega \neq \omega_0) \tag{5}$$

where $\omega_0 = \sqrt{(C/m)}$, a, φ are constants given by the initial conditions and

$$Y = \frac{P}{m(\omega_0^2 - \omega^2)}. \tag{6}$$

For $\omega = \omega_0$ (the case of *resonance*),

$$y = -\frac{P}{2m\omega_0} t \cos \omega_0 t + a \sin (\omega_0 t + \varphi). \tag{7}$$

As a rule, m, ω_0, P are fixed constants. The dependence of Y on ω expressed in (6) is illustrated in Fig. 4.45 (the *resonance curve*).

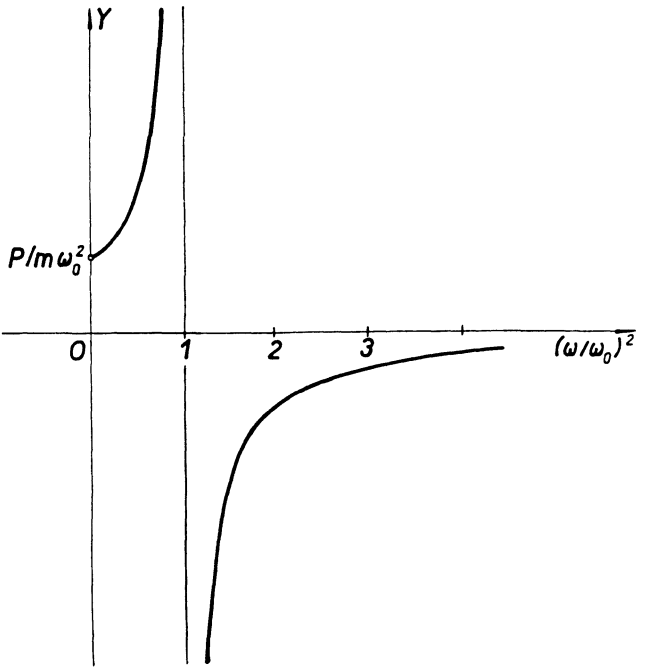


Fig. 4.45

(b) Damped oscillations. The motion is retarded by a force F proportional to the velocity ($F = -k\dot{y}$).

(α) *Free damped oscillations.* The differential equation of the motion is

$$\ddot{y} + 2b\dot{y} + \omega_0^2 y = 0 \tag{8}$$

where $\omega_0 = \sqrt{(C/m)}$, $b = k/(2m)$. The solution depends on the roots of the auxiliary equation (cf. § 17.13)

$$\alpha^2 + 2b\alpha + \omega_0^2 = 0\,. \tag{9}$$

1. If $b = \omega_0$, then $\alpha_1 = \alpha_2 = -\omega_0$. The general solution is

$$y = \mathrm{e}^{-\omega_0 t}(C_1 + C_2 t)\,.$$

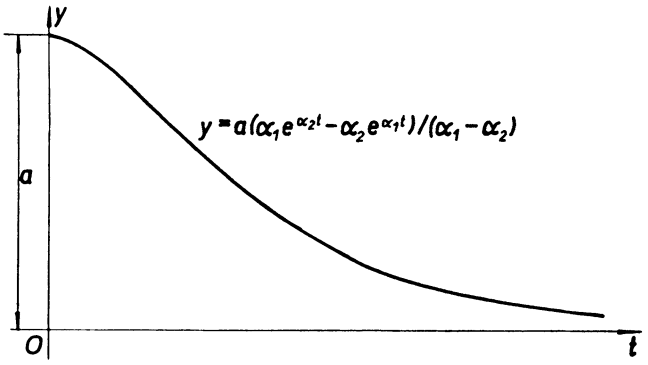


Fig. 4.46.

If, for $t = 0$, the initial conditions are $y = a$, $\dot{y} = 0$, then the solution is

$$y = a\mathrm{e}^{-\omega_0 t}(1 + \omega_0 t)\,.$$

This is the case of *critical damping*. For $t \to +\infty$ we have $y \to 0$ (Fig. 4.46).

2. If $b > \omega_0$, then the roots

$$\alpha_1 = -b + \sqrt{(b^2 - \omega_0^2)}\,, \quad \alpha_2 = -b - \sqrt{(b^2 - \omega_0^2)}$$

are real and distinct. The general solution is

$$y = C_1\mathrm{e}^{\alpha_1 t} + C_2\mathrm{e}^{\alpha_2 t}\,;$$

if for $t = 0$,

$$y = a\,, \quad \dot{y} = 0\,,$$

then

$$y = \frac{a}{\alpha_1 - \alpha_2}(\alpha_1\mathrm{e}^{\alpha_2 t} - \alpha_2\mathrm{e}^{\alpha_1 t})\,.$$

This case is referred to as *supercritical damping* (Fig. 4.47). The motions 1, 2 are called *aperiodic*.

3. If $b < \omega_0$, then, writing

$$\omega_1 = \sqrt{(\omega_0^2 - b^2)}\,,$$

the general solution is

$$y = e^{-bt}(C_1 \cos \omega_1 t + C_2 \sin \omega_1 t) .$$

If, for $t = 0$, the initial conditions are $y = a$, $\dot{y} = 0$, then

$$y = e^{-bt}\left(a \cos \omega_1 t + \frac{ab}{\omega_1} \sin \omega_1 t\right) = Ae^{-bt} \sin (\omega_1 t + \varphi) \tag{10}$$

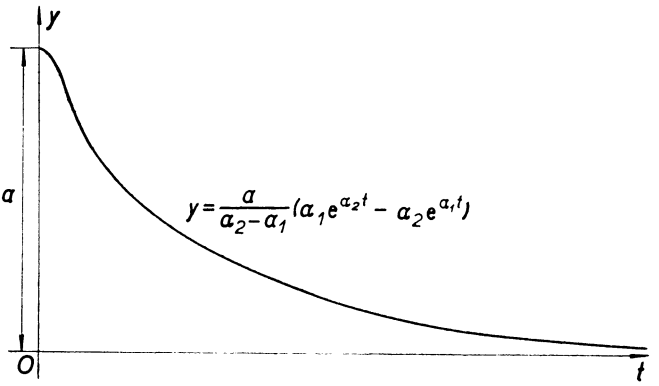


Fig. 4.47.

(cf. equation (4.12.2)). Here, the period $T = 2\pi/\omega_1$ is longer than in the case of an undamped oscillation. The ratio of the displacements y_1, y_2 at instants t_1, $t_1 + T$ is

$$\frac{y_1}{y_2} = e^{2\pi b/\omega_1} ;$$

its natural logarithm $\vartheta = 2\pi b/\omega_1 = bT$ is called the *logarithmic decrement* of the motion.

The zero points of the curve (10) are obtained for $t_n = (n\pi - \varphi)/\omega_1$, the vertices for

$$t_n = \frac{\arctan (\omega_1/b) - \varphi + n\pi}{\omega_1}, \quad y_n = \pm \frac{A\omega_1}{\sqrt{(\omega_1^2 + b^2)}} e^{-bt_n} .$$

The curve (10) can be constructed (Fig. 4.48) by means of the enveloping curves

$$y_1 = Ae^{-bt} \quad \text{and} \quad \bar{y}_1 = -Ae^{-bt}$$

(see Construction 4.9.1) and by the curve

$$y_2 = \sin (\omega_1 t + \varphi) ,$$

using the proportion

$$1 : y_1 = y_2 : y .$$

(β) *Forced damped oscillations.* The differential equation is

$$\ddot{y} + 2b\dot{y} + \omega_0^2 y = \frac{P}{m} \sin \omega t \tag{11}$$

(the notation as in equation (8)). The general solution (provided equation (9) has complex roots) is

$$\begin{aligned} y &= A_1 \sin \omega t + A_2 \cos \omega t + e^{-bt}(C_1 \cos \omega_1 t + C_2 \sin \omega_1 t) = \\ &= A \sin (\omega t + \varphi) + e^{-bt}(C_1 \cos \omega_1 t + C_2 \sin \omega_1 t) \end{aligned} \tag{12}$$

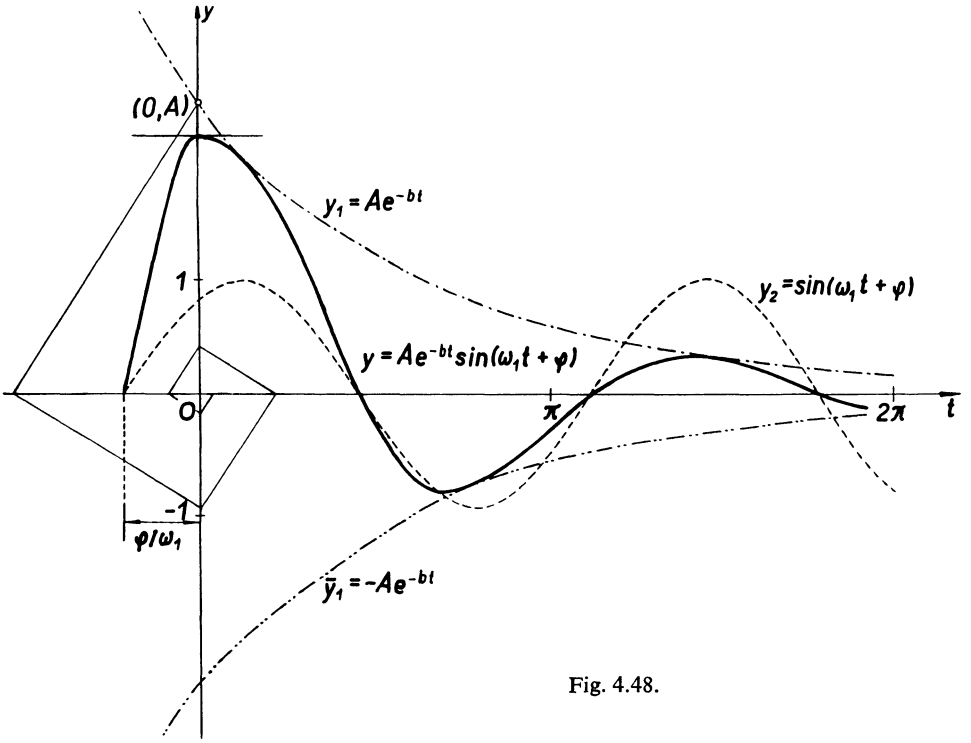


Fig. 4.48.

(cf. (4.12.2)). Here

$$\begin{aligned} \omega_1 = \sqrt{(\omega_0^2 - b^2)}, \quad A_1 &= \frac{P(\omega_0^2 - \omega^2)}{m[(\omega_0^2 - \omega^2)^2 + 4b^2\omega^2]}, \\ A_2 &= \frac{-2Pb\omega}{m[(\omega_0^2 - \omega^2)^2 + 4b^2\omega^2]}, \\ A &= \frac{P}{m\sqrt{[(\omega_0^2 - \omega^2)^2 + 4b^2\omega^2]}}. \end{aligned} \tag{13}$$

Since $b > 0$, the second term of (12) becomes negligible, after a certain time, with regard to the first one, so that the motion is characterized only by the first term with the amplitude A. The magnitude of this amplitude depends on ω (P, m, ω_0, b are constants) and is illustrated in Fig. 4.49 (the *resonance curve*). The case $\omega = \omega_0$ is said to give *resonance*. For small b ($b \ll \omega_0$) A attains its maximum for $\omega \approx \omega_0$ (more precisely, if $\omega_0^2 > 2b^2$, for $\omega = \sqrt{(\omega_0^2 - 2b^2)}$). The first and second terms of the oscillation (12) are often called the *steady-state* and *transient oscillations*, respectively.

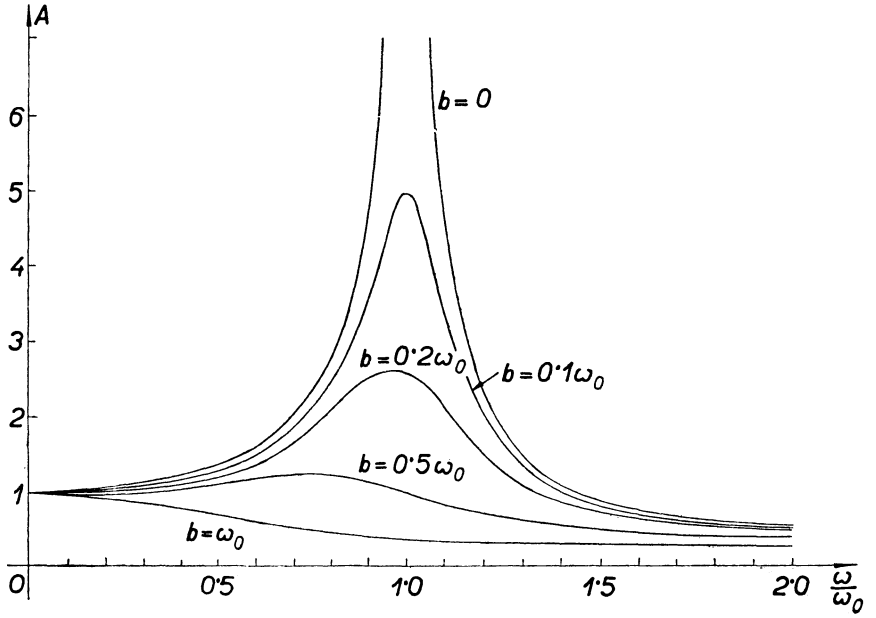


Fig. 4.49.

4.14. Growth Curves

Definition 1. A solution $x = F(t)$ of the differential equation

$$\frac{dx}{dt} = f(x) \tag{1}$$

is called the *law of growth* which is assigned to any phenomenon observed to satisfy the equation.

We make the following assumptions:

(a) the necessary parameters involved in $f(x)$ have been established for the phenomenon under consideration on the basis of statistical data;

(b) the growth of the quantity x in time t takes place without any external intervention;

(c) the initial condition $t = 0$, $x = x_0$ holds.

Example 1. If $f(x) = m = \text{const.}$, the solution of the differential equation (1) is

$$x = mt + x_0 \tag{2}$$

and the growth curve is the straight line of gradient $\tan \alpha = m$; this line passes through the point $(0, x_0)$.

Example 2. If $f(x) = ax + b$, $a \neq 0$, the solution is

$$x = -\frac{b}{a} + \left(x_0 + \frac{b}{a}\right) e^{at} \tag{3}$$

and the law of growth is given by the exponential curve passing through the point $(0, x_0)$.

If $b = 0$, $x_0 \neq 0$, the law (3) assumes the form (Fig. 4.50)

$$x = x_0 e^{at}, \quad \text{or} \quad (\text{writing} \quad c = -a \quad \text{when} \quad a < 0) \quad x = x_0 e^{-ct}. \tag{4}$$

If $b \neq 0$, $x_0 = 0$, the law (3) takes the form (Fig. 4.51)

$$x = A(e^{at} - 1) \quad \text{or} \quad x = A(1 - e^{-ct}), \tag{5}$$

where $c = -a$ (if $a < 0$) and $A = b/a$ or $A = b/c$, respectively.

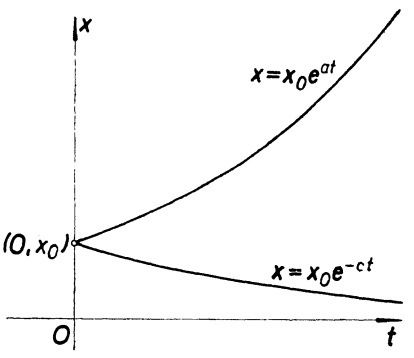


Fig. 4.50.

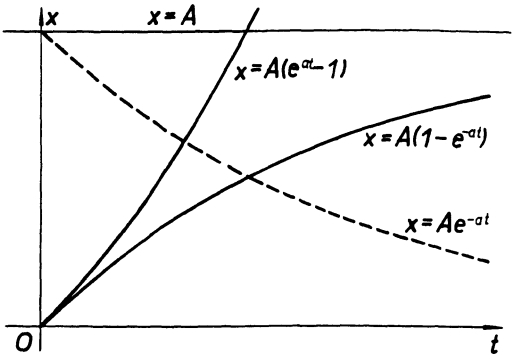


Fig. 4.51.

In Fig. 4.51 the auxiliary curve $x = Ae^{-ct}$ is also shown. The line $x = A = b/c$ is the asymptote of the second of curves (5) and it determines the limit of the evolution.

Example 3. If $f(x) = m + ax - bx^2 = b(x - x_2)(x_1 - x)$, $a > 0$, $b > 0$, $m > 0$ so that x_1, x_2 are the roots of the quadratic equation $f(x) = 0$, then the general

solution of equation (1) is

$$\frac{x - x_2}{x_1 - x} = e^{(x_1 - x_2)(bt + c)}, \quad c = \text{const}. \tag{6}$$

The solution can be written in the form

$$x = x_2 + \frac{x_1 - x_2}{1 + Ce^{-(x_1 - x_2)bt}} \tag{7}$$

or

$$x = \frac{x_1 + x_2}{2} + \frac{x_1 - x_2}{2} \tanh \frac{x_1 - x_2}{2}(bt + c). \tag{8}$$

The straight lines $x = x_1$, $x = x_2$ are the asymptotes of integral curves (6). The ordinate of the point of inflexion of the curve is $\xi = \frac{1}{2}(x_1 + x_2)$; the abscissa τ of this point satisfies $b\tau + c = 0$. Using the given initial condition $x = x_0$, we get

$$\tau = \frac{1}{b(x_1 - x_2)} \ln \frac{x_1 - x_0}{x_0 - x_2}.$$

By translation of the origin to the point of inflexion (by means of the equations $X = x - \xi$, $T = t - \tau$) equation (8) assumes the form

$$X = \frac{x_1 - x_2}{2} \tanh \frac{x_1 - x_2}{2} bT. \tag{9}$$

The gradient of the tangent at the point of inflexion is

$$\tan \alpha = \frac{b}{4}(x_1 - x_2)^2.$$

Hence the law of growth in this case takes the form of a hyperbolic tangent, sometimes called a *logistic curve.* The curve is symmetrical about the point of inflexion and its graph lies within the strip bounded by the asymptotes $x = x_1$, $x = x_2$.

Example 4. If, in Example 3, $m = 0$, then the solution of equation (1) is called *Robertson's law of growth.* Here $f(x) = x(a - bx)$. The given initial condition $t = 0$, $x = x_0$ yields the solution

$$x = \frac{a}{b(1 + Ce^{-at})}, \quad \text{where} \quad C = \frac{a - bx_0}{bx_0}. \tag{10}$$

Using the coordinates $\xi = a/2b$, $\tau = (1/a) \ln [(a - bx_0)/bx_0]$ of the point of inflexion we get the form

$$x = \frac{a}{2b} + \frac{a}{2b} \tanh \frac{a}{2}(t - \tau); \tag{11}$$

on translating the origin we get

$$X = \frac{a}{2b} \tanh \frac{a}{2} T. \tag{12}$$

The asymptotes are $x = a/b$, $x = 0$; the gradient of the tangent at the point of inflexion is

$$\tan \alpha = \frac{a^2}{4b}.$$

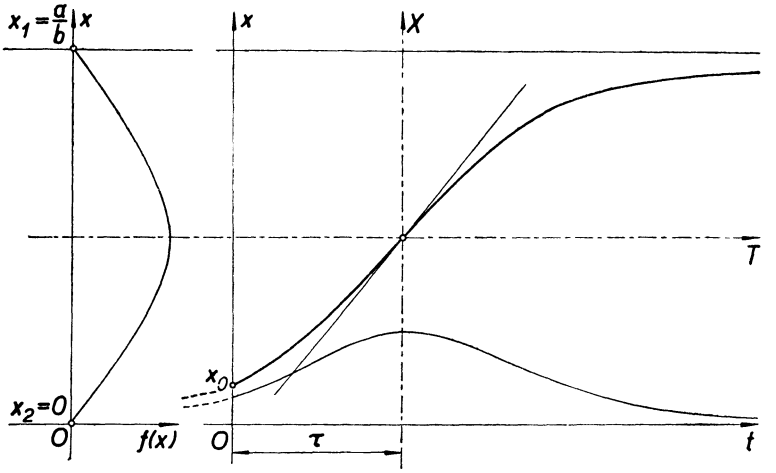


Fig. 4.52.

In Fig. 4.52 Robertson's law of growth together with the parabola $f(x) = ax - bx^2$ and the determination of the coordinates of the point of inflexion is illustrated; the curve is constructed in the coordinates T, X. The curve

$$\frac{dX}{dT} = \frac{a^2}{4b \cosh^2 \frac{1}{2} aT},$$

(also illustrated), shows the speed of growth.

4.15. Some Approximate Constructions

Construction 1 of the tangent t at a given point T of a curve k (Fig. 4.53): We draw secants through the point T meeting the curve in the points ..., -2, -1, 1, 2, ... near T, and an arbitrary straight line q (not passing through T) in the points ..., $-2'$, $-1'$, $1'$, $2'$, ... Now, on the secants corresponding to the points 1, 2, ... we determine the points at the distances $T1$, $T2$, ... from $1'$, $2'$, ... on one side of the straight line q;

in a similar manner, we determine the points on the secants corresponding to -1, $-2, \ldots$ on the other side of q. These points determine the curve s intersecting the straight line q at the point O'. The tangent t is then the line $O'T$.

Construction 2 of the point of contact T of a tangent t constructed (by means of a ruler) from a point R to a curve k (Fig. 4.54): From the point R we construct a pencil of secants, which intersect the curve k in the neighbourhood of the required point T

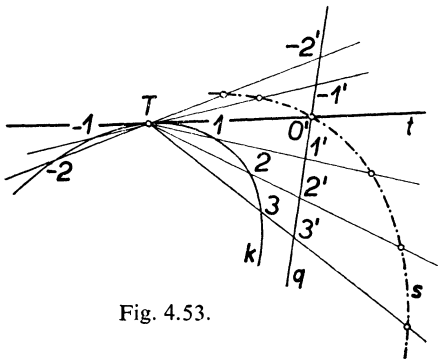


Fig. 4.53.

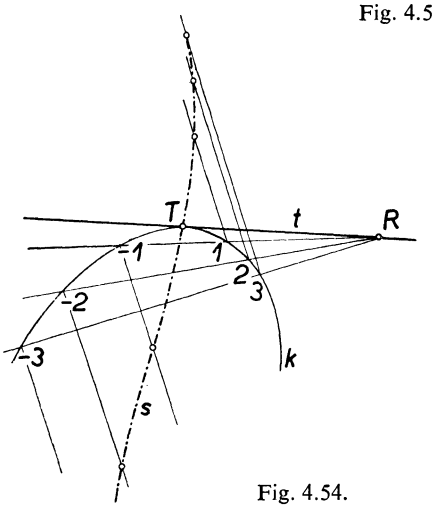


Fig. 4.54.

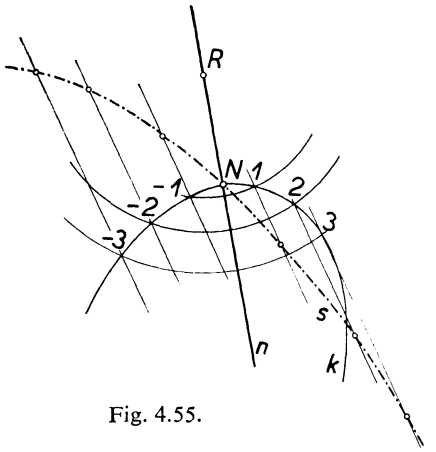


Fig. 4.55.

in pairs of points $-1, +1; -2, +2; \ldots$. On parallel lines drawn through these points we mark off points at distances equal to the length of the corresponding chords; on one side of the curve for the points marked by $+$ and on the other side for the points marked by $-$. If we join those points, the resulting curve s intersects the tangent t (and thus, also the curve k) at the point of contact T of the tangent t.

Construction 3 of a normal n from a point R to a curve k (Fig. 4.55): From the point R we describe concentric circles, each intersecting the curve at two points $-1, +1; -2, +2; \ldots$ in the neighbourhood of the foot N of the required normal. The curve s constructed in the same way as in Construction 2 intersects the curve k at the foot N of the required normal n from the point R.

5. PLANE ANALYTIC GEOMETRY

By MILOSLAV ZELENKA

References: [9], [12], [20], [22], [27], [46], [55], [59], [62], [98], [101], [103], [117], [145], [150], [156], [162], [165], [175], [186].

5.1. Coordinates of a Point on a Straight Line and in a Plane. Distance between Two Points

Definition 1. Let us divide a straight line p by a point O into two half-lines $+p$ and $-p$ (Fig. 5.1). Let us choose on p a unit of length. The *coordinate* x of a point M is defined to be the distance of the point M from the point O prefixed by a sign, plus or minus (the so-called *directed distance*) according as M belongs to $+p$ or $-p$, respectively. We write $M(x)$.

REMARK 1. The position of a point M on a line p is uniquely determined by the coordinate x (and vice versa). We say that a *coordinate system* has been introduced on the line p. The point O is called the *origin of the coordinate system.*

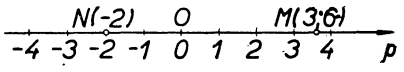


Fig. 5.1.

Theorem 1. *The distance d between two points $A(x_1)$ and $B(x_2)$ on a line is equal to*

$$d = |x_2 - x_1| . \tag{1}$$

REMARK 2. In a similar way, a coordinate system can be introduced in a plane (Fig. 5.2): We select units of length on two intersecting lines, called *axes of coordinates*; the intersection of the lines is taken as the origin on each of them; we denote it by O and call it the *origin of the coordinate system in the plane.* A point M in the plane is then uniquely determined by its coordinates x, y (see Fig. 5.2), and vice versa.

If the axes x, y are mutually perpendicular, the coordinate system is called *rectangular.* As in the case of coordinates on a line, the *coordinate* x or y gives the *directed distance* of the point $M(x, y)$ from the coordinate axis y or x, respectively.

A rectangular system in which both axes have the same unit of length is called *cartesian*. Throughout this chapter — unless otherwise stated — we use the cartesian coordinate system.

The plane is divided by the coordinate axes into four parts called the *quadrants* (Fig. 5.2).

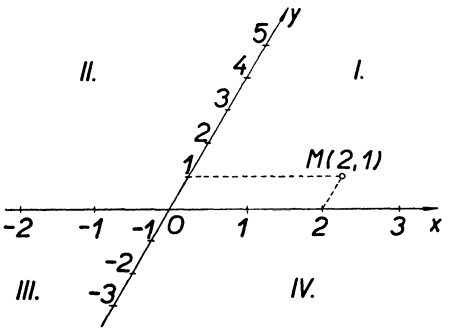


Fig. 5.2.

Theorem 2. *The distance between two points* $A(x_1, y_1)$ *and* $B(x_2, y_2)$ *in a cartesian coordinate system is equal to*

$$d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}. \tag{2}$$

REMARK 3. In § 7.1, the concept of a three-dimensional (real) vector is introduced. Similarly, a two-dimensional vector $\boldsymbol{a}$ with two components a_1, a_2 can be defined; the notation $\boldsymbol{a}(a_1, a_2)$ or $\boldsymbol{a} = (a_1, a_2)$ is used. As in the case of three-dimensional vectors, two-dimensional vectors can also be represented by directed line segments. If two-dimensional vectors are used in problems of analytical geometry in a plane, they are, of course, represented by directed segments lying in the plane.

A two-dimensional vector represented by a directed line segment in a plane xy is often considered as a special case of a three-dimensional vector, the third component of which is zero (although this is not written explicitly). Then we can, without any alterations, apply the definitions and theorems of Chap. 7 concerning operations on vectors in three-dimensional space. For example, the formula

$$\boldsymbol{a} \,.\, \boldsymbol{b} = a_1 b_1 + a_2 b_2$$

for the scalar product of two vectors $\boldsymbol{a}(a_1, a_2)$ and $\boldsymbol{b}(b_1, b_2)$ holds.

The number $\sqrt{(a_1^2 + a_2^2)}$ is called the *length* (or *magnitude*) *of the vector* $\boldsymbol{a}$ and is denoted by either $|\boldsymbol{a}|$ or a.

5.2. Division of a Line Segment in a Given Ratio. Area of a Triangle and Polygon

Theorem 1. *An arbitrary point $M(x, y)$ on the line segment with the end points $M_1(x_1, y_1)$ and $M_2(x_2, y_2)$ can be represented in vector form by*

$$\mathbf{m} = \mathbf{m}_1 + t(\mathbf{m}_2 - \mathbf{m}_1) \quad (0 \leqq t \leqq 1), \tag{1}$$

where $\mathbf{m}$, $\mathbf{m}_1$ and $\mathbf{m}_2$ are the radius vectors of the points M, M_1 and M_2, respectively, or in coordinate form by

$$\begin{aligned} x &= x_1 + t(x_2 - x_1), \\ y &= y_1 + t(y_2 - y_1) \end{aligned} \tag{2}$$

$(0 \leqq t \leqq 1)$.

If the point $M(x, y)$ divides the line segment M_1M_2 in the ratio $M_1M : M_2M = \lambda < 0$, then, putting $t = -\lambda/(1 - \lambda)$, we obtain

$$x = \frac{x_1 - \lambda x_2}{1 - \lambda}, \quad y = \frac{y_1 - \lambda y_2}{1 - \lambda}. \tag{3}$$

If $\lambda = -1$, M is the midpoint of the line segment M_1M_2, and formulae (3) become

$$x = \frac{x_1 + x_2}{2}, \quad y = \frac{y_1 + y_2}{2}. \tag{4}$$

Theorem 2. *The area of a polygon with vertices $A_1(x_1, y_1)$, $A_2(x_2, y_2)$, ..., ..., $A_n(x_n, y_n)$ occuring in that order, is*

$$P = \tfrac{1}{2}\left|\begin{vmatrix} x_1, & x_2 \\ y_1, & y_2 \end{vmatrix} + \begin{vmatrix} x_2, & x_3 \\ y_2, & y_3 \end{vmatrix} + \ldots + \begin{vmatrix} x_n, & x_1 \\ y_n, & y_1 \end{vmatrix}\right|. \tag{5}$$

In particular, the area of the triangle with vertices $A_1(x_1, y_1)$, $A_2(x_2, y_2)$ and $A_3(x_3, y_3)$ is

$$P = \tfrac{1}{2}\left|\begin{vmatrix} x_1, & y_1, & 1 \\ x_2, & y_2, & 1 \\ x_3, & y_3, & 1 \end{vmatrix}\right|. \tag{6}$$

5.3. The Equation of a Curve as the Locus of a Point

Definition 1. The *equation of a curve* is the name given to the relation (equation) which is satisfied by the coordinates x, y of all the points lying on the given curve (and only those points).

In order to obtain the equation of a curve as a locus of a point having a given property, we proceed as follows:

1. we choose an arbitrary point M of the curve and denote its coordinates by (x, y);
2. we express the required property of points on the locus by an equation between x and y;
3. we arrange the equation in a simpler form, if possible, at the same time expressing all the quantities involved in terms of x, y and the given elements (constants).

Example 1. Let us obtain the equation of the locus of the point in a plane, which is always at a distance $d = 3$ from the point $S(-2, 1)$.

Thus,

1. $$M(x, y)\,; \tag{1}$$

2. $$\sqrt{[(x + 2)^2 + (y - 1)^2]} = 3\,; \tag{2}$$

3. we square equation (2) and obtain the equation $(x + 2)^2 + (y - 1)^2 = 9$, which needs no further rearrangement; it is the equation of a circle.

5.4. The Gradient, Intercept, General and Vector Forms of the Equation of a Straight Line. Parametric Equations of a Straight Line. Equation of the Straight Line through Two Given Points. The Point of Intersection of Two Straight Lines. Equation of a Pencil of Lines

$$y = kx + q \quad (\textit{the gradient form of the equation of a straight line})\,; \tag{1}$$

$$\frac{x}{p} + \frac{y}{q} = 1 \quad (p \neq 0, q \neq 0) \quad (\textit{the intercept form of the equation of a straight line})\,; \tag{2}$$

$$ax + by + c = 0 \quad (a^2 + b^2 > 0) \quad (\textit{the general equation of a straight line})\,; \tag{3}$$

$$\mathbf{m} = \mathbf{m}_1 + t\mathbf{a} \quad (\mathbf{a} \neq \mathbf{0}) \quad (\textit{the vector equation of a straight line})\,; \tag{4}$$

$$\left.\begin{aligned} x &= x_1 + a_1 t \\ y &= y_1 + a_2 t \end{aligned}\right\} \quad (\textit{the parametric equations of a straight line})\,. \tag{5}$$

The geometrical meaning of the constants involved in these equations can be seen from Fig. 5.3, 5.4, 5.5, 5.6; $k = \tan\varphi$ is the so-called *slope* (or *gradient*) of the line. For $k = 0$ the line is parallel to the x-axis. For $q = 0$ the line passes through the origin. The equation of the y-axis or a line parallel to the y-axis (i.e. if $\varphi = \frac{1}{2}\pi$) cannot be written in the form (1).

The numbers, p, q, in equation (2) (which may be positive or negative) are the so-called *intercepts on the axes*. A straight line which passes through the origin or is

parallel to a coordinate axis cannot be written in the form (2). The constants a, b in equation (3) determine the vector $\boldsymbol{n}(a, b)$ perpendicular to the line (3). If $a = 0$, the line is parallel to the x-axis; if $b = 0$, it is parallel to the y-axis. The third parameter c is related to the distance of the line from the origin (see § 5.6); if $c = 0$, the line passes through the origin.

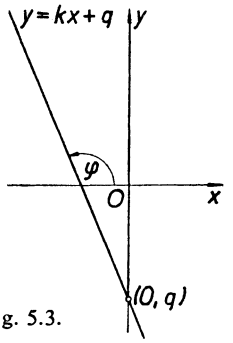


Fig. 5.3.

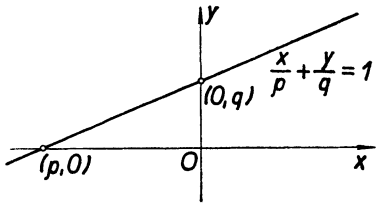


Fig. 5.4.

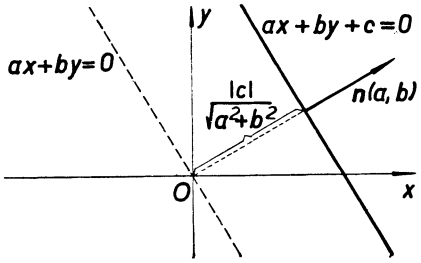


Fig. 5.5.

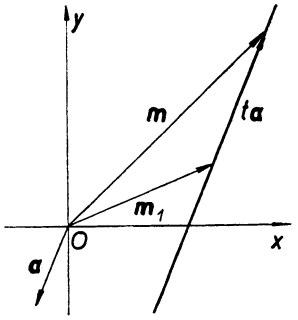


Fig. 5.6.

The line (4) passes through the point $M_1(x_1, y_1)$, the radius vector of which is denoted by $\boldsymbol{m}_1$; its direction is determined by the vector $\boldsymbol{a}(a_1, a_2)$ and t is a variable parameter $(-\infty < t < +\infty)$. To each particular value of t there corresponds a particular point $M(x, y)$ whose radius vector is $\boldsymbol{m}$ (Fig. 5.6). The vector equation is, in fact, a more concise version of the parametric equations (5).

Example 1. The straight line given by the parametric equations

$$x = 3 + 2t,$$

$$y = 1 - 3t$$

is to be expressed in the form (3).

Eliminating t from the parametric equations we obtain the required relation between x and y: adding three times the first equation to twice the second we obtain

$$3x + 2y = 11,$$

i.e.

$$3x + 2y - 11 = 0 .$$

Example 2. Find the equation of the straight line whose segment AB, intercepted by the positive semi-axes x and y, is bisected by the point $P(4, 3)$.

Similarity of the triangles PCB and AOB (Fig. 5.7) implies that $p = 8$, $q = 6$ and thus, by (2) the required equation is

$$\frac{x}{8} + \frac{y}{6} = 1 .$$

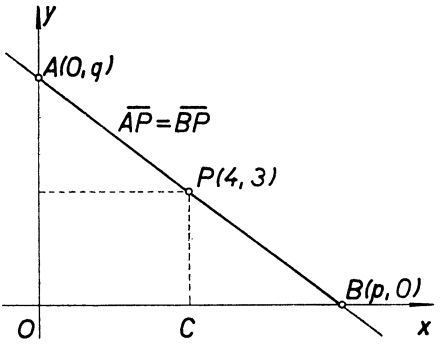


Fig. 5.7.

Theorem 1. *The equation of the straight line passing through two given points* $A(x_1, y_1)$ *and* $B(x_2, y_2)$ *is*

$$\frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1} ,$$

i.e.

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) .$$

If $x_2 = x_1$, or $y_2 = y_1$, then the equation of the straight line is $x = x_1$, or $y = y_1$, respectively.

Example 3. The line passing through the points $A(-1, 5)$, $B(3, 7)$ has the equation

$$y - 5 = \frac{7 - 5}{3 + 1}(x + 1) , \quad \text{i.e.} \quad x - 2y + 11 = 0 .$$

The line passing through the points $A(3, 3)$, $B(3, 8)$ has, of course, the equation $x = 3$ (and is a line parallel to the y-axis).

Theorem 2. *The point of intersection* $P(x_0, y_0)$ *of two intersecting straight lines given by equations* $a_1x + b_1y + c_1 = 0$ *and* $a_2x + b_2y + c_2 = 0$ *can be found*

by solving these equations simultaneously; hence,

$$x_0 = \frac{\begin{vmatrix} -c_1, & b_1 \\ -c_2, & b_2 \end{vmatrix}}{\begin{vmatrix} a_1, & b_1 \\ a_2, & b_2 \end{vmatrix}}, \quad y_0 = \frac{\begin{vmatrix} a_1, & -c_1 \\ a_2, & -c_2 \end{vmatrix}}{\begin{vmatrix} a_1, & b_1 \\ a_2, & b_2 \end{vmatrix}}, \quad \text{provided that} \quad \begin{vmatrix} a_1, & b_1 \\ a_2, & b_2 \end{vmatrix} \neq 0. \tag{6}$$

If $\begin{vmatrix} a_1, & b_1 \\ a_2, & b_2 \end{vmatrix} = 0$, *the lines are parallel or they coincide.*

Definition 1. The set of straight lines in a plane, all of which pass through one point S, is called a *pencil of lines.* The point S is called the *centre* (or *vertex*) *of the pencil of lines.*

Theorem 3. *The straight lines belonging to the pencil with centre* $S(x_0, y_0)$ *have as their equation either*

$$y - y_0 = k_1(x - x_0) \tag{7}$$

or

$$x - x_0 = k_2(y - y_0), \tag{7'}$$

where k_i $(i = 1, 2)$ *is a variable parameter; to each line of the pencil, there corresponds a unique value of* k_1 *or* k_2, *and conversely.*

Theorem 4. *The straight lines belonging to the pencil determined by the intersecting lines*

$$a_1x + b_1y + c_1 = 0 \quad \text{and} \quad a_2x + b_2y + c_2 = 0$$

have as their equations:

$$\lambda_1(a_1x + b_1y + c_1) + \lambda_2(a_2x + b_2y + c_2) = 0, \tag{8}$$

where λ_i $(i = 1, 2)$ *are variable parameters not simultaneously equal to zero; to each line of the pencil there corresponds a unique ratio* λ_1/λ_2 or λ_2/λ_1, *and conversely.*

Definition 2. Equations (7), (7') or (8) are called the *equations of a pencil of lines.*

Example 4. Let us find the equation of the straight line which passes through the point of intersection of the lines

$$2x - y + 3 = 0, \quad x + 3y - 1 = 0$$

and through the point $P(2, 1)$.

The equation of the line will be, by (8), of the form

$$\lambda_1(2x - y + 3) + \lambda_2(x + 3y - 1) = 0. \tag{9}$$

The point $P(2, 1)$ must lie on the line (9) and hence

$$\lambda_1(2 \cdot 2 - 1 + 3) + \lambda_2(2 + 3 \cdot 1 - 1) = 0,$$
$$6\lambda_1 + 4\lambda_2 = 0; \tag{10}$$

in order to satisfy equation (10), it suffices to put $\lambda_1 = 2$, $\lambda_2 = -3$. Then, by (9) the required equation is

$$2(2x - y + 3) - 3(x + 3y - 1) = 0, \quad \text{i.e.} \quad x - 11y + 9 = 0.$$

Check: We can calculate the coordinates of Q, the point of intersection of the two lines (by Theorem 2) and then verify that the points P and Q satisfy the equation $x - 11y + 9 = 0$.

5.5. Directed (Oriented) Straight Line. Direction Cosines. The Angle between Two Straight Lines

Definition 1. A straight line p is said to be *directed* (or *oriented*), if, for every pair of points $A, B\,(A \neq B)$ on this line, one can decide by means of a given rule, whether A lies before B (notation $A \prec B$) or B lies before A (while the relations $A \prec B$ and $B \prec C$ together imply $A \prec C$). We say that on p the so-called *positive sense* and *negative sense* of orientation are given. It is customary to mark the direction (orientation) of a line in diagrams by an arrow showing its positive sense.

In a similar way, a *directed half-line* and *directed line segment* are defined. In the case of a directed half-line we speak of its initial point; in the case of a directed line segment we speak of its initial and end points. If we choose a point O on a directed line, we divide it into the so-called *positive part* (*positive half-line*) $+p$ and the *negative part* (*negative half-line*) $-p$ by this point.

Definition 2. If $A(a_1, a_2)$ and $B(b_1, b_2)$ are two points on a directed line p such that A lies before B, then the expressions

$$\frac{b_1 - a_1}{\sqrt{[(b_1 - a_1)^2 + (b_2 - a_2)^2]}}$$

and

$$\frac{b_2 - a_2}{\sqrt{[(b_1 - a_1)^2 + (b_2 - a_2)^2]}}$$

are called the *direction cosines* of the directed line p; we denote them by $\cos \alpha_1$, $\cos \alpha_2$. (The unit-vector with components $\cos \alpha_1$, $\cos \alpha_2$ lies on the line p.)

Theorem 1. *The expressions introduced in Definition 2 and denoted by* $\cos \alpha_1$, $\cos \alpha_2$ *are cosines of the undirected angles* α_1, α_2 $(0 \leqq \alpha_1, \alpha_2 \leqq 180°)$ *between the*

positive part of the line p and the positive parts of the coordinate axes x and y, respectively.

Example 1. Let us choose the points $A(1, 2)$ and $B(0, -1)$ on the straight line $y = 3x - 1$ (Fig. 5.8); if this line is directed from A to B then its direction cosines are

$$\cos \alpha_1 = \frac{0 - 1}{\sqrt{[(0 - 1)^2 + (-1 - 2)^2]}} = \frac{-1}{\sqrt{10}},$$

$$\cos \alpha_2 = \frac{-1 - 2}{\sqrt{10}} = \frac{-3}{\sqrt{10}}.$$

The corresponding direction angles are

$$\alpha_1 \doteq 108°26',$$

$$\alpha_2 \doteq 161°34'.$$

Theorem 2. *If φ is the acute angle between the lines $y = k_1x + q_1$ and $y = k_2x + q_2$ then*

$$\tan \varphi = \left| \frac{k_1 - k_2}{1 + k_1k_2} \right| \quad \left(k_1 \neq -\frac{1}{k_2} ; \text{ see Theorem 4} \right). \tag{1}$$

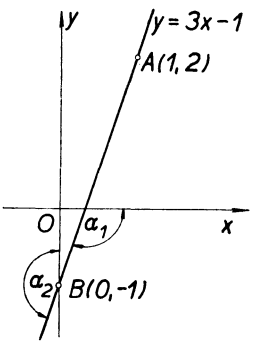


Fig. 5.8.

If the lines are given in the form $a_1x + b_1y + c_1 = 0$ and $a_2x + b_2y + c_2 = 0$, then

$$\tan \varphi = \left| \frac{a_1b_2 - a_2b_1}{a_1a_2 + b_1b_2} \right| \quad (a_1a_2 + b_1b_2 \neq 0 ; \text{ see Theorem 4}). \tag{1'}$$

Theorem 3. *The condition for the straight lines of Theorem 2 to be parallel is*

$$k_1 = k_2, \quad \text{or} \quad a_1b_2 - a_2b_1 = 0.$$

Theorem 4. *The condition for the straight lines of Theorem 2 to be perpendicular is*

$$k_2 = -\frac{1}{k_1}, \quad \text{or} \quad a_1a_2 + b_1b_2 = 0 .$$

Example 2. Let us determine the angles between the following straight lines: a) $y = 3x - 1$, $y = -x + 7$; b) $2x + 3y - 5 = 0$, $3x - 2y + 1 = 0$; c) $y = 2x + 3$, $4x - 2y + 1 = 0$.

Solution: a) $k_1 = 3$, $k_2 = -1$. Thus, $\tan \varphi = -4/(1 - 3) = 2$, $\varphi \doteq 63°26'$. b) $a_1a_2 + b_1b_2 = 2 . 3 - 3 . 2 = 0$; the lines are perpendicular. c) First, we put the equations of both lines in the same form: $2x - y + 3 = 0$, $4x - 2y + 1 = 0$. Clearly, $a_1b_2 - a_2b_1 = 2 . (-2) - (-1) . 4 = 0$; the lines are parallel.

REMARK 1. The equation of any line perpendicular to the straight line

$$a_1x + b_1y + c_1 = 0 \tag{2}$$

can be written in the form

$$b_1x - a_1y + c_2 = 0 \tag{3}$$

(since, the coefficients of the variables x and y in equations (2) and (3) satisfy the conditions of Theorem 4, namely $a_1b_1 + b_1(-a_1) = 0$).

Example 3. Find the equation of the line p passing through the point $P(1, 4)$ and perpendicular to the line

$$2x + 3y + 5 = 0 . \tag{4}$$

By (3), the equation of the line p can be written in the form

$$3x - 2y + c_2 = 0 . \tag{5}$$

Substituting the coordinates 1 and 4 of the point P for x and y respectively into (5), we obtain

$$3 . 1 - 2 . 4 + c_2 = 0 ,$$

i.e. $c_2 = 5$ and thus the equation of the line p is

$$3x - 2y + 5 = 0 .$$

Example 4. Find the equation of the straight line which passes through the point of intersection of the lines

$$x - 2y + 3 = 0 ,$$

$$3x + 5y - 2 = 0$$

and which is perpendicular to the line

$$4x + y - 7 = 0 . \tag{6}$$

By (5.4.8), p. 173, the equation of the required line is of the form

$$\lambda_1(x - 2y + 3) + \lambda_2(3x + 5y - 2) = 0 ,$$

i.e.

$$(\lambda_1 + 3\lambda_2)\, x + (-2\lambda_1 + 5\lambda_2)\, y + (3\lambda_1 - 2\lambda_2) = 0 . \tag{7}$$

The condition for the lines (6) and (7) to be perpendicular is, by Theorem 4,

$$4(\lambda_1 + 3\lambda_2) + (-2\lambda_1 + 5\lambda_2) = 0 ,$$

i.e.

$$2\lambda_1 + 17\lambda_2 = 0 .$$

Hence, it suffices to choose $\lambda_2 = 2$, $\lambda_1 = -17$. Substituting these values into (7), we obtain the required equation in the form

$$-11x + 44y - 55 = 0 ,$$

i.e.

$$x - 4y + 5 = 0 .$$

5.6. The Normal Equation of a Straight Line. Distance of a Point from a Straight Line. The Equations of the Bisectors of the Angles between Two Straight Lines

Definition 1. The equation

$$\frac{a}{\pm\sqrt{(a^2 + b^2)}}\, x + \frac{b}{\pm\sqrt{(a^2 + b^2)}}\, y + \frac{c}{\pm\sqrt{(a^2 + b^2)}} = 0 , \tag{1}$$

where a, b, c are three arbitrary numbers ($a^2 + b^2 > 0$) and the sign of the denominators is the opposite of that of the number c, is called the *normal equation of a straight line.*

The geometrical meanings of the coefficients are:

1. $\left(\frac{a}{\pm\sqrt{(a^2 + b^2)}}, \frac{b}{\pm\sqrt{(a^2 + b^2)}}\right)$ is a unit-vector perpendicular to the straight line (directed from the origin of coordinates to the line);

2. $\left|\frac{c}{\pm\sqrt{(a^2 + b^2)}}\right|$ is the length d of the perpendicular from the origin to the straight line.

Denoting the direction cosines of the above-mentioned vector by $\cos\alpha$ and $\cos\beta$, (α, β are the magnitudes of the angles which it makes with the positive parts of the axes x, y; see Theorem 5.5.1, p. 174), then

$$\cos\alpha = \frac{a}{\pm\sqrt{(a^2 + b^2)}}, \quad \cos\beta = \frac{b}{\pm\sqrt{(a^2 + b^2)}},$$

and the equation (1) can be rewritten in the form

$$x\cos\alpha + y\cos\beta - d = 0.$$

Theorem 1. *The distance d of a point $A(x_0, y_0)$ from a straight line $ax + by + c = 0$ is given by*

$$d = \left|\frac{ax_0 + by_0 + c}{\sqrt{(a^2 + b^2)}}\right|. \tag{2}$$

Theorem 2. *The equations of the bisectors of the angles between the two straight lines $a_1x + b_1y + c_1 = 0$ and $a_2x + b_2y + c_2 = 0$ are*

$$\frac{a_1x + b_1y + c_1}{\sqrt{(a_1^2 + b_1^2)}} + \frac{a_2x + b_2y + c_2}{\sqrt{(a_2^2 + b_2^2)}} = 0 \tag{3}$$

and

$$\frac{a_1x + b_1y + c_1}{\sqrt{(a_1^2 + b_1^2)}} - \frac{a_2x + b_2y + c_2}{\sqrt{(a_2^2 + b_2^2)}} = 0. \tag{3'}$$

REMARK 1. In order to decide which of the two bisectors (3), (3') passing through a given vertex of a triangle is the internal bisector of the angle, it is sufficient to find which one meets the opposite side of the triangle.

5.7. Polar Coordinates

The position of a point M may be determined by polar coordinates ϱ, φ: the coordinate ϱ is the distance of the point M from the origin (or *pole*) O, the coordinate φ is the directed angle between the segment OM and a fixed half-line p (with initial

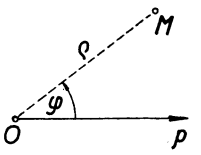


Fig. 5.9.

point O) called the *polar semi-axis* (or *initial-line* (Fig. 5.9.)). Here $\varrho \geqq 0$, $0 \leqq \varphi < 2\pi$. It is necessary to restrict the coordinates ϱ and φ in some way in order to establish a one-to-one correspondence between the points of a plane and the pairs of numbers (ϱ, φ) (with the exception of the pole), and we choose this particular

way. However, sometimes φ is not restricted to this, or even any interval (and occasionally even $\varrho < 0$ is used, especially in the equations of spirals etc.).

The relations between the cartesian and polar coordinates in the case where the pole is at the origin of the cartesian system and the polar semi-axis coincides with the positive part of the x-axis are:

$$x = \varrho \cos \varphi , \quad y = \varrho \sin \varphi ; \tag{1}$$

conversely

$$\begin{aligned}
&\varrho = \sqrt{(x^2 + y^2)}, \\
&\varphi = \arctan \frac{y}{x} && \text{for } x > 0, y > 0 ; \\
&\varphi = \tfrac{1}{2}\pi && \text{for } x = 0, y > 0 ; \\
&\varphi = \pi + \arctan \frac{y}{x} && \text{for } x < 0, y \text{ arbitrary} ; \\
&\varphi = \tfrac{3}{2}\pi && \text{for } x = 0, y < 0 ; \\
&\varphi = 2\pi + \arctan \frac{y}{x} && \text{for } x > 0, y < 0 .
\end{aligned} \tag{1'}$$

Example 1. The point M, the cartesian coordinates of which are $(-2, 1)$, has polar coordinates $(\sqrt{5}, 2{\cdot}68)$ (the angle being measured in radians). The point N, the polar coordinates of which are $(4, \frac{5}{3}\pi)$, has cartesian coordinates $(2, -2\sqrt{3})$ (Fig. 5.10).

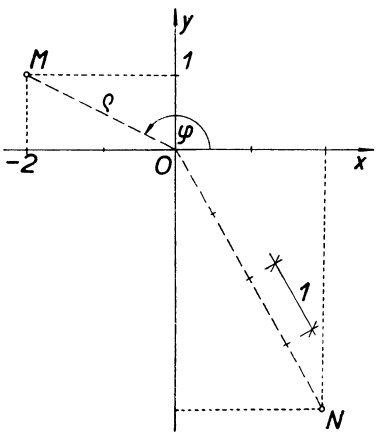


Fig. 5.10.

The equation of a curve in polar coordinates is — as in the case of cartesian coordinates — a relation which is satisfied by the coordinates of all the points of the curve (and only those points). The equations of some curves have a particularly simple form in polar coordinates.

Example 2. The cartesian equation of the circle whose centre is at the origin and whose radius is 7 is $x^2 + y^2 = 49$. If the coordinates are changed to polars using (1), the equation, after a small simplification, takes the form $\varrho = 7$, which is the equation of the same circle. (This is obvious geometrically.)

Example 3. If we choose as pole the focus of an ellipse, hyperbola, or parabola, and if that part of the focal axis of symmetry which does not contain the nearer vertex, be chosen as polar semi-axis, then all these curves have an equation of the form:

$$\varrho = \frac{p}{1 - e \cos \varphi}, \tag{2}$$

where e is the eccentricity (see §§ 5.10, 5.11, 5.12) and $2p$ is the *latus rectum* (i.e. the focal chord perpendicular to the focal axis of symmetry). In the case of the ellipse and hyperbola, $p = b^2/a$; in the case of the parabola, $e = 1$.

5.8. Parametric Equations of a Curve in a Plane

The equations $x = x(t)$, $y = y(t)$, where t is a variable *parameter*, are called the *parametric equations of a curve.*

Here $x(t)$, $y(t)$ are, as a rule, differentiable functions of t within an interval I. If t ranges over this interval, the point $M(x, y)$ moves along the curve. See Chap. 9.

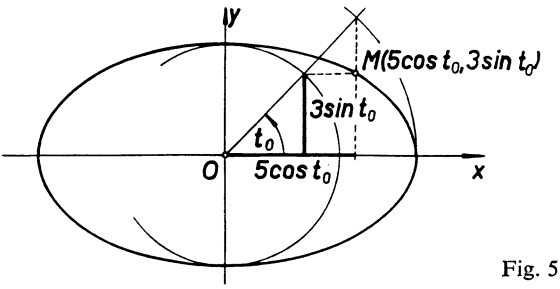


Fig. 5.11.

Example 1. a) The equations $x = 5 \cos t$, $y = 3 \sin t$ for $0 \leqq t < 2\pi$ are the parametric equations of an ellipse, the axes of which coincide with the coordinate axes, the lengths of the semi-axes being $a = 5$, $b = 3$ (Fig. 5.11).

b) If we eliminate t from these equations, we obtain the equation of the ellipse in cartesian coordinates (see §5.10):

$$\frac{x}{5} = \cos t, \quad \frac{y}{3} = \sin t;$$

hence

$$\left(\frac{x}{5}\right)^2 + \left(\frac{y}{3}\right)^2 = \cos^2 t + \sin^2 t = 1 ,$$

i.e.

$$\frac{x^2}{25} + \frac{y^2}{9} = 1 .$$

REMARK 1. Some curves can only be expressed in a simple form parametrically, (either we cannot eliminate t from the parametric equations or it is inconvenient to do so; for example, in the case of a cycloid $x = t - \sin t$, $y = 1 - \cos t$); some curves can be expressed simply in both ways. Obviously, we use whichever form is more convenient.

5.9. The Circle (see also § 4.1, p. 112)

Definition 1. A *circle* is the locus of a point $X(x, y)$ in a plane which moves so that its distance from a fixed point — the *centre* S — is constant.

Theorem 1. *The equation of the circle, whose centre in cartesian coordinates is* $S(x_0, y_0)$ *and whose radius is* r, *is*

$$(x - x_0)^2 + (y - y_0)^2 = r^2 . \tag{1}$$

Example 1. The circle with centre $S(-2, 1)$ and radius $r = 3$ has the equation

$$(x + 2)^2 + (y - 1)^2 = 9 .$$

If we remove the brackets in equation (1), we obtain an equation of the form

$$x^2 + y^2 + mx + ny + p = 0. \tag{2}$$

If we want to obtain an equation of the form (1) from equation (2), we "complete the squares" on the left-hand side of equation (2) and obtain

$$\left(x + \frac{m}{2}\right)^2 + \left(y + \frac{n}{2}\right)^2 = \frac{m^2}{4} + \frac{n^2}{4} - p . \tag{3}$$

Comparing this with equation (1) we can see that the expression on the right-hand side of equation (3) must be positive in order to get a real circle; also $(-\frac{1}{2}m, -\frac{1}{2}n)$ are the coordinates of the centre of this circle.

Theorem 2. *The parametric equations of a circle are*

$$x = x_0 + r \cos t ,$$
$$y = y_0 + r \sin t ,$$

where the point $S(x_0, y_0)$ is the centre, r is the radius, (x, y) are the coordinates of a general point X on the circle and t $(0 \leqq t < 2\pi)$ is a variable parameter, the geometrical significance of which is that it is the angle formed by the half-line SX and the positive semi-axis $+x$.

Theorem 3. *The equation of a circle of radius a in polar coordinates is $\varrho = a$, if S coincides with O* (see Example 5.7.2, p. 180), *and $\varrho = 2a \cos \varphi \left(-\frac{1}{2}\pi < \varphi \leqq \frac{1}{2}\pi\right)$, if S lies on the polar semi-axis and the circle passes through the pole.*

Example 2. Find the coordinates of the points of intersection P_1, P_2 of the line

$$4x - 3y + 4 = 0 \tag{4}$$

with the circle whose centre is at the point (2, 4) and whose radius r is 5.

The equation of the circle is, by (1),

$$(x - 2)^2 + (y - 4)^2 = 25 . \tag{5}$$

The coordinates of the common points of the line and the circle satisfy simultaneously equations (4) and (5); hence, they are given by solving the equations (4) and (5). From (4), it follows that

$$y = \tfrac{4}{3}(x + 1) . \tag{6}$$

Substituting (6) in (5), we obtain the quadratic equation

$$25(x^2 - 4x + 4) = 225 ,$$

i.e.

$$x^2 - 4x - 5 = 0 ,$$

for the x-coordinates of the points of intersection, the roots of this equation being

$$x_1 = 5 , \quad x_2 = -1 . \tag{7}$$

The corresponding values for y_1, y_2 are found by substituting (7) into (6) (not into (5)!):

$$y_1 = 8 , \quad y_2 = 0 .$$

The required points of intersection are $P_1(5, 8)$, $P_2(-1, 0)$.

REMARK 1. The problem of finding the points of intersection of a straight line and a circle reduces therefore to the solution of a quadratic equation. If this equation possesses two real roots, or a double root, or two conjugate complex roots, then the straight line is a secant (chord) of the circle, or a tangent to the circle, or it does not intersect the circle at all, respectively.

We proceed in the same way (and the same conclusion holds) when finding the points of intersection of a straight line and other conics. The only exceptions are the lines parallel to the axis of a parabola and to the asymptotes of a hyperbola.

5.10. The Ellipse (see also § 4.2, p. 114)

Definition 1. An *ellipse* is the locus of a point $X(x, y)$ which moves in a plane such that the sum of its distances from two fixed points $F_1(x_1, y_1)$ and $F_2(x_2, y_2)$ – the *foci* – is equal to a constant which is usually denoted by $2a$ if the foci both lie on the x-axis.

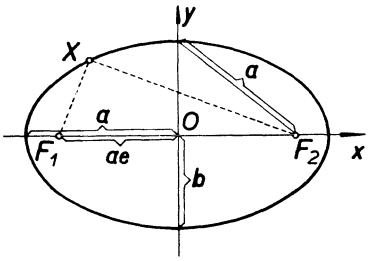


Fig. 5.12.

Clearly $\overline{F_1F_2} < 2a$ (i.e. $\overline{F_1F_2} = 2ae$ where e is some positive number less than unity).

Definition 2. The number e is called the *eccentricity of the ellipse.*

Theorem 1. *The standard equation of an ellipse (for the case where the axes of the ellipse coincide with the coordinate axes, the foci lying on the x-axis* (Fig. 5.12)) *is*

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1\,, \tag{1}$$

where $b^2 = a^2(1 - e^2)$.

In fact a is the length of the semi-major axis and b the length of the semi-minor axis of the ellipse.

REMARK 1. If the axes of an ellipse are parallel to the cooordinate axes and if the centre is at the point $S(x_0, y_0)$, then (1) becomes

$$\frac{(x - x_0)^2}{a^2} + \frac{(y - y_0)^2}{b^2} = 1\,. \tag{1'}$$

REMARK 2. If the foci lie on the y-axis the sum of the focal distances is denoted by $2b$, $\overline{F_1F_2} = 2be$ and a is now defined by $a^2 = b^2(1 - e^2)$. The equation is the same as (1).

For the equation of an ellipse in polar coordinates see Example 5.7.3, p. 180; for parametric equations of an ellipse see Example 5.8.1, p. 180.

5.11. The Hyperbola (see also § 4.3, p. 119)

Definition 1. A hyperbola is the locus of a point $X(x, y)$ which moves in a plane such that the difference of its distances from two fixed points $F_1(x_1, y_1)$ and $F_2(x_2, y_2)$ — the foci — is in absolute value equal to a constant which is usually denoted by $2a$ if the foci both lie on the x-axis.

Clearly $\overline{F_1F_2} > 2a$ (i.e. $\overline{F_1F_2} = 2ae$ where e is some number greater than unity).

Definition 2. The number e is called the *eccentricity of the hyperbola.*

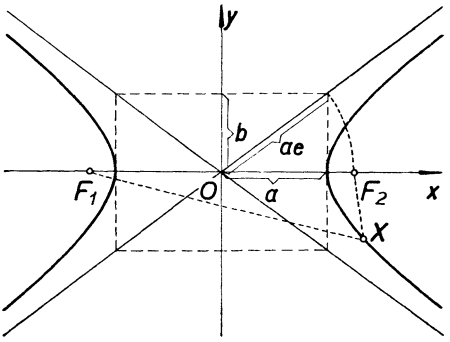


Fig. 5.13.

Theorem 1. *The standard equation of a hyperbola (for the case where the axes of the hyperbola coincide with the coordinate axes, the foci lying on the x-axis* (Fig. 5.13)) *is*

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1\,, \tag{1}$$

where $b^2 = a^2(e^2 - 1)$.

In fact a is the length of the (real) semi-major axis and b the length of the (imaginary) semi-minor axis of the hyperbola.

REMARK 1. If the foci lie on the y-axis the absolute value of the difference of the focal distances is denoted by $2b$, $F_1F_2 = 2be$ and a is now defined by $a^2 = b^2(e^2 - 1)$. Equation (1) becomes

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = -1\,. \tag{1'}$$

REMARK 2. If the axes of the hyperbola are parallel to the coordinate axes and the centre is at the point $S(x_0, y_0)$, then (1) and (1') become

$$\frac{(x - x_0)^2}{a^2} - \frac{(y - y_0)^2}{b^2} = \pm 1\,. \tag{1''}$$

Theorem 2. *The lines* $y = \pm bx/a$ *are the asymptotes of the hyperbolas* (1) *and* (1'). *Their equations can be combined thus:*

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 0 .$$

REMARK 3. If $a = b$ the hyperbola is called *rectangular*; its equation is $x^2 - y^2 = a^2$ (or $y^2 - x^2 = a^2$), the equations of its asymptotes being $y = \pm x$.

REMARK 4. The hyperbolas $x^2/a^2 - y^2/b^2 = 1$ and $y^2/b^2 - x^2/a^2 = 1$ are called *conjugate*. They have the same asymptotes; the first of these hyperbolas has real points of intersection with the x-axis, the second one with the y-axis.

5.12. The Parabola (see also § 4.4, p. 122)

Definition 1. A *parabola* is the locus of a point $X(x, y)$ in a plane, equidistant from a fixed point $F(x_1, y_1)$ – the *focus* – and from a fixed line d – the *directrix*.

Theorem 1. *The parabola whose vertex is at the origin of the coordinate system* (Fig. 5.14) *and whose axis coincides with the x-axis, or y-axis, has the cartesian equation*

$$y^2 = 2px , \quad \text{or} \quad x^2 = 2py , \quad \textit{respectively} . \tag{1}$$

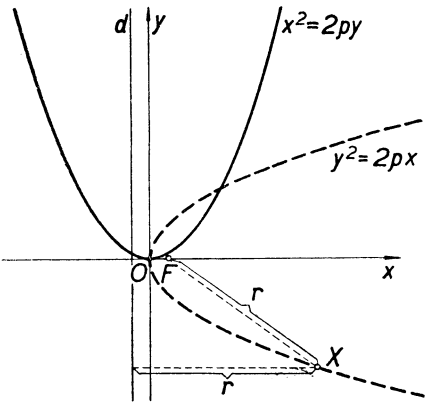


Fig. 5.14.

Theorem 2. *The parabola* $y^2 = 2px$ *has the focus* $F(\frac{1}{2}p, 0)$, *and the directrix* $x = -\frac{1}{2}p$. *The parabola* $x^2 = 2py$ *has the focus* $F_1(0, \frac{1}{2}p)$ *and the directrix* $y = -\frac{1}{2}p$.

(In the English literature the standard form of the equation of a parabola is generally taken as $y^2 = 4ax$, so that the focus is $(a, 0)$ and the directrix is $x = -a$.)

REMARK 1. If the axis of the parabola is parallel to the x-axis, or to the y-axis and its vertex is at the point $V(x_0, y_0)$, equations (1) become

$$(y - y_0)^2 = 2p(x - x_0), \quad \text{or} \quad (x - x_0)^2 = 2p(y - y_0), \quad \text{respectively}. \tag{2}$$

5.13. Congruent Transformations of Cartesian Coordinates in a Plane

We assume that all the different cartesian coordinate systems considered in this paragraph have the same unit of length.

Theorem 1. *Any change from one cartesian coordinate system to another can be performed by one translation and one rotation provided that both systems have the same orientation (i.e. both are right-handed or both left-handed).*

Theorem 2. *The transformation of the coordinates of a point X when passing from the cartesian system $(O; x, y)$ to the similarly oriented cartesian system $(O'; x', y')$, in which the axes x', y', are parallel to the axes x, y, respectively, is given by the formulae:*

$$\begin{aligned} x' &= x - m, \\ y' &= y - n, \end{aligned} \tag{1}$$

where (m, n) are the coordinates of the new origin O' in the original system $(O; x, y)$.

Conversely:

$$\begin{aligned} x &= x' + m, \\ y &= y' + n. \end{aligned} \tag{1'}$$

Theorem 3. *The transformation of the coordinates of a point X when passing from the cartesian system $(O; x, y)$ to the similarly oriented cartesian system $(O; x', y')$ which is obtained from $(O; x, y)$ by a rotation about the common origin through an angle α, is given by the formulae:*

$$\begin{aligned} x' &= x \cos\alpha + y \sin\alpha, \\ y' &= -x \sin\alpha + y \cos\alpha. \end{aligned} \tag{2}$$

Conversely:

$$\begin{aligned} x &= x' \cos\alpha - y' \sin\alpha, \\ y &= x' \sin\alpha + y' \cos\alpha. \end{aligned} \tag{2'}$$

Theorem 4. *In any change from a cartesian coordinate system $(O; x, y)$ to a similarly oriented cartesian coordinate system $(O'; x', y')$ the coordinates are trans-*

formed according to the formulae:

$$\begin{aligned} x' &= \quad x\cos\alpha + y\sin\alpha - m', \\ y' &= -x\sin\alpha + y\cos\alpha - n'. \end{aligned} \tag{3}$$

Conversely:

$$\begin{aligned} x &= x'\cos\alpha - y'\sin\alpha + m, \\ y &= x'\sin\alpha + y'\cos\alpha + n. \end{aligned} \tag{3'}$$

The numbers m, n are the coordinates of the origin O' in the system $(O; x, y)$, $m' = m\cos\alpha + n\sin\alpha$, $n' = -m\sin\alpha + n\cos\alpha$, and α is the angle of rotation of the coordinate axes.

REMARK 1. The determinant consisting of the coefficients of x, y in equations (2) or (3) equals unity. This fact is characteristic of congruent transformations of cartesian coordinates in a plane; it is also a sufficient condition for the equations (2) or (3) to be solvable in x, y (to give the inverse transformations (2') and (3'); the derivation of (1') from (1) is obvious).

REMARK 2. Equations (1), (2), (3) represent the relationship between the coordinates of a fixed point in a plane with respect to two different coordinate systems; but they may also be interpreted as showing the relationship between the coordinates of two different points in a plane with respect to the same coordinate system.

5.14. Homogeneous Coordinates

Definition 1. The three ordered numbers (ξ_0, ξ_1, ξ_2) $(\xi_0 \neq 0)$ are called the *rectangular homogeneous coordinates* of a point M in a plane, if $\xi_1/\xi_0 = x$, $\xi_2/\xi_0 = y$, where (x, y) *are the cartesian coordinates of the point M. We write $M(\xi_0, \xi_1, \xi_2)$.* [The coordinate ξ_0 is frequently placed last in the group of these numbers viz. (ξ_1, ξ_2, ξ_0)].

Theorem 1. *If we use homogeneous coordinates then the equations of algebraic curves in a plane are homogeneous.*

Example 1. If we transform the cartesian equation of a line $ax + by + c = 0$ by means of the formulae in Definition 1, we obtain the linear homogeneous equation:

$$a\xi_1 + b\xi_2 + c\xi_0 = 0. \tag{1}$$

REMARK 1. In contrast to the case of the general equation of a line in cartesian coordinates, the numbers a and b in equation (1) may both be equal to zero, so that the equation can be of the form

$$\xi_0 = 0;$$

this is the equation of the so-called *line at infinity* (*improper line*), by which the plane has been extended due to the introduction of homogeneous coordinates.

Example 2. The equation of an ellipse

$$\frac{(x-\bar{x})^2}{a^2}+\frac{(y-\bar{y})^2}{b^2}=1$$

[whose centre is at the point $(\bar{x}, \bar{y})$] becomes, on substitution of homogeneous coordinates and after some simplification:

$$(a^2\bar{\xi}_2^2+b^2\bar{\xi}_1^2)\,\xi_0^2+b^2\bar{\xi}_0^2\xi_1^2+a^2\bar{\xi}_0^2\xi_2^2-2b^2\bar{\xi}_0\bar{\xi}_1\xi_0\xi_1-2a^2\bar{\xi}_0\bar{\xi}_2\xi_0\xi_2= \\ =a^2b^2\bar{\xi}_0^2\xi_0^2\,;$$

this is a homogeneous equation of the second degree in the variables ξ_0, ξ_1, ξ_2.

5.15. General Equation of a Conic

Theorem 1. *The general equation of a curve of the second degree in cartesian coordinates is*

$$a_{11}x^2+2a_{12}xy+a_{22}y^2+2a_{13}x+2a_{23}y+a_{33}=0\,; \tag{1}$$

this equation may represent an ellipse (*a circle as a special case*), *a hyperbola, a parabola, a pair of straight lines* (*which may coincide*), *a point, or it may not be satisfied by any* (*real*) *point at all.*

Dafinition 1. Let us form two determinants from the coefficients of equation (1):

$$\Delta=\begin{vmatrix} a_{11}, & a_{12}, & a_{13} \\ a_{12}, & a_{22}, & a_{23} \\ a_{13}, & a_{23}, & a_{33} \end{vmatrix}, \qquad \delta=\begin{vmatrix} a_{11}, & a_{12} \\ a_{12}, & a_{22} \end{vmatrix};$$

the number Δ is called the *discriminant* of the conic section (1), δ – the *discriminant of the quadratic members.*

Theorem 2. *The curves of the second degree can be classified in terms of* Δ *and* δ, *as shown in Tab.* 5.1, p. 189.

Theorem 3. *By means of a rotation of the coordinate system through an angle* φ *it is possible to make the axes* (*or the axis*) *of a regular conic parallel to the coordinate axes; the angle* φ *can be found from the relation*

$$\tan 2\varphi=\frac{2a_{12}}{a_{11}-a_{22}}\,. \tag{2}$$

If $a_{11}=a_{22}$ *we can choose* $\varphi=\frac{1}{4}\pi$ (see Example 5.17.1, p. 193).

TABLE 5.1

	Regular (non-singular) conic sections ($\Delta \neq 0$)	Singular conic sections ($\Delta = 0$)
$\delta > 0$	an ellipse (real or imaginary)	two imaginary lines with a real point of intersection
$\delta < 0$	a hyperbola	two intersecting lines
$\delta = 0$	a parabola	two parallel lines (real or imaginary, different or coincident)

Theorem 4. *If the position of a regular conic is such that its axes are parallel to the coordinate axes (or its axis is parallel to one of the coordinate axes), then its equation does not contain the term involving xy, i.e. $a_{12} = 0$ (and conversely). The nature of the conic can then be determined as follows:*

a) $a_{11}a_{22} > 0$ — *an ellipse (a circle if $a_{11} = a_{22}$)*,
b) $a_{11}a_{22} < 0$ — *a hyperbola (a rectangular hyperbola if $a_{11} = -a_{22}$)*,
c) $a_{11}a_{22} = 0$ — *a parabola*.

REMARK 1. In the case mentioned in Theorem 4, i.e. when the equation does not contain any xy term, we can easily find the type of the conic section and at the same time find its centre (or vertex) and semi-axes by the method of "completing the squares", as in the case of the circle (Equation (5.9.3), p. 181).

5.16. Affine and Projective Transformations

Definition 1. *The affine (position) ratio* of a point M on a straight line with respect to two base points P, Q of the line is the ratio of the distances of the point M from the two points P, Q; if M is an inner point of the line-segment PQ, the ratio is negative, if it is an external point, the ratio is positive. We use the notation

$$(PQM) = \frac{PM}{QM} \quad (M \neq Q).$$

Definition 2. *The cross ratio of four points P, Q, M, N* on a line (the order in which they are written is important) is the quotient of the affine ratios of the points M and N

with respect to P and Q. We write

$$(PQMN) = \frac{(PQM)}{(PQN)} = \frac{PM}{QM} \cdot \frac{QN}{PN} \quad (M \neq Q, N \neq P).$$

Definition 3. By an *affine transformation* of a plane we mean a transformation which carries the point $M(x, y)$ into the point $M'(x', y')$ according to the equations

$$\begin{aligned} x' &= a_1 x + b_1 y + c_1 , \\ y' &= a_2 x + b_2 y + c_2 \end{aligned} \tag{1}$$

where

$$\begin{vmatrix} a_1, & b_1 \\ a_2, & b_2 \end{vmatrix} \neq 0 .$$

Theorem 1. *An affine transformation preserves the affine ratio of a point on a line with respect to any two points on the line; the line at infinity is transformed into itself (i.e. parallelism is preserved).*

Theorem 2. *An affine transformation of a plane into which a homogeneous coordinate system is introduced, is given by the equations:*

$$\begin{aligned} \xi_0' &= a_0 \xi_0 , \\ \xi_1' &= a_1 \xi_0 + b_1 \xi_1 + c_1 \xi_2 , \\ \xi_2' &= a_2 \xi_0 + b_2 \xi_1 + c_2 \xi_2 , \end{aligned} \tag{2}$$

where

$$\begin{vmatrix} a_0, & 0, & 0 \\ a_1, & b_1, & c_1 \\ a_2, & b_2, & c_2 \end{vmatrix} \neq 0 .$$

Theorem 3. *Every congruent transformation is a particular case of an affine transformation.*

Theorem 4. *By an affine transformation, a conic section is transformed into a conic section of the same type, i.e. an ellipse into an ellipse, a hyperbola into a hyperbola and a parabola into a parabola.*

Definition 4. By a *projective transformation* of a plane we mean a transformation which carries the point $M(x, y)$ into the point $M'(x', y')$ according to the equations:

$$\begin{aligned} x' &= \frac{a_{11} x + a_{12} y + a_{13}}{a_{31} x + a_{32} y + a_{33}} , \\ y' &= \frac{a_{21} x + a_{22} y + a_{23}}{a_{31} x + a_{32} y + a_{33}} , \end{aligned} \tag{3}$$

where

$$\begin{vmatrix} a_{11}, & a_{12}, & a_{13} \\ a_{21}, & a_{22}, & a_{23} \\ a_{31}, & a_{32}, & a_{33} \end{vmatrix} \neq 0 .$$

Theorem 5. *A projective transformation preserves the cross ratio of any four points on a line.*

Theorem 6. *A projective transformation of a plane into which a homogeneous coordinate system is introduced is given by the equations:*

$$\begin{aligned} \xi_0' &= a_{11}\xi_0 + a_{12}\xi_1 + a_{13}\xi_2 , \\ \xi_1' &= a_{21}\xi_0 + a_{22}\xi_1 + a_{23}\xi_2 , \\ \xi_2' &= a_{31}\xi_0 + a_{32}\xi_1 + a_{33}\xi_2 , \end{aligned} \tag{4}$$

where

$$\begin{vmatrix} a_{11}, & a_{12}, & a_{13} \\ a_{21}, & a_{22}, & a_{23} \\ a_{31}, & a_{32}, & a_{33} \end{vmatrix} \neq 0 .$$

Theorem 7. *Every affine transformation is a particular case of a projective transformation.*

Theorem 8. *By a projective transformation a regular conic section is transformed into a regular conic section (not necessarily of the same type), a singular conic section is transformed into a singular conic section (of the same type in the projective sense; i.e. the properties of being real, imaginary, distinct or coincident are preserved).*

REMARK 1. Since the determinants of the systems (1)–(4) are different from zero, the undashed coordinates can be expressed by means of the dashed coordinates in each of the systems, i.e. there exists an inverse transformation for each of the transformations under consideration.

5.17. Pole, Polar, Centre, Conjugate Diameters and Tangents of a Conic Section

Definition 1. If the cross ratio of four points A, B, C, D is equal to -1, i.e. $(ABCD) = -1$, we say that these points form a *harmonic set (range)*.

Theorem 1. *Let us consider a pencil of lines passing through a point P chosen in the plane of a regular conic, the individual lines intersecting the conic in pairs of points M_1, N_1; M_2, N_2 etc.* (Fig. 5.15). *Then, the locus of a point Q_i, which forms a harmonic set with the point P and the points M_i, N_i on every line of the pencil*

(i.e. $M_iN_iPQ_i) = -1$), *is a straight line p called the polar of the point P with respect to the conic. The point P is called the pole of the line p with respect to the conic.*

Theorem 2. *The equation of the polar p of a point* $P(x_0, y_0)$ *with respect to the regular conic*

$$a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + 2a_{13}x + 2a_{23}y + a_{33} = 0 \tag{1}$$

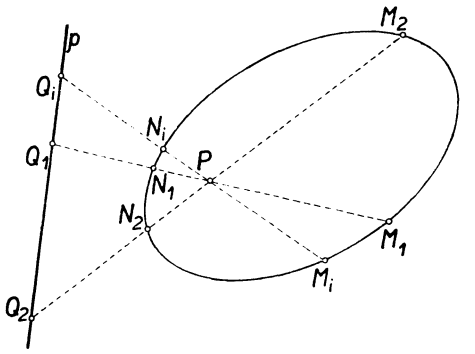


Fig. 5.15.

is

$$(a_{11}x_0 + a_{12}y_0 + a_{13})\,x + (a_{12}x_0 + a_{22}y_0 + a_{23})\,y + (a_{13}x_0 + a_{23}y_0 + a_{33}) = 0\,. \tag{2}$$

Theorem 3. *The polar of a point* $T(x_0, y_0)$ *of a regular conic with respect to this conic passes through the point T and is the tangent to the conic at this point. Its equation is*

$$\frac{xx_0}{a^2} + \frac{yy_0}{b^2} = 1 \quad \textit{for an ellipse whose equation is in standard form},$$

$$\frac{xx_0}{a^2} - \frac{yy_0}{b^2} = 1 \quad \textit{for a hyperbola whose equation is in standard form},$$

$$yy_0 = p(x + x_0) \quad \textit{for the parabola } y^2 = 2px\,.$$

Theorem 4. *The tangents from a point P to a regular conic (if they exist) pass through the points of intersection of the polar of P and the conic.*

Theorem 5. *The mid-points of all parallel chords of a regular conic lie on a line.*

Definition 2. The line on which all the mid-points of parallel chords of a regular conic lie is called a *diameter of the conic.*

Definition 3. The common *direction* of parallel chords of a regular conic is said to be *conjugate to the direction* of the diameter which passes through the mid-points of those chords.

Theorem 6. a) *All diameters of a parabola are parallel.*

b) *All diameters of an ellipse (or hyperbola) pass through a common point called the centre of the ellipse* (or *hyperbola*).

Theorem 7. *The coordinates of the centre of the conic* (1) *are given by the solution of the equations*:

$$\begin{aligned} a_{11}x + a_{12}y + a_{13} &= 0\,, \\ a_{12}x + a_{22}y + a_{23} &= 0\,; \end{aligned} \tag{3}$$

(*the left-hand sides of equations* (3) *are in fact half the partial derivatives of the left-hand side of equation* (1) *with respect to x and y, respectively.*)

Theorem 8. *If the direction* (s_1) *is conjugate to the direction* (s_2) *with respect to a regular central conic, then the direction* (s_2) *is conjugate to the direction* (s_1) *with respect to the same conic.* Thus, such directions are called *conjugate directions with respect to the conic.*

Definition 4. Two diameters of a regular conic whose directions are conjugate with respect to this conic are said to be *conjugate diameters* of this conic.

Definition 5. The two conjugate diameters of a central conic which are perpendicular are called the *axes of the conic.*

Theorem 9. *The slopes* k_1, k_2 *of conjugate directions satisfy the relation*

$$a_{11} + a_{12}(k_1 + k_2) + a_{22}k_1k_2 = 0\,, \tag{4}$$

i.e.

$$k_1k_2 = -\frac{b^2}{a^2} \quad \textit{for an ellipse whose equation is in standard form}\,,$$

$$k_1k_2 = \frac{b^2}{a^2} \quad \textit{for a hyperbola whose equation is in standard form}\,.$$

Theorem 10. *The equation of the diameter conjugate to the direction whose slope is k, is*

$$a_{11}x + a_{12}y + a_{13} + k(a_{12}x + a_{22}y + a_{23}) = 0\,, \tag{5}$$

i.e.

$$y = -\frac{b^2}{a^2k}x \quad \textit{for an ellipse whose equation is in standard form}\,,$$

$$y = \frac{b^2}{a^2k}x \quad \textit{for a hyperbola whose equation is in standard form}\,,$$

$$y = \frac{p}{k} \quad \textit{for the parabola } y^2 = 2px\,.$$

Example 1. Let us investigate the curve of the second degree which is given by the equation

$$3x^2 - 2xy + 3y^2 + 4x + 4y - 4 = 0 \tag{6}$$

and draw the tangents from the point $P(3, 1)$ to it.

Solution: Since

$$\Delta = \begin{vmatrix} 3, & -1, & 2 \\ -1, & 3, & 2 \\ 2, & 2, & -4 \end{vmatrix} = -64 \neq 0\,, \quad \delta = \begin{vmatrix} 3, & -1 \\ -1, & 3 \end{vmatrix} = 8 > 0\,,$$

either the curve (6) is an ellipse or it contains no real points (see Theorem 5.15.2). Further, from the formula (5.15.2) we can see that $\varphi = \frac{1}{4}\pi$; thus, $\sin\varphi = \cos\varphi = \frac{1}{2}\sqrt{2}$ so that by successive substitution into equations (5.13.2′) and from them into (6), we obtain the equation

$$x'^2 + 2y'^2 + 2\sqrt{(2)}\,x' - 2 = 0\,,$$

i.e.

$$\frac{(x' + \sqrt{2})^2}{4} + \frac{y'^2}{2} = 1\,.$$

Hence, (6) is the ellipse whose centre is at the point $(-\sqrt{2}, 0)$ and whose semi-axes a (of length 2) and b (of length $\sqrt{2}$) make an angle $\varphi = \frac{1}{4}\pi$ with the coordinate axes. The coordinates of the centre S are expressed in the transformed coordinates; its coordinates in the original system can be found, for instance, by equations (3):

$$3x - y + 2 = 0\,,$$
$$-x + 3y + 2 = 0\,;$$

hence $x_0 = -1$, $y_0 = -1$, and so S is the point $(-1, -1)$.

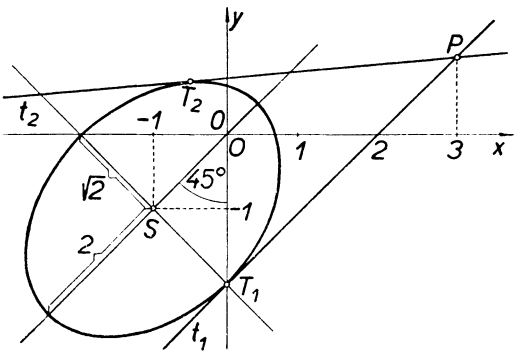


Fig. 5.16.

In order to find the equation of a tangent from the point P to the ellipse, we first find its polar; by (2) the equation of the polar is $5x + y + 2 = 0$. The coordinates of the points of contact T_1, T_2 are given by the simultaneous solution of this equation and equation (6): $T_1(0, -2)$, $T_2(-\frac{6}{11}, \frac{8}{11})$. Then, by means of the coordinates of P, T_1 and T_2, we can easily find the equations of the tangent lines: $x - y - 2 = 0$, $x - 13y + 10 = 0$.

6. SOLID ANALYTIC GEOMETRY

By FRANTIŠEK KEJLA

References: [14], [27], [59], [62], [67], [124], [163].

6.1. Coordinate Systems

The position of an arbitrary point in three dimensional space is usually determined so that, to each point of the space an ordered triplet of real numbers (called *coordinates*) is assigned and conversely, to each ordered triplet of real numbers there corresponds a certain unique point of the space.

Definition 1. Surfaces consisting of those points which have one particular coordinate constant are called *coordinate surfaces.*

Various coordinate systems can be established, the most important being:

a) Rectangular coordinate system. This system is introduced in a manner similar to that for a rectangular coordinate system in a plane (see Remark 5.1.2, p. 167): We choose three mutually perpendicular directed lines x, y, z in space passing through a common point O and three units of length, one on each line. The directed distances x, y, z (cf. Remark 5.1.2, p. 167) of an arbitrary point M from the planes yz, xz, xy respectively, are called the *rectangular coordinates* of the point M. We write $M(x, y, z)$ to denote this. If the units of length are identical, the coordinates are called *cartesian.* In what follows we shall always be referring to these coordinates (unless otherwise stated).

The point $M_1(x, y, 0)$ is the orthogonal projection (top view) of the point M onto the plane xy, so that x, y are the cartesian coordinates of the projection of the point M in the system $(O; x, y)$ in the sense of plane analytical geometry (see Remark 5.1.2, p. 167). The situation is similar for the point $M_2(0, y, z)$ – front view of the point M – and for the point $M_3(x, 0, z)$ – side view of the point M.

The lines x, y, z are called the *coordinate axes*, the point O – the *origin of the coordinate system*, and the planes yz, xz, xy – the *coordinate planes.*

Definition 2. If when viewed from an arbitrary point on the positive semi-axis $+z$, the positive semi-axis $+x$ is carried by counter-clockwise rotation through a right

angle into the positive semi-axis $+y$, the coordinate system $(O; x, y, z)$ is said to be *positively oriented* (*right-handed*). Otherwise the system is said to be negatively *oriented* (*left-handed*) – see Fig. 6.1a,b.

Theorem 1. *The coordinate surfaces in a cartesian coordinate system are planes parallel to the coordinate planes* (*perpendicular to the corresponding coordinate axes*).

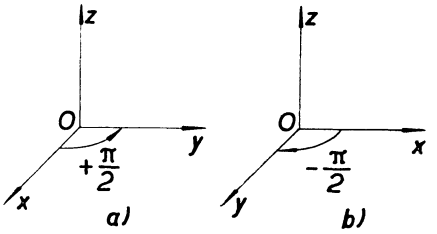


Fig. 6.1.

b) Cylindrical (semi-polar) coordinate system. This system is determined by a coordinate plane xy into which polar coordinates ϱ, φ are introduced (see p. 178) and by a directed z-axis, passing through the pole of the system, perpendicular to the plane. An arbitrary point M is determined by an ordered triplet of numbers (ϱ, φ, z), where ϱ, φ are the polar coordinates of the orthogonal projection M_1 of the point M onto the plane xy and z is the directed distance of the point M from the plane xy (see Fig. 6.2).

Theorem 2. *The coordinate surfaces in a cylindrical system are*

a) *half-planes passing through the z-axis* (φ = const.);

b) *cylinders of revolution with axis coinciding with the z-axis* (ϱ = const.; $\varrho = 0$ corresponds to the z-axis);

c) *planes perpendicular to the z-axis* (z = const.).

c) Spherical (polar) coordinate system. This system is determined by a coordinate plane xy into which polar coordinates ϱ, φ are introduced (see p. 178) and by a directed half-line, passing through the pole of the system perpendicular to the plane. The following coordinates determine the position of an arbitrary point M in this coordinate system:

a) the distance r of the point M from the origin (pole) O of the system;

b) the magnitude φ of the angle between the half-line OM_1 and the positive semi-axis $+x$, where M_1 is the orthogonal projection of the point M onto the plane xy;

c) the magnitude ϑ of the angle between the half-line OM and the positive semi-axis $+z$ (the axes x, y, z are at the same time the axes of a cartesian coordinate system – the so-called *adjoined system*; see Fig. 6.3).

The coordinate r is never negative ($r \geqq 0$), the coordinate φ ranges over the interval $[0, 2\pi)$, the coordinate ϑ ranges over the interval $[0, \pi]$. (Sometimes the interval $(-\pi, \pi]$ is used for φ.)

Theorem 3. *The coordinate surfaces in a spherical system are*

a) *spheres whose centres are at the pole of the system* ($r = \text{const.}$);
b) *half-planes passing through the z-axis* ($\varphi = \text{const.}$);
c) *cones* (or *half-cones*, more precisely) *of revolution, the vertices of which are at the pole and the axes of which coincide with the z-axis* ($\vartheta = \text{const.}$).

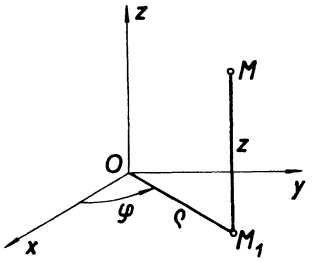


Fig. 6.2.

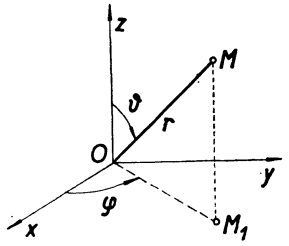


Fig. 6.3.

In particular, $r = 0$ gives one point (the pole), $\vartheta = \frac{1}{2}\pi$ – the plane xy, $\vartheta = 0$, π – the half-lines $+z$, $-z$, respectively.

Theorem 4. *The cartesian coordinates x, y, z, the cylindrical coordinates* ϱ, φ, z *and the spherical coordinates r,* ϑ, φ *of the same point satisfy the following relations*:

a) $x = \varrho \cos \varphi$, $y = \varrho \sin \varphi$, $z = z$,

$$\varrho = \sqrt{(x^2 + y^2)}, \quad \sin \varphi = \frac{y}{\sqrt{(x^2 + y^2)}}, \quad \cos \varphi = \frac{x}{\sqrt{(x^2 + y^2)}}, \quad \tan \varphi = \frac{y}{x};$$

b) $x = r \sin \vartheta \cos \varphi$, $y = r \sin \vartheta \sin \varphi$, $z = r \cos \vartheta$,

$$r = \sqrt{(x^2 + y^2 + z^2)}, \quad \sin \vartheta = \sqrt{\frac{x^2 + y^2}{x^2 + y^2 + z^2}},$$

$$\cos \vartheta = \frac{z}{\sqrt{(x^2 + y^2 + z^2)}}, \quad \tan \vartheta = \frac{\sqrt{(x^2 + y^2)}}{z},$$

$$\sin \varphi = \frac{y}{\sqrt{(x^2 + y^2)}}, \quad \cos \varphi = \frac{x}{\sqrt{(x^2 + y^2)}}, \quad \tan \varphi = \frac{y}{x}.$$

REMARK 1. The correspondence between the sets of coordinates and the points themselves is one-to-one without exception, only in the case of a rectangular system.

It is not so in the other two systems. All points on the z-axis are so-called *singular points* of those systems – the coordinate φ may be chosen quite arbitrarily; however these points are uniquely determined by their remaining two coordinates. In calculations though, care is sometimes required.

Theorem 5 (*Transformation of a Cartesian Coordinate System*).

a) Translation. *If x, y, z denote the coordinates in the original system, X, Y, Z – the coordinates of the same point in the new system, x_0, y_0, z_0 – the coordinates of the new origin in the original system, then*

$$X = x - x_0\,, \quad Y = y - y_0\,, \quad Z = z - z_0\,.$$

b) Rotation and reflection. *If the cosines of the angles formed by the new axes X, Y, Z and the original axes x, y, z are as shown in the following scheme*

	X	Y	Z
x	a_1	a_2	a_3
y	b_1	b_2	b_3
z	c_1	c_2	c_3

then the following relations hold between the original and new coordinates of the same point:

$$x = a_1X + a_2Y + a_3Z\,, \quad X = a_1x + b_1y + c_1z\,,$$
$$y = b_1X + b_2Y + b_3Z\,, \quad Y = a_2x + b_2y + c_2z\,,$$
$$z = c_1X + c_2Y + c_3Z\,; \quad Z = a_3x + b_3y + c_3z\,.$$

Theorem 6. *The cosines listed in the table above satisfy the relations:*

$$a_1^2 + a_2^2 + a_3^2 = 1\,, \quad a_1^2 + b_1^2 + c_1^2 = 1\,,$$
$$b_1^2 + b_2^2 + b_3^2 = 1\,, \quad a_2^2 + b_2^2 + c_2^2 = 1\,,$$
$$c_1^2 + c_2^2 + c_3^2 = 1\,; \quad a_3^2 + b_3^2 + c_3^2 = 1\,;$$

$$a_1b_1 + a_2b_2 + a_3b_3 = 0\,, \quad a_1a_2 + b_1b_2 + c_1c_2 = 0\,,$$
$$a_1c_1 + a_2c_2 + a_3c_3 = 0\,, \quad a_1a_3 + b_1b_3 + c_1c_3 = 0\,,$$
$$b_1c_1 + b_2c_2 + b_3c_3 = 0\,; \quad a_2a_3 + b_2b_3 + c_2c_3 = 0\,.$$

Theorem 7.

$$\Delta = \begin{vmatrix} a_1, & a_2, & a_3 \\ b_1, & b_2, & b_3 \\ c_1, & c_2, & c_3 \end{vmatrix} = \pm 1 \quad \text{(the so-called } \textit{determinant of the transformation}\text{)},$$

where the $\begin{Bmatrix} upper \\ lower \end{Bmatrix}$ *sign is valid according as the two systems have* $\begin{Bmatrix} the\ same \\ different \end{Bmatrix}$ *orientations* (see Definition 2).

If the orientation of the two systems is the same, then each element of the determinant above is equal to its complement (cofactor). In the case of different orientations of the systems each element of the determinant of transformation is equal to minus its complement.

6.2. Linear Concepts

In § 6.2 we confine ourselves to relations expressed in cartesian coordinates. (For the meaning at any vector terminology mentioned see Chap. 7.)

Theorem 1. *The distance d between two points* $M_1(x_1, y_1, z_1)$, $M_2(x_2, y_2, z_2)$ *is equal to the length of the vector* $\overrightarrow{M_1M_2}$; i.e.

$$d = |\overrightarrow{M_1M_2}| = |\mathbf{r}_2 - \mathbf{r}_1| = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2]},$$

where $\mathbf{r}_1$, $\mathbf{r}_2$ *are radius vectors of the points* M_1, M_2.

Theorem 2. *A point M lying on the line which joins the points* M_1, M_2 *is determined by the ratio* $\lambda = \overline{M_1M}/\overline{M_2M}$ (cf. § 5.2). *If* $\mathbf{r}_1$, $\mathbf{r}_2$ *are the radius vectors of the points* M_1, M_2, *then the radius vector* $\mathbf{r}$ *of the point* $M \neq M_2$ *is given by*

$$\mathbf{r} = \frac{\mathbf{r}_1 - \lambda \mathbf{r}_2}{1 - \lambda},$$

i.e. the coordinates x, y, z of the point M are given by

$$x = \frac{x_1 - \lambda x_2}{1 - \lambda}, \quad y = \frac{y_1 - \lambda y_2}{1 - \lambda}, \quad z = \frac{z_1 - \lambda z_2}{1 - \lambda}.$$

In particular, if M is the mid-point of the line segment, we have

$$\mathbf{r} = \frac{\mathbf{r}_1 + \mathbf{r}_2}{2}, \quad \text{i.e.} \quad x = \frac{x_1 + x_2}{2}, \quad y = \frac{y_1 + y_2}{2}, \quad z = \frac{z_1 + z_2}{2}.$$

Theorem 3. *The centroid (centre of mass — if the mass is uniformly distributed) of a triangle is given by*

$$\mathbf{r} = \frac{\mathbf{r}_1 + \mathbf{r}_2 + \mathbf{r}_3}{3}, \quad \text{i.e.} \quad x = \frac{x_1 + x_2 + x_3}{3}, \quad y = \frac{y_1 + y_2 + y_3}{3},$$

$$z = \frac{z_1 + z_2 + z_3}{3},$$

where the $\mathbf{r}_k$ $(k = 1, 2, 3)$ *denote the radius vectors, and the* x_k, y_k, z_k $(k = 1, 2, 3)$ *the coordinates, of the vertices of the triangle.*

REMARK 1. The radius vector of the centre of mass of a system of particles $M_k(\mathbf{r}_k)$ with masses m_k $(k = 1, 2, \ldots, n)$ is given by

$$\mathbf{r} = \frac{\sum_{k=1}^{n} m_k \mathbf{r}_k}{\sum_{k=1}^{n} m_k}.$$

Theorem 4. *The volume* V_4, *of a tetrahedron with vertices* $M_k(x_k, y_k, z_k)$ $(k = 1, 2, 3, 4)$ *is given by*

$$\pm V_4 = \frac{1}{6}\begin{vmatrix} x_2 - x_1, & y_2 - y_1, & z_2 - z_1 \\ x_3 - x_1, & y_3 - y_1, & z_3 - z_1 \\ x_4 - x_1, & y_4 - y_1, & z_4 - z_1 \end{vmatrix} = \frac{1}{6}\begin{vmatrix} x_1, & y_1, & z_1, & 1 \\ x_2, & y_2, & z_2, & 1 \\ x_3, & y_3, & z_3, & 1 \\ x_4, & y_4, & z_4, & 1 \end{vmatrix}.$$

(Obviously the positive value of V_4 is taken).

REMARK 2. All the four points M_k lie in the same plane if and only if the above determinant is equal to zero.

Theorem 5. *The equation of a plane can be written:*

a) *in general form:* $Ax + By + Cz + D = 0$ *(at least one of the numbers A, B, C being non-zero);*

b) *in vector form:* $\mathbf{r} \cdot \mathbf{n} + D = 0$
[the vector $\mathbf{n} = (A, B, C)$ *is perpendicular to the plane,* and is a so-called *normal vector to the plane];*

c) *in normal form:* $x \cos \alpha + y \cos \beta + z \cos \gamma - d = 0$, i.e. $\mathbf{r} \cdot \mathbf{n}^0 - d = 0$ *($d \geqq 0$ is the distance of the plane from the origin; α, β, γ are the magnitudes of the angles formed by that normal to the plane which is directed from the origin, and the coordinate axes; $\mathbf{n}^0$ is the unit vector in the direction of the normal; if $d = 0$ then the orientation of the normal is not uniquely determined);*

d) *in intercept form*:

$$\frac{x}{p} + \frac{y}{q} + \frac{z}{r} = 1$$

(*p, q, r are the intercepts cut on the coordinate axes by the plane, due regard being paid to their orientation*; for instance $p = -3$ means that the plane cuts the negative semi-axis $-x$ at a distance 3 from the origin, i.e. at the point $(-3, 0, 0)$).

REMARK 3. The equation of any plane can be written in the forms a), b), c) but it is impossible to write the equation of planes which are parallel to one of the coordinate axes or which pass through the origin in the intercept form d).

Theorem 6. *The general equation of a plane can be turned into the normal form by dividing throughout by the number* $\pm\sqrt{(A^2 + B^2 + C^2)}$. *The sign in front of the root is the opposite of that of the constant term in the general equation.* (*The normal of the plane is thereby directed from the origin to the plane.*)

Example 1. $2x - 3y - 6z + 21 = 0 \,|:[-\sqrt{(2^2 + 3^2 + 6^2)}] = -7$, and so we obtain

$$\frac{2}{-7}x + \frac{-3}{-7}y + \frac{-6}{-7}z - 3 = 0\,, \quad \text{i.e.} \quad \frac{-2}{7}x + \frac{3}{7}y + \frac{6}{7}z - 3 = 0\,.$$

To construct the given plane, first construct the vector

$$\boldsymbol{n} = 3\boldsymbol{n}^0 = 3\left(-\tfrac{2}{7}, \tfrac{3}{7}, \tfrac{6}{7}\right) = \left(-\tfrac{6}{7}, \tfrac{9}{7}, \tfrac{18}{7}\right)$$

whose initial point is at the origin, and then the plane which is perpendicular to, and which contains the terminal point of, $\boldsymbol{n}$.

Theorem 7. *The general equation of a plane can be turned into the intercept form by dividing throughout by minus the constant term* (i.e. by the number $-D$).

Example 2. $2x - 3y - 6z + 21 = 0 \,|:(-21)$,

$$\frac{x}{-\frac{21}{2}} + \frac{y}{7} + \frac{z}{\frac{7}{2}} = 1\,.$$

Theorem 8. *The equation of the plane which is perpendicular to the vector* $\boldsymbol{a}(a_1, a_2, a_3)$ *and which passes through the point* $M(x_1, y_1, z_1)$, *is*

$$a_1(x - x_1) + a_2(y - y_1) + a_3(z - z_1) = 0\,.$$

Theorem 9. *The equation of the plane containing the three non-collinear points* $M_k(x_k, y_k, z_k)$ $(k = 1, 2, 3)$, *is*

$$\begin{vmatrix} x, & y, & z, & 1 \\ x_1, & y_1, & z_1, & 1 \\ x_2, & y_2, & z_2, & 1 \\ x_3, & y_3, & z_3, & 1 \end{vmatrix} = 0\,;$$

or, in vector form:

$$[(\mathbf{r} - \mathbf{r}_1)(\mathbf{r}_2 - \mathbf{r}_1)(\mathbf{r}_3 - \mathbf{r}_1)] = 0 .$$

($[\mathbf{a}\,\mathbf{b}\,\mathbf{c}]$ is the mixed or triple scalar product of the vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$; see Definition 7.1.13, p. 230.)

Theorem 10. *The equation of the plane which contains the two points* $M_k(x_k, y_k, z_k)$ $(k = 1, 2)$, *and which is parallel to the vector* $\mathbf{a}(a_1, a_2, a_3)$, *is*

$$\begin{vmatrix} x - x_1, & y - y_1, & z - z_1 \\ x_2 - x_1, & y_2 - y_1, & z_2 - z_1 \\ a_1, & a_2, & a_3 \end{vmatrix} = 0 ;$$

or, in vector form:

$$[\mathbf{a}(\mathbf{r} - \mathbf{r}_1)(\mathbf{r}_2 - \mathbf{r}_1)] = 0 .$$

Theorem 11. *The equation of the plane which contains the point* $M(x_1, y_1, z_1)$ *and which is parallel to the two vectors* $\mathbf{a}(a_1, a_2, a_3)$, $\mathbf{b}(b_1, b_2, b_3)$ *is*

$$\begin{vmatrix} x - x_1, & y - y_1, & z - z_1 \\ a_1, & a_2, & a_3 \\ b_1, & b_2, & b_3 \end{vmatrix} = 0 ;$$

or, in vector form:

$$[\mathbf{a}\,\mathbf{b}(\mathbf{r} - \mathbf{r}_1)] = 0 .$$

Theorem 12. *The distance of the point* $M(x_1, y_1, z_1)$ *from the plane* $Ax + By + Cz + D = 0$ *is equal to the absolute value of the result of substituting the coordinates of the point in the left-hand side of the normal equation of the plane, i.e.*

$$d = \left| \frac{Ax_1 + By_1 + Cz_1 + D}{\sqrt{(A^2 + B^2 + C^2)}} \right| .$$

Theorem 13. *The angle between the two planes* $A_1x + B_1y + C_1z + D_1 = 0$, $A_2x + B_2y + C_2z + D_2 = 0$ *is equal to the angle between their normals (which are represented by the vectors* $\mathbf{n}_1(A_1, B_1, C_1)$, $\mathbf{n}_2(A_2, B_2, C_2)$*). Thus considering* $0 \leqq \varphi \leqq \frac{1}{2}\pi$,

$$\cos\varphi = \frac{|\mathbf{n}_1 \cdot \mathbf{n}_2|}{|\mathbf{n}_1|\,|\mathbf{n}_2|} = \frac{|A_1A_2 + B_1B_2 + C_1C_2|}{\sqrt{[(A_1^2 + B_1^2 + C_1^2)(A_2^2 + B_2^2 + C_2^2)]}} .$$

In particular, a necessary and sufficient condition for the two planes to be perpendicular is that

$$\mathbf{n}_1 \cdot \mathbf{n}_2 = A_1A_2 + B_1B_2 + C_1C_2 = 0 .$$

Further, since the normals to two parallel planes are parallel, we have the following theorem:

Theorem 14. *A necessary and sufficient condition for the two planes*

$$A_1x + B_1y + C_1z + D_1 = 0\,, \quad A_2x + B_2y + C_2z + D_2 = 0$$

to be parallel is that

$$A_1 : A_2 = B_1 : B_2 = C_1 : C_2\,, \quad \text{i.e.} \quad \mathbf{n}_1 = \lambda \mathbf{n}_2\,.$$

REMARK 4. The equations of two parallel planes can, therefore, always be modified so that the coefficients of the variables are the same for both planes and the equations may differ only in the constant term. This is especially useful when calculating the distance between two parallel planes:

Theorem 15. *The distance d between the two parallel planes $Ax + By + Cz + D_1 = 0$, $Ax + By + Cz + D_2 = 0$ is given by*

$$d = \left| \frac{D_2 - D_1}{\sqrt{(A^2 + B^2 + C^2)}} \right|.$$

Example 3. The distance between the planes $4x - 2y - 4z + 11 = 0$, $-2x + y + 2z + 5 = 0$ can be calculated by the foregoing formula thus: we first multiply the second equation by the number -2. Then, $A = 4$, $B = -2$, $C = -4$, $D_1 = 11$, $D_2 = -10$ and so

$$d = \left| \frac{-10 - 11}{\sqrt{(16 + 4 + 16)}} \right| = \left| \frac{-21}{6} \right| = 3{\cdot}5\,.$$

Definition 1. The set of all planes which pass through a fixed line or the set of all planes parallel to a particular plane is called an (*axial*) *pencil of planes* (*sheaf of planes*).

REMARK 5. We often speak in geometry about *points, lines and planes at infinity.* Accordingly, in the preceding definition it is sufficient to refer to the set of all planes which have a line in common (at infinity in the case of parallel planes).

Theorem 16. *The planes belonging to a pencil, two of whose members are the planes $A_1x + B_1y + C_1z + D_1 = 0$ and $A_2x + B_2y + C_2z + D_2 = 0$, have as their equations:*

$$\lambda_1(A_1x + B_1y + C_1z + D_1) + \lambda_2(A_2x + B_2y + C_2z + D_2) = 0\,, \qquad (1)$$

where λ_1, λ_2 are variable parameters, at least one of them being non-zero; clearly, only their ratio is significant.

REMARK 6. The above equation is especially useful when solving problems in which the equation of a plane passing through the line of intersection of two given planes and satisfying some additional condition is to be found.

Example 4. A plane is determined by the point $M(2, -1, 3)$ and the line of intersection of the planes whose equations are $6x + 2y - z - 3 = 0$, $3x + 4y - 2z - 2 = 0$. Find its equation.

If we substitute the coordinates of the given point into equation (1) which represents this particular pencil, we obtain the condition $4\lambda_1 - 6\lambda_2 = 0$ for λ_1, λ_2. This condition will be satisfied, if we choose, for instance $\lambda_1 = 3$, $\lambda_2 = 2$. The equation of the required plane is then

$$3(6x + 2y - z - 3) + 2(3x + 4y - 2z - 2) = 0,$$

$$\text{i.e.} \quad 24x + 14y - 7z - 13 = 0.$$

Theorem 17. *The equations of the planes which bisect the angles between two intersecting planes $\varrho_1 = 0$, $\varrho_2 = 0$ can be obtained if we add and subtract the normal equations of the two given planes:*

$$\frac{A_1x + B_1y + C_1z + D_1}{\sqrt{(A_1^2 + B_1^2 + C_1^2)}} \pm \frac{A_2x + B_2y + C_2z + D_2}{\sqrt{(A_2^2 + B_2^2 + C_2^2)}} = 0.$$

Example 5. $\varrho_1 \equiv 2x - y - 2z + 3 = 0$, $\varrho_2 \equiv 3x + 2y + 6z - 1 = 0$. The normal equations of these planes are

$$\frac{2x - y - 2z + 3}{-3} = 0 \quad \text{and} \quad \frac{3x + 2y + 6z - 1}{7} = 0,$$

respectively. Adding both equations and simplifying we obtain

$$5x - 13y - 32z + 24 = 0.$$

Subtracting we obtain, similarly,

$$23x - y + 4z + 18 = 0.$$

Definition 2. The set of all planes which pass through a fixed point (see Remark 5) is called a *bundle of planes* (*star of planes*).

Theorem 18. *The planes belonging to a bundle, three of whose members are the planes $A_1x + B_1y + C_1z + D_1 = 0$, $A_2x + B_2y + C_2z + D_2 = 0$, $A_3x + B_3y + C_3z + D_3 = 0$, have as their equations:*

$$\lambda_1(A_1x + B_1y + C_1z + D_1) + \lambda_2(A_2x + B_2y + C_2z + D_2) +$$

$$+ \lambda_3(A_3x + B_3y + C_3z + D_3) = 0,$$

where λ_1, λ_2, λ_3 are variable parameters, at least one of them being non-zero; clearly, only their ratios are significant.

REMARK 7. The relative positions of several planes can be decided by a detailed analysis of the solution of the system of linear equations which represent these planes (see § 1.18).

The position of two points relative to each other with respect to a given plane can easily be decided on the basis of the result of substituting the coordinates of these points into the equation of the plane:

Theorem 19. *If the results of substituting the coordinates of two points into the left-hand side of the general equation of a plane are of the same sign, then both points lie in the same half-space determined by the plane (i.e. on the same side of the plane); if they are of different sign, then the two points lie in different half-spaces (i.e. on different sides of the plane). (If the result equals zero, the point lies in the plane, of course.)*

Theorem 20. *The equations of a straight line:*

a) *the general equations are*

$$A_1x + B_1y + C_1z + D_1 = 0\,,$$
$$A_2x + B_2y + C_2z + D_2 = 0$$

provided that the two planes represented by these equations intersect, i.e.

$$A_1 : B_1 : C_1 \neq A_2 : B_2 : C_2\,;$$

b) *the vector equation is*

$$\mathbf{r} = \mathbf{r}_0 + t\mathbf{a}$$

where $\mathbf{r}_0$ is the radius vector of a fixed point on the line, $\mathbf{a}$ — the direction of the line, i.e. a vector parallel to the given line, and t is a variable parameter;

c) *the parametric equations (merely a paraphrase of the vector equation):*

$$x = x_0 + a_1t\,,\quad y = y_0 + a_2t\,,\quad z = z_0 + a_3t\,,$$

where (x_0, y_0, z_0) is a fixed point on the line, $\mathbf{a} = (a_1, a_2, a_3)$ — the direction vector of the line (a_1, a_2, a_3 are so-called direction parameters of the line);

d) *the reduced equations:*

$$x = mz + p\,,\quad y = nz + q\,.$$

The equations of lines parallel to the plane xy cannot be written in this form. These equations are a particular case of the general equations, the reference planes being

those planes which contain the line and its projection on the xz and yz planes respectively. They are, however, also a particular case of the parametric equations, in which the coordinate z is chosen as the parameter, i.e. $z = t$. The point (p, q) is then the point of intersection of the given line and the plane xy. Choosing the coordinate x, or y, as the parameter we obtain the other pairs of reduced equations of the line – provided that the line is not parallel to the plane yz, or xz, respectively.

Theorem 21. *The equation of a line determined by*

a) *a point* (x_0, y_0, z_0) *and a direction vector* $\mathbf{a} = (a_1, a_2, a_3)$:
in canonical form

$$\frac{x - x_0}{a_1} = \frac{y - y_0}{a_2} = \frac{z - z_0}{a_3}; \tag{2}$$

in parametric form

$$x = x_0 + a_1 t, \quad y = y_0 + a_2 t, \quad z = z_0 + a_3 t; \tag{3}$$

b) *two points* (x_1, y_1, z_1), (x_2, y_2, z_2):

$$\frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1} = \frac{z - z_1}{z_2 - z_1}$$

or (*in parametric form*)

$$x = x_1 + (x_2 - x_1)\, t, \quad y = y_1 + (y_2 - y_1)\, t, \quad z = z_1 + (z_2 - z_1)\, t.$$

REMARK 8. Equations (2) can be obtained from the parametric equations (3) by eliminating the parameter. If zero occurs in the denominator of some of the above fractions we consider those equations which are involved as a mere formal notation and, as a rule, use another form of the equations.

Theorem 22. *The direction vector of a line is parallel to the vector product of the normal vectors of any two planes which contain the line,* i.e.

$$a_1 : a_2 : a_3 = \begin{vmatrix} B_1, & C_1 \\ B_2, & C_2 \end{vmatrix} : \begin{vmatrix} C_1, & A_1 \\ C_2, & A_2 \end{vmatrix} : \begin{vmatrix} A_1, & B_1 \\ A_2, & B_2 \end{vmatrix}.$$

Example 6. Obtain the parametric equations of the line given by the general equations

$$3x + 4y + 5z - 3 = 0,$$
$$x - 2y - 3z + 4 = 0.$$

The direction vector of the line can be found by Theorem 22:

$$a_1 : a_2 : a_3 = \begin{vmatrix} 4, & 5 \\ -2, & -3 \end{vmatrix} : \begin{vmatrix} 5, & 3 \\ -3, & 1 \end{vmatrix} : \begin{vmatrix} 3, & 4 \\ 1, & -2 \end{vmatrix} =$$
$$= (-2) : 14 : (-10) = 1 : (-7) : 5.$$

In order to find the parametric equations of this line we determine one of its points. We choose $z_0 = 0$ say and find x_0, y_0, from the equations

$$3x + 4y - 3 = 0,$$
$$x - 2y + 4 = 0.$$

We obtain $x_0 = -1$, $y_0 = 1{\cdot}5$. Thus, the required equations are

$$x = -1 + t, \quad y = 1{\cdot}5 - 7t, \quad z = 5t.$$

Theorem 23. *The distance d of a point $P(x_0, y_0, z_0)$ from the line*

$$\frac{x - x_1}{a_1} = \frac{y - y_1}{a_2} = \frac{z - z_1}{a_3}$$

is given by

$$d = \frac{|\boldsymbol{u} \times \boldsymbol{a}|}{|\boldsymbol{a}|}, \quad \text{where} \quad \boldsymbol{u} = (x_0 - x_1, y_0 - y_1, z_0 - z_1), \quad \boldsymbol{a} = (a_1, a_2, a_3).$$

Example 7. Find the distance of the point $M(3, -1, 2)$ from the line

$$\frac{x - 2}{2} = \frac{y}{1} = \frac{z + 1}{-2}.$$

Here,

$$\boldsymbol{u} = (1, -1, 3), \quad \boldsymbol{a} = (2, 1, -2), \quad \boldsymbol{u} \times \boldsymbol{a} = \left(\begin{vmatrix} -1, & 3 \\ 1, & -2 \end{vmatrix}, \begin{vmatrix} 3, & 1 \\ -2, & 2 \end{vmatrix}, \begin{vmatrix} 1, & -1 \\ 2, & 1 \end{vmatrix}\right) =$$
$$= (-1, 8, 3),$$
$$|\boldsymbol{u} \times \boldsymbol{a}| = \sqrt{[(-1)^2 + 8^2 + 3^2]} = \sqrt{74}, \quad |\boldsymbol{a}| = \sqrt{[2^2 + 1^2 + (-2)^2]} = 3;$$
$$d = \frac{\sqrt{74}}{3} = 2{\cdot}867\ldots.$$

REMARK 9. In the same way, the distance between two parallel lines can be calculated: we choose a particular point on one of them and then find the distance of this point from the other line.

Theorem 24. *The distance d between two skew lines ${}^1p \equiv \boldsymbol{r} = \boldsymbol{r}_1 + \boldsymbol{at}$, ${}^2p \equiv \boldsymbol{r} = \boldsymbol{r}_2 + \boldsymbol{bt}'$ is given by*

$$d = \frac{|[(\boldsymbol{r}_2 - \boldsymbol{r}_1)\, \boldsymbol{ab}]|}{|\boldsymbol{a} \times \boldsymbol{b}|}.$$

Example 8. Find the distance between the skew lines

$${}^1p \equiv \frac{x - 1}{2} = \frac{y + 2}{2} = \frac{z + 3}{-1}, \quad {}^2p \equiv \frac{x - 2}{1} = \frac{y + 1}{-2} = \frac{z - 1}{-2}.$$

Here $\boldsymbol{r}_1 = (1, -2, -3)$, $\boldsymbol{r}_2 = (2, -1, 1)$, $\boldsymbol{a} = (2, 2, -1)$, $\boldsymbol{b} = (1, -2, -2)$, and thus $\boldsymbol{r}_2 - \boldsymbol{r}_1 = (1, 1, 4)$,

$$\boldsymbol{a} \times \boldsymbol{b} = \left(\begin{vmatrix} 2, & -1 \\ -2, & -2 \end{vmatrix}, \begin{vmatrix} -1, & 2 \\ -2, & 1 \end{vmatrix}, \begin{vmatrix} 2, & 2 \\ 1, & -2 \end{vmatrix} \right) = (-6, 3, -6),$$

$$|\boldsymbol{a} \times \boldsymbol{b}| = 9, \quad [(\boldsymbol{r}_2 - \boldsymbol{r}_1)\boldsymbol{a}\boldsymbol{b}] = (\boldsymbol{r}_2 - \boldsymbol{r}_1) \,.\, (\boldsymbol{a} \times \boldsymbol{b}) = -6 + 3 - 24 = -27.$$

Hence

$$d = \frac{|-27|}{9} = 3.$$

REMARK 10. Two lines ${}^1p \equiv \boldsymbol{r} = \boldsymbol{r}_1 + \boldsymbol{a}t$, ${}^2p \equiv \boldsymbol{r} = \boldsymbol{r}_2 + \boldsymbol{b}t'$ lie in the same plane if and only if the mixed product $[(\boldsymbol{r}_2 - \boldsymbol{r}_1)\boldsymbol{a}\boldsymbol{b}]$ equals zero. If, in addition, $\boldsymbol{a}$ is not parallel to $\boldsymbol{b}$, then they intersect.

Theorem 25. *The angle between two lines is equal to the angle between their direction vectors* $\boldsymbol{a}$, $\boldsymbol{b}$, . *Thus considering* $0 \leq \varphi \leq \frac{1}{2}\pi$,

$$\cos \varphi = \frac{\boldsymbol{a} \,.\, \boldsymbol{b}}{|\boldsymbol{a}| \,.\, |\boldsymbol{b}|} = \frac{a_1 b_1 + a_2 b_2 + a_3 b_3}{\sqrt{(a_1^2 + a_2^2 + a_3^2)}\sqrt{(b_1^2 + b_2^2 + b_3^2)}}.$$

Theorem 26. a) *A necessary and sufficient condition for two straight lines whose direction vectors are* $\boldsymbol{a}$, $\boldsymbol{b}$ *to be perpendicular is*

$$\boldsymbol{a} \,.\, \boldsymbol{b} = 0; \quad \text{i.e.} \quad a_1 b_1 + a_2 b_2 + a_3 b_3 = 0.$$

b) *A necessary and sufficient condition for two straight lines whose direction vectors are* $\boldsymbol{a}$, $\boldsymbol{b}$ *to be parallel is*

$$\boldsymbol{a} \parallel \boldsymbol{b}, \quad \text{i.e.} \quad a_1 : a_2 : a_3 = b_1 : b_2 : b_3.$$

Theorem 27. *The angle* φ *between a line and a plane is equal to the complement of the angle between the direction vector of the line and a normal vector to the plane; thus*

$$\sin \varphi = \frac{|\boldsymbol{a} \,.\, \boldsymbol{n}|}{|\boldsymbol{a}| \,.\, |\boldsymbol{n}|},$$

i.e. *if* $\boldsymbol{r} = \boldsymbol{r}_0 + \boldsymbol{a}t$ *and* $Ax + By + Cz + D = 0$ *are the equations of the line and the plane, respectively, then*

$$\sin \varphi = \frac{|a_1 A + a_2 B + a_3 C|}{\sqrt{(a_1^2 + a_2^2 + a_3^2)}\sqrt{(A^2 + B^2 + C^2)}}.$$

A necessary and sufficient condition for a line and a plane to be perpendicular is

$$\boldsymbol{a} \parallel \boldsymbol{n}; \quad \text{i.e.} \quad a_1 : a_2 : a_3 = A : B : C.$$

A necessary and sufficient condition for a line and a plane to be parallel is

$$\boldsymbol{a} \,.\, \boldsymbol{n} = 0 \,; \quad \text{i.e.} \quad a_1A + a_2B + a_3C = 0 \,.$$

6.3. Quadrics (Surfaces of the Second Order)

REMARK 1. In this section, a *surface* is defined as the locus of a point whose rectangular coordinates satisfy the equation $F(x, y, z) = 0$, where F is a function having continuous partial derivatives of at least the first order at every point. The points of a surface at which at least one of these partial derivatives differs from zero are called *regular points* of the surface, whereas the points at which all the first partial derivatives vanish are called *singular points* of the surface (for example the vertex of a cone).

(For a more detailed treatment see Chap. 9.)

Theorem 1. *The equation of the sphere with centre* $S(x_0, y_0, z_0)$ *and radius* r *is*

$$(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 = r^2 \,.$$

If we perform the operations indicated in this equation, we obtain the *general equation of a sphere* in the form

$$x^2 + y^2 + z^2 + mx + ny + pz + q = 0 \,.$$

It should be noticed that the products xy, xz, yz do not occur and that the coefficients of the squared variables are all equal.

The coordinates of the centre, and the radius, of a sphere given by the general equation can be found by completing the squares:

$$\left(x + \frac{m}{2}\right)^2 + \left(y + \frac{n}{2}\right)^2 + \left(z + \frac{p}{2}\right)^2 = \frac{m^2 + n^2 + p^2}{4} - q \,.$$

If the right-hand side of this modified equation is a positive number, then the general equation represents the so-called *real sphere* with centre $S(-\frac{1}{2}m, -\frac{1}{2}n, -\frac{1}{2}p)$ and radius $r = \sqrt{[\frac{1}{4}(m^2 + n^2 + p^2) - q]}$; if the right-hand side equals zero, then only one real point (the centre of the sphere of zero radius) satisfies the general equation; if the right-hand side is a negative number, then no real point in space satisfies the general equation. (In this case we speak about a *virtual sphere.*)

Theorem 2. *The equation of the general ellipsoid with centre at the origin and the semi-axes* a, b, c, *coincident with the* x, y *and* z *axes, respectively, is*

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \,.$$

Particular cases are:

a) $a = b > c$ (*an oblate spheroid*; Fig. 6.4) ;
b) $a = b < c$ (*a prolate spheroid*; Fig. 6.5) ;
c) $a = b = c$ (*a sphere of radius a*) .

In cases a,) b) *the z-axis is the axis of revolution.*

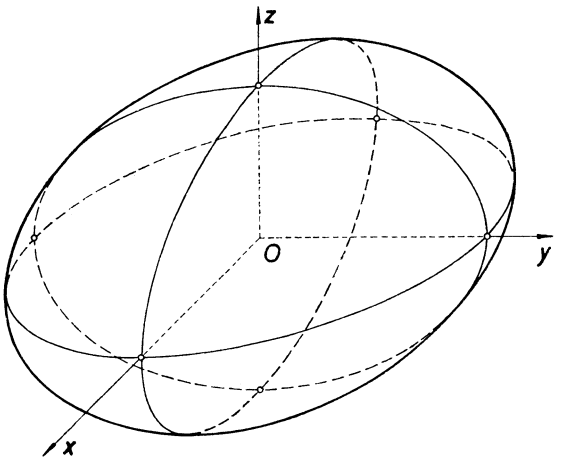


Fig. 6.4.

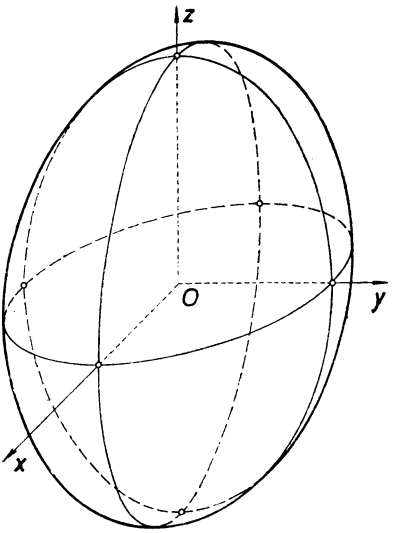


Fig. 6.5.

Theorem 3. *Hyperboloids with centre at the origin and semi-axes a, b, c coincident with the axes x, y, z, respectively are of two types*:

a) *a hyperboloid of one sheet* (Fig. 6.6), *having the equation*

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$$

(*a, b being its real semi-axes, and c its imaginary semi-axis*),

b) *a hyperboloid of two sheets* (Fig. 6.7), *having the equation*

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = -1$$

(*a, b being its imaginary semi-axes, and c the real semi-axis*).

If, in either case, a = b, then the hyperboloid is a hyperboloid of revolution and the z-axis is the axis of revolution.

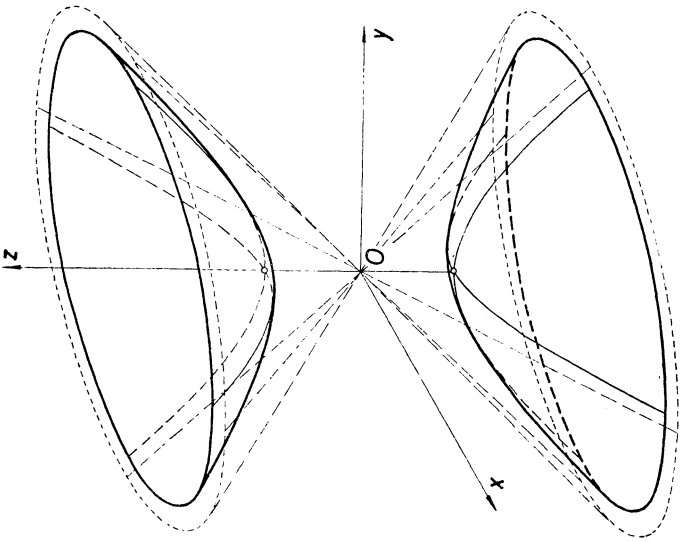


Fig. 6.7.

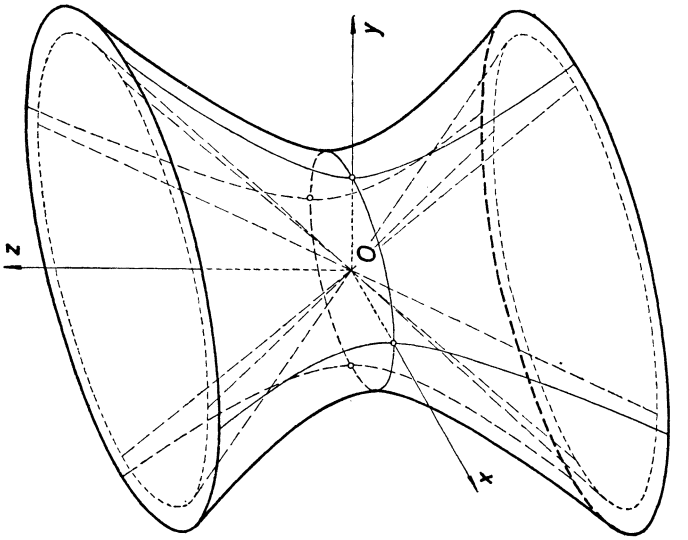


Fig. 6.6.

Theorem 4. *On a hyperboloid of one sheet there exist two sets of straight lines. Every line of the first set intersects every line of the other set, while no two lines of the same set intersect. The equations of the two sets of lines on the hyperboloid*

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$$

are

a)
$$k_1\left(\frac{x}{a} + \frac{z}{c}\right) = k_2\left(1 - \frac{y}{b}\right),$$
$$k_2\left(\frac{x}{a} - \frac{z}{c}\right) = k_1\left(1 + \frac{y}{b}\right);$$

b)
$$k_1\left(\frac{x}{a} + \frac{z}{c}\right) = k_2\left(1 + \frac{y}{b}\right),$$
$$k_2\left(\frac{x}{a} - \frac{z}{c}\right) = k_1\left(1 - \frac{y}{b}\right),$$

where k_1, k_2 are arbitrary real numbers (not both equal to zero); clearly, only their ratio $k_1 : k_2$ is significant.

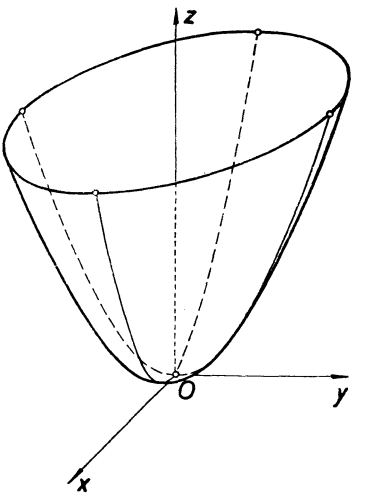


Fig. 6.8.

Theorem 5. a) *The equation of an elliptic paraboloid* (Fig. 6.8) *with vertex at the origin and whose planes of symmetry (the so-called principal sections) coincide with the planes $x = 0$, $y = 0$, is*

$$z = \frac{x^2}{2p} + \frac{y^2}{2q}, \quad pq > 0\,.$$

b) *The equation of a hyperbolic paraboloid* (Fig. 6.9) *with vertex at the origin and whose planes of symmetry* (*principal sections*) *coincide with the planes* $x = 0$, $y = 0$, *is*

$$z = \frac{x^2}{2p} - \frac{y^2}{2q}, \quad pq > 0 .$$

If, in the case of an elliptic paraboloid, $p = q$, then the paraboloid is a surface of revolution and the z-axis is the axis of revolution.

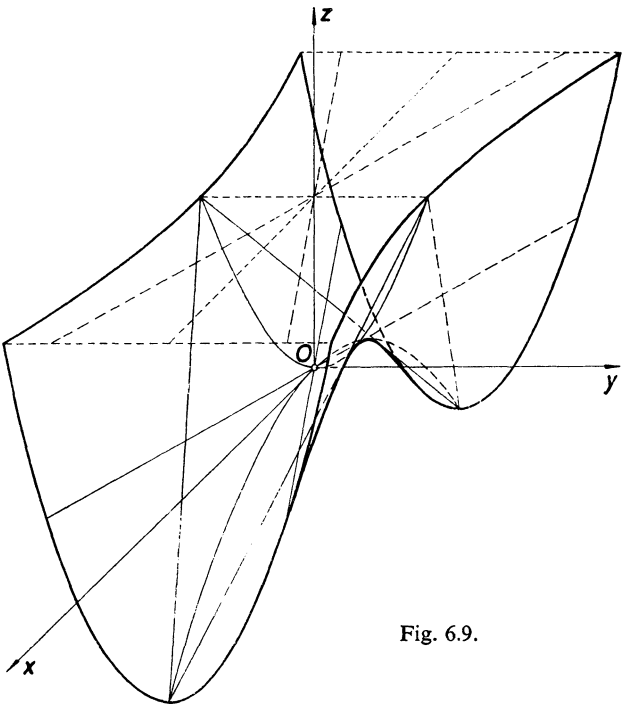


Fig. 6.9.

REMARK 2. If, in the equations of surfaces of Theorems 2 and 3, we replace x, y, z by $x - x_0$, $y - y_0$, $z - z_0$ respectively, we obtain the equations of the same surfaces with their centres translated to the point (x_0, y_0, z_0) and their axes parallel to the coordinate axes.

The same change in the equations of Theorem 5 gives the equations of paraboloids whose vertices are at the point (x_0, y_0, z_0) and whose planes of symmetry are parallel to the planes $x = 0$, $y = 0$.

In all these equations (after removing brackets) the products xy, xz, yz of the variables are missing. By similar re-arrangements as in the case of a sphere (see p. 209) the position of the centre, or the vertex, as well as the quantities a, b, c or p, q, can be found.

Theorem 6. *On a hyperbolic paraboloid there exist two sets of straight lines. Every line of the first set intersects every line of the other set, while no two lines of the same set intersect. The equations of the two sets of lines on the hyperbolic paraboloid*

$$z = \frac{x^2}{2p} - \frac{y^2}{2q} \quad (pq > 0)$$

are

$$\text{a)}\ k_1\left(\frac{x}{\sqrt{(2|p|)}} + \frac{y}{\sqrt{(2|q|)}}\right) = \frac{p}{|p|}k_2 z\,, \quad \text{b)}\ k_1\left(\frac{x}{\sqrt{(2|p|)}} - \frac{y}{\sqrt{(2|q|)}}\right) = \frac{p}{|p|}k_2 z\,,$$

$$k_2\left(\frac{x}{\sqrt{(2|p|)}} - \frac{y}{\sqrt{(2|q|)}}\right) = k_1\,; \qquad k_2\left(\frac{x}{\sqrt{(2|p|)}} + \frac{y}{\sqrt{(2|q|)}}\right) = k_1\,,$$

where k_1, k_2 are arbitrary real numbers (not both equal to zero); clearly only their ratio $k_1 : k_2$ is significant.

Theorem 7. *The equation of the quadric cone with vertex at the origin and whose directrix* is the ellipse*

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

in the plane $z = c$, is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{z^2}{c^2}\,.$$

This cone is also the asymptotic cone of the two hyperboloids

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = \pm 1\,.$$

If $a = b$, the cone is a cone of revolution and the z-axis is the axis of revolution.

Theorem 8. *Quadric cylinders are of three types*:

a) *elliptic*:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

(*its directrix is an ellipse of the same equation in the plane $z = 0$ and the generating lines are parallel to the z-axis; if $a = b$, the cylinder is a cylinder of revolution and the z-axis is the axis of revolution*);

* For the meaning of the term *directrix*, as used here, see Definition 6.4.4, p. 221.

b) *hyperbolic:*

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

(*its directrix is a hyperbola of the same equation in the plane* $z = 0$ *and the generating lines are parallel to the z-axis*);

c) *parabolic:*

$$y^2 = 2px$$

(*its directrix is a parabola of the same equation in the plane* $z = 0$ *and the generating lines are parallel to the z-axis*).

Theorem 9. *The general equation of a quadric is*

$$a_{11}x^2 + a_{22}y^2 + a_{33}z^2 + 2a_{12}xy + 2a_{13}xz + 2a_{23}yz + 2a_{14}x + 2a_{24}y + \\ + 2a_{34}z + a_{44} = 0\,;$$

it has four so-called orthogonal invariants (*i.e. functions of its coefficients whose values do not alter under translation or rotation of the coordinate system*):

a) *the discriminant of the quadric*

$$A = \begin{vmatrix} a_{11}, & a_{12}, & a_{13}, & a_{14} \\ a_{12}, & a_{22}, & a_{23}, & a_{24} \\ a_{13}, & a_{23}, & a_{33}, & a_{34} \\ a_{14}, & a_{24}, & a_{34}, & a_{44} \end{vmatrix};$$

b) *the minor* A_{44} *of the discriminant*

$$A_{44} = \begin{vmatrix} a_{11}, & a_{12}, & a_{13} \\ a_{12}, & a_{22}, & a_{23} \\ a_{13}, & a_{23}, & a_{33} \end{vmatrix};$$

c) *the quadratic invariant*

$$I_2 = \begin{vmatrix} a_{11}, & a_{12} \\ a_{12}, & a_{22} \end{vmatrix} + \begin{vmatrix} a_{11}, & a_{13} \\ a_{13}, & a_{33} \end{vmatrix} + \begin{vmatrix} a_{22}, & a_{23} \\ a_{23}, & a_{33} \end{vmatrix};$$

d) *the linear invariant*

$$I_1 = a_{11} + a_{22} + a_{33}\,.$$

In addition, the above equation has two so-called semi-invariants whose values do not alter under rotation of the coordinate system:

e)

$$S_2 = \begin{vmatrix} a_{11}, & a_{14} \\ a_{14}, & a_{44} \end{vmatrix} + \begin{vmatrix} a_{22}, & a_{24} \\ a_{24}, & a_{44} \end{vmatrix} + \begin{vmatrix} a_{33}, & a_{34} \\ a_{34}, & a_{44} \end{vmatrix};$$

TABLE 6.1

*Determination of the type of a quadric given by the general equation in rectangular coordinates**

		Type of surface	Transformed equation	Canonical equation
$A_{44} \neq 0$	$A < 0$ $I_2 > 0$ $I_1 A_{44} > 0$	real ellipsoid	$k_1 x^2 + k_2 y^2 + k_3 z^2 + \frac{A}{A_{44}} = 0$	$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$
	$A > 0$ $I_2 > 0$ $I_1 A_{44} > 0$	virtual ellipsoid		$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = -1$
	$A = 0$ $I_2 > 0$ $I_1 A_{44} > 0$	virtual cone (real point)		$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 0$
	$A > 0$; at least one of the numbers I_2, $I_1 A_{44}$ negative	hyperboloid of one sheet		$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$
	$A < 0$, at least one of the numbers I_2, $I_1 A_{44}$ negative	hyperboloid of two sheets		$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = -1$
	$A = 0$, at least one of the numbers I_2, $I_1 A_{44}$ negative	real cone		$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 0$

* The entries of the first and second columns of Table 6.1 are necessary and sufficient conditions for the type of the surface stated in the third column. This table thus replaces a number of theorems, e.g.: a necessary and sufficient condition for a quadric to be a paraboloid is $A \neq 0$, $A_{44} = 0$; or, a necessary and sufficient condition for a quadric to be degenerate is $A = 0$, $A_{44} = 0$, $S_3 = 0$.

TABLE 6.1

		Type of surface	Transformed equation	Canonical equation
$A_{44} = 0$	$A < 0$	elliptic paraboloid	$k_1x^2 + k_2y^2 \pm \pm 2\sqrt{\left(-\frac{A}{I_2}\right)}z = 0$	$\frac{x^2}{p} + \frac{y^2}{q} - 2z = 0$
$A \neq 0$	$A > 0$	hyperbolic paraboloid		$\frac{x^2}{p} - \frac{y^2}{q} - 2z = 0$
$A_{44} = 0$ $A = 0$ $I_2 \neq 0$	$I_2 > 0$ $I_1S_3 < 0$	real elliptic cylinder	$k_1x^2 + k_2y^2 + \frac{S_3}{I_2} = 0$	$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$
	$I_2 > 0$ $I_1S_3 > 0$	virtual elliptic cylinder		$\frac{x^2}{a^2} + \frac{y^2}{b^2} = -1$
	$I_2 > 0$ $S_3 = 0$	two intersecting virtual planes (a real line)		$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 0$
	$I_2 < 0$ $S_3 \neq 0$	hyperbolic cylinder		$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$
	$I_2 < 0$ $S_3 = 0$	two intersecting real planes		$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 0$
$A_{44} = 0$ $A = 0$ $I_2 = 0$	$S_3 \neq 0$	parabolic cylinder	$k_1x^2 \pm 2\sqrt{\left(-\frac{S_3}{I_1}\right)}y = 0$	$x^2 - 2py = 0$
$A_{44} = 0$ $A = 0$ $I_2 = 0$ $S_3 = 0$	$S_2 < 0$	two parallel real planes	$k_1x^2 + \frac{S_2}{I_1} = 0$	$x^2 - a^2 = 0$
	$S_2 > 0$	two parallel virtual planes		$x^2 + a^2 = 0$
	$S_2 = 0$	single plane (two coinciding planes)		$x^2 = 0$

f)

$$S_3 = \begin{vmatrix} a_{11}, & a_{12}, & a_{14} \\ a_{12}, & a_{22}, & a_{24} \\ a_{14}, & a_{24}, & a_{44} \end{vmatrix} + \begin{vmatrix} a_{11}, & a_{13}, & a_{14} \\ a_{13}, & a_{33}, & a_{34} \\ a_{14}, & a_{34}, & a_{44} \end{vmatrix} + \begin{vmatrix} a_{22}, & a_{23}, & a_{24} \\ a_{23}, & a_{33}, & a_{34} \\ a_{24}, & a_{34}, & a_{44} \end{vmatrix}.$$

Using these invariants and semi-invariants, the type of a quadric given by the general equation introduced above can be determined (Table 6.1). It should be noticed that a general quadratic equation in three variables with real coefficients need not always represent an equation of a quadric in the proper sense, i.e. ellipsoids, hyperboloids, paraboloids, cones or cylinders. For example, the equation $x^2 - y^2 = 0$ is satisfied by the points of two planes $x - y = 0$ and $x + y = 0$ (in this case, we speak about a *degenerate quadric*); similarly there is no point in space satisfying the equation $x^2 + y^2 + z^2 = -1$ and in this case we speak about a *virtual quadric.*

Theorem 10. *The coordinates of the centre of a quadric given by the general equation satisfy the system of linear equations*

$$\begin{aligned} a_{11}x + a_{12}y + a_{13}z + a_{14} &= 0, \\ a_{12}x + a_{22}y + a_{23}z + a_{24} &= 0, \\ a_{13}x + a_{23}y + a_{33}z + a_{34} &= 0. \end{aligned}$$

Depending upon the number of solutions of the above system, the quadric has no centre (quadric without centre), a single centre, an entire line of centres, or, finally, an entire plane of centres. (For a detailed analysis of the solution of a system of linear equations see § 1.18.)

To convert the general equation of a quadric to the standard form (so-called *canonical equation*) used in Theorems 1, 2, 3, 5, 7 and 8, the so-called *discriminating cubic* is important.

Definition 1. The equation $k^3 - I_1k^2 + I_2k - A_{44} = 0$ is called the *discriminating cubic of the quadric.* Its roots k_1, k_2, k_3 are all real.

Example 1. Determine the type and the canonical equation of the surface $2xy - 2xz + 2yz - 4x + 1 = 0$.

$$A = \begin{vmatrix} 0, & 1, & -1, & -2 \\ 1, & 0, & 1, & 0 \\ -1, & 1, & 0, & 0 \\ -2, & 0, & 0, & 1 \end{vmatrix} = 2 > 0; \quad A_{44} = \begin{vmatrix} 0, & 1, & -1 \\ 1, & 0, & 1 \\ -1, & 1, & 0 \end{vmatrix} = -2 \neq 0.$$

It follows immediately (according to Table 6.1) that the surface is either a virtual ellipsoid or a hyperboloid of one sheet. Also,

$$I_1 = 0; \quad I_2 = \begin{vmatrix} 0, & 1 \\ 1, & 0 \end{vmatrix} + \begin{vmatrix} 0, & -1 \\ -1, & 0 \end{vmatrix} + \begin{vmatrix} 0, & 1 \\ 1, & 0 \end{vmatrix} = -3 < 0.$$

Thus, the given surface is a hyperboloid of one sheet. Solving the discriminating cubic

$$k^3 - 0 \cdot k^2 + (-3) \cdot k - (-2) = 0,$$

i.e.

$$k^3 - 3k + 2 = 0,$$

we find that the roots are $k_1 = k_2 = 1$, $k_3 = -2$. Further, since $A/A_{44} = -1$, the equation of the surface after translation and rotation of the coordinate system is

$$x^2 + y^2 - 2z^2 - 1 = 0,$$

i.e.

$$x^2 + y^2 - 2z^2 = 1.$$

Hence, it is a hyperboloid of revolution. Finally, let us find the position of the centre of the surface in the original coordinate system:

$$\begin{aligned} y - z - 2 &= 0, \\ x + z &= 0, \\ -x + y &= 0. \end{aligned}$$

Solving these equations we obtain $x = 1$, $y = 1$, $z = -1$; these are the coordinates of the centre of the given surface.

6.4. Surfaces of Revolution and Ruled Surfaces

Definition 1. In this section a *curve* is defined as a (one-parameter) set of points such that their rectangular coordinates satisfy equations of the form $x = x(t)$, $y = y(t)$, $z = z(t)$, where $x(t)$, $y(t)$, $z(t)$ are functions which are defined in a given domain (for example, in an interval I). These functions are assumed to possess, at every point of the domain considered, first derivatives which are not all zero. The above equations can be replaced by a single vector equation $\boldsymbol{r} = \boldsymbol{r}(t)$.

(For a more detailed treatment see Chap. 9.)

REMARK 1. A curve is often given in space as the intersection of two surfaces (not having a common two-parameter part), for example $z = f(x, y)$ and $z = g(x, y)$, or $F(x, y, z) = 0$ and $G(x, y, z) = 0$. It is usually possible, by a suitable choice of the parameter, to obtain the parametric equations of Definition 1.

Definition 2. A surface generated by a curve rotating about a fixed straight line (axis of revolution) is called a *surface of revolution.*

Theorem 1. *The equation of the surface generated by the rotation about the z-axis of the curve* $y = f(z)$ in the plane $x = 0$ *is*

$$x^2 + y^2 = [f(z)]^2 .$$

Theorem 2. *The equation of the surface generated by the rotation about the z-axis of the curve* $f(y, z) = 0$ $(y \geqq 0)$ *in the plane* $x = 0$ *is*

$$f(\sqrt{(x^2 + y^2)}, z) = 0 .$$

Example 1. The equation of the torus (anchor-ring) generated by the circle $x = 0$, $(y - a)^2 + z^2 = r^2$ $(0 < r \leqq a)$ rotating about the z-axis is

$$(\sqrt{(x^2 + y^2)} - a)^2 + z^2 = r^2 ,$$

which can be put (on removing the radicals) into the form

$$(x^2 + y^2 + z^2 + a^2 - r^2)^2 = 4a^2(x^2 + y^2) .$$

REMARK 2. The condition $y \geqq 0$ stated in Theorem 2 is usually not essential, since the equation $f(y, z) = 0$ can often be put into the form $g(y^2, z) = 0$. Then, the equation of the corresponding surface of revolution is $g(x^2 + y^2, z) = 0$. For example, the equation of the spheroid generated by the ellipse

$$x = 0 , \quad \frac{y^2}{b^2} + \frac{z^2}{c^2} - 1 = 0$$

rotating about the z-axis is

$$\frac{x^2 + y^2}{b^2} + \frac{z^2}{c^2} - 1 = 0 .$$

REMARK 3. In the general case, when the surface of revolution is generated by the curve $x = x(t)$, $y = y(t)$, $z = z(t)$ rotating about the straight line

$$\frac{x - x_0}{a} = \frac{y - y_0}{b} = \frac{z - z_0}{c} ,$$

we derive the equation of the surface in the following way:

We write down the equations of the circle traced by a general point of the rotating curve and then eliminate the parameter of this general point from these equations. This circle is the intersection of the plane in which the general point moves (this plane being perpendicular to the axis of rotation) and a sphere whose centre is at a point on the axis of rotation – for example the point (x_0, y_0, z_0) – and passing through the general point of the rotating curve, as shown in Example 2.

Example 2. The line

$$x = t\sqrt{2}, \quad y = \tfrac{1}{2} + t, \quad z = -1 + t(2 + \sqrt{2})$$

rotates about the axis

$$\frac{x-1}{1} = \frac{y-1}{-1} = \frac{z+1}{1}.$$

Derive the equation of the surface of revolution.

1. The plane of rotation passes through the point $M(t\sqrt{2}, \frac{1}{2} + t, -1 + t(2 + \sqrt{2}))$ and is perpendicular to the vector $(1, -1, 1)$. Its equation is

$$x - y + z - t(2\sqrt{(2)} + 1) + \tfrac{3}{2} = 0.$$

2. The equation of the sphere with centre at the point $S(1, 1, -1)$ passing through the point M is

$$x^2 + y^2 + z^2 - 2x - 2y + 2z + \tfrac{7}{4} = (2\sqrt{(2)} + 1)^2 t^2 - t(2\sqrt{(2)} + 1).$$

In order to eliminate the parameter t we obtain an explicit expression for t from the first equation

$$t(2\sqrt{(2)} + 1) = x - y + z + \tfrac{3}{2}$$

and substitute it into the second. On simplification we obtain

$$2xy - 2xz + 2yz - 4x + 1 = 0;$$

this is the equation of a hyperboloid of revolution of one sheet (see Example 6.3.1).

Definition 3. A surface through every point of which it is possible to draw a straight line lying entirely on the surface is called a *ruled surface.*

Particular examples of ruled surfaces are the (general) conical and (general) cylindrical surfaces.

Definition 4. The set of all straight lines passing through a fixed point V and intersecting a fixed curve c is called a (*general*) *conical surface* (with the exception of the case when c is a straight line passing through the point V). The point V is the *vertex,* the curve c is *the directrix*, the lines of the surface are the *ruling* (*generating*) *lines,* or *generators.*

Definition 5. The aggregate of all straight lines parallel to a given direction (i.e. to a vector or a line) and intersecting a fixed curve c is called a (*general*) *cylindrical surface.* The curve c is the *directrix,* the lines of the surface are the *ruling* (*generating*) *lines* or *generators.*

The equations of conical and cylindrical surfaces can be derived, for example, in the following way: express the directrix parametrically and obtain the equation

of the line joining the vertex (at infinity in the case of a cylinder) to a general point of the directrix. Then, eliminate from these equations the parameter of the general point.

Example 3. A conical surface has directrix

$$c \equiv (x^2 + y^2 + z^2 = r^2\,,\ x^2 + y^2 = rx)$$

and its vertex is at the origin of the coordinate system.

The parametric equations of the directrix can be written in the form $x = r\cos^2 t$, $y = r \sin t \cos t$, $z = r \sin t$. A general line of the surface is therefore

$$\frac{x}{r\cos^2 t} = \frac{y}{r \sin t \cos t} = \frac{z}{r \sin t}$$

and can be re-written in the form

$$x \sin t = y \cos t\,,$$
$$y = z \cos t\,.$$

The easiest way to eliminate the parameter is to express $\cos t$ and $\sin t$ in terms of x, y, z from the latter equations, and then to square and add, giving

$$\cos t = \frac{y}{z}\,,$$

$$\sin t = \frac{y^2}{xz}$$

$$1 = \frac{y^2}{z^2} + \frac{y^4}{x^2z^2}\,.$$

Hence we obtain, on simplification, the required equation of the surface:

$$x^2z^2 = y^2(x^2 + y^2)\,.$$

Theorem 3. a) *The equation of the cylindrical surface whose directrix in the plane* $z = 0$ *is* $f(x, y) = 0$ *and whose ruling lines are parallel to the z-axis is*

$$f(x, y) = 0\,.$$

b) *If, in the equation of a surface, one of the variables is missing, then this equation represents a cylindrical surface such that the ruling lines are parallel to the axis which is denoted by the missing variable, and the equation of the directrix is identical with that of the given surface and lies in the coordinate plane perpendicular to the ruling lines.*

Example 4. The equation $(x^2 + y^2)^2 = a^2(x^2 - y^2)$ is the equation of the cylindrical surface whose lines are parallel to the z-axis and whose directrix is the lemniscate of Bernoulli (see § 4.11) in the plane xy.

Theorem 4. *The equation of a cylindrical surface whose ruling lines are parallel to the vector* (a_1, a_2, a_3) *can always be put into the form*

$$F(a_3x - a_1z, a_3y - a_2z) = 0.$$

Theorem 5. *The equation of a conical surface with vertex* $V(x_0, y_0, z_0)$ *can always be put into the form*

$$F\left(\frac{x - x_0}{z - z_0}, \frac{y - y_0}{z - z_0}\right) = 0.$$

The equation of the conical surface derived in Example 3 can, for example, be written in the form

$$\frac{y^2}{z^2} + \frac{y^4}{z^4}\frac{z^2}{x^2} - 1 = 0.$$

More general ruled surfaces are usually determined by three directrices and by the condition that the surface is formed by lines intersecting all the directrices. The equations of such surfaces can be derived in a way similar to that used in the case of the equations of cylindrical and conical surfaces: express one of the directrices parametrically, project the other two from a general point of this directrix and obtain, in this way, the equations of two conical surfaces with a common vertex. From these equations, eliminate the parameter of the variable point of the first directrix and so obtain the required equation of the surface.

In some ruled surfaces, one of the directrices may be a *plane*; the ruling lines are parallel to it.

Definition 6. A ruled surface determined by three directrices which consist of a curve, a straight line and a plane, is called a *conoid.*

Example 5. The conoid determined by the directrices: the curve $c \equiv (x = a\cos t, y = a\sin t, z = bt)$, the line $p \equiv (x = 0, y = 0)$ and the plane $z = 0$ is a so-called *helicoid.*)

Through a general point of the directrix-curve we draw on the one hand a plane parallel to the directrix-plane and on the other a plane containing the directrix-line. Eliminating the parameter, we obtain the required equation of the surface in the form

$$y = x\tan\frac{z}{b}.$$

Theorem 6. *The equation of the circular conoid determined by the directrix-curve* $x = 0$, $y^2 + z^2 = r^2$, *the directrix-line* $z = 0$, $x = a$ *and the directrix-plane* $y = 0$ *is*

$$a^2z^2 = (r^2 - y^2)(a - x)^2 .$$

Theorem 7. *The equation of Plücker's conoid determined by the directrix curve* $x^2 + y^2 = rx$, $z = x \tan \alpha$, *the directrix-line* $x = y = 0$ *and the directrix-plane* $z = 0$ *is*

$$z(x^2 + y^2) = rx^2 \tan \alpha .$$

Theorem 8. *The equation of Küpper's conoid determined by the directrix-curve* $x^2 + y^2 = rx$, $z = 0$, *the directrix-line* $x = y = 0$ *and the directrix-plane* $z = x$ *is*

$$rx^2 = (x^2 + y^2)(x - z) .$$

Theorem 9. *The equation of the Montpellier conoid determined by the directrix-circle* $y^2 + z^2 = r^2$, $x = 0$ *and two directrix-lines* ${}^1p \equiv (y = z = 0)$, ${}^2p \equiv (x = a, z = b)$ *is*

$$r^2z^2(a - x)^2 = (az - bx)^2 (y^2 + z^2) .$$

7. VECTOR CALCULUS

References: [4], [16], [23], [35], [68], [90], [91], [92], [122], [134], [147], [149], [167], [181].

A. VECTOR ALGEBRA

By FRANTIŠEK KEJLA

7.1. Vector Algebra; Scalar (Inner), Vector (Cross), Mixed and Triple Products

There is a great advantage in using vector calculus when solving various problems in applied mathematics. This advantage consists on the one hand in a special notation facilitating a very simple description of many relations which would otherwise be expressed by awkward and incomprehensible formulae, on the other in the possibility of expressing many laws and formulae in a form independent of the coordinate system.

Convention 1. Throughout this chapter, by the term "vector" a three-component vector, i.e. a vector in three-dimensional space will be understood (for general definition of a vector see § 1.15).

Definition 1. Ordered triplets of real numbers for which

a) equality: $(a_1, a_2, a_3) = (b_1, b_2, b_3)$ if and only if $a_1 = b_1$, $a_2 = b_2$, $a_3 = b_3$;

b) sum: $(a_1, a_2, a_3) + (b_1, b_2, b_3) = (a_1 + b_1, a_2 + b_2, a_3 + b_3)$;

c) product of a triple and a number: $k(a_1, a_2, a_3) = (ka_1, ka_2, ka_3)$ are defined, are called *vectors*. We denote them usually by bold letters, i.e. $\mathbf{a} = (a_1, a_2, a_3)$ or $\mathbf{a}(a_1, a_2, a_3)$. The numbers a_1, a_2, a_3 are said to be the *components of the vector* $\mathbf{a}$.

Definition 2. The vector (0, 0, 0) is called the *zero vector* or *null vector;* we denote it by $\mathbf{0}$.

Definition 3. By the *vector opposite* to a vector $\mathbf{a}(a_1, a_2, a_3)$ we mean the vector $(-a_1, -a_2, -a_3)$ and denote it by $-\mathbf{a}$.

Theorem 1. *Vectors satisfy*

(i) *the commutative law:* $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$;
(ii) *the associative law:* $(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c})$;
(iii) *distributive laws:* $k(\mathbf{a} + \mathbf{b}) = k\mathbf{a} + k\mathbf{b}$, $(k_1 + k_2)\mathbf{a} = k_1\mathbf{a} + k_2\mathbf{a}$, where k, k_1, k_2 are numbers.

For further properties of vectors see § 1.15.

Some quantities in physics (force, velocity etc.) are known to be of vector character. They are customarily represented by *directed segments,* i.e. by a segment having a certain length and direction. In what follows, we assume that a fixed cartesian coordinate system in space has been chosen. In this system, we establish the above-mentioned representation of vectors by the following definition:

Definition 4. Every ordered pair of points $A(a_1, a_2, a_3)$, $B(b_1, b_2, b_3)$ determines a vector $(b_1 - a_1, b_2 - a_2, b_3 - a_3)$. This vector is denoted by $\overrightarrow{AB}$. By the arrow the direction of the vector is marked; A is called the *initial (starting) point, B the end (terminal) point of the vector* $\overrightarrow{AB}$.

In figures a vector $\overrightarrow{AB}$ is illustrated by a segment AB with an arrow at the end point B (see Fig. 7.1).

Convention 2. In the following text we shall use the term "vector" also for the graphical illustration of a vector.

Theorem 2. *If a point $A(a_1, a_2, a_3)$ and a vector $\mathbf{u}(u_1, u_2, u_3)$ are given, then there exists a unique point B such that $\overrightarrow{AB} = \mathbf{u}$. The coordinates of the point B are $a_1 + u_1, a_2 + u_2, a_3 + u_3$.*

Theorem 3. *Two vectors $\overrightarrow{AB}$, $\overrightarrow{CD}$ are equal if and only if the equations $b_1 - a_1 = d_1 - c_1$, $b_2 - a_2 = d_2 - c_2$, $b_3 - a_3 = d_3 - c_3$ hold.*

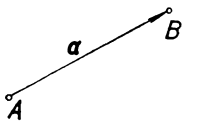


Fig. 7.1.

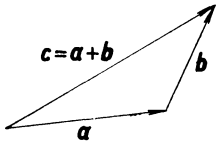


Fig. 7.2.

REMARK 1. Theorem 3 states that a vector can be arbitrarily placed in space by a choice of its initial point. Its end point is then determined uniquely. We speak of so-called *free vectors.*

A vector with its initial point at the origin and end point at the given point P is called the *radius (position) vector of the point P.*

Theorem 4. *Let* $\boldsymbol{a}$, $\boldsymbol{b}$ *be two vectors. If we place the vector* $\boldsymbol{b}$ *so that its initial point coincides with the end point of the vector* $\boldsymbol{a}$, *then the vector* $\boldsymbol{c}$ *determined by the initial point of the vector* $\boldsymbol{a}$ *and by the end point of the vector* $\boldsymbol{b}$ (*and directed in this sense*) *equals the sum of the vectors* $\boldsymbol{a}$, $\boldsymbol{b}$, *i.e.* $\boldsymbol{c} = \boldsymbol{a} + \boldsymbol{b}$ (see Fig. 7.2).

Definition 5. By the *length of a vector* $\boldsymbol{a}(a_1, a_2, a_3)$ we understand the non-negative number $\sqrt{(a_1^2 + a_2^2 + a_3^2)}$; we denote it by $|\boldsymbol{a}|$ or a. A vector whose length equals unity is called a *unit vector.*

REMARK 2. Instead of "the length of a vector" the terms *modulus, magnitude, norm* or *absolute value of a vector* are also used. *The length of a vector is equal to the length of a segment representing the given vector.*

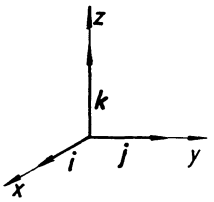
Theorem 5. *For every non-zero vector* $\boldsymbol{a}$ *there exists a unit vector* $\boldsymbol{a}^0$ *conformably parallel to the vector* $\boldsymbol{a}$ (Theorem 7). *It is* (*uniquely*) *determined by the relation*

$$\boldsymbol{a}^0 = \frac{\boldsymbol{a}}{|\boldsymbol{a}|}.$$

(The notation $\hat{\boldsymbol{a}}$ is often used for a unit vector.)

Definition 6. The linearly independent vectors $\boldsymbol{i}(1, 0, 0)$, $\boldsymbol{j}(0, 1, 0)$, $\boldsymbol{k}(0, 0, 1)$ (Fig. 7.3) are called *principal* or *coordinate vectors.* (For the concept of linear dependence and independence see Definition 1.15.3, p. 24.)

Fig. 7.3.



Theorem 6. *Any four vectors are linearly dependent. Thus every vector* $\boldsymbol{a}(a_1, a_2, a_3)$ *can be expressed as a linear combination of three linearly independent vectors, in particular as a linear combination of the principal vectors* $\boldsymbol{i}$, $\boldsymbol{j}$, $\boldsymbol{k}$:

$$\boldsymbol{a} = a_1\boldsymbol{i} + a_2\boldsymbol{j} + a_3\boldsymbol{k}.$$

Definition 7. Two linearly dependent vectors are called *collinear* (*parallel*); three linearly dependent vectors are called *coplanar.*

Theorem 7. *Two vectors* $\boldsymbol{a}$, $\boldsymbol{b}$ *are linearly dependent* (*parallel*) *if and only if one of them is a multiple of the other, i.e. if there is a number* k *such that either* $\boldsymbol{a} = k\boldsymbol{b}$ or $\boldsymbol{b} = k\boldsymbol{a}$. *In a graphical illustration, they are represented by two parallel segments* (*unless one of them is the zero vector*) *whose lengths satisfy* $|\boldsymbol{a}| = |k|\,|\boldsymbol{b}|$, or $|\boldsymbol{b}| = |k|\,|\boldsymbol{a}|$. If $k > 0$, the vectors $\boldsymbol{a}$, $\boldsymbol{b}$ are said to be *conformably collinear* (*con-*

formably parallel); if $k < 0$, they are said to be *unconformably collinear* (*unconformably parallel*).

Theorem 8. *Coplanar vectors are parallel to a common plane* (*they can be placed in the same plane*).

Definition 8. The *angle between two non-zero vectors* $\boldsymbol{a}$, $\boldsymbol{b}$ is the angle φ $(0 \leqq \varphi \leqq \pi)$ between the directed segments representing both vectors.

Definition 9. The *scalar* (*inner, dot*) *product* ($\boldsymbol{a} \,.\, \boldsymbol{b}$ or $\boldsymbol{ab}$, in symbols) of vectors $\boldsymbol{a}(a_1, a_2, a_3)$, $\boldsymbol{b}(b_1, b_2, b_3)$ is the (scalar) number $a_1b_1 + a_2b_2 + a_3b_3$.

Theorem 9. *If* $\boldsymbol{a}$, $\boldsymbol{b}$ *are non-zero vectors and* φ *the angle between them, then the following relation holds for their scalar product:*

$$\boldsymbol{a} \,.\, \boldsymbol{b} = |\boldsymbol{a}|\,|\boldsymbol{b}| \cos \varphi \,.$$

Example 1.

$$\boldsymbol{i} \,.\, \boldsymbol{i} = 1\,, \quad \boldsymbol{j} \,.\, \boldsymbol{j} = 1\,, \quad \boldsymbol{k} \,.\, \boldsymbol{k} = 1\,,$$

$$\boldsymbol{i} \,.\, \boldsymbol{j} = 0\,, \quad \boldsymbol{j} \,.\, \boldsymbol{k} = 0\,, \quad \boldsymbol{k} \,.\, \boldsymbol{i} = 0\,;$$

these results can easily be verified by Definition 9 or Theorem 9 (see Fig. 7.3).

REMARK 3. Theorem 9 is very often used to compute the angle between two vectors which are given by their components.

Example 2. For $\boldsymbol{a}(2, 1, 2)$, $\boldsymbol{b}(1, -1, 4)$

$$\cos \varphi = \frac{\boldsymbol{a} \,.\, \boldsymbol{b}}{|\boldsymbol{a}|\,|\boldsymbol{b}|} = \frac{2 \,.\, 1 + 1 \,.\, (-1) + 2 \,.\, 4}{\sqrt{(2^2 + 1^2 + 2^2)}\,\sqrt{[1^2 + (-1)^2 + 4^2]}} = \frac{9}{3 \,.\, \sqrt{18}} = \frac{1}{\sqrt{2}}\,,$$

and thus $\varphi = \frac{1}{4}\pi$.

Theorem 10. *Two non-zero vectors* $\boldsymbol{a}$, $\boldsymbol{b}$ *are perpendicular if and only if*

$$\boldsymbol{a} \,.\, \boldsymbol{b} = 0\,.$$

Theorem 11. *The scalar product of vectors satisfies the relations*:

a) $\boldsymbol{a} \,.\, \boldsymbol{b} = \boldsymbol{b} \,.\, \boldsymbol{a}$;

b) $(\boldsymbol{a} + \boldsymbol{b}) \,.\, \boldsymbol{c} = \boldsymbol{a} \,.\, \boldsymbol{c} + \boldsymbol{b} \,.\, \boldsymbol{c}$;

c) $\boldsymbol{a} \,.\, \boldsymbol{a} = |\boldsymbol{a}|^2$.

REMARK 4. Instead of $\boldsymbol{a} \,.\, \boldsymbol{a}$ we often write $\boldsymbol{a}^2$.

Definition 10. The angles which a non-zero vector makes with the principal vectors (and thus with the coordinate axes) are called the *direction angles* and their cosines the *direction cosines* of the given vector.

Theorem 12. *Denoting by α, β, γ the direction angles of a non-zero vector $\boldsymbol{a}(a_1, a_2, a_3)$, the following relations hold for the direction cosines of the vector $\boldsymbol{a}$:*

a) $\cos\alpha = \frac{a_1}{|\boldsymbol{a}|}$; $\cos\beta = \frac{a_2}{|\boldsymbol{a}|}$; $\cos\gamma = \frac{a_3}{|\boldsymbol{a}|}$;

b) $\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1$.

Definition 11. Three linearly independent vectors $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ with a common initial point determine a trihedral angle $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$. It is said to be *positively oriented* if the determinant of the coordinates of the vectors (in the given order), i.e. the determinant

$$\begin{vmatrix} a_1, & a_2, & a_3 \\ b_1, & b_2, & b_3 \\ c_1, & c_2, & c_3 \end{vmatrix}$$

is positive. If it is negative, then we say that the trihedral angle is *negatively oriented.*

Example 3. The trihedral angle defined by the vectors $\boldsymbol{i}$, $\boldsymbol{j}$, $\boldsymbol{k}$ (in the given order) is positively oriented, since

$$\begin{vmatrix} 1, & 0, & 0 \\ 0, & 1, & 0 \\ 0, & 0, & 1 \end{vmatrix} = 1 > 0 .$$

Definition 12. The *vector (cross, outer) product* ($\boldsymbol{a} \times \boldsymbol{b}$, or $\boldsymbol{a} \wedge \boldsymbol{b}$, in symbols) of vectors $\boldsymbol{a}(a_1, a_2, a_3)$, $\boldsymbol{b}(b_1, b_2, b_3)$ is the vector

$$\boldsymbol{w}\left(\begin{vmatrix} a_2, & a_3 \\ b_2, & b_3 \end{vmatrix}, \begin{vmatrix} a_3, & a_1 \\ b_3, & b_1 \end{vmatrix}, \begin{vmatrix} a_1, & a_2 \\ b_1, & b_2 \end{vmatrix}\right).$$

Theorem 13. *The vector product satisfies the relations*:

a) $\boldsymbol{a} \times \boldsymbol{b} = -(\boldsymbol{b} \times \boldsymbol{a})$;

b) $k\boldsymbol{a} \times \boldsymbol{b} = k(\boldsymbol{a} \times \boldsymbol{b})$;

c) $\boldsymbol{a} \times (\boldsymbol{b} + \boldsymbol{c}) = \boldsymbol{a} \times \boldsymbol{b} + \boldsymbol{a} \times \boldsymbol{c}$.

REMARK 5. Hence, the commutative law does not hold for the vector product.

Theorem 14. *The vector product of two linearly dependent vectors is the zero vector.*

Theorem 15. *The vector product $\boldsymbol{w} = \boldsymbol{a} \times \boldsymbol{b}$ of two linearly independent vectors $\boldsymbol{a}$, $\boldsymbol{b}$ possesses the following properties*:

a) *It is perpendicular to both given vectors $\boldsymbol{a}$, $\boldsymbol{b}$,* i.e. $\boldsymbol{w} \cdot \boldsymbol{a} = 0$, $\boldsymbol{w} \cdot \boldsymbol{b} = 0$;

b) *its length is numerically equal to the area of the parallelogram of which the vectors $\boldsymbol{a}$, $\boldsymbol{b}$ are concurrent sides*: i.e. $|\boldsymbol{w}| = |\boldsymbol{a}| \cdot |\boldsymbol{b}| \cdot \sin\varphi$, *where φ is the angle between the vectors $\boldsymbol{a}$, $\boldsymbol{b}$*;

c) *the trihedral angle* $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{w})$ *is positively oriented.*

Example 4. In view of Definition 12, or Theorem 15, the following relations can readily be established:

$$\boldsymbol{i} \times \boldsymbol{j} = \boldsymbol{k}, \quad \boldsymbol{j} \times \boldsymbol{k} = \boldsymbol{i}, \quad \boldsymbol{k} \times \boldsymbol{i} = \boldsymbol{j},$$

$$\boldsymbol{i} \times \boldsymbol{i} = \boldsymbol{0}, \quad \boldsymbol{j} \times \boldsymbol{j} = \boldsymbol{0}, \quad \boldsymbol{k} \times \boldsymbol{k} = \boldsymbol{0}$$

(see Fig. 7.3).

REMARK 6. The properties listed in Theorem 15 are sometimes used (in physics, for example) to define the vector product.

Theorem 16. *The vector product of vectors* $\boldsymbol{a}, \boldsymbol{b}$ *can be written by means of the principal vectors in the form*

$$\boldsymbol{a} \times \boldsymbol{b} = \begin{vmatrix} \boldsymbol{i}, & \boldsymbol{j}, & \boldsymbol{k} \\ a_1, & a_2, & a_3 \\ b_1, & b_2, & b_3 \end{vmatrix}.$$

Definition 13. *The mixed product or triple scalar product of three vectors* $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ is the number $\boldsymbol{a} \,.\, (\boldsymbol{b} \times \boldsymbol{c})$, denoted by $[\boldsymbol{abc}]$ or $\boldsymbol{abc}$.

REMARK 7. The mixed product of three vectors is sometimes called also a *trivector.*

Theorem 17. *The following relations hold for the mixed product of vectors* $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$:

$$[\boldsymbol{abc}] = [\boldsymbol{bca}] = [\boldsymbol{cab}] = -[\boldsymbol{acb}] = -[\boldsymbol{cba}] = -[\boldsymbol{bac}] = \begin{vmatrix} a_1, & a_2, & a_3 \\ b_1, & b_2, & b_3 \\ c_1, & c_2, & c_3 \end{vmatrix}.$$

Theorem 18. *The absolute value of the mixed product is equal to the volume of the parallelepiped of which* $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ *are concurrent edges.*

Theorem 19. *Three vectors* $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ *are coplanar if and only if*

$$[\boldsymbol{abc}] = 0.$$

Definition 14. The vector $\boldsymbol{a} \times (\boldsymbol{b} \times \boldsymbol{c})$ is said to be the *triple vector product of vectors* $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ (in the given order).

Theorem 20. *The triple vector product of vectors* $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ *can be expressed without using vector multiplication:* $\boldsymbol{a} \times (\boldsymbol{b} \times \boldsymbol{c}) = (\boldsymbol{a} \,.\, \boldsymbol{c})\, \boldsymbol{b} - (\boldsymbol{a} \,.\, \boldsymbol{b})\, \boldsymbol{c}$.

REMARK 8. In general, $\boldsymbol{a} \times (\boldsymbol{b} \times \boldsymbol{c}) \neq (\boldsymbol{a} \times \boldsymbol{b}) \times \boldsymbol{c}$.

Theorem 21. $(\boldsymbol{a} \times \boldsymbol{b}) \,.\, (\boldsymbol{c} \times \boldsymbol{d}) = (\boldsymbol{a} \,.\, \boldsymbol{c})\, (\boldsymbol{b} \,.\, \boldsymbol{d}) - (\boldsymbol{b} \,.\, \boldsymbol{c})\, (\boldsymbol{a} \,.\, \boldsymbol{d})$ (the so-called *Lagrange identity*).

REMARK 9. In particular, Theorem 21 yields: $(\boldsymbol{a} \times \boldsymbol{b})^2 = \boldsymbol{a}^2\boldsymbol{b}^2 - (\boldsymbol{a} \cdot \boldsymbol{b})^2$.

Theorem 22.

$$[\boldsymbol{abc}] \cdot [\boldsymbol{def}] = \begin{vmatrix} \boldsymbol{a} \cdot \boldsymbol{d}, & \boldsymbol{a} \cdot \boldsymbol{e}, & \boldsymbol{a} \cdot \boldsymbol{f} \\ \boldsymbol{b} \cdot \boldsymbol{d}, & \boldsymbol{b} \cdot \boldsymbol{e}, & \boldsymbol{b} \cdot \boldsymbol{f} \\ \boldsymbol{c} \cdot \boldsymbol{d}, & \boldsymbol{c} \cdot \boldsymbol{e}, & \boldsymbol{c} \cdot \boldsymbol{f} \end{vmatrix}.$$

B. VECTOR ANALYSIS

By KAREL REKTORYS

7.2. Derivative of a Vector. Scalar and Vector Fields. Gradient, Divergence, Curl (Rotation). Operator ∇, Laplace Operator. Transformation to Cylindrical and Spherical Coordinates

In physical and geometrical considerations we often have to deal with the case where the components of a vector are functions of a scalar variable t,

$$\boldsymbol{a}(t) = a_1(t)\,\boldsymbol{i} + a_2(t)\,\boldsymbol{j} + a_3(t)\,\boldsymbol{k}. \tag{1}$$

Thus, for every t of the domain under consideration we get, in general, a different vector; we speak about the *vector field*, or briefly about the *vector* $\boldsymbol{a}(t)$. The components of a vector can also be functions of several variables.

We define the derivative of a vector $\boldsymbol{a}(t)$. (The definition of partial derivatives, when the components of $\boldsymbol{a}$ are functions of several variables, follow in a similar way.)

Derivative of a vector $\boldsymbol{a}(t)$:

$$\boldsymbol{a}'(t) = \frac{\mathrm{d}\boldsymbol{a}(t)}{\mathrm{d}t} = \lim_{\Delta \to 0} \frac{\boldsymbol{a}(t + \Delta t) - \boldsymbol{a}(t)}{\Delta t} = a_1'(t)\,\boldsymbol{i} + a_2'(t)\,\boldsymbol{j} + a_3'(t)\,\boldsymbol{k}. \tag{2}$$

Similarly

$$\boldsymbol{a}''(t) = \lim_{\Delta \to 0} \frac{\boldsymbol{a}'(t + \Delta t) - \boldsymbol{a}'(t)}{\Delta t} = a_1''(t)\,\boldsymbol{i} + a_2''(t)\,\boldsymbol{j} + a_3''(t)\,\boldsymbol{k} \tag{3}$$

etc.

Theorem 1.

$$(\boldsymbol{a} \cdot \boldsymbol{b})' = \boldsymbol{a}' \cdot \boldsymbol{b} + \boldsymbol{a} \cdot \boldsymbol{b}', \quad (\boldsymbol{a} \cdot \boldsymbol{b})'' = \boldsymbol{a}'' \cdot \boldsymbol{b} + 2\boldsymbol{a}' \cdot \boldsymbol{b}' + \boldsymbol{a} \cdot \boldsymbol{b}'', \tag{4}$$

$$(\boldsymbol{a} \times \boldsymbol{b})' = \boldsymbol{a}' \times \boldsymbol{b} + \boldsymbol{a} \times \boldsymbol{b}', \quad (\boldsymbol{a} \times \boldsymbol{b})'' = \boldsymbol{a}'' \times \boldsymbol{b} + 2\boldsymbol{a}' \times \boldsymbol{b}' + \boldsymbol{a} \times \boldsymbol{b}''. \tag{5}$$

Example 1. If the length of a vector $\boldsymbol{a}(t)$ is constant and equal to k then from the equation

$$\boldsymbol{a}^2(t) = \boldsymbol{a}(t) \,.\, \boldsymbol{a}(t) = k^2$$

the relation

$$\boldsymbol{a}'(t)\, \boldsymbol{a}(t) + \boldsymbol{a}(t)\, \boldsymbol{a}'(t) = 2\boldsymbol{a}'(t)\, \boldsymbol{a}(t) = 0$$

follows and thus (for every t at which the derivative $\boldsymbol{a}'(t)$ exists and $\boldsymbol{a}(t) \neq \boldsymbol{0}$, $\boldsymbol{a}'(t) \neq \boldsymbol{0}$)

$$\boldsymbol{a}'(t) \perp \boldsymbol{a}(t) \,.$$

REMARK 1. A case of particular importance is that of the space curve described by the end-point of the radius vector

$$\boldsymbol{r}(s) = x(s)\, \boldsymbol{i} + y(s)\, \boldsymbol{j} + z(s)\, \boldsymbol{k} \,;$$

here s denotes the length of the curve, measured from a fixed point on the curve (see § 9.2, p. 265). Then, the vector

$$\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}s} = \boldsymbol{t} \tag{6}$$

is a unit vector and is called the *tangent (unit) vector* of the curve under consideration. The unit vector

$$\frac{\dfrac{\mathrm{d}\boldsymbol{t}}{\mathrm{d}s}}{\left|\dfrac{\mathrm{d}\boldsymbol{t}}{\mathrm{d}s}\right|} = \boldsymbol{n} \tag{7}$$

is called the *principal normal (unit) vector* of the curve. The vector

$$\boldsymbol{b} = \boldsymbol{t} \times \boldsymbol{n} \tag{8}$$

is called the *binormal vector* of this curve. The vectors $\boldsymbol{t}$, $\boldsymbol{n}$, $\boldsymbol{b}$ are mutually orthogonal and form the so-called *moving trihedral* of the curve (for more details see § 9.3).

GRADIENT. By means of a function

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

a *scalar field* is given in the region O in which the function is defined. The surfaces $u = \text{const.}$ are the *level (equipotential) surfaces* of this scalar field.

Definition 1. The vector

$$\operatorname{grad} u = \frac{\partial u}{\partial x}\, \boldsymbol{i} + \frac{\partial u}{\partial y}\, \boldsymbol{j} + \frac{\partial u}{\partial z}\, \boldsymbol{k} \tag{9}$$

is said to be the *gradient* of the given scalar field.

REMARK 2. Thus, the gradient of a scalar field is a vector. *At a fixed point* $(x_0, y_0, z_0) \in O$, *this vector is perpendicular to the level surface passing through this point* .

An example of a scalar field is an electrostatic potential field. Its level surfaces are called *equipotential surfaces*. The gradient defines the *vector field* characterizing (at every point) the intensity of the given electrostatic field. The curves touching this field at every point (i.e. curves such that the gradient at every point of a curve is a tangent vector of this curve) are called the *lines of force*.

If a vector (vector field) $\boldsymbol{a}(x, y, z)$ is given in a region O and if there is a (univalent) function $u = f(x, y, z)$ in O such that this vector is the gradient of the function u in O, i.e.

$$\boldsymbol{a}(x, y, z) = \operatorname{grad} u(x, y, z) , \tag{10}$$

then this vector field is called *potential (conservative)* and u is the *scalar potential*. In a potential field the work done by the force $\mathbf{a} = \operatorname{grad} u$ along a curve c lying in O and connecting two points A, B of this field does not depend on the form of this curve. In particular, the work along a closed curve is zero:

$$\oint_c \operatorname{grad} u \,.\, \mathrm{d}\boldsymbol{s} = 0 \tag{11}$$

(see equation (7.3.4)); $\mathrm{d}\boldsymbol{s} = \boldsymbol{i}\,\mathrm{d}x + \boldsymbol{j}\,\mathrm{d}y + \boldsymbol{k}\,\mathrm{d}z$.

Theorem 2. *The relation*

$$\mathrm{d}u = \operatorname{grad} u \,.\, \mathrm{d}\boldsymbol{s} \tag{12}$$

holds.

Roughly speaking (replacing the increment by the differential): *The increment of the potential along the path characterized by a small vector* $\mathrm{d}\boldsymbol{s}$ *is given by the scalar product* (12).

REMARK 3. At a fixed point and for a fixed length $\mathrm{d}s$ of the vector $\mathrm{d}\boldsymbol{s}$ the increment (differential, more precisely) of the potential u is (in accordance to Theorem 7.1.9) the greatest in the direction of the gradient. Thus, *the gradient determines at every point the greatest descent in the field.*

Theorem 3.

$$\operatorname{grad}(u_1 + u_2 + \ldots + u_n) = \operatorname{grad} u_1 + \operatorname{grad} u_2 + \ldots + \operatorname{grad} u_n , \tag{13}$$

$$\operatorname{grad}(uv) = u \operatorname{grad} v + v \operatorname{grad} u , \tag{14}$$

$$\operatorname{grad} r = \frac{\boldsymbol{r}}{r} = \hat{\boldsymbol{r}} \quad (\boldsymbol{r} \text{ is the radius vector of the point } (x, y, z), \hat{\boldsymbol{r}} \text{ is the unit vector in the direction of } \mathbf{r}), \tag{15}$$

$$\operatorname{grad} f(u) = f'(u) \operatorname{grad} u ; \tag{16}$$

in particular

$$\operatorname{grad}\frac{1}{r} = -\frac{1}{r^2}\operatorname{grad} r = -\frac{\boldsymbol{r}}{r^3} \tag{17}$$

(*field of force of a unit charge lying at the origin of the coordinate system*);

$$|\operatorname{grad} u| = \sqrt{\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial u}{\partial z}\right)^2\right]}. \tag{18}$$

REMARK 4. For the gradient the following notation is used:

$$\operatorname{grad} u = \nabla u = \left(\boldsymbol{i}\frac{\partial}{\partial x} + \boldsymbol{j}\frac{\partial}{\partial y} + \boldsymbol{k}\frac{\partial}{\partial z}\right)u, \tag{19}$$

where

$$\nabla = \boldsymbol{i}\frac{\partial}{\partial x} + \boldsymbol{j}\frac{\partial}{\partial y} + \boldsymbol{k}\frac{\partial}{\partial z} \tag{20}$$

is the so-called *Hamilton nabla operator*, often called "del". Formula (12) is then written in the form

$$\mathrm{d}u = \nabla u \,.\, \mathrm{d}\boldsymbol{s}\,. \tag{21}$$

DIVERGENCE AND CURL OF A VECTOR FIELD. Consider a vector field given by the vector

$$\boldsymbol{a}(x, y, z) = a_1(x, y, z)\,\boldsymbol{i} + a_2(x, y, z)\,\boldsymbol{j} + a_3(x, y, z)\,\boldsymbol{k} \tag{22}$$

which is thus a vector function of a point (x, y, z).

Definition 2. The *divergence of the vector* $\boldsymbol{a}$ is the scalar

$$\operatorname{div}\boldsymbol{a} = \nabla\boldsymbol{a} = \frac{\partial a_1}{\partial x} + \frac{\partial a_2}{\partial y} + \frac{\partial a_3}{\partial z}. \tag{23}$$

In English mathematical literature it is more common to write $\nabla\,.\,\boldsymbol{a}$, rather than $\nabla\boldsymbol{a}$, for div $\boldsymbol{a}$.

REMARK 5. Let us consider a steady flow of fluid characterized (in a region O) by a velocity vector $\boldsymbol{a}(x, y, z)$. The divergence of the vector $\boldsymbol{a}$ measures the (volume) quantity of fluid produced in a unit volume in unit time.

For an incompressible fluid div $\boldsymbol{a} = 0$. Such a vector field (i.e. a field for which div $\boldsymbol{a} = 0$) is called *solenoidal* (*sourceless*). The flux of such a field through a closed surface equals zero; the quantity leaving the surface is the same as that entering it (see equation (7.3.7)).

Theorem 4. *The following relations hold:*

$$\operatorname{div}(\boldsymbol{a} + \boldsymbol{b}) = \operatorname{div}\boldsymbol{a} + \operatorname{div}\boldsymbol{b}\,, \quad \operatorname{div}(u\boldsymbol{a}) = u \operatorname{div}\boldsymbol{a} + \boldsymbol{a} \operatorname{grad} u\,, \tag{24}$$

$$\operatorname{div}\boldsymbol{r} = 3\,, \quad \operatorname{div}\hat{\boldsymbol{r}} = \frac{2}{r}\,, \quad \operatorname{div}\frac{\hat{\boldsymbol{r}}}{r^2} = 0\,, \tag{25}$$

where $\hat{\boldsymbol{r}} = \dfrac{\boldsymbol{r}}{r}$ (*$\boldsymbol{r}$ is the radius vector of the point* (x, y, z)).

Definition 3. The *curl of a vector* $\boldsymbol{a}$ is the vector

$$\operatorname{curl}\boldsymbol{a} = \nabla \times \boldsymbol{a} = \left(\frac{\partial a_3}{\partial y} - \frac{\partial a_2}{\partial z}\right)\boldsymbol{i} + \left(\frac{\partial a_1}{\partial z} - \frac{\partial a_3}{\partial x}\right)\boldsymbol{j} + \left(\frac{\partial a_2}{\partial x} - \frac{\partial a_1}{\partial y}\right)\boldsymbol{k}$$

$$= \begin{vmatrix} \boldsymbol{i}, & \boldsymbol{j}, & \boldsymbol{k} \\ \dfrac{\partial}{\partial x}, & \dfrac{\partial}{\partial y}, & \dfrac{\partial}{\partial z} \\ a_1, & a_2, & a_3 \end{vmatrix}. \tag{26}$$

The symbol rot $\boldsymbol{a}$ is often used instead of curl $\boldsymbol{a}$.

REMARK 6. If $\boldsymbol{a}$ is the velocity of a fluid, then the direction of curl $\boldsymbol{a}$ indicates the direction of the axis about which the fluid rotates in a "small" neighbourhood of the point under consideration. The length of the vector $\frac{1}{2}$ curl $\boldsymbol{a}$ determines the speed of rotation (in circular measure).

Theorem 5. *The following relations hold:*

$$\operatorname{curl}(\boldsymbol{a} + \boldsymbol{b}) = \operatorname{curl}\boldsymbol{a} + \operatorname{curl}\boldsymbol{b}\,, \quad \operatorname{curl}(u\boldsymbol{a}) = u \operatorname{curl}\boldsymbol{a} - \boldsymbol{a} \times \operatorname{grad} u\,, \tag{27}$$

$$\operatorname{curl}\boldsymbol{r} = \boldsymbol{0}\,, \quad \operatorname{curl} f(r)\,\boldsymbol{r} = 0 \quad (\boldsymbol{r} \text{ is the radius vector of the point } (x, y, z))\,. \tag{28}$$

REMARK 7. The field in which curl $\boldsymbol{a} = \boldsymbol{0}$ holds is called *irrotational.* A vector field constructed as the gradient of a scalar field $u(x, y, z)$, is irrotational. Conversely, every irrotational field in a simply connected region can be represented as the gradient of a scalar field (and is thus a potential field).

REMARK 8. The scalar product of the operator ∇ with itself gives the so-called *Laplacian operator* Δ (*delta*):

$$\Delta = \nabla\nabla = \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\,. \tag{29}$$

Theorem 6. *The above-defined operations expressed in terms of*

a) *cylindrical (polar) coordinates* $x = \varrho \cos \varphi$, $y = \varrho \sin \varphi$, $z = z$ give

$$\operatorname{div} \boldsymbol{a} = \frac{1}{\varrho} \frac{\partial}{\partial \varrho}(\varrho a_\varrho) + \frac{1}{\varrho} \frac{\partial a_\varphi}{\partial \varphi} + \frac{\partial a_z}{\partial z}, \tag{30}$$

$$\Delta u = \frac{\partial^2 u}{\partial \varrho^2} + \frac{1}{\varrho} \frac{\partial u}{\partial \varrho} + \frac{1}{\varrho^2} \frac{\partial^2 u}{\partial \varphi^2} + \frac{\partial^2 u}{\partial z^2}; \tag{31}$$

the components of the vector grad u *in the directions of* ϱ, φ, z:

$$\frac{\partial u}{\partial \varrho}, \quad \frac{1}{\varrho} \frac{\partial u}{\partial \varphi}, \quad \frac{\partial u}{\partial z}: \tag{32}$$

the components of the vector curl $\boldsymbol{a}$ *in the directions of* ϱ, φ, z:

$$\frac{1}{\varrho} \frac{\partial a_z}{\partial \varphi} - \frac{\partial a_\varphi}{\partial z}, \quad \frac{\partial a_\varrho}{\partial z} - \frac{\partial a_z}{\partial \varrho}, \quad \frac{1}{\varrho} \frac{\partial}{\partial \varrho}(\varrho a_\varphi) - \frac{1}{\varrho} \frac{\partial a_\varrho}{\partial \varphi}; \tag{33}$$

b) *spherical coordinates* $x = r \sin \vartheta \cos \varphi$, $y = r \sin \vartheta \sin \varphi$, $z = r \cos \vartheta$ give

$$\operatorname{div} \boldsymbol{a} = \frac{1}{r^2} \frac{\partial}{\partial r}(r^2 a_r) + \frac{1}{r \sin \vartheta} \frac{\partial}{\partial \vartheta}(a_\vartheta \sin \vartheta) + \frac{1}{r \sin \vartheta} \frac{\partial a_\varphi}{\partial \varphi}, \tag{34}$$

$$\Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \vartheta^2} + \frac{1}{r^2} \cot \vartheta \frac{\partial u}{\partial \vartheta} + \frac{1}{r^2 \sin^2 \vartheta} \frac{\partial^2 u}{\partial \varphi^2}; \tag{35}$$

the components of the vector grad u *in the directions of* r, ϑ, φ:

$$\frac{\partial u}{\partial r}, \quad \frac{1}{r} \frac{\partial u}{\partial \vartheta}, \quad \frac{1}{r \sin \vartheta} \frac{\partial u}{\partial \varphi};$$

the components of the vector curl $\boldsymbol{a}$ *in the directions of* r, ϑ, φ:

$$-\frac{1}{r \sin \vartheta}\left[\frac{\partial a_\vartheta}{\partial \varphi} - \frac{\partial}{\partial \vartheta}(a_\varphi \sin \vartheta)\right], \quad -\frac{1}{r}\left[\frac{\partial}{\partial r}(r a_\varphi) - \frac{1}{\sin \vartheta} \frac{\partial a_r}{\partial \varphi}\right],$$
$$-\left[\frac{1}{r} \frac{\partial a_r}{\partial \vartheta} - \frac{1}{r} \frac{\partial}{\partial r}(r a_\vartheta)\right]. \tag{36}$$

Theorem 7 (*Some Formulae for Calculation with the Operators* ∇ *and* Δ).

1. $\nabla(uv) = \operatorname{grad}(uv) = u \operatorname{grad} v + v \operatorname{grad} u$,
2. $\nabla(u\boldsymbol{a}) = \operatorname{div}(u\boldsymbol{a}) = u \operatorname{div} \boldsymbol{a} + \boldsymbol{a} \operatorname{grad} u$,
3. $\nabla \times (u\boldsymbol{a}) = \operatorname{curl}(u\boldsymbol{a}) = u \operatorname{curl} \boldsymbol{a} - \boldsymbol{a} \times \operatorname{grad} u$,

4. $\nabla(\boldsymbol{a} \cdot \boldsymbol{b}) = \operatorname{grad}(\boldsymbol{a} \cdot \boldsymbol{b}) = \boldsymbol{a} \times \operatorname{curl} \boldsymbol{b} + \boldsymbol{b} \times \operatorname{curl} \boldsymbol{a} +$
$+ \left(a_1 \frac{\partial \boldsymbol{b}}{\partial x} + a_2 \frac{\partial \boldsymbol{b}}{\partial y} + a_3 \frac{\partial \boldsymbol{b}}{\partial z}\right) + \left(b_1 \frac{\partial \boldsymbol{a}}{\partial x} + b_2 \frac{\partial \boldsymbol{a}}{\partial y} + b_3 \frac{\partial \boldsymbol{a}}{\partial z}\right),$
5. $\nabla(\boldsymbol{a} \times \boldsymbol{b}) = \operatorname{div}(\boldsymbol{a} \times \boldsymbol{b}) = \boldsymbol{b} \cdot \operatorname{curl} \boldsymbol{a} - \boldsymbol{a} \cdot \operatorname{curl} \boldsymbol{b},$
6. $\nabla \times (\boldsymbol{a} \times \boldsymbol{b}) = \operatorname{curl}(\boldsymbol{a} \times \boldsymbol{b}) = \boldsymbol{a} \operatorname{div} \boldsymbol{b} - \boldsymbol{b} \operatorname{div} \boldsymbol{a} +$
$+ \left(b_1 \frac{\partial \boldsymbol{a}}{\partial x} + b_2 \frac{\partial \boldsymbol{a}}{\partial y} + b_3 \frac{\partial \boldsymbol{a}}{\partial z}\right) - \left(a_1 \frac{\partial \boldsymbol{b}}{\partial x} + a_2 \frac{\partial \boldsymbol{b}}{\partial y} + a_3 \frac{\partial \boldsymbol{b}}{\partial z}\right),$
7. $\nabla^2 u = \nabla\nabla u = \operatorname{div} \operatorname{grad} u = \Delta u,$
8. $\nabla \times (\nabla u) = \operatorname{curl} \operatorname{grad} u = \boldsymbol{0},$
9. $\nabla(\nabla \boldsymbol{a}) = \operatorname{grad} \operatorname{div} \boldsymbol{a} = \operatorname{curl} \operatorname{curl} \boldsymbol{a} + \Delta \boldsymbol{a},$
10. $\nabla(\nabla \times \boldsymbol{a}) = \operatorname{div} \operatorname{curl} \boldsymbol{a} = 0.$

Theorem 8 (*Some Properties of the Laplacian Operator*).

1. $\Delta(u + v) = \Delta u + \Delta v, \quad \Delta(uv) = u\,\Delta v + v\,\Delta u + 2 \operatorname{grad} u \cdot \operatorname{grad} v,$
2. $\Delta(\boldsymbol{a} + \boldsymbol{b}) = \Delta \boldsymbol{a} + \Delta \boldsymbol{b}, \quad \Delta \operatorname{grad} u = \operatorname{grad}(\Delta u), \quad \Delta \operatorname{curl} \boldsymbol{a} = \operatorname{curl} \Delta \boldsymbol{a},$
3. $\Delta \frac{1}{r} = 0.$

REMARK 9. In accordance with the definition of the gradient of a *scalar*, the divergence of a *vector* and the curl of a *vector*, we read, of course,

$$\nabla u = \operatorname{grad} u, \quad \nabla \boldsymbol{a} = \operatorname{div} \boldsymbol{a}, \quad \nabla \times \boldsymbol{a} = \operatorname{curl} \boldsymbol{a}$$

and not e.g. $\nabla u = \operatorname{div} u$, for the operator "divergence" can be applied to a *vector* only, not to a *scalar*, etc.

The formulae stated in Theorem 7 can often easily be formally deduced if we note that the operator ∇ is given in vector form by (20); for example,

$$\nabla^2 = \nabla \cdot \nabla = \left(\boldsymbol{i} \frac{\partial}{\partial x} + \boldsymbol{j} \frac{\partial}{\partial y} + \boldsymbol{k} \frac{\partial}{\partial z}\right) \cdot \left(\boldsymbol{i} \frac{\partial}{\partial x} + \boldsymbol{j} \frac{\partial}{\partial y} + \boldsymbol{k} \frac{\partial}{\partial z}\right) =$$

$$= \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} = \Delta.$$

Similarly, also

$$\nabla(u\boldsymbol{a}) = (\nabla u) \cdot \boldsymbol{a} + u\,\nabla \boldsymbol{a} = \operatorname{grad} u \cdot \boldsymbol{a} + u \operatorname{div} \boldsymbol{a}.$$

The last two expressions in formula 4 of Theorem 7 can be symbolically written as

$$\left(a_1 \frac{\partial \boldsymbol{b}}{\partial x} + a_2 \frac{\partial \boldsymbol{b}}{\partial y} + a_3 \frac{\partial \boldsymbol{b}}{\partial z}\right) + \left(b_1 \frac{\partial \boldsymbol{a}}{\partial x} + b_2 \frac{\partial \boldsymbol{a}}{\partial y} + b_3 \frac{\partial \boldsymbol{a}}{\partial z}\right) = (\boldsymbol{a}\nabla)\,\boldsymbol{b} + (\boldsymbol{b}\nabla)\,\boldsymbol{a},$$

where

$$\boldsymbol{a}\nabla = \boldsymbol{a}\,\mathrm{grad} = (a_1\boldsymbol{i} + a_2\boldsymbol{j} + a_3\boldsymbol{k}) \cdot \left(\boldsymbol{i}\frac{\partial}{\partial x} + \boldsymbol{j}\frac{\partial}{\partial y} + \boldsymbol{k}\frac{\partial}{\partial z}\right) =$$

$$= a_1\frac{\partial}{\partial x} + a_2\frac{\partial}{\partial y} + a_3\frac{\partial}{\partial z}, \text{ etc.}$$

The corresponding expressions in formula 6 can be written in a similar way.

7.3. Curvilinear and Surface Integrals of a Vector. Vector Notation for the Theorems of Stokes, Gauss and Green

Let c be a sectionally smooth oriented curve in space (see §14.1). Denote $\boldsymbol{t}\,ds$ by $d\boldsymbol{s}$, where s is arc-length and $\boldsymbol{t}$ the tangent unit vector at the point of the curve under consideration. We define

$$\int_c \boldsymbol{a} \cdot d\boldsymbol{s} = \int_c \boldsymbol{a} \cdot \boldsymbol{t}\,ds = \int_c (a_1\,dx + a_2\,dy + a_3\,dz)\,. \tag{1}$$

If c is a closed curve, we usually write

$$\oint_c \boldsymbol{a} \cdot d\boldsymbol{s} \tag{2}$$

(this is the *circulation of the vector* $\boldsymbol{a}$ *along the closed curve* c).

If the vector $\boldsymbol{a}$ denotes a force, then integral (1) represents the work done by this force along the curve c.

If $\boldsymbol{a} = \mathrm{grad}\,u$, then

$$\int_A^B \boldsymbol{a} \cdot d\boldsymbol{s} = u(B) - u(A)\,, \tag{3}$$

where A is the initial and B the end point of the curve. Thus, in a potential field the integral (1) depends on the initial and end points of the curve and not on the shape of the curve. In particular, the integral along a closed curve in a potential field $\boldsymbol{a}$ is equal to zero:

$$\oint_c \boldsymbol{a} \cdot d\boldsymbol{s} = \oint_c \mathrm{grad}\,u \cdot d\boldsymbol{s} = 0\,. \tag{4}$$

If a vector (vector field) $\boldsymbol{a}$ is irrotational in a simply connected region O, i.e. if $\mathrm{curl}\,\boldsymbol{a} = \boldsymbol{0}$ in O, then the vector $\boldsymbol{a}$ can be expressed as the gradient of a scalar u (cf. Remark 7.2.7); the integral (1) does not depend on the shape of the curve but merely on its initial and end points; the integral along a closed curve is equal to zero.

In a similar way the surface integral of a vector $\boldsymbol{a}$ (over a sectionally smooth oriented surface, with unit normal vector $\boldsymbol{n}$) can be defined:

$$\iint_S \boldsymbol{a} \,.\, \mathrm{d}\boldsymbol{S} = \int_S \boldsymbol{a} \,.\, \boldsymbol{n}\, \mathrm{d}S\,. \tag{5}$$

In vector notation some theorems of integral calculus (see § 14·8) can be written in a simple form:

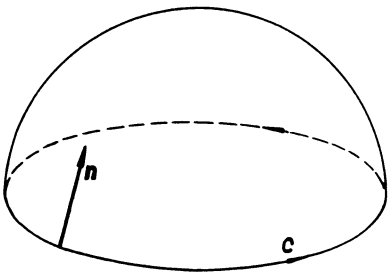


Fig. 7.4.

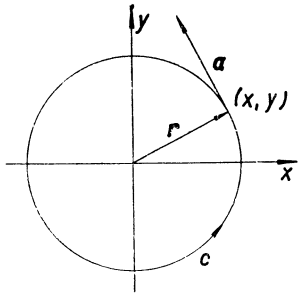


Fig. 7.5.

1. *Stokes's Theorem*:

$$\iint_S \operatorname{curl} \boldsymbol{a} \,.\, \mathrm{d}\boldsymbol{S} = \oint_c \boldsymbol{a} \,.\, \mathrm{d}\boldsymbol{s} \tag{6}$$

where S is a surface bounded by a curve c (Fig. 7.4); the orientation of the surface and the curve can be seen in Fig. 7.4 (see also Theorem 14.8.6).

The physical interpretation of equation (6): The flux of the vector curl $\boldsymbol{a}$ through a surface S equals the circulation of the vector $\boldsymbol{a}$ along the bounding curve c of S.

Example 1. Consider a two dimensional vector field (in the xy plane) given by the vector

$$\boldsymbol{a} = -y\boldsymbol{i} + x\boldsymbol{j}$$

(at every point (x, y) the vector $\boldsymbol{a}$ is perpendicular to the corresponding radius vector $\boldsymbol{r}$ of this point and its length is r (see Fig. 7.5)). Let c be the circle of radius r with centre at the origin, positively oriented with respect to its interior S. Thus, the normal $\boldsymbol{n}$ of the surface S is directed upwards (in the positive sense of the z-axis). Evidently

$$\int_c \boldsymbol{a} \,.\, \mathrm{d}\boldsymbol{s} = r \,.\, 2\pi r = 2\pi r^2\,.$$

Further

$$\operatorname{curl} \boldsymbol{a} = \begin{vmatrix} \boldsymbol{i}, & \boldsymbol{j}, & \boldsymbol{k} \\ \dfrac{\partial}{\partial x}, & \dfrac{\partial}{\partial y}, & \dfrac{\partial}{\partial z} \\ -y, & x, & 0 \end{vmatrix} = 2\boldsymbol{k}\,,$$

and thus

$$\iint_S \operatorname{curl} \boldsymbol{a} \,.\, \mathrm{d}\boldsymbol{S} = 2 \,.\, \pi r^2 = 2\pi r^2$$

in accordance with (6).

If $\operatorname{curl} \boldsymbol{a} = \boldsymbol{0}$, then equation (6) yields $\oint_c \boldsymbol{a} \,.\, \mathrm{d}\boldsymbol{s} = 0$ in accordance with (4).

2. *Gauss's Theorem*:

$$\iiint_V \operatorname{div} \boldsymbol{a} \, \mathrm{d}V = \iint_S \boldsymbol{a} \,.\, \mathrm{d}\boldsymbol{S}\,, \tag{7}$$

where the integral on the right-hand side is the surface integral over a closed surface whose interior is V; $\mathrm{d}\boldsymbol{S} = \boldsymbol{n}\,\mathrm{d}S$, *where* $\boldsymbol{n}$ *is the outward normal vector* (see Theorem 14.8.5).

Physical meaning: The flux of a vector $\boldsymbol{a}$ through a closed surface is equal to the volume integral of the divergence of the vector $\boldsymbol{a}$.

Similarly, the relations

$$\iiint_V \operatorname{grad} u \, \mathrm{d}V = \iint_S u \, \mathrm{d}\boldsymbol{S}\,, \quad \iiint_V \operatorname{curl} \boldsymbol{a} \, \mathrm{d}V = -\iint_S \boldsymbol{a} \times \mathrm{d}\boldsymbol{S} \tag{8}$$

hold.

REMARK 1. On the basis of the relations (7), (8), the operators grad, div, curl can be defined, without use of a special coordinate system:

$$\operatorname{grad} u = \lim_{V \to 0} \frac{1}{V} \iint_S u \, \mathrm{d}\boldsymbol{S}\,, \quad \operatorname{div} \boldsymbol{a} = \lim_{V \to 0} \frac{1}{V} \iint_S \boldsymbol{a} \,.\, \mathrm{d}\boldsymbol{S}\,, \tag{9}$$

$$\operatorname{curl} \boldsymbol{a} = -\lim_{V \to 0} \frac{1}{V} \iint_S \boldsymbol{a} \times \mathrm{d}\boldsymbol{S}\,. \tag{10}$$

3. *Green's Theorems*:

$$\iiint_V (\operatorname{grad} u \,.\, \operatorname{grad} v) \, \mathrm{d}V + \iiint_V u \, \Delta v \, \mathrm{d}V = \iint_S u \frac{\partial v}{\partial n} \, \mathrm{d}S\,, \tag{11}$$

$$\iiint_V (u \, \Delta v - v \, \Delta u) \, \mathrm{d}V = \iint_S \left(u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \right) \mathrm{d}S\,, \tag{12}$$

where n is the outward unit normal (see Theorem 14.8.9). If we put $v = u$ in (11) and if, moreover, u is harmonic ($\Delta u = 0$), then

$$\iiint_V (\nabla u)^2 \, dV = \iiint_V \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial u}{\partial z}\right)^2 \right] dV = \iint_V u \frac{\partial u}{\partial n} \, dS . \tag{13}$$

4. *Let the point $Q(x_0, y_0, z_0)$ be inside S, u be harmonic inside S, $\partial u/\partial n$ continuously extensible on S and $r = \sqrt{[(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2]}$. Then*

$$\iint_S \left[u \frac{\partial}{\partial n}\left(\frac{1}{r}\right) - \frac{1}{r}\frac{\partial u}{\partial n} \right] dS = -4\pi u_0 , \tag{14}$$

where u_0 is the value of the function u at the point Q.

8. TENSOR CALCULUS

By VÁCLAV VILHELM

References: [16], [23], [35], [42], [64], [90], [114], [116], [128], [154], [157], [161], [166], [171], [174], [181].

8.1. Contravariant and Covariant Coordinates of a Vector and their Transformation by a Change of the Coordinate System

If $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ are three arbitrary non-coplanar vectors then they define a coordinate system $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ in space in the sense that every vector $\boldsymbol{a}$ can be written uniquely in the form

$$\boldsymbol{a} = a^1\mathbf{e}_1 + a^2\mathbf{e}_2 + a^3\mathbf{e}_3\,, \tag{1}$$

where a^1, a^2, a^3 are real numbers. (We mention explicitly that a^i does *not* denote the i-th power of a but a number a^i with the upper index i.)

Definition 1. The numbers a^1, a^2, a^3 are called the *contravariant coordinates of the vector* $\boldsymbol{a}$ in the coordinate system $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$.

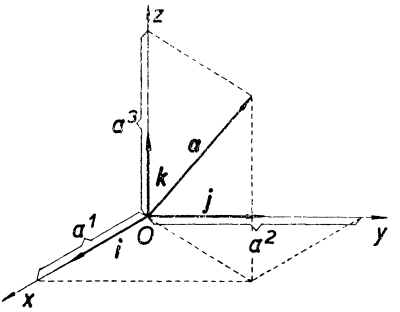


Fig. 8.1.

Example 1. Let us choose three mutually perpendicular unit vectors in space and denote them by $\mathbf{e}_1 = \boldsymbol{i}$, $\mathbf{e}_2 = \boldsymbol{j}$, $\mathbf{e}_3 = \boldsymbol{k}$ (Fig. 8.1). If $\boldsymbol{a}$ is an arbitrary vector, then $\boldsymbol{a} = a^1\boldsymbol{i} + a^2\boldsymbol{j} + a^3\boldsymbol{k} = a^i\mathbf{e}_i$ (see Remark 2); the numbers a^1, a^2, a^3 are its contravariant coordinates in the coordinate system $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k})$.

REMARK 1. From (1), a vector is obviously uniquely determined by its contravariant coordinates in the given coordinate system.

REMARK 2. Equation (1) can be written in the form $\mathbf{a} = \sum_{i=1}^{3} a^i \mathbf{e}_i$. It is customary in tensor algebra to adopt the "summation convention" by which we omit the sum symbol $\sum$ and write simply $\mathbf{a} = a^i \mathbf{e}_i$. In this convention it is understood that whenever an index is *repeated* (as in $a^i \mathbf{e}_i$) we sum over the values $i = 1, 2, 3$. Thus, $a^i \mathbf{e}_i$ stands for the sum $a^1 \mathbf{e}_1 + a^2 \mathbf{e}_2 + a^3 \mathbf{e}_3$. In what follows we shall normally use this brief notation.

If $\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3$ denote three other *non-coplanar* vectors in space, then

$$\begin{aligned} \mathbf{e}'_1 &= e_1^1 \mathbf{e}_1 + e_1^2 \mathbf{e}_2 + e_1^3 \mathbf{e}_3 , \\ \mathbf{e}'_2 &= e_2^1 \mathbf{e}_1 + e_2^2 \mathbf{e}_2 + e_2^3 \mathbf{e}_3 , \\ \mathbf{e}'_3 &= e_3^1 \mathbf{e}_1 + e_3^2 \mathbf{e}_2 + e_3^3 \mathbf{e}_3 , \end{aligned} \tag{2}$$

briefly this may be written as

$$\mathbf{e}'_i = e_i^j \mathbf{e}_j \quad (i = 1, 2, 3) ,$$

since j is a repeatable index.

Definition 2. The matrix $\mathbf{A} = (e_j^i)$ (the upper index refers to the columns, the lower to the rows) is called the *transformation matrix of the coordinate system* $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ *to the coordinate system* $(\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)$.

Theorem 1. *The determinant of the transformation matrix is different from zero; hence we may write*

$$\mathbf{e}_i = f_i^j \mathbf{e}'_j \quad (i = 1, 2, 3) , \tag{3}$$

where the matrix (f_j^i) *is the inverse of* (e_j^i) (see Example 2).

Theorem 2 (*Transformation of the Contravariant Coordinates of a Vector*). *If the contravariant coordinates of a vector* $\mathbf{a}$ *in the coordinate system* $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ *are* a^1, a^2, a^3 *and those in the coordinate system* $(\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)$ *are* a'^1, a'^2, a'^3, *then the following relation holds between these coordinates:*

$$a'^i = f_j^i a^j , \quad a^i = e_j^i a'^j \quad (i = 1, 2, 3) . \tag{4}$$

Here the matrix

$$\begin{bmatrix} e_1^1, & e_2^1, & e_3^1 \\ e_1^2, & e_2^2, & e_3^2 \\ e_1^3, & e_2^3, & e_3^3 \end{bmatrix}$$

is the transpose $\mathbf{A}'$ *of the transformation matrix* $\mathbf{A}$ *of the system* $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ *to the system* $(\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)$ *and the matrix*

$$\begin{bmatrix} f_1^1, & f_2^1, & f_3^1 \\ f_1^2, & f_2^2, & f_3^2 \\ f_1^3, & f_2^3, & f_3^3 \end{bmatrix}$$

is the inverse of the matrix $\mathbf{A}'$, *i.e.*

$$\sum_{i=1}^{3} e_j^i f_i^k = \delta_j^k \quad (\textit{briefly} \quad e_j^i f_i^k = \delta_j^k)\,, \tag{5}$$

where $\delta_j^k = \left\{\begin{matrix} 1 \textit{ for } k = j \\ 0 \textit{ for } k \neq j \end{matrix}\right\}$ (see Example 2).

REMARK 3. δ_j^k in Theorem 2 is known as the *Kronecker delta*.

Theorem 3. *Let the contravariant coordinates of vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *in the coordinate system* $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ *be* a^i *and* b^i, *respectively. Then, for the scalar product* $\mathbf{a} \,.\, \mathbf{b}$ *we have the relation*

$$\mathbf{a} \,.\, \mathbf{b} = \sum_{i=1}^{3} \sum_{j=1}^{3} g_{ij} a^i b^j$$

or briefly

$$\mathbf{a} \,.\, \mathbf{b} = g_{ij} a^i b^j\,, \tag{6}$$

where

$$g_{ij} = \mathbf{e}_i \,.\, \mathbf{e}_j \quad (i, j = 1, 2, 3)$$

(see Example 2 below).

Definition 3. The numbers

$$\sum_{j=1}^{3} g_{1j} a^j\,, \quad \sum_{j=1}^{3} g_{2j} a^j\,, \quad \sum_{j=1}^{3} g_{3j} a^j$$

(see Theorem 4) are called the *covariant coordinates* of the vector $\mathbf{a}$ in the coordinate system $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ and are denoted by a_1, a_2, a_3. Thus, $a_i = g_{ij} a^j$.

REMARK 4. Since the numbers a^1, a^2, a^3 can be determined from the equations $a_i = g_{ij} a^j$ uniquely (because the determinant of the matrix (g_{ij}) is different from zero), a vector is, according to Remark 1, uniquely determined by its covariant coordinates in the given coordinate system. In the system $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ of Example 1, $a^i = a_i$.

Theorem 4 (*Transformation of the Covariant Coordinates of a Vector*). *If the covariant coordinates of a vector* $\mathbf{a}$ *in the coordinate system* $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ *are* a_1, a_2, a_3 *and those in the coordinate system* $(\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)$ *are* a'_1, a'_2, a'_3, *the the following relation holds between these coordinates:*

$$a'_i = e_i^j a_j\,, \quad a_i = f_i^j a'_j\,; \tag{7}$$

the numbers e_i^j, f_i^j *having the same meaning as in Theorem 2.*

Consider the vector $\boldsymbol{a}$, the contravariant coordinates of which in the coordinate system $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_2)$ are a^1, a^2, a^3, i.e. $\boldsymbol{a} = a^i\mathbf{e}_i$. Hence, $\boldsymbol{a} = a^1(\mathbf{e}'_1 \cos\alpha - \mathbf{e}'_2 \sin\alpha) + a^2(\mathbf{e}'_1 \sin\alpha + \mathbf{e}'_2 \cos\alpha) + a^3 \frac{1}{2}\mathbf{e}'_3 = (a^1 \cos\alpha + a^2 \sin\alpha)\,\mathbf{e}'_1 + (-a^1 \sin\alpha + a^2 \cos\alpha)\,\mathbf{e}'_2 + \frac{1}{2}a^3\mathbf{e}'_3$. Thus, the contravariant coordinates of the vector $\boldsymbol{a}$ in the coordinate system $(\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)$ are

$$\begin{aligned} a'^1 &= a^1 \cos\alpha + a^2 \sin\alpha, \\ a'^2 &= -a^1 \sin\alpha + a^2 \cos\alpha, \\ a'^3 &= \tfrac{1}{2}a^3. \end{aligned}$$

The matrix of this transformation (see equation (4)) is

$$\begin{bmatrix} \cos\alpha, & \sin\alpha, & 0 \\ -\sin\alpha, & \cos\alpha, & 0 \\ 0, & 0, & \frac{1}{2} \end{bmatrix},$$

which is the inverse of the matrix $\boldsymbol{A}'$ (see Theorem 2).

The covariant coordinates of the vector $\boldsymbol{a}$ in the coordinate system $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ are (Theorem 3) $a_i = (\mathbf{e}_i \,.\, \mathbf{e}_j)\, a^j = \delta^i_j a^j = a^i$ (i.e. the same as the contravariant coordinates), while the covariant coordinates of the vector $\boldsymbol{a}$ in the coordinate system $(\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)$ are $a'_1 = (\mathbf{e}'_1 .\ \mathbf{e}'_j)\, a'^j = a'^1$, $a'_2 = (\mathbf{e}'_2 \,.\, \mathbf{e}'_j)\, a'^j = a'^2$, $a'_3 = (\mathbf{e}'_3 \,.\, \mathbf{e}'_j)\, a'^j = 4a'^3$ (since $\mathbf{e}'_i \,.\, \mathbf{e}'_j = 0$ for $i \neq j$, $\mathbf{e}'_1 \,.\, \mathbf{e}'_1 = \mathbf{e}'_2 \,.\, \mathbf{e}'_2 = 1$, $\mathbf{e}'_3 \,.\, \mathbf{e}'_3 = 4$).

The scalar product $\boldsymbol{a} \,.\, \boldsymbol{a}$ is (Theorem 5)

$$\boldsymbol{a} \,.\, \boldsymbol{a} = a^i a_i = (a_1)^2 + (a_2)^2 + (a_3)^2 = a'^i a'_i = (a^1 \cos\alpha + a^2 \sin\alpha)^2 + (-a^1 \sin\alpha + a^2 \cos\alpha)^2 + \tfrac{1}{2}a^3 \,.\, 2a^3 .$$

8.2. The Concept of a Tensor in Space

We have shown in § 1 that in every coordinate system, a vector is determined by an ordered triplet of numbers – by its contravariant or covariant coordinates. In changing from one coordinate system to another, this system of numbers defining the vector transforms in a certain way. The transformation formulae for contravariant and covariant coordinates are different (see Theorem 8.1.2 and Theorem 8.1.4). On the other hand, if to every coordinate system we assign three numbers a^1, a^2, a^3 or b_1, b_2, b_3 in such a way that when changing from one coordinate system to another, these numbers are transformed according to the formulae $a'^i = f^i_j a^j$ (or $b'_i = e^j_i b_j$), where (e^i_j) is the corresponding transformation matrix and (f^i_j) is the transpose of the inverse of the matrix (e^j_i), then these numbers can be understood to be the contravariant, or covariant coordinates of the vector $\boldsymbol{a}$, or $\boldsymbol{b}$, respectively. This follows from Theorems 8.1.2 and 8.1.4 and, thus, these numbers define the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. This idea is exploited in the following definition of a tensor.

Theorem 5. *Let a^i, b^i be the contravariant and a_i, b_i the covariant coordinates of vectors* $\boldsymbol{a}$, $\boldsymbol{b}$ *in the given coordinate system, respectively. Then the scalar product* $\boldsymbol{a} \,.\, \boldsymbol{b}$ *satisfies*

$$\boldsymbol{a} \,.\, \boldsymbol{b} = a^i b_i = a_i b^i$$

(where again, $a^i b_i = a^1 b_1 + a^2 b_2 + a^3 b_3$, $a_i b^i = a_1 b^1 + a_2 b^2 + a_3 b^3$; see Example 2).

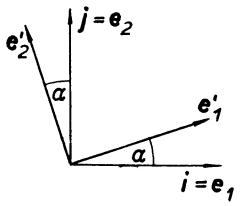


Fig. 8.2.

Example 2. Let $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = (\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k})$ be the coordinate system of Example 1 and let us choose another three linearly independent vectors $\mathbf{e}'_1$ $\mathbf{e}'_2$, $\mathbf{e}'_3$ such that the vector $\mathbf{e}'_i$ is obtained from the vector $\mathbf{e}_i$ $(i = 1, 2)$ by rotation through an angle α in the plane of the vectors $\mathbf{e}_1$, $\mathbf{e}_2$ (see Fig. 8.2), and $\mathbf{e}'_3 = 2\mathbf{e}_3$. Then equations (2) take the form

$$\begin{aligned} \mathbf{e}'_1 &= \mathbf{e}_1 \cos\alpha + \mathbf{e}_2 \sin\alpha\,, \\ \mathbf{e}'_2 &= -\mathbf{e}_1 \sin\alpha + \mathbf{e}_2 \cos\alpha\,, \\ \mathbf{e}'_3 &= 2\mathbf{e}_3\,. \end{aligned}$$

The transformation matrix of the coordinate system $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ to the coordinate system $(\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)$ is

$$\boldsymbol{A} = \begin{bmatrix} \cos\alpha, & \sin\alpha, & 0 \\ -\sin\alpha, & \cos\alpha, & 0 \\ 0, & 0, & 2 \end{bmatrix}.$$

We can easily show that equations (3) have the form

$$\begin{aligned} \mathbf{e}_1 &= \mathbf{e}'_1 \cos\alpha - \mathbf{e}'_2 \sin\alpha\,, \\ \mathbf{e}_2 &= \mathbf{e}'_1 \sin\alpha + \mathbf{e}'_2 \cos\alpha\,, \\ \mathbf{e}_3 &= \tfrac{1}{2}\mathbf{e}'_3 \end{aligned}$$

and, thus, the transformation matrix of the coordinate system $(\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)$ to the system $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ is

$$\boldsymbol{B} = \begin{bmatrix} \cos\alpha, & -\sin\alpha, & 0 \\ \sin\alpha, & \cos\alpha, & 0 \\ 0, & 0, & \frac{1}{2} \end{bmatrix},$$

i.e. the inverse of $\boldsymbol{A}$.

Definition 1. We say that a *tensor* is defined in space if, to every coordinate system, there correspond 3^{p+q} numbers $a^{ijk\ldots}_{rst\ldots}$ (the number of upper indices is p, the number of lower indices q) such that they are transformed according to the formulae

$$a'^{ijk\ldots}_{rst\ldots} = a^{lmn\ldots}_{uvw\ldots} e^u_r e^v_s e^w_t \ldots f^i_l f^j_m f^k_n \ldots \tag{1}$$

by any change from one coordinate system to another (in the right-hand side of formulae (1) we sum (from one to three) over all indices which appear twice there). Here, (e^i_j) is the transformation matrix and (f^i_j) the transpose of the inverse of the matrix (e^i_j). The tensor so defined is said to be *p-times contravariant* and *q-times covariant*. The number $p + q$ is called the *rank of the tensor*, the numbers $a^{ijk\ldots}_{rst\ldots}$ are called the *coordinates of the tensor*.

REMARK 1. Instead of "tensor of rank two" the term "*quadratic tensor*" is used. A quadratic tensor once covariant and once contravariant is called a *mixed quadratic tensor*. A tensor satisfying $q = 0$, or $p = 0$, is called a *contravariant*, or *covariant*, *tensor*, respectively.

Example 1 (*a scalar*). If to every coordinate system $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ there corresponds the same number a, a tensor of rank zero ($p = q = 0$), called a *scalar*, is defined.

Example 2 (*a contravariant vector*). If a^i are contravariant coordinates of a vector, then, by a change of the coordinate system, they are transformed according to the formulae $a'^i = f^i_j a^j$; this is a particular case of the formulae (1) for $p = 1$, $q = 0$. Thus a^i are the coordinates of a contravariant tensor of rank 1, called a *contravariant vector*.

Example 3 (*a covariant vector*). If a_i are covariant coordinates of a vector, then by a change of the coordinate system, they are transformed (see Theorem 8.1.4) according to formulae (1), where $p = 0$, $q = 1$. Thus, a_i are the coordinates of a covariant tensor of rank 1, called a *covariant vector*.

Example 4. The coordinates of a contravariant tensor a^{ij} are transformed, using (1), as follows:

$$a'^{ij} = a^{lm} f^i_l f^j_m .$$

Hence, in the transformation formulae for a contravariant (or covariant) tensor only the elements of the matrix (f^i_j) (or (e^i_j)) appear.

Example 5 (*a metric tensor of the space*). If to every coordinate system $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ we assign the numbers $g_{ij} = \mathbf{e}_i . \mathbf{e}_j$ (see Theorem 8.1.3), we can easily check that these numbers are transformed, by a change of the coordinate system, according to the formulae $g'_{ij} = g_{lm} e^l_i e^m_j$. Thus, g_{ij} are the coordinates of a double covariant tensor of rank 2 (i.e. a quadratic double covariant tensor), called the (*covariant*) *metric*

tensor. The coordinates g_{ij} can be written in the form of a matrix

$$\begin{bmatrix} g_{11}, & g_{12}, & g_{13} \\ g_{21}, & g_{22}, & g_{23} \\ g_{31}, & g_{32}, & g_{33} \end{bmatrix}.$$

If

$$\begin{bmatrix} g^{11}, & g^{12}, & g^{13} \\ g^{21}, & g^{22}, & g^{23} \\ g^{31}, & g^{32}, & g^{33} \end{bmatrix}$$

is its inverse (i.e. $g_{ij}g^{jk} = \delta_j^k$ – see Remark 8.1.3), then the numbers g^{ij} are the coordinates of a double contravariant tensor of rank 2, called the (*contravariant*) *metric tensor.* If the contravariant (or covariant) coordinates of vectors $\boldsymbol{a}$, $\boldsymbol{b}$ in the given coordinate system are a^i, b^i (or a_i, b_i) and the covariant (or contravariant) coordinates of a metric tensor in this system are g_{ij} (or g^{ij}), then $\boldsymbol{a} \,.\, \boldsymbol{b} = g_{ij}a^ib^j = g^{ij}a_ib_j$. This justifies the term "metric tensor".

Example 6. If to every coordinate system we assign the numbers

$$\delta_j^i = \begin{cases} 1 \text{ for } i = j, \\ 0 \text{ for } i \neq j, \end{cases}$$

then $\delta_j'^i = \delta_s^r e_j^s f_r^i = e_j^r f_r^i = \delta_j^i$ (see Theorem 8.1.2). Thus, δ_j^i are the coordinates of a once covariant and once contravariant tensor of rank 2 (i.e. a mixed quadratic tensor). These coordinates are the same in all coordinate systems.

Example 7. Let us choose a coordinate system in space and assign to every vector $\boldsymbol{a}$, the contravariant coordinates of which are a^i, the vector $\boldsymbol{b}$, the contravariant coordinates b^i of which are defined by the equations

$$b^i = c_j^i a^j, \tag{2}$$

i.e.

$$\begin{aligned} b^1 &= c_1^1 a^1 + c_2^1 a^2 + c_3^1 a^3, \\ b^2 &= c_1^2 a^1 + c_2^2 a^2 + c_3^2 a^3, \\ b^3 &= c_1^3 a^1 + c_2^3 a^2 + c_3^3 a^3. \end{aligned}$$

If we change the given coordinate system to a new one in which the coordinates of the vector $\boldsymbol{a}$, or the vector $\boldsymbol{b}$, are a'^i, or b'^i, respectively, then the following relation between these coordinates holds:

$$b'^i = c_j'^i a'^j,$$

where $c_j'^i = c_r^s e_j^r f_s^i$ ((e_j^i) is the transformation matrix of the original coordinate system to the new one). Thus, c_j^i are the coordinates of a mixed quadratic tensor.

In particular, considering the so-called *small deformations of a solid* whereby the vector a^i is transformed into the vector $\bar{a}^i$, then relations (2) hold between the vector a^i and the vector $b^i = \bar{a}^i - a^i$. The coefficients c^i_j are the coordinates of the so-called *deformation tensor* (see Example 8.4.4, p. 256).

8.3. A Tensor on a Surface

Definition 1. Let π be a smooth surface in space defined by the radius vector $\mathbf{r}(u_1, u_2)$ (see equations (9.11.1), (9.11.6) where u, v are written instead of u_1, u_2).

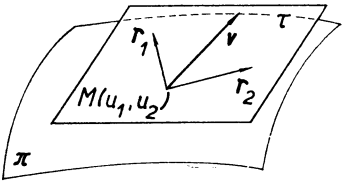


Fig. 8.3.

If, to every point M of the surface π, the coordinates of which are u_1, u_2, there corresponds a vector $\mathbf{v}(u_1, u_2)$ having initial point $M(u_1, u_2)$ and lying in the tangent plane of the surface at this point, we say that a (*tangent*) *vector field*, or briefly a (*tangent*) *vector* $\mathbf{v}(u_1, u_2)$ is given on the surface π (see Fig. 8.3).

REMARK 1. It is known (see § 9·12) that the vectors

$$\mathbf{r}_1(u_1, u_2) = \frac{\partial}{\partial u_1}\mathbf{r}(u_1, u_2),$$

$$\mathbf{r}_2(u_1, u_2) = \frac{\partial}{\partial u_2}\mathbf{r}(u_1, u_2)$$

lie in the tangent plane of the surface $\mathbf{r}(u_1, u_2)$ at the point $M(u_1, u_2)$ and are non-collinear. Therefore, they can be taken as the coordinate vectors in the tangent plane at the point $M(u_1, u_2)$. Every vector $\mathbf{v}(u_1, u_2)$ on the surface π can then be uniquely written in the form

$$\mathbf{v}(u_1, u_2) = v^1(u_1, u_2)\,\mathbf{r}_1(u_1, u_2) + v^2(u_1, u_2)\,\mathbf{r}_2(u_1, u_2), \tag{1}$$

briefly

$$\mathbf{v} = v^i\mathbf{r}_i.$$

Definition 2. $v^1(u_1, u_2)$, $v^2(u_1, u_2)$ (briefly v^i) are the so-called *contravariant coordinates of the vector* $\mathbf{v}(u_1, u_2)$ *on the surface* $\mathbf{r}(u_1, u_2)$ (in the local coordinate system $(\mathbf{r}_1(u_1, u_2), \mathbf{r}_2(u_1, u_2))$ of the point $M(u_1, u_2)$ of the surface).

REMARK 2. If $\mathbf{r}'(u_1', u_2')$ is another parametric expression of the surface π of Definition 1 in which the point M with original coordinates u_1, u_2 has coordinates u_1', u_2', then we shall always assume that u_1', u_2' are continuously differentiable functions of the variables u_1, u_2:

$$\begin{aligned} u_1' &= u_1'(u_1, u_2), \\ u_2' &= u_2'(u_1, u_2). \end{aligned} \tag{2}$$

Similarly, we shall assume that u_1, u_2 are continuously differentiable functions of the variables u_1', u_2';

$$\begin{aligned} u_1 &= u_1(u_1', u_2'), \\ u_2 &= u_2(u_1', u_2'). \end{aligned} \tag{3}$$

Here, equations (3) represent the solution of equations (2) with respect to the variables u_1, u_2. The determinant of the matrix

$$\begin{bmatrix} \dfrac{\partial u_1'}{\partial u_1}, & \dfrac{\partial u_1'}{\partial u_2} \\ \dfrac{\partial u_2'}{\partial u_1}, & \dfrac{\partial u_2'}{\partial u_2} \end{bmatrix} \quad \left(\text{briefly } \left(\frac{\partial u_i'}{\partial u_j}\right)\right)$$

is different from zero and the matrix $\left(\dfrac{\partial u_i}{\partial u_j'}\right)$ (formed by the partial derivatives of the functions (3)) is the inverse of the matrix $\left(\dfrac{\partial u_i'}{\partial u_j}\right)$, hence

$$\frac{\partial u_i'}{\partial u_j}\frac{\partial u_j}{\partial u_k'} = \delta_k^i, \quad \frac{\partial u_i}{\partial u_j'}\frac{\partial u_j'}{\partial u_k} = \delta_k^i,$$

where δ_k^i is the Kronecker delta (see Remark 8.1.3).

Theorem 1. *If we transform the parametric expression* $\mathbf{r}(u_1, u_2)$ *of a surface to a new parametric expression* $\mathbf{r}'(u_1', u_2')$ *by means of equations* (3), *then the following relations hold between the local coordinate vectors* $\mathbf{r}_1(u_1, u_2)$, $\mathbf{r}_2(u_1, u_2)$ *in the original expression and the local coordinate vectors* $\mathbf{r}_1'(u_1', u_2')$, $\mathbf{r}_2'(u_1', u_2')$:

$$\mathbf{r}_i'(u_1', u_2') = \sum_{j=1}^{2} \frac{\partial u_j}{\partial u_i'}(u_1', u_2')\, \mathbf{r}_j(u_1, u_2) \quad \left(\textit{briefly } \mathbf{r}_i' = \frac{\partial u_j}{\partial u_i'}\, r_j\right), \tag{4}$$

$$\mathbf{r}_i = \frac{\partial u_j'}{\partial u_i}\, \mathbf{r}_j'. \tag{5}$$

REMARK 3. From (4), we see that the transformation matrix of the coordinate system $(\mathbf{r}_1, \mathbf{r}_2)$ to the coordinate system $(\mathbf{r}'_1, \mathbf{r}'_2)$ is $\left(\frac{\partial u_j}{\partial u'_i}\right)$ (cf. Definition 8.1.2).

Theorem 2 (*Transformation of the Contravariant Coordinates of a Vector on a Surface*). *If the contravariant coordinates of a tangent vector* $\mathbf{v}$ *at a point* M *on a surface* π *in the local coordinate system* $(\mathbf{r}_1, \mathbf{r}_2)$ *are* v^1, v^2 *and those in the local coordinate system* $(\mathbf{r}'_1, \mathbf{r}'_2)$ (*which has resulted from the original system by the change of the parametric expression of the surface defined by equations* (2) *and* (3)) *are* v'^1, v'^2 *then*

$$v'^i = \frac{\partial u'_i}{\partial u_j} v^j, \quad v^i = \frac{\partial u_i}{\partial u'_j} v'^j.$$

REMARK 4. The coordinates v^1, v^2 in Theorem 2 are naturally functions of the variables u_1, u_2; similarly the coordinates v'^1, v'^2 are functions of the variables u'_1, u'_2.

Theorem 3. *Let the contravariant coordinates of vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *on a surface in the local coordinate system* $(\mathbf{r}_1, \mathbf{r}_2)$ *be* a^i *and* b^i, *respectively. Then*

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{2} \sum_{j=1}^{2} g_{ij} a^i b^j = g_{ij} a^i b^j, \tag{6}$$

where $g_{ij} = g_{ij}(u_1, u_2) = \mathbf{r}_i(u_1, u_2) \cdot \mathbf{r}_j(u_1, u_2)$.

Definition 3. The numbers $a_i = g_{ij} a^j$ are said to be the *covariant coordinates of the vector* $\mathbf{a}$ in the coordinate system $(\mathbf{r}_1, \mathbf{r}_2)$.

Theorem 4 (*Transformation of the Covariant Coordinates of a Vector on a Surface*). *If the covariant coordinates of a vector* $\mathbf{v}$ *on a surface* $\mathbf{r}(u_1, u_2)$ *in the local coordinate system* $(\mathbf{r}_1, \mathbf{r}_2)$ *are* v_i *and those in the local coordinate system* $(\mathbf{r}'_1, \mathbf{r}'_2)$ (*which has resulted from the original system by the change of the parametric expression of the surface according to equations* (2) *and* (3)) *are* v'_i, *then*

$$v'_i = \frac{\partial u_j}{\partial u'_i} v_j, \quad v_i = \frac{\partial u'_j}{\partial u_i} v'_j.$$

Theorem 5. *Let* a^i, b^i *be the contravariant and* a_i, b_i *the covariant coordinates of vectors* $\mathbf{a}$, $\mathbf{b}$ *on a surface* $\mathbf{r}(u_1, u_2)$, *respectively. Then* $\mathbf{a} \cdot \mathbf{b} = a^i b_i = a_i b^i$ (*here,* i *runs from* 1 *to* 2).

Definition 4 (*Definition of a Tensor on a Surface*; cf. Definition 8.2.1). We say a *tensor field* (briefly a *tensor*) is defined on a surface π if, to every local coordinate system $(\mathbf{r}_1(u_1, u_2), \mathbf{r}_2(u_1, u_2))$ defined by the corresponding parametric expression $\mathbf{r}(u_1, u_2)$ of the surface π, there correspond 2^{p+q} numbers (depending on the point

of the surface) $a^{ij\ldots}_{kl\ldots}$ (the number of upper indices is p, the number of lower indices q; $i, j, k, l, \ldots = 1, 2$) such that they are transformed according to the formulae

$$a'^{ij\ldots}_{rs\ldots} = a^{lm\ldots}_{tv\ldots} \frac{\partial u'_i}{\partial u_l} \frac{\partial u'_j}{\partial u_m} \ldots \frac{\partial u_t}{\partial u'_r} \frac{\partial u_v}{\partial u'_s} \ldots \tag{7}$$

by any change from the coordinate system $(\mathbf{r}_1(u_1, u_2), \mathbf{r}_2(u_1, u_2))$ to the coordinate system $(\mathbf{r}'_1(u'_1, u'_2), \mathbf{r}'_2(u'_1, u'_2))$ which has resulted from the original system by the change of the parametric expression of the surface according to equations (2) and (3). This tensor is said to be p-times contravariant and q-times covariant. The number $p + q$ is called the *rank of the tensor*, the numbers $a^{ij\ldots}_{rs\ldots}$ are called the *coordinates of the tensor.*

REMARK 5. The coordinates $a^{ij\ldots}_{rs\ldots}$ of a tensor on a surface π in the local coordinate system $(\mathbf{r}_1, \mathbf{r}_2)$ defined by the parametric expression $\mathbf{r}(u_1, u_2)$ of the surface π are evidently functions of the variables u_1, u_2 (see Remark 4).

Example 1 (*A Scalar on a Surface*). If to every point of a surface π there corresponds a certain fixed number a, then a tensor field of rank zero ($p = q = 0$), called a *scalar field*, briefly a *scalar*, is determined. In the coordinate system defined by the parametric expression $\mathbf{r}(u_1, u_2)$ of the surface π, a is a function of the variables u_1, u_2: $a = a(u_1, u_2)$. For a different expression $\mathbf{r}'(u'_1, u'_2)$ of the surface π in which the point with original curvilinear coordinates u_1, u_2 has curvilinear coordinates u'_1, u'_2, we naturally have $a = a(u'_1, u'_2) = a(u_1, u_2)$.

Example 2 (*A Contravariant and Covariant Vector on a Surface*). Let v^i and v_i be respectively the contravariant and covariant coordinates of a vector $\mathbf{v}$ on a surface. Then, comparing the transformation formulae of Theorems 2 and 4 and the formulae of Definition 4, v^i and v_i are easily seen to be the coordinates of a once contravariant, or a once covariant tensor of rank 1, the so-called *contravariant*, or *covariant, vector on a surface*, respectively.

Example 3 (*The Metric Tensor of a Surface* (*the first fundamental tensor of a surface*)). If we assign to every local coordinate system $(\mathbf{r}_1, \mathbf{r}_2)$ defined by the expression $\mathbf{r}(u_1, u_2)$ the numbers $g_{ij}(u_1, u_2) = \mathbf{r}_1(u_1, u_2) \,.\, \mathbf{r}_2(u_1, u_2)$, we easily verify that these numbers under any change of the coordinate system satisfy the transformation formulae

$$g'_{ij}(u'_1, u'_2) = \frac{\partial u_a}{\partial u'_i} \frac{\partial u_b}{\partial u'_j} g_{ab}(u_1, u_2) \,.$$

Hence, g_{ij} are the coordinates of a twice covariant quadratic tensor, the so-called (*covariant*) *metric* (or the *first fundamental*) *tensor of the surface.* The determinant of the matrix (g_{ij}) is different from zero and thus there exists the inverse matrix

(g^{ij}) (i.e. $g_{ij}g^{jk} = \delta_i^k$; see Remark 8.1.3). The numbers $g^{ij}(u_1, u_2)$ are the coordinates of a twice contravariant quadratic tensor, the so-called *(contravariant) metric tensor of the surface.* If a^i, b^i, or a_i, b_i, are the contravariant, or covariant, coordinates, respectively, of the vectors $\boldsymbol{a}$, $\boldsymbol{b}$ on a surface in the coordinate system in which the coordinates of the metric tensor are g_{ij}, then $\boldsymbol{a} \cdot \boldsymbol{b} = g_{ij}a^i b^j = g^{ij}a_i b_j$, $\boldsymbol{a} \cdot \boldsymbol{a} = g_{ij}a^i a^j = g^{ij}a_i a_j$. This justifies the term "metric tensor": by means of it we measure the lengths of vectors on a surface, and angles between them. The coordinates $g_{ij}(u_1, u_2)$ are called the *coefficients of the first fundamental form of the surface* (§ 9.14).

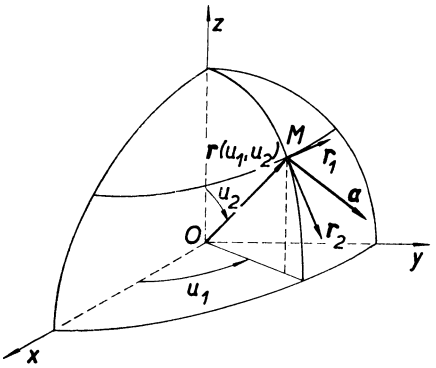


Fig. 8.4.

Example 4 (*The Second Fundamental Tensor of a Surface*). Let π be a surface given by the parametric expression $\boldsymbol{r}(u_1, u_2)$ and $h_{11}(u_1, u_2)\,du_1^2 + 2h_{12}(u_1, u_2) \cdot du_1\,du_2 + h_{22}(u_1, u_2)\,du_2^2$, briefly $h_{ij}\,du_i\,du_j$, be its second fundamental form (see § 9.15). Then h_{ij} are the coordinates of a quadratic twice covariant tensor, the so-called *second fundamental tensor of the surface.*

Example 5. In a rectangular coordinate system, let a spherical surface with centre at the origin of the coordinate system and radius r be given (see Fig. 8.4). Let us choose the parametric expression $\boldsymbol{r}(u_1, u_2)$ in such a way that the coordinates of $\boldsymbol{r}(u_1, u_2)$ are

$$x = r \cos u_1 \sin u_2\,,$$

$$y = r \sin u_1 \sin u_2\,, \quad 0 \leqq u_1 < 2\pi\,, \quad 0 < u_2 < \pi\,.$$

$$z = r \cos u_2\,,$$

The coordinates of the local coordinate vectors $\boldsymbol{r}_1(u_1, u_2)$, $\boldsymbol{r}_2(u_1, u_2)$ (see Remark 1) are

$$\boldsymbol{r}_1(u_1, u_2) = (-r \sin u_1 \sin u_2, r \cos u_1 \sin u_2, 0)\,,$$

$$\boldsymbol{r}_2(u_1, u_2) = (\ r \cos u_1 \cos u_2, r \sin u_1 \cos u_2, -r \sin u_2)\,.$$

The covariant coordinates $g_{ij}(u_1, u_2)$ of the metric tensor (see Example 3) are

$$g_{11} = \mathbf{r}_1 \cdot \mathbf{r}_1 = r^2 \sin^2 u_1 \sin^2 u_2 + r^2 \cos^2 u_1 \sin^2 u_2 = r^2 \sin^2 u_2 ,$$
$$g_{12} = g_{21} = \mathbf{r}_1 \cdot \mathbf{r}_2 = -r^2 \sin u_1 \sin u_2 \cos u_1 \cos u_2 +$$
$$+r^2 \cos u_1 \sin u_2 \sin u_1 \cos u_2 = 0 ,$$
$$g_{22} = \mathbf{r}_2 \cdot \mathbf{r}_2 = r^2 \cos^2 u_1 \cos^2 u_2 + r^2 \sin^2 u_1 \cos^2 u_2 + r^2 \sin^2 u_2 = r^2 .$$

The contravariant coordinates $g^{ij}(u_1, u_2)$ of the metric tensor satisfy $g_{ij}g^{jk} = \delta_i^k$, i.e.

$$g_{11}g^{11} + g_{12}g^{21} = g^{11}r^2 \sin^2 u_2 = 1 ,$$
$$g_{11}g^{12} + g_{12}g^{22} = g^{12}r^2 \sin^2 u_2 = 0 ,$$
$$g_{21}g^{11} + g_{22}g^{21} = g^{21}r^2 = 0 ,$$
$$g_{21}g^{12} + g_{22}g^{22} = g^{22}r^2 = 1 .$$

Thus, $g^{11} = 1/(r^2 \sin^2 u_2)$, $g^{12} = g^{21} = 0$, $g^{22} = 1/r^2$. Let a vector $\mathbf{a}$ at a point M of the spherical surface be given, the contravariant coordinates of which are a^1, a^2. Then its covariant coordinates are (see Definition 3)

$$a_1 = g_{1j}a^j = a^1 r^2 \sin^2 u_2 , \quad a_2 = g_{2j}a^j = a^2 r^2 .$$

The scalar product $\mathbf{a} \cdot \mathbf{a}$ has the form (see Example 3)

$$\mathbf{a} \cdot \mathbf{a} = g_{ij}a^i a^j = (a^1)^2 r^2 \sin^2 u_2 + (a^2)^2 r^2 .$$

8.4. Basic Algebraic Operations on Tensors

REMARK 1. By the term "tensor" we understand here both tensor in space and tensor on a surface. It is necessary to bear in mind that the indices of the coordinates of a tensor in space assume the values 1, 2, 3 while those of the coordinates of a tensor on a surface only the values 1, 2.

Definition 1 (*Equality of Tensors*). We say that *two tensors are equal* if they are both p-times contravariant and q-times covariant and their coordinates are equal in at least one coordinate system. (Then, the coordinates are equal in every coordinate system.)

Definition 2 (*Addition of Tensors*). If $a_{rs\ldots}^{ij\ldots}$, $b_{rs\ldots}^{ij\ldots}$ are the coordinates of two tensors of the same type (i.e. if both are p-times contravariant and q-times covariant), then the numbers

$$c_{rs\ldots}^{ij\ldots} = a_{rs\ldots}^{ij\ldots} + b_{rs\ldots}^{ij\ldots}$$

are the coordinates of a tensor which is said to be the *sum* of these tensors (and is of the same type).

Definition 3 (*Multiplication of Tensors*). If $a^{ij\cdots}_{pq\cdots}$ are the coordinates of a p_1-times contravariant and q_1-times covariant tensor and $b^{kl\cdots}_{rs\cdots}$ the coordinates of a p_2-times contravariant and q_2-times covariant tensor, then the numbers

$$c^{ij\ldots kl\ldots}_{pq\ldots rs\ldots} = a^{ij\cdots}_{pq\ldots} b^{kl\cdots}_{rs\ldots}$$

are the coordinates of a $(p_1 + p_2)$-times contravariant and $(q_1 + q_2)$-times covariant tensor which is said to be the *product* of these tensors.

Definition 4 (*Contraction of Tensors*). Let $a^{ij\cdots}_{rs\cdots}$ be the coordinates of a p-times contravariant and q-times covariant tensor. Consider the sums

$$c^{j\cdots}_{r\cdots} = \sum_i a^{ij\cdots}_{ir\cdots} = a^{ij\cdots}_{ir\cdots}.$$

Then $c^{j\cdots}_{r\cdots}$ are the coordinates of a $(p-1)$-times contravariant and $(q-1)$-times covariant tensor. The tensor $c^{j\cdots}_{r\cdots}$ is called a *contraction of the tensor* $a^{ij\cdots}_{rs\cdots}$. Contraction can be performed not only on the first upper and the first lower indices but also on arbitrary k upper and k lower indices. For example, the contraction of a tensor a^{ij}_{rs} performed on both upper and both lower indices is the scalar $a = a^{ij}_{ij} = \sum_i \sum_j a^{ij}_{ij}$.

Example 1. Let v^i be a contravariant vector, g_{ij} a (covariant) metric tensor. By multiplication, we get the tensor $g_{ij}v^k$ of rank three; the covariant vector $v_i = g_{ij}v^j$ is its contraction.

Definition 5 (*Lowering and Raising of Indices*). For every p-times contravariant and q-times covariant tensor $a^{ij\cdots}_{rs\cdots}$ a new $(p+q)$-times covariant tensor $a_{ij\ldots rs\ldots} = g_{ki}g_{lj\ldots}a^{kl\cdots}_{rs\cdots}$ can be constructed, where g_{ij} are the coordinates of the metric tensor. We say that the tensor $a_{ij\ldots rs\ldots}$ was obtained from the tensor $a^{ij\cdots}_{rs\cdots}$ by *lowering of indices*.

Similarly a new $(p+q)$-times contravariant tensor $a^{rs\ldots ij\cdots} = g^{rk}g^{sl}\ldots a^{ij\cdots}_{kl\ldots}$ can be constructed from a tensor $a^{ij\cdots}_{rs\cdots}$. The tensor $a^{rs\ldots ij\cdots}$ was obtained from $a^{ij\cdots}_{rs\cdots}$ by *raising of indices*.

REMARK 2. By raising some of the indices of a covariant tensor we again get a tensor; however, it is necessary to indicate those indices which have been raised. This may be done by means of dots which indicate the place of the raised indices, as illustrated in the following examples:

$$a^{\cdot i}_{jk} = g^{il}a_{ljk}, \quad a^{j}_{i\cdot k} = g^{lj}a_{ilk}, \quad a^{ij}_{\cdot\cdot k} = g^{il}g^{jp}a_{lpk}.$$

A similar notation is used when lowering indices, e.g. $a^{i\cdot k}_{\ j} = g_{lj}a^{ilk}$.

Example 2. By lowering the contravariant coordinates v^i of a vector we get its covariant coordinates $v_i = g_{ij}v^j$.

Definition 6. A tensor is said to be *symmetric with respect to given upper* (*or lower*) *indices* if its coordinates do not alter by an arbitrary permutation of these

indices. For example, a tensor a^l_{ijk} is symmetric with respect to the first two lower indices if $a^l_{ijk} = a^l_{jik}$.

Example 3. The metric tensor g_{ij} is symmetric for $g_{ij} = \mathbf{r}_i \cdot \mathbf{r}_j = \mathbf{r}_j \cdot \mathbf{r}_i = g_{ji}$. Also the tensor g^{ij} and the second fundamental tensor h_{ij} of a surface (see Example 8.3.4) are symmetric.

Definition 7. A tensor is said to be *skew-symmetric (alternating) with respect to a given group of upper (lower) indices* if the sign changes with every interchange of two arbitrary indices of the group. For example, a tensor a_{ij} is skew-symmetric if $a_{ij} = -a_{ji}$.

Definition 8 (*Operation of Symmetrization*). For every tensor, a tensor symmetric with respect to a given group of indices can be constructed. For example, by symmetrization of a tensor $a_{ijkl\ldots}$ with respect to the first three indices we get the tensor

$$a_{(ijk)l\ldots} = \frac{1}{3!}(a_{ijkl\ldots} + a_{ikjl\ldots} + a_{jikl\ldots} + a_{jkil\ldots} + a_{kijl\ldots} + a_{kjil\ldots}). \tag{1}$$

The tensor $a_{(ijk)}$ is the so-called *symmetric part of the tensor* a_{ijk}.

Definition 9 (*Operation of Skew-symmetrization*). For every tensor, a tensor skew-symmetric with respect to a given group of indices can be constructed. For example, by skew-symmetrization of a tensor $a_{ijkl\ldots}$ with respect to the first three indices we get the tensor

$$a_{[ijk]l\ldots} = \frac{1}{3!}(a_{ijkl\ldots} - a_{ikjl\ldots} - a_{jikl\ldots} + a_{jkil\ldots} + a_{kijl\ldots} - a_{kjil\ldots}). \tag{2}$$

(Here, we choose the plus sign with an even and minus with an odd permutation of the indices i, j, k.)

The tensor $a_{[ijk]}$ is the so-called *skew-symmetric part of the tensor* a_{ijk}.

REMARK 3. A quadratic tensor is the sum of its symmetric and skew-symmetric parts: $a_{ij} = a_{(ij)} + a_{[ij]}$.

Example 4. If $a_{ij} = c^k_j g_{ik}$ are the covariant coordinates of the deformation tensor of Example 8.2.7, then its symmetric part $a_{(ij)} = \frac{1}{2}(a_{ij} + a_{ji})$ is the so-called *tensor of a pure deformation*, its skew-symmetric part $a_{[ij]} = \frac{1}{2}(a_{ij} - a_{ji})$ is the so-called *tensor of rotation* (it represents, roughly speaking, the rotation of the body).

8.5. Symmetric Quadratic Tensors

Definition 1. On a surface defined by a parametric expression $\mathbf{r}(u_1, u_2)$, let a quadratic symmetric (non-zero) tensor be given. According to Definition 8.4.5 we can

assume that it is covariant and its coordinates are $a_{ij}(u_1, u_2)$ $(a_{ij} = a_{ji})$. Let us choose a point O on the surface whose curvilinear coordinates are u_1, u_2 and construct in the tangent plane at this point the locus of terminal points of vectors t^i on the surface with the initial point at O which satisfy whichever of the equations

$$a_{ij}t^it^j = a_{11}(t^1)^2 + 2a_{12}t^1t^2 + a_{22}(t^2)^2 = \pm 1\,. \tag{1}$$

This locus is called the *indicatrix of the tensor* a_{ij} *at the point* (u_1, u_2).

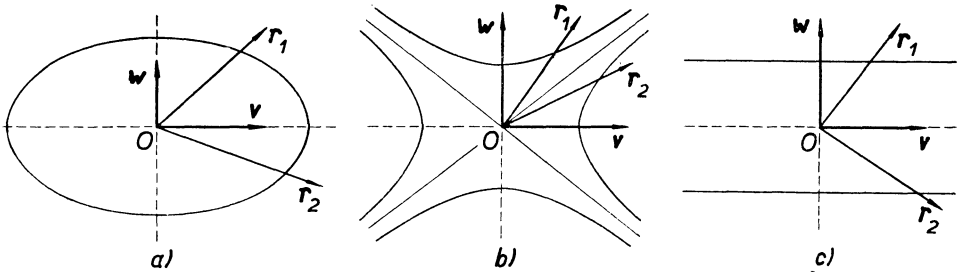


Fig. 8.5.

Theorem 1. *The indicatrix of the tensor* a_{ij} *at a point* O *is*
1. *an ellipse (or a circle), if the determinant*

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

is positive (i.e. the form $a_{ij}t^it^j$ *is definite)* (see Fig. 8.5a);

2. *a pair of hyperbolas with common asymptotes and centre at the point* O (see Fig. 8.5b), *if the determinant* (2) *is negative (i.e. the form* $a_{ij}t^it^j$ *is indefinite)*;

3. *a pair of parallel lines, if the determinant* (2) *equals zero* (see Fig. 8.5c)

REMARK 1. In what follows we shall restrict our considerations to the case when the determinant (2) is non-zero.

Definition 2. The directions of conjugate diameters of the indicatrix of a tensor are called *conjugate directions* of the tensor; the directions of the axes are the *principal directions* of the tensor.

Theorem 2. *Vectors* v^i, w^j *lie in conjugate directions of a tensor* a_{ij} *if and only if* $a_{ij}v^iw^j = 0$.

REMARK 2. *Determination of the principal directions of a quadratic tensor:*

1. If a_{ij} is a multiple of the metric tensor (i.e. $a_{ij} = \lambda g_{ij}$), then the indicatrix is a circle and any direction is principal;

2. Let a_{ij} not be a multiple of g_{ij}; let $\mathbf{v}$, $\mathbf{w}$ be vectors lying in the principal directions of the tensor a_{ij} (see Fig. 8.5a,b), v^i, w^i their contravariant coordinates. Then

$$\begin{aligned} a_{ij}v^iw^j &= 0\,, \\ \mathbf{v}\,.\,\mathbf{w} = g_{ij}v^iw^j &= 0\,. \end{aligned} \tag{3}$$

Re-write the equations (3) in the form

$$\begin{aligned} v^i(a_{ij}w^j) &= 0\,, \\ v^i(g_{ij}w^j) &= 0\,. \end{aligned}$$

In order that there exist a non-zero solution v^i, the determinant

$$\begin{vmatrix} a_{1j}w^j, & a_{2j}w^j \\ g_{1j}w^j, & g_{2j}w^j \end{vmatrix}$$

must be equal to zero, i.e. $a_{ij}w^j = \lambda g_{ij}w^j$ and thus

$$(a_{ij} - \lambda g_{ij})\, w^j = 0\,. \tag{4}$$

In order that equations (4) have a non-zero solution for w^j, it is necessary that

$$\begin{vmatrix} a_{11} - \lambda g_{11}, & a_{12} - \lambda g_{12} \\ a_{21} - \lambda g_{21}, & a_{22} - \lambda g_{22} \end{vmatrix} = 0\,. \tag{5}$$

The roots λ_1, λ_2 of the quadratic equation (5) are the so-called *characteristic numbers of the tensor*. Substituting them successively into equations (4) we can determine the required vectors w^i, v^i (see Example 1).

REMARK 3. For a quadratic symmetric tensor in space we can obtain results similar to those just introduced. However, the indicatrix is then a quadratic surface (or a pair of such surfaces) and there are, in general, three principal directions.

Example 1. Let the tensor of the membrane stresses in the middle surface of a spherical shell of radius r be given by the covariant coordinates $\sigma_{ij}(u_1, u_2)$, with respect to the coordinate system of Example 8.3.5. Let us find the directions of the principal stresses in the middle surface.

The tensor of the membrane stresses is symmetric and its principal directions coincide with the directions of the principal stresses. In order to determine the principal directions of the tensor σ_{ij} we substitute in equation (5) above (where $g_{11} = r^2 \sin^2 u_2$, $g_{12} = g_{21} = 0$, $g_{22} = r^2$), giving

$$\begin{vmatrix} \sigma_{11} - \lambda r^2 \sin^2 u_2, & \sigma_{12} \\ \sigma_{12}, & \sigma_{22} - \lambda r^2 \end{vmatrix} = 0\,,$$

i.e.

$$\lambda^2 r^4 \sin^2 u_2 - \lambda r^2(\sigma_{11} + \sigma_{22} \sin^2 u_2) + \sigma_{11}\sigma_{22} - \sigma_{12}^2 = 0\,.$$

Denoting the roots of this equation by λ_1, λ_2, we find vectors v^i, w^i lying in the principal directions, from equations $(a_{ij} - \lambda_1 g_{ij})v^j = 0$, $(a_{ij} - \lambda_2 g_{ij})w^j = 0$, i.e.

$$(\sigma_{11} - \lambda_1 r^2 \sin^2 u_2)\, v^1 + \sigma_{12} v^2 = 0\,, \quad (\sigma_{11} - \lambda_2 r^2 \sin^2 u_2)\, w^1 + \sigma_{12} w^2 = 0\,.$$

Thus, $v^1/v^2 = \sigma_{12}/(\lambda_1 r^2 \sin^2 u_2 - \sigma_{11})$ and similarly $w^1/w^2 = \sigma_{12}/(\lambda_2 r^2 \sin^2 u_2 - \sigma_{11})$. Here, the ratio v^1/v^2 (or w^1/w^2) represents the tangent of the angle which the principal direction (axis of the indicatrix) makes with the corresponding parallel line on the middle surface.

REMARK 4. In tensor calculus, tensors may also be introduced by means of the concept of a dyad (see e.g. [90]). For *tensor analysis (covariant derivative* etc.) see e.g. [181]).

9. DIFFERENTIAL GEOMETRY

By Bořivoj Kepr

References: [19], [45], [65], [104], [110], [132], [138], [153], [173].

9.1. Introduction

Differential geometry is the study of curves (both plane and space curves) and surfaces by means of the calculus. When investigating geometric configurations (on the basis of their equations) in differential geometry, we aim mostly at the study of *invariant properties*, i.e., properties independent of the choice of the coordinate system and so belonging directly to the curve or surface (e.g. the points of inflexion, the curvature and so on). But we also study those properties of geometric configurations that depend on the choice of the coordinate system (e.g. the sections of a surface by the coordinate planes, the slope of the tangent and so on). Differential geometry studies mostly the *local properties* of curves and surfaces, i.e. those which pertain to sufficiently small portions of the curve or the surface; so it is essentially "geometry in the small". But differential geometry also investigates those properties of curves and surfaces which pertain to the configuration as a whole (e.g. the length of a curve, the number of vertices and so on).

A. CURVES

9.2. Definition and Equations of a Curve, Length of Arc and Tangent Line

Definition 1. A *piecewise smooth space curve, defined parametrically*, is a set of points (x, y, z) given by the equations

$$x = x(t), \quad y = y(t), \quad z = z(t), \tag{1}$$

where the functions $x(t)$, $y(t)$, $z(t)$, defined in some interval I (most often in a closed interval $[a, b]$ or in the interval $(-\infty, +\infty)$),

1. are continuous in the interval I,

2. have, in I, piecewise continuous derivatives $\dot{x}(t)$, $\dot{y}(t)$, $\dot{z}(t)$ (we write $dx/dt = \dot{x}$, etc.), while with the exception of at most a finite number of points t_k from the interval I the relation $\dot{x}^2 + \dot{y}^2 + \dot{z}^2 > 0$ holds (i.e. at least one of the derivatives $\dot{x}$, $\dot{y}$, $\dot{z}$ is non-zero).

REMARK 1. If $I = [a, b]$ and if

$$x(a) = x(b), \quad y(a) = y(b), \quad z(a) = z(b)$$

the curve is said to be *closed*. If the functions $\dot{x}$, $\dot{y}$, $\dot{z}$ are continuous in I (cf. Definition 1) (and in the case of a closed curve the values of the right-hand derivatives of functions x, y, z at the point a and the left-hand derivatives of these functions at the point b are equal) and if everywhere in I (in the case of a closed curve also at the points a, b) $\dot{x}^2 + \dot{y}^2 + \dot{z}^2 > 0$, the curve is said to be *smooth. In the following text the word curve will stand for a smooth or piecewise smooth curve. As is customary in differential geometry, we shall suppose that the functions* $x(t)$, $y(t)$, $z(t)$ *also possess* (*continuous*) *derivatives of an order, r, higher than one* (*according as the problem under investigation may require*) *without stating explicitly this condition.*
(A similar remark holds for Definition 2.)

The argument t in equations (1) is called the *parameter of the curve* and equations (1) are called the *parametric equations of the curve.* Every number t from the interval I is called a *point of the curve* (to every value $t \in I$ there corresponds a point (x, y, z) on the curve), namely a *regular point* when $\dot{x}^2 + \dot{y}^2 + \dot{z}^2 > 0$ and when no other value of $t \in I$ corresponds to the considered point (x, y, z). (An exception may occur at the points a, b in the case of a closed curve.) Every other point is called a *singular point* of the curve.

REMARK 2. Whether a point t is a regular or singular point of a curve may, in the general case, depend on the chosen parametric representation of the curve. For the parabola, represented parametrically by the equations $x = t$, $y = t^2$, $z \equiv 0$, $t \in (-\infty, +\infty)$, the point $t = 0$ is not a singular point; but if it is represented parametrically by equations $x = t^3$, $y = t^6$, $z \equiv 0$, $t \in (-\infty, +\infty)$ (both pairs of equations define the same set of points in the cartesian coordinate system $(O; x, y, z)$), then the point $t = 0$ is a singular point with this parametric representation because $\dot{x}^2 + \dot{y}^2 + \dot{z}^2 = 0$. In these cases we speak of a *removable singular point* of the given curve.

REMARK 3. Of course, we obtain the curve as the set of all points (1) for every t from the interval I. If to several different values of t from I there corresponds a single point $P(x_0, y_0, z_0)$, then such a singular point is called a *multiple point* of the curve (Fig. 9.1).

In the following exposition the word "point" will mean a regular point of the curve. At such a point there exists a unique tangent.

REMARK 4. If $\boldsymbol{r}$ is the *radius vector* of a point on the curve (1) (where the coordinates are $x(t)$, $y(t)$, $z(t)$ and the starting point is at the origin), then we can use a single symbolic equation of the curve (the *vector equation*)

$$\boldsymbol{r} = \boldsymbol{r}(t) \equiv \boldsymbol{i}\, x(t) + \boldsymbol{j}\, y(t) + \boldsymbol{k}\, z(t) \tag{2}$$

where $\boldsymbol{i}$, $\boldsymbol{j}$ and $\boldsymbol{k}$ are the unit vectors along the positive axes; $\boldsymbol{r}$ is the so-called *radius vector* (the *position vector*) of the point (x, y, z) on the curve. As t runs through the interval I, the end point of the radius vector describes the given curve (Fig. 9.2).

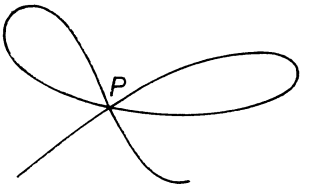


Fig. 9.1.

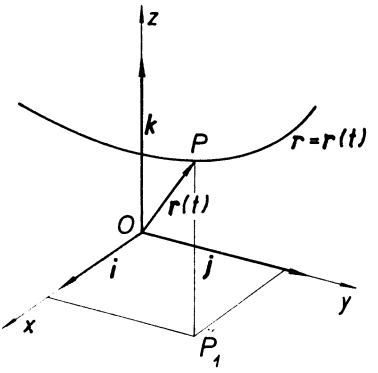


Fig. 9.2.

Example 1. For the *circular helix* (see Example 9.3.1)

$$x = a\cos t\,, \quad y = a\sin t\,, \quad z = bt\,, \quad t \in (-\infty, \infty)\,,$$

the vector equation is of the form

$$\boldsymbol{r} = \boldsymbol{i}a\cos t + \boldsymbol{j}a\sin t + \boldsymbol{k}bt\,, \quad t \in (-\infty, \infty)\,.$$

Definition 2. Let the functions $\varphi(x, y, z)$, $\psi(x, y, z)$, defined in the (three-dimensional) domain O, have continuous partial derivatives of the first order. By a *curve, defined in space implicitly by the equations*

$$\varphi(x, y, z) = 0\,, \quad \psi(x, y, z) = 0\,, \tag{3}$$

is meant the set of points whose cartesian coordinates (x, y, z) satisfy simultaneously both equations (3); we suppose that the matrix

$$\begin{bmatrix} \dfrac{\partial\varphi}{\partial x}\,, & \dfrac{\partial\varphi}{\partial y}\,, & \dfrac{\partial\varphi}{\partial z} \\[2ex] \dfrac{\partial\psi}{\partial x}\,, & \dfrac{\partial\psi}{\partial y}\,, & \dfrac{\partial\psi}{\partial z} \end{bmatrix} \tag{4}$$

is of rank 2 at every such point (x, y, z), with the exception of at most a finite number of points.

The curve so defined is the *intersection* of the surfaces (3). Equations (3) are called the *implicit equations of this curve.*

Example 2. The circle with centre at the origin and radius a which lies in the plane $z = x$ may be expressed as the intersection of the considered plane and a sphere with the centre at the origin in the form (3) as follows:

$$z - x = 0\,, \quad x^2 + y^2 + z^2 - a^2 = 0\,.$$

REMARK 5. If the equations (3) are algebraic (or if they can be transformed so as to become algebraic; for example, φ and ψ are polynomials), the curve is called *algebraic*; if the equations (3) are not algebraic (and if they cannot be transformed so that they become algebraic), the curve is said to be *transcendent.* If the functions (1) are rational functions of the variable t (or if they can be so reshaped), then the curve (1) is said to be *rational.* The parameter t is the *coordinate on the curve.* The form (1) is often used in mechanics in the study of the movement of a particle, and in this case t represents the time.

Definition 3. A curve (1) is said to be a *plane curve* if constants A, B, C, D (at least one of which is non-zero) can be found such that

$$A\,x(t) + B\,y(t) + C\,z(t) + D = 0 \tag{5}$$

holds for every t from I. (The whole curve lies in the plane $Ax + By + Cz + D = 0$.) In the contrary case the curve (1) is a *space curve* (a *skew* or *twisted curve*).

REMARK 6. The curve (3) is said to be a *plane curve* if at least one of equations (3) is the equation of a plane or may be replaced by the equation of a plane (cf. Example 2). Otherwise the curve (3) is a *space curve.*

Example 3. The curve

$$x = t\,, \quad y = t^2\,, \quad z = t^3 \tag{6}$$

is a space curve because, as is well known from algebra, the equation

$$At + Bt^2 + Ct^3 + D = 0$$

can be satisfied identically only in the case where $A = B = C = D = 0$.

If we eliminate t from equations (6), we obtain an implicit representation of the curve

$$y = x^2\,, \quad z = x^3\,.$$

Example 4. If the equations

$$x = t\,, \quad y = y(t)\,, \quad z = 0 \tag{7}$$

represent a curve, then it is a plane curve since the identity

$$At + B\,y(t) + C\,.\,0 + D = 0$$

is satisfied by $A = B = D = 0$, C being an arbitrary number. The plane curve (7) is often expressed in the short form

$$y = y(x) \quad \text{or} \quad y = f(x)\,. \tag{8}$$

(The equation $z = 0$ is assumed.) Equation (8) is the so-called *explicit equation of a (plane) curve.*

Example 5. If the equations

$$\varphi(x, y) = 0\,, \quad z = 0 \tag{9}$$

represent a curve, then it is a plane curve because it lies in the coordinate plane xy. We write the plane curve (9) in the form

$$F(x, y) = 0 \text{ or } f(x, y) = 0 \text{ or } \varphi(x, y) = 0 \text{ or in some similar way}\,. \tag{10}$$

(The equation $z = 0$ is assumed.) Equations (10) are called the *implicit equations of a (plane) curve.*

Definition 4. A point (x_0, y_0) on a plane curve $F(x, y) = 0$ is called a *regular (ordinary) point of the curve* if at least one of

$$\frac{\partial F(x_0, y_0)}{\partial x}\,, \quad \frac{\partial F(x_0, y_0)}{\partial y}$$

is non-zero. Every other point on this curve is said to be *singular.* (Cf., however, Remark 2.)

Definition 5. An equation

$$t = t(\bar{t}) \tag{11}$$

expresses a so-called *admissible transformation of the parameter* in an interval $\bar{I}$ if the function (11), defined in $\bar{I}$, possesses the following properties:

1. it is a continuous function and has a continuous derivative (or continuous derivatives up to the order r. cf. Remark 1),

2. $\mathrm{d}t/\mathrm{d}\bar{t} \neq 0$.

The parameter $\bar{t}$ introduced in place of the parameter t by transformation (11) is called an *admissible parameter.*

REMARK 7. All admissible parameters (and these only) are mutually equivalent. A very convenient transformation of the parameter is that one which introduces

the arc s of the curve instead of a general parameter t (in terms of the arc s many results often become of a much simpler form):

Definition 6. The expression

$$s = s(t) = \int_{t_0}^{t} \mathrm{d}s = \int_{t_0}^{t} \sqrt{(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)}\,\mathrm{d}t =$$

$$= \int_{t_0}^{t} \sqrt{\left(\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} \cdot \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t}\right)}\,\mathrm{d}t = \int_{t_0}^{t} \sqrt{(\dot{\boldsymbol{r}} \cdot \dot{\boldsymbol{r}})}\,\mathrm{d}t = \int_{t_0}^{t} \sqrt{(\mathrm{d}\boldsymbol{r} \cdot \mathrm{d}\boldsymbol{r})} \tag{12}$$

is called the *arc of the curve* from the point $t_0 \in I$ to the point $t \in I$.

REMARK 8. It is known, from Integral Calculus, that the expression (12) is the *length of the arc of the curve* between the points t_0 and t. The differential

$$\mathrm{d}s = \sqrt{\left(\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} \cdot \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t}\right)}\,\mathrm{d}t = \sqrt{(\dot{\boldsymbol{r}} \cdot \dot{\boldsymbol{r}})}\,\mathrm{d}t = \sqrt{(\mathrm{d}\boldsymbol{r} \cdot \mathrm{d}\boldsymbol{r})} =$$

$$= \sqrt{(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)}\,\mathrm{d}t = \sqrt{(\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2)}$$

is called the *element of length*, or *linear element, of the curve*. Instead of "the length of the arc s" we say only "the arc s". If $\dot{x}^2 + \dot{y}^2 + \dot{z}^2 > 0$ everywhere in I, the arc s is an admissible parameter of the curve. In this case we can write the equation of the curve in the form

$$\boldsymbol{r} = \boldsymbol{r}(s)\,. \tag{13}$$

The *length of the curve* in the interval $[a, b]$ is

$$l = \int_{a}^{b} \sqrt{(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)}\,\mathrm{d}t\,. \tag{14}$$

Theorem 1. *The parameter t represents the arc length s of the curve if the value $t = 0$ corresponds to the starting point of the curve and the relation*

$$\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} \cdot \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} = \dot{\boldsymbol{r}} \cdot \dot{\boldsymbol{r}} = \dot{x}^2 + \dot{y}^2 + \dot{z}^2 = 1 \tag{15}$$

holds good for every $t \in I$.

REMARK 9. We shall denote derivatives with respect to the arc s by primes,

$$x' = \frac{\mathrm{d}x}{\mathrm{d}s}, \quad x'' = \frac{\mathrm{d}^2x}{\mathrm{d}s^2} \quad \text{etc.}$$

The radius vector $\boldsymbol{r}$ of a point on a curve corresponding to the value of the parameter $t = t_0$ is denoted by $\boldsymbol{r}_0$ (we say shortly the point $\boldsymbol{r}_0$) and so on.

Definition 7. The *tangent vector* of the curve $\mathbf{r} = \mathbf{r}(t)$ at its point $\mathbf{r}_0$ (that is, at the point $t = t_0$ or (x_0, y_0, z_0)) is the vector

$$\dot{\mathbf{r}}_0 = \left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)_0 \tag{16}$$

the coordinates of which are $\dot{x}_0$, $\dot{y}_0$, $\dot{z}_0$ and its starting point at the point $\mathbf{r}_0$ (i.e. the point (x_0, y_0, z_0)). The straight line that contains this vector is called the *tangent to the curve* at the point $\mathbf{r}_0$, which is called the *point of contact* (*contact point*)

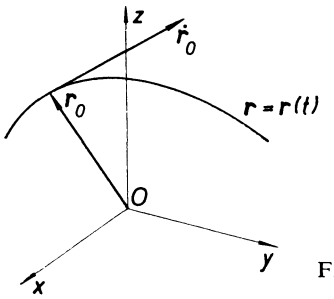


Fig. 9.3.

of the tangent (Fig. 9.3). (The tangent defined in this way is the limiting position of the secant line when its two points of intersection with the curve coincide in a single point of contact.)

REMARK 10. To *norm a tangent vector* $\dot{\mathbf{r}}_0$ means to represent it in terms of a unit vector of the same direction and sense (orientation) and with the starting point $\mathbf{r}_0$.

Theorem 2. *The vector*

$$\mathbf{t} = c\dot{\mathbf{r}}, \quad \textit{where} \quad c = \frac{1}{\sqrt{(\dot{\mathbf{r}} \cdot \dot{\mathbf{r}})}}, \tag{17}$$

with the coordinates (*the direction cosines of the tangent*)

$$t_x = c\dot{x}, \quad t_y = c\dot{y}, \quad t_z = c\dot{z} \tag{18}$$

is the so-called unit tangent vector (its length is equal to 1).

REMARK 11. If the parameter is the arc s, then the modulus of the tangent vector $\mathbf{r}'$ is 1. We write

$$\mathbf{r}' = \mathbf{t}, \tag{19a}$$

where $\mathbf{t}$ is the unit tangent vector. Its coordinates

$$t_x = x', \quad t_y = y', \quad t_z = z' \tag{19b}$$

are the direction cosines of the tangent vector.

For a general parameter t we obtain the direction cosines of the tangent to the curve in the form

$$t_x = \frac{\dot{x}}{\dot{s}}, \quad t_y = \frac{\dot{y}}{\dot{s}}, \quad t_z = \frac{\dot{z}}{\dot{s}} \quad \text{or} \quad \mathbf{t} = \frac{\dot{\mathbf{r}}}{\dot{s}} \quad \left(\dot{s} = \frac{\mathrm{d}s}{\mathrm{d}t} = \sqrt{(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)}\right) \tag{20}$$

and the following relations hold

$$\mathbf{r}' = \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}s} = \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} \cdot \frac{\mathrm{d}t}{\mathrm{d}s} = \dot{\mathbf{r}} \frac{\mathrm{d}t}{\mathrm{d}s} = \dot{\mathbf{r}} \frac{1}{\mathrm{d}s/\mathrm{d}t} = \frac{\dot{\mathbf{r}}}{\dot{s}} = \frac{\dot{\mathbf{r}}}{\sqrt{(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)}}. \tag{21}$$

Let us denote by $\mathbf{R}$ the radius-vector of a current point (X, Y, Z) in space.

Theorem 3. *The equations of the tangent of the curve* $\mathbf{r} = \mathbf{r}(t)$ *at its point* $\mathbf{r}_0$ *are*:

$$\mathbf{R} = \mathbf{r}_0 + u\dot{\mathbf{r}}_0, \text{ i.e. } X = x_0 + u\dot{x}_0, \; Y = y_0 + u\dot{y}_0, \; Z = z_0 + u\dot{z}_0 \tag{22}$$

or

$$\frac{X - x_0}{\dot{x}_0} = \frac{Y - y_0}{\dot{y}_0} = \frac{Z - z_0}{\dot{z}_0} \quad \text{or} \quad \frac{X - x_0}{(\mathrm{d}x)_0} = \frac{Y - y_0}{(\mathrm{d}y)_0} = \frac{Z - z_0}{(\mathrm{d}z)_0} \tag{23}$$

(*where* (x_0, y_0, z_0) *is the point of contact,* $(\dot{x}_0, \dot{y}_0, \dot{z}_0)$ *are the direction ratios (direction parameters) of the tangent,* $\dot{x}_0 = \dot{x}(t_0)$ *etc. and* u *is the variable parameter on the tangent*).

REMARK 12. If a curve is given implicitly by the equations $F(x, y, z) = 0$, $G(x, y, z) = 0$, then the *direction ratios* $(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z)$ of its *tangent* satisfy the relation

$$\mathrm{d}x : \mathrm{d}y : \mathrm{d}z = \begin{vmatrix} \frac{\partial F}{\partial y}, & \frac{\partial F}{\partial z} \\ \frac{\partial G}{\partial y}, & \frac{\partial G}{\partial z} \end{vmatrix} : \begin{vmatrix} \frac{\partial F}{\partial z}, & \frac{\partial F}{\partial x} \\ \frac{\partial G}{\partial z}, & \frac{\partial G}{\partial x} \end{vmatrix} : \begin{vmatrix} \frac{\partial F}{\partial x}, & \frac{\partial F}{\partial y} \\ \frac{\partial G}{\partial x}, & \frac{\partial G}{\partial y} \end{vmatrix}. \tag{24}$$

REMARK 13. The *equation of the tangent* to a plane curve $x = \varphi(t)$, $y = \psi(t)$ at its point (x_0, y_0) can be written in the form

$$(X - \varphi_0)\dot{\psi}_0 = (Y - \psi_0)\dot{\varphi}_0$$

or

$$\frac{X - \varphi_0}{\dot{\varphi}_0} = \frac{Y - \psi_0}{\dot{\psi}_0} \quad (\dot{\varphi}_0 \neq 0, \dot{\psi}_0 \neq 0). \tag{25}$$

The *equation of the tangent* to a plane curve $y = f(x)$ at its point (x_0, y_0) is of the form

$$Y - y_0 = \dot{y}_0(X - x_0) \quad \left(\dot{y}_0 = \left(\frac{\mathrm{d}f}{\mathrm{d}x}\right)_0\right). \tag{26}$$

The *equation of the tangent* to a plane curve $F(x, y) = 0$ at its point (x_0, y_0) is of the form

$$(X - x_0) F_x^0 + (Y - y_0) F_y^0 = 0 \quad \left(F_x^0 = \frac{\partial F}{\partial x}(x_0, y_0), \quad F_y^0 = \frac{\partial F}{\partial y}(x_0, y_0)\right). \quad (27)$$

(The derivatives

$$\dot{y} = \frac{dy}{dx} \text{ (see (26))}, \quad \frac{dy}{dx} = \frac{\dot{\psi}(t)}{\dot{\varphi}(t)} \text{ (see (25))}, \quad \frac{dy}{dx} = -\frac{\partial F/\partial x}{\partial F/\partial y} \text{ (see (27))}$$

denote the tangent of the angle between the tangent at the given point of the curve and the positive x-axis.)

9.3. The Moving Trihedron and the Frenet Formulae

Definition 1. The unit vector $\boldsymbol{n}$ with the same direction and sense as the vector $\mathbf{t}' = d\mathbf{t}/ds$ and with starting point at the point $\boldsymbol{r}$ on the curve is called the *principal normal unit vector* or briefly the *principal normal* at the considered point.

REMARK 1. The straight line containing the vector $\boldsymbol{n}$ is also called the *principal normal* of the curve at its point. A principal normal is defined at a point of the curve if $\boldsymbol{r}'' \neq 0$ (i.e. if the coordinates (x'', y'', z'') of this vector are not all simultaneously equal to zero).

Theorem 1. *We have that*

$$\boldsymbol{n} = \frac{\mathbf{t}'}{k_1} = \frac{\boldsymbol{r}''}{k_1}, \quad \textit{its coordinates being} \quad n_x = \frac{x''}{k_1}, \; n_y = \frac{y''}{k_1}, \; n_z = \frac{z''}{k_1} \qquad (1)$$

where

$$k_1 = \sqrt{(\mathbf{t}' \,.\, \mathbf{t}')} = \sqrt{(\boldsymbol{r}'' \,.\, \boldsymbol{r}'')} = \sqrt{(x''^2 + y''^2 + z''^2)}, \quad x'' = \frac{d^2x}{ds^2} \; \textit{etc.}$$

The vectors $\mathbf{t}$ *and* $\boldsymbol{n}$ *are perpendicular* $(\mathbf{t} \,.\, \boldsymbol{n} = 0)$.

REMARK 2. If the curve is given by its parametric representation with a general parameter t, then

$$\boldsymbol{n} = \frac{\ddot{\boldsymbol{r}}c + \dot{\boldsymbol{r}}\dot{c}}{\sqrt{[(\ddot{\boldsymbol{r}}c + \dot{\boldsymbol{r}}\dot{c}) \,.\, (\ddot{\boldsymbol{r}}c + \dot{\boldsymbol{r}}\dot{c})]}} \quad \left(c = \frac{1}{\sqrt{(\dot{\boldsymbol{r}} \,.\, \dot{\boldsymbol{r}})}} = \frac{1}{\dot{s}}, \quad \dot{\boldsymbol{r}} \neq \mathbf{0}, \quad \ddot{\boldsymbol{r}}c + \dot{\boldsymbol{r}}\dot{c} \neq \mathbf{0}\right).$$

The *direction cosines of the principal normal* are

$$n_x = \frac{\dot{s}(\dot{x}/\dot{s})^{\cdot}}{\sqrt{(\ddot{x}^2 + \ddot{y}^2 + \ddot{z}^2 - \ddot{s}^2)}}, \quad n_y = \frac{\dot{s}(\dot{y}/\dot{s})^{\cdot}}{\sqrt{(\ddot{x}^2 + \ddot{y}^2 + \ddot{z}^2 - \ddot{s}^2)}},$$

$$n_z = \frac{\dot{s}(\dot{z}/\dot{s})^{\cdot}}{\sqrt{(\ddot{x}^2 + \ddot{y}^2 + \ddot{z}^2 - \ddot{s}^2)}} \quad \left(\dot{s} = \frac{\mathrm{d}s}{\mathrm{d}t}, \quad \ddot{s} = \frac{\mathrm{d}^2 s}{\mathrm{d}t^2}\right) \tag{2a}$$

or

$$n_x = \frac{(\dot{x}/\dot{s})^{\cdot}}{k_1 \dot{s}}, \quad n_y = \frac{(\dot{y}/\dot{s})^{\cdot}}{k_1 \dot{s}}, \quad n_z = \frac{(\dot{z}/\dot{s})^{\cdot}}{k_1 \dot{s}} \quad \left(k_1 = \frac{1}{\dot{s}^2}\sqrt{(\ddot{x}^2 + \ddot{y}^2 + \ddot{z}^2 - \ddot{s}^2)}\right). \tag{2b}$$

REMARK 3. The direction ratios of the principal normal (when the parameter is the arc s of the curve) are given by

$$n_x : n_y : n_z = x'' : y'' : z''$$

or (when the parameter of the curve is a general parameter t) by

$$n_x : n_y : n_z = \left(\frac{\dot{x}}{\dot{s}}\right)^{\cdot} : \left(\frac{\dot{y}}{\dot{s}}\right)^{\cdot} : \left(\frac{\dot{z}}{\dot{s}}\right)^{\cdot}.$$

REMARK 4. The equation of the principal normal of a plane curve $F(x, y) = 0$, at its point (x_0, y_0) is

$$(X - x_0) F_y^0 - (Y - y_0) F_x^0 = 0 \quad \text{or} \quad \frac{X - x_0}{F_x^0} = \frac{Y - y_0}{F_y^0} \tag{3}$$

$$(\text{provided } F_x^0 \neq 0, \ F_y^0 \neq 0),$$

for the curve $x = \varphi(t)$, $y = \psi(t)$ at its point t_0

$$(X - \varphi_0)\dot{\varphi}_0 + (Y - \psi_0)\dot{\psi}_0 = 0, \tag{4}$$

and for the curve $y = f(x)$

$$Y - y_0 = -\left(\frac{\mathrm{d}x}{\mathrm{d}y}\right)_0 (X - x_0). \tag{5}$$

Definition 2. A unit vector $\boldsymbol{b}$ with its starting point at the point $\boldsymbol{r}$ on the curve and oriented so that it forms, with the vectors $\mathbf{t}$ and $\boldsymbol{n}$, a positively oriented normed rectangular trihedron (hence $\boldsymbol{b} = \mathbf{t} \times \boldsymbol{n}$, see Theorem 7.1.15), is called the *binormal unit vector* (or for short the *binormal*) to the curve at its point.

REMARK 5. The straight line containing the vector $\boldsymbol{b}$ is also called the *binormal* to the curve at its point. All perpendiculars to the tangent line at its point of contact are called *normals* to the curve. Among them the principal normal and the

binormal are of fundamental importance. In the case of a plane curve the principal normal is that normal which lies in the plane of the curve. This principal normal is briefly called the *normal*.

Theorem 2. *The relations*

$$\boldsymbol{b} = \mathbf{t} \times \boldsymbol{n}, \quad \boldsymbol{n} = \boldsymbol{b} \times \mathbf{t}, \quad \mathbf{t} = \boldsymbol{n} \times \boldsymbol{b} \tag{6}$$

hold (*where* $\mathbf{t} \times \boldsymbol{n}$ *etc. are vector products*, see § 7.1, p. 229).

The following relations hold for the direction cosines (b_x, b_y, b_z) *of the binormal* $\boldsymbol{b}$ *at a point of the curve* $\boldsymbol{r} = \boldsymbol{r}(s)$ (*if the parameter is the arc* s):

$$b_x = \frac{\begin{vmatrix} y', & z' \\ y'', & z'' \end{vmatrix}}{k_1}, \quad b_y = \frac{\begin{vmatrix} z', & x' \\ z'', & x'' \end{vmatrix}}{k_1}, \quad b_z = \frac{\begin{vmatrix} x', & y' \\ x'', & y'' \end{vmatrix}}{k_1} \quad \text{(for } k_1 \text{ see Theorem 1)}, \tag{7a}$$

in the case of a general parameter t $(\boldsymbol{r} = \boldsymbol{r}(t))$:

$$b_x = \frac{\begin{vmatrix} \dot{y}, & \dot{z} \\ \ddot{y}, & \ddot{z} \end{vmatrix}}{\dot{s}^3 k_1}, \quad b_y = \frac{\begin{vmatrix} \dot{z}, & \dot{x} \\ \ddot{z}, & \ddot{x} \end{vmatrix}}{\dot{s}^3 k_1}, \quad b_z = \frac{\begin{vmatrix} \dot{x}, & \dot{y} \\ \ddot{x}, & \ddot{y} \end{vmatrix}}{\dot{s}^3 k_1} \quad \text{(for } k_1 \text{ see (2b))}. \tag{7b}$$

REMARK 6. The *equations of the binormal* to the curve $\boldsymbol{r} = \boldsymbol{r}(t)$ at its point (x, y, z) are

$$\frac{X - x}{[\dot{y}, \ddot{z}]} = \frac{Y - y}{[\dot{z}, \ddot{x}]} = \frac{Z - z}{[\dot{x}, \ddot{y}]} \tag{8}$$

(X, Y, Z are the orthogonal coordinates of the running point on the binormal, $[\dot{y}, \ddot{z}] = \begin{vmatrix} \dot{y}, & \dot{z} \\ \ddot{y}, & \ddot{z} \end{vmatrix}$ etc.)

Definition 3. The normed orthogonal and right-handed (positively oriented) trihedron formed by the vectors $\mathbf{t}$, $\boldsymbol{n}$, $\boldsymbol{b}$ at a point of a curve is called the *moving trihedron* (*moving trihedral*) of the curve.

Theorem 3 (*The Frenet or Serret-Frenet Formulae*).

a) *For the curve* $\boldsymbol{r} = \boldsymbol{r}(s)$ (*the parameter is the arc* s)

$$\begin{aligned} \mathbf{t}' &= +k_1 \boldsymbol{n} \\ \boldsymbol{n}' &= -k_1 \mathbf{t} + k_2 \boldsymbol{b} \\ \boldsymbol{b}' &= -k_2 \boldsymbol{n} \end{aligned}$$

(for k_1 see Theorem 1,

$$k_2 = \frac{[x', y'', z''']}{x''^2 + y''^2 + z''^2},$$

b) *For the curve* $\boldsymbol{r} = \boldsymbol{r}(t)$ (*with a general parameter* t)

$$\begin{aligned} \dot{\mathbf{t}} &= +k_1 \dot{s} \boldsymbol{n} \\ \dot{\boldsymbol{n}} &= -k_1 \dot{s} \mathbf{t} + k_2 \dot{s} \boldsymbol{b} \\ \dot{\boldsymbol{b}} &= -k_2 \dot{s} \boldsymbol{n} \end{aligned} \tag{9}$$

(for k_1 see (2b),

$$k_2 = \frac{[\dot{x}, \ddot{y}, \dddot{z}]}{\dot{s}^2(\ddot{x}^2 + \ddot{y}^2 + \ddot{z}^2 - \ddot{s}^2)},$$

where

$$[x', y'', z'''] = \begin{vmatrix} x', & y', & z' \\ x'', & y'', & z'' \\ x''', & y''', & z''' \end{vmatrix}).$$

where

$$[\dot{x}, \ddot{y}, \dddot{z}] = \begin{vmatrix} \dot{x}, & \dot{y}, & \dot{z} \\ \ddot{x}, & \ddot{y}, & \ddot{z} \\ \dddot{x}, & \dddot{y}, & \dddot{z} \end{vmatrix}).$$

REMARK 7. The Frenet formulae define the relations between the direction cosines (or the direction ratios) of the tangent, the principal normal and the binormal at a general point of the curve, and their derivatives. The numbers k_1 and k_2 in Theorem 1, Remark 2 and Theorem 3 are called the *first curvature* (briefly the *curvature*), and the *second curvature* (briefly the *torsion*), respectively. For a detailed treatment see § 9.4.

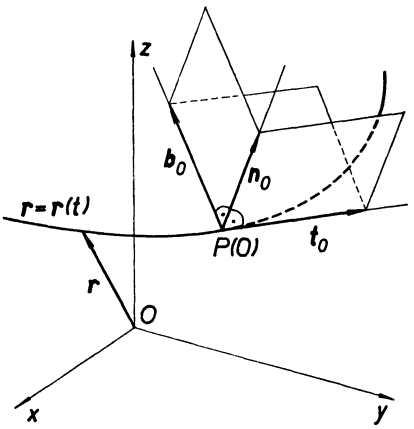


Fig. 9.4.

Definition 4. The plane determined by the principal normal and the binormal (at the point considered) is called the *normal plane*, the plane determined by the binormal and the tangent is called the *rectifying plane* and the plane determined by the tangent and the principal normal is the *osculating plane* (Fig. 9.4).

Theorem 4. *The equations of the normal plane, the rectifying plane and the osculating plane, respectively, at a point* $\mathbf{r}_0$ *of a curve* $\mathbf{r} = \mathbf{r}(t)$ *(or* $\mathbf{r} = \mathbf{r}(s)$*):*

The normal plane (perpendicular to the tangent line):

$$(\mathbf{R} - \mathbf{r}_0) \,.\, \mathbf{t}_0 = 0 \text{ or } (\mathbf{R} - \mathbf{r}_0) \,.\, \dot{\mathbf{r}}_0 = 0 \text{ or } (\mathbf{R} - \mathbf{r}_0) \,.\, \mathbf{r}'_0 = 0\,, \tag{10}$$

for example,

$$(X - x_0)\,(\mathrm{d}x)_0 + (Y - y_0)\,(\mathrm{d}y)_0 + (Z - z_0)\,(\mathrm{d}z)_0 = 0 \tag{11}$$

((x_0, y_0, z_0) being a point on the curve).

The rectifying plane (perpendicular to the principal normal):

$$(\mathbf{R} - \mathbf{r}_0) \,.\, \mathbf{n}_0 = 0 \quad \text{or} \quad (\mathbf{R} - \mathbf{r}_0) \,.\, \mathbf{r}''_0 = 0\,, \tag{12}$$

for example,

$$(X - x_0)(n_x)_0 + (Y - y_0)(n_y)_0 + (Z - z_0)(n_z)_0 = 0 .$$

The osculating plane (perpendicular to the binormal):

$$(\mathbf{R} - \mathbf{r}_0) \,.\, \mathbf{b}_0 = 0 \tag{13}$$

(**R** *being the radius vector of the running point* (X, Y, Z) *of the plane*).

REMARK 8. If the curve is given by the equations $F(x, y, z) = 0$, $G(x, y, z) = 0$, the *equation of the normal plane* at the point (x, y, z) of the curve is

$$\begin{vmatrix} X - x, & Y - y, & Z - z \\ \dfrac{\partial F}{\partial x}, & \dfrac{\partial F}{\partial y}, & \dfrac{\partial F}{\partial z} \\ \dfrac{\partial G}{\partial x}, & \dfrac{\partial G}{\partial y}, & \dfrac{\partial G}{\partial z} \end{vmatrix} = 0 . \tag{14}$$

Theorem 5. *The equation of the osculating plane at the point* (x, y, z) *with the radius vector* **r**

a) *if the parameter is the arc s:*

$$[\mathbf{R} - \mathbf{r}, \mathbf{r}', \mathbf{r}''] = 0 ,$$

i.e.

$$\begin{vmatrix} X - x, & Y - y, & Z - z \\ x', & y', & z' \\ x'', & y'', & z'' \end{vmatrix} = 0$$

or

$$\begin{vmatrix} X - x, & Y - y, & Z - z \\ \mathrm{d}x, & \mathrm{d}y, & \mathrm{d}z \\ \mathrm{d}^2x, & \mathrm{d}^2y, & \mathrm{d}^2z \end{vmatrix} = 0 . \tag{15b}$$

b) *with a general parameter t:*

$$[\mathbf{R} - \mathbf{r}, \dot{\mathbf{r}}, \ddot{\mathbf{r}}] = 0 ,$$

i.e.

$$\begin{vmatrix} X - x, & Y - y, & Z - z \\ \dot{x}, & \dot{y}, & \dot{z} \\ \ddot{x}, & \ddot{y}, & \ddot{z} \end{vmatrix} = 0 \tag{15a}$$

REMARK 9. Every plane passing through the tangent of a (space) curve is called a *tangent plane* to the curve at the corresponding point of contact. At this point the contact of the tangent plane with the curve is at least a two-point contact. The osculating plane of the curve at its point is an important tangent plane because its contact with the curve at this point is at least a three-point contact. At a point of the curve where the equation $\mathbf{r}'' = \mathbf{0}$ (with the parameter s) or $\ddot{\mathbf{r}} = \lambda\dot{\mathbf{r}}$ (with general parameter t; λ being a real number) is satisfied, the osculating plane is indefinite. At such points on the curve, e.g. at so-called *points of inflexion*, the equation of the osculating plane is satisfied identically. In the case of a plane curve, the osculating plane at each of its points is the plane of the curve.

Example 1. The curve given by the equations

$$x = a\cos t\,,\quad y = a\sin t\,,\quad z = bt\ (a > 0,\ b \neq 0 \text{ being real constants}) \tag{16}$$

is called a *circular helix*. It is one of the most important curves in applications. It lies on the circular cylinder (Fig. 9.5)

$$x^2 + y^2 = a^2 \tag{17}$$

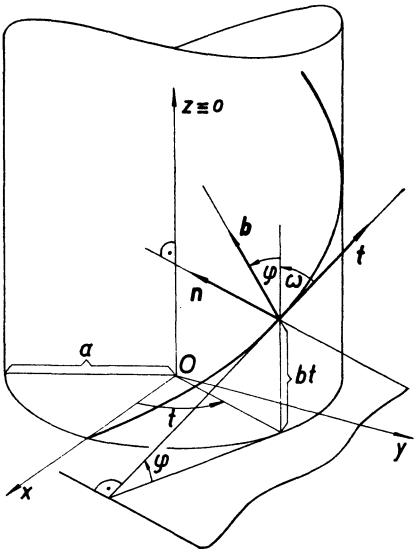


Fig. 9.5.

(as we can find by elimination of t from the first two equations (16)), and the axis of the cylinder is the coordinate z-axis. The axis of the cylinder on which the circular helix lies is called the *axis of the helix*. From the first and second of equations (16) we obtain in the first quadrant (for $0 \leqq t < \frac{1}{2}\pi$) $t = \tan^{-1}(y/x)$ and by the third of equations (16)

$$z = b\tan^{-1}\frac{y}{x}\,, \tag{18}$$

i.e. the equation of the *right helicoid* (see Example 9.12.2, p. 316). For $\frac{1}{2}\pi < t < \frac{3}{2}\pi$ we have $z = b[\tan^{-1}(y/x) + \pi]$ etc. The given helix is the intersection of the conoid (18) and the circular cylinder (17). From (16) it follows that

$$\dot{x} = -a\sin t\,,\quad \dot{y} = a\cos t\,,\quad \dot{z} = b\,, \tag{19}$$

hence

$$\frac{1}{\sqrt{(\dot{\boldsymbol{r}}\,.\,\dot{\boldsymbol{r}})}} = \frac{1}{\sqrt{(a^2 + b^2)}}$$

and for the direction cosines (t_x, t_y, t_z) of the tangent line we obtain (see Theorem 9.2.2)

$$t_x = -\frac{a \sin t}{\sqrt{(a^2 + b^2)}}, \quad t_y = \frac{a \cos t}{\sqrt{(a^2 + b^2)}}, \quad t_z = \frac{b}{\sqrt{(a^2 + b^2)}}. \tag{20}$$

Thus, the tangents to a circular helix make a constant angle ω with its axis, where $\cos \omega = b/\sqrt{(a^2 + b^2)}$, and therefore the tangent makes a constant angle φ (the *gradient of the helix*), with every plane which is perpendicular to the axis of the helix. Moreover, $\sin \varphi = \cos \omega$, so that $\tan \varphi = b/a$ (the *slope of the helix*). By (20) and Theorem 9.2.3 the equations of the tangent at the point t of a helix are of the form

$$-\frac{X - a \cos t}{a \sin t} = \frac{Y - a \sin t}{a \cos t} = \frac{Z - bt}{b}. \tag{21}$$

If the circular cylinder, on which the helix lies, is developed upon a plane, then at the same time, every turn of the helix and every circle on the cylinder are developed into two segments that intersect at an angle φ. Further, from (19), we have that

$$\dot{s}^2 = \left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^2 = \dot{\mathbf{r}} \cdot \dot{\mathbf{r}} = \dot{x}^2 + \dot{y}^2 + \dot{z}^2 = a^2 + b^2$$

(see Definition 9.2.6) so that the length of the arc of one turn of the helix is

$$s = \int_0^{2\pi} \sqrt{(a^2 + b^2)}\, \mathrm{d}t = 2\pi \sqrt{(a^2 + b^2)}$$

(as can be seen immediately from the development in Fig. 9.6). The length of the arc of the helix from the point $t = 0$ to the point t, is $s(t) = t\sqrt{(a^2 + b^2)} = ct$. If we substitute in equation (16) the length of arc s as parameter instead of the general parameter t, we obtain the equations of the helix in the form

$$x = a \cos \frac{s}{c}, \quad y = a \sin \frac{s}{c}, \quad z = b\frac{s}{c} \quad (c = \sqrt{(a^2 + b^2)}).$$

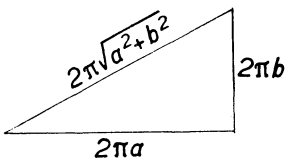


Fig. 9.6.

From (16) we obtain

$$\ddot{x} = -a \cos t, \quad \ddot{y} = -a \sin t, \quad \ddot{z} = 0, \quad \ddot{x}^2 + \ddot{y}^2 + \ddot{z}^2 = a^2 \tag{22}$$

and further $\ddot{s} = 0$. Substituting in (2) we obtain for the direction cosines of the

principal normal the expressions

$$n_x = -\cos t\,, \quad n_y = -\sin t\,, \quad n_z = 0\,. \tag{23}$$

Thus, the principal normals of the helix are perpendicular to the axis of the helix (its direction cosines are (0, 0, 1)). From (7) we obtain for the direction cosines of the binormal the expressions

$$b_x = \frac{b \sin t}{\sqrt{(a^2 + b^2)}}, \quad b_y = -\frac{b \cos t}{\sqrt{(a^2 + b^2)}}, \quad b_z = \frac{a}{\sqrt{(a^2 + b^2)}}. \tag{24}$$

Thus, the binormals of the helix are inclined at a constant angle φ to the axis of the helix, such that $\cos \varphi = a/\sqrt{(a^2 + b^2)}$.

By (15a), (19) and (22) we obtain the equation of the osculating plane at the point t of the helix

$$\begin{vmatrix} X - a\cos t, & Y - a\sin t, & Z - bt \\ -a\sin t, & a\cos t, & b \\ -a\cos t, & -a\sin t, & 0 \end{vmatrix} = Xb\sin t - Yb\cos t + Za - abt = 0\,.$$

Further, it follows from (24) that

$$\dot{b}_x = \frac{b \cos t}{\sqrt{(a^2 + b^2)}}, \quad \dot{b}_y = \frac{b \sin t}{\sqrt{(a^2 + b^2)}}, \quad \dot{b}_z = 0\,.$$

Putting $\dot{s} = \sqrt{(a^2 + b^2)}$ in the third of the Frenet formulae (9) and using (23), we obtain for the torsion k_2 at the point t of the helix, the expression

$$k_2 = \frac{b}{a^2 + b^2}\,. \tag{25}$$

Similarly, from the first of the Frenet formulae (9) it follows that at the point t of the helix the curvature k_1 is given by the expression

$$k_1 = \frac{a}{a^2 + b^2}\,. \tag{26}$$

Thus, the circular helix (and the circular helix alone from among all space curves) has both curvatures constant (and not vanishing at all its points). The curvature k_1 and the torsion k_2 can be computed, of course, by formulae (9.4.1) and (9.4.4). The sign of k_2 agrees with the sign of the constant b. The helix is *right-handed* or *left-handed* according as $b > 0$ (and then $k_2 > 0$) or $b < 0$ respectively.

If k_1 and k_2 in (25) and (26) stand for the curvature and the torsion, respectively, at any point of a space curve and if (25) and (26) are solved for a and b, then these numbers a and b define a circular helix (generally turned round the axis with regard

to the helix (16)); this helix has the same curvature and the same torsion as the given curve at the point considered. Such a helix plays a similar role for the given curve as the osculating circle does for a plane curve, and the contact of this helix with the given space curve at the point considered is of the fourth order at least (cf. § 9.5).

Example 2 (*Components of the Vector of Acceleration*). Let a particle move along a space curve

$$\boldsymbol{r} = \boldsymbol{i}\, x(t) + \boldsymbol{j}\, y(t) + \boldsymbol{k}\, z(t)$$

(t denoting time).

The velocity vector is

$$\boldsymbol{v} = \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} = \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}s} \cdot \frac{\mathrm{d}s}{\mathrm{d}t} = \frac{\mathrm{d}s}{\mathrm{d}t}\, \boldsymbol{r}' = \frac{\mathrm{d}s}{\mathrm{d}t}\, \boldsymbol{t}\,.$$

The acceleration vector is

$$\boldsymbol{a} = \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = \frac{\mathrm{d}^2 s}{\mathrm{d}t^2}\, \boldsymbol{t} + \frac{\mathrm{d}s}{\mathrm{d}t} \frac{\mathrm{d}\boldsymbol{t}}{\mathrm{d}t} = \frac{\mathrm{d}^2 s}{\mathrm{d}t^2}\, \boldsymbol{t} + \frac{\mathrm{d}s}{\mathrm{d}t} \frac{\mathrm{d}\boldsymbol{t}}{\mathrm{d}s} \frac{\mathrm{d}s}{\mathrm{d}t} = \frac{\mathrm{d}^2 s}{\mathrm{d}t^2}\, \boldsymbol{t} + \left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^2 \boldsymbol{t}'\,.$$

Using the first Frenet formula $\boldsymbol{t}' = k_1 \boldsymbol{n}$ and writing

$$k_1 = \frac{1}{r_1}$$

(r_1 is the so-called *radius of curvature*), we obtain

$$\boldsymbol{a} = \frac{\mathrm{d}^2 s}{\mathrm{d}t^2}\, \boldsymbol{t} + \frac{\left(\dfrac{\mathrm{d}s}{\mathrm{d}t}\right)^2}{r_1}\, \boldsymbol{n}\,.$$

We can see from the result that the acceleration vector resolves into two components, one in the direction of the tangent vector and the other in the direction of the principal normal vector. They therefore lie in the osculating plane. The normal component,

$$\frac{\left(\dfrac{\mathrm{d}s}{\mathrm{d}t}\right)^2}{r_1}\, \boldsymbol{n}\,,$$

is called the *normal acceleration*. For the case of circular motion, this represents a well-known formula of physics.

REMARK 10. When considering plane curves, we often speak of a *subtangent* s_t or a *subnormal* s_n; they are the (oriented) orthogonal projections on the x-axis of the segments of the tangent t, or the normal n, from the contact point P up to the point

of intersection of the tangent, or the normal, with the x-axis, respectively. If we suppose that the derivative $\dot{y} = \mathrm{d}y/\mathrm{d}x \neq 0$ at the point P is finite, then it follows from Fig. 9.7 that

$$s_t = -\frac{y}{\dot{y}}, \quad s_n = y\dot{y}.$$

(Often only the absolute values of s_t and s_n are considered.)

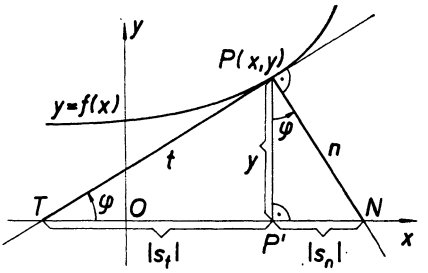


Fig. 9.7.

For the *length of the tangent* t *and the normal* n we obtain

$$t = \left|\frac{y\dot{s}}{\dot{y}}\right|, \quad n = |y\dot{s}| \quad \left(\dot{s} = \frac{\mathrm{d}s}{\mathrm{d}x} = \sqrt{(1 + \dot{y}^2)}\right).$$

9.4. First and Second Curvatures, Natural Equations of a Curve

Definition 1. The expression k_1 given by the equation

a) for the parameter s:

$$k_1 = \sqrt{(\mathbf{t}' \,.\, \mathbf{t}')} = \sqrt{(\mathbf{r}'' \,.\, \mathbf{r}'')} = \sqrt{(x''^2 + y''^2 + z''^2)}$$

(see (9.3.1) and (9.3.9))

b) for a general parameter t:

$$k_1 = \frac{1}{\dot{s}^2}\sqrt{(\ddot{x}^2 + \ddot{y}^2 + \ddot{z}^2 - \ddot{s}^2)} \qquad (1)$$

(see (9.3.2b) and (9.3.9))

is called the *first curvature* (briefly: the *curvature*) of the curve at the point considered; its reciprocal value $r_1 = 1/k_1$ is the so-called *radius of curvature* (cf. Examples 9.3.1 and 9.3.2).

REMARK 1. A necessary and sufficient condition that a curve be a straight line is that the equation $k_1 = 0$ (i.e. $\mathbf{r}'' = \mathbf{0}$) holds at every point of the curve. The case where the curvature vanishes only at individual points of the curve (the case of points of inflexion) is considered in § 9.5.

Theorem 1 (*The Geometric Interpretation of the Curvature* k_1 *of a Curve*). *Let* $\varphi = \varphi(s, 0)$ *be the angle between the tangent lines* $\mathbf{t}(s)$ *and* $\mathbf{t}_0$ $(s > 0)$ *at the*

points $Q(s)$ and $P(0)$, *respectively on the curve* $\mathbf{r} = \mathbf{r}(s)$. *Then* (Fig. 9.8)

$$\lim_{s\to 0} \frac{\varphi}{s} = \left(\frac{\mathrm{d}\varphi}{\mathrm{d}s}\right)_0 = (k_1)_0 . \tag{2}$$

REMARK 2. If $\Delta s = \overset{\frown}{PQ}$ is the arc between two "neighbouring" points P and Q on curve k and if $\Delta\varphi$ is the acute angle between the tangents to the curve constructed at these points, then the ratio $\Delta\varphi/\Delta s$ is called the *mean curvature* of the curve on the

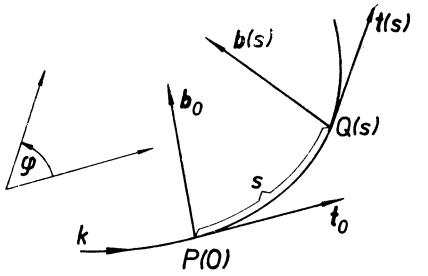


Fig. 9.8.

arc Δs. The acute angle between the tangents $\mathbf{t}_0$ and $\mathbf{t}(\Delta s)$ at the points $P(0)$ and $Q(\Delta s)$ is called the *angle of contingence.*

REMARK 3. In the case of a plane curve $y = f(x)$ the curvature at the point (x, y) is

$$k_1 = \left|\frac{\ddot{y}}{(1+\dot{y}^2)^{3/2}}\right| \quad \text{or} \quad r_1 = \left|\frac{(1+\dot{y}^2)^{3/2}}{\ddot{y}}\right| \quad (\ddot{y} \neq 0) . \tag{3}$$

For example, for the circle

$$y = \sqrt{(r^2 - x^2)}$$

we obtain $k_1 = 1/r$ and so the radius of curvature of the circle is, at all its points, constant and has the same value as its radius (this property holds only for the circle).

Definition 2. The expression k_2, given by the equation

a) for the parameter s:

$$k_2 = \frac{\begin{vmatrix} x', & y', & z' \\ x'', & y'', & z'' \\ x''', & y''', & z''' \end{vmatrix}}{x''^2 + y''^2 + z''^2}$$

b) for a general parameter t:

$$k_2 = \frac{\begin{vmatrix} \dot{x}, & \dot{y}, & \dot{z} \\ \ddot{x}, & \ddot{y}, & \ddot{z} \\ \dddot{x}, & \dddot{y}, & \dddot{z} \end{vmatrix}}{\dot{s}^2(\ddot{x}^2 + \ddot{y}^2 + \ddot{z}^2 - \ddot{s}^2)} \tag{4}$$

(see (9.3.9)), is called the *second curvature* (the *torsion*) of the curve at the point considered, and its reciprocal value $r_2 = 1/k_2$ is called the *radius of torsion* (cf. Example 9.3.1).

Theorem 2. *A necessary and sufficient condition that a curve be a plane curve is that*

$$k_2 = 0\,, \quad \text{i.e.} \quad [x', y'', z'''] \text{ or } [\dot{x}, \ddot{y}, \dddot{z}] = 0 \quad \text{(see (4))}$$

at all points of the curve.

REMARK 4. Space curves ($k_1 \neq 0, k_2 \neq 0$) are sometimes called curves of two curvatures or curves with torsion. In the case of plane curves the Frenet formulae reduce (see (9.3.9)) to

$$\mathbf{t}' = k_1\mathbf{n}\,, \quad \mathbf{n}' = -k_1\mathbf{t} \text{ (for the parameter } s) \qquad \dot{\mathbf{t}} = k_1\dot{s}\mathbf{n}\,, \quad \dot{\mathbf{n}} = -k_1\dot{s}\mathbf{t} \text{ (for the general parameter } t).$$

Theorem 3. (*The Geometric Interpretation of the Torsion k_2 of a Curve*). *Let $\psi = \psi(s, 0)$ be the angle between the binormals $\mathbf{b}(s)$ and $\mathbf{b}_0$ $(s > 0)$ at the points $Q(s)$ and $P(0)$, respectively, on the curve $\mathbf{r} = \mathbf{r}(s)$. Then* (Fig. 9.8)

$$\lim_{s\to 0} \frac{\psi}{s} = \left(\frac{\mathrm{d}\psi}{\mathrm{d}s}\right)_0 = |(k_2)_0|\,. \tag{5}$$

REMARK 5. If $\Delta s = \overset{\frown}{PQ}$ is the arc between two neighbouring points $P(0)$ and $Q(\Delta s)$ on the curve and $\Delta\psi$ is the acute angle between the binormals at these points, then the ratio $\Delta\psi/\Delta s$ is called the *mean torsion* of the curve on the arc Δs. For plane curves, $\mathbf{b}'(s) = \dot{\mathbf{b}}(t) = \mathbf{0}$ identically.

REMARK 6. If we introduce a system of coordinates such that at the point $s = 0$ the vectors $\mathbf{t}$, $\mathbf{n}$, $\mathbf{b}$ of the moving trihedron correspond to the half-axes $+x$, $+y$, $+z$ respectively, then we can write, for the cartesian coordinates $x(s)$, $y(s)$, $z(s)$ of a point on the curve in a neighbourhood of the point $s = 0$, the series

$$\begin{aligned} x(s) &= s - \frac{s^3}{3!}(k_1^2)_0 - \dots\,, \\ y(s) &= \frac{s^2}{2!}(k_1)_0 + \frac{s^3}{3!}(k_1')_0 + \dots\,, \\ z(s) &= \frac{s^3}{3!}(k_1k_2)_0 + \dots\,. \end{aligned} \tag{6}$$

Equations (6) are called the *canonical equations* (or the *canonical representation*) *of the curve*. We obtain from (6) (when we use only the first term of each series) the equations of the simplest algebraic space curve (the so-called *cubical parabola*),

$$x(s) = s\,, \quad y(s) = \frac{(k_1)_0}{2}s^2\,, \quad z(s) = \frac{(k_1k_2)_0}{6}s^3\,.$$

This parabola approximates the given curve at the point $s = 0$.

REMARK 7. A curve is said to be *right-handed* or *left-handed* if at any point of the curve $k_2 > 0$ or $k_2 < 0$, respectively (cf. Example 9.3.1).

REMARK 8. If two continuous functions $k_1 = k_1(s) > 0$, and $k_2 = k_2(s)$ are given, then it is possible to construct a curve for which k_1 is its curvature, k_2 its torsion and the parameter s its arc. This curve is thus determined uniquely except for its position in space (if no definite special conditions are given).

Theorem 4. *Let two continuous functions*

$$k_1 = k_1(s) > 0\,, \quad k_2 = k_2(s) \tag{7}$$

be given (k_1 having a continuous second derivative and k_2 having a continuous first derivative). Then there exists a unique curve having the following properties:

1. *its arc is s, its curvature is k_1 and its torsion is k_2;*
2. *it passes through an arbitrary given point s_0;*
3. *three arbitrary mutually perpendicular unit vectors $\mathbf{t}_0$, $\mathbf{n}_0$, $\mathbf{b}_0$ are the tangent unit vector, the principal normal unit vector, the binormal unit vector, respectively, at the point s_0.*

Theorem 5. *It is always possible to write the equations of a plane curve with given curvature $k_1(s)$ (the torsion k_2 being equal to zero identically) in the form*

$$\left.\begin{aligned} x &= \int \cos\left(\int k_1\,ds + c\right) ds + a\,, \\ y &= \pm \int \sin\left(\int k_1\,ds + c\right) ds + b\,, \\ z &= 0 \end{aligned}\right\} \tag{8}$$

(a, b, c are arbitrary real constants).

REMARK 9. All plane curves (in the plane $z = 0$) with the same curvature $k_1(s)$ at the point s may be obtained from the rotation of the curve

$$x = \int \cos\left(\int k_1\,ds\right) ds\,, \quad y = \pm \int \sin\left(\int k_1\,ds\right) ds\,, \quad z = 0$$

through an angle c and by the translation of the curve along the oriented segment given by the vector with the coordinates (a, b) (for the constants a, b, c see Theorem 5).

Definition 3. The quantities s, k_1 and k_2 are called the *natural coordinates* and the relations

$$k_1 = k_1(s)\,, \quad k_2 = k_2(s) \tag{9}$$

between k_1, k_2 and s are called the *natural equations (intrinsic equations) of the curve.*

REMARK 10. The natural coordinates of a curve are independent of any coordinate system. The natural equations of a curve, which express the curve independently of the choice of the coordinate system, are suitable for the investigation of those properties of curves that are not dependent on the coordinates. For example, the natural equation

$$k_1 = \frac{1}{a} \quad (k_2 = 0,\ a \text{ is a positive constant})$$

is the equation of *all circles* of radius $r = a$ in the plane $z = 0$, the implicit equation of which is

$$(x - m)^2 + (y - n)^2 = a^2 \quad (m, n \text{ being arbitrary real constants}).$$

The plane curve having the property that its curvature k_1 is directly proportional to the length of the arc s is called a *clothoid* (cf. § 4.8). This property is expressed by its natural equation

$$k_1 = \frac{1}{a^2} s \quad (a \text{ is a real constant}).$$

By the proper choice of the coordinate system, the cartesian coordinates of the points of a clothoid may be expressed with the help of the Fresnel integrals,

$$x = \frac{a}{\sqrt{2}} \int_0^{\varphi} \frac{\cos \varphi}{\sqrt{\varphi}} \, \mathrm{d}\varphi, \quad y = \frac{a}{\sqrt{2}} \int_0^{\varphi} \frac{\sin \varphi}{\sqrt{\varphi}} \, \mathrm{d}\varphi,$$

where $\varphi = s^2/2a^2$ is the angle between the tangent line at its point (x, y) and the half-axis $+x$.

9.5. Contact of Curves, Osculating Circle

Let

$$^1\mathbf{r} = {}^1\mathbf{r}(s), \quad {}^2\mathbf{r} = {}^2\mathbf{r}(s) \tag{1}$$

be the equations of two curves ^1k and ^2k represented in terms of the same parameter s, which is the length of the arc on both curves. Let the curves have a common (regular) point $s = 0$ (i.e. $^1\mathbf{r}_0 = {}^2\mathbf{r}_0$) from which we shall measure the common parameter along both curves. Let us consider a point on each curve corresponding to the same value of the parameter s (Fig. 9.9) and let us investigate the mutual position of both curves (1) in a sufficiently small neighbourhood of their common point $s = 0$.

Definition 1. Two curves $^1\mathbf{r} = {}^1\mathbf{r}(s)$, $^2\mathbf{r} = {}^2\mathbf{r}(s)$ are said to have, at their common point $^1\mathbf{r}_0 = {}^2\mathbf{r}_0$, *contact of order* q *at least* i.e. *at least* $(q + 1)$*-point contact*,

provided that the equations

$$\lim_{s\to 0} \frac{\mathbf{d}(s)}{s^p} = \mathbf{0} \quad (p = 0, 1, 2, \ldots, q), \tag{2}$$

are satisfied, where

$$\mathbf{d}(s) = {}^1\mathbf{r}(s) - {}^2\mathbf{r}(s).$$

Theorem 1. *The necessary and sufficient condition that the curves* (1) *have contact of order q at least, i.e. at least* $(q + 1)$*-point contact, at their common point is that the equations*

$${}^1\mathbf{r}_0 = {}^2\mathbf{r}_0, \quad {}^1\mathbf{r}_0' = {}^2\mathbf{r}_0', \quad \ldots, \quad {}^1\mathbf{r}_0^{(q)} = {}^2\mathbf{r}_0^{(q)}. \tag{3}$$

are satisfied. (*The existence of a sufficient number of derivatives is assumed.*)

REMARK 1. The phrase "at least $(q + 1)$-point contact" corresponds to the conception that at the common point of contact of two curves, both curves have at least $(q + 1)$ coincident points of intersection. For example if ${}^1\mathbf{r}_0 = {}^2\mathbf{r}_0$, and ${}^1\mathbf{r}_0' \neq {}^2\mathbf{r}_0'$, then at the point $s = 0$ the curves have contact of order 0 exactly, i.e. exactly one-point contact (they intersect at the point $s = 0$).

REMARK 2. For the plane curves ${}^1y = {}^1y(x)$ and ${}^2y = {}^2y(x)$ we may take, instead of equations (2), the equations

$$\lim_{x\to 0} \frac{d(x)}{x^p} = 0 \quad (p = 0, 1, 2, \ldots, q), \quad d(x) = {}^1y(x) - {}^2y(x)$$

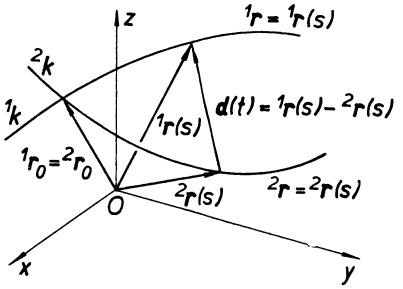


Fig. 9.9.

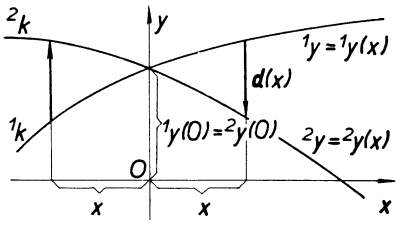


Fig. 9.10.

assuming that no curve possesses the tangent $x = 0$ at their common point $x = 0$ (see Fig. 9.10, where ${}^1y(0) = {}^2y(0)$, but ${}^1\dot{y}(0) \neq {}^2\dot{y}(0)$). Instead of equations (3) we have, in the case of the curves ${}^1y = {}^1y(x)$ and ${}^2y = {}^2y(x)$,

$${}^1y_0 = {}^2y_0, \quad {}^1\dot{y}_0 = {}^2\dot{y}_0, \quad \ldots, \quad {}^1y_0^{(q)} = {}^2y_0^{(q)}. \tag{4}$$

(We use the notation

$$^{i}\dot{y}(x) = \frac{\mathrm{d}^{i}y}{\mathrm{d}x}$$

etc.)

REMARK 3. If two curves have, at their common point, the same tangent (in the case of space curves also the same tangent plane) then at that point they have contact of order 1 at least, i.e. at least two-point contact, and conversely. If at their common point two curves have the same tangent, principal normal and curvature, they have contact of order 2 at least, i.e. at least three-point contact, and conversely. Two space curves that have, at their common point, contact at least of order 2 (at least three-point contact) have at this point a common osculating plane (provided that the osculating planes exist at this point). The plane that has contact of the first order (two-point contact) or contact of the second order (three-point contact) with a given curve is the tangent plane or the osculating plane, respectively, at the point considered.

REMARK 4. If $^{1}\mathbf{r}_0 = {}^{2}\mathbf{r}_0$, $^{1}\mathbf{r}'_0 = {}^{2}\mathbf{r}'_0$, but $^{1}\mathbf{r}''_0 \neq {}^{2}\mathbf{r}''_0$ (or $^{1}y_0 = {}^{2}y_0$, $^{1}\dot{y}_0 = {}^{2}\dot{y}_0$, but $^{1}\ddot{y}_0 \neq {}^{2}\ddot{y}_0$), then we say that both curves have *contact of exactly the first order* or *exactly two-point contact.* Similarly for the contact of any order. In the example considered, both curves touch each other at their common point and their contact is a so-called ordinary one. The curve $^{2}y = {}^{2}y(x)$ lies on the same side of the curve $^{1}y = {}^{1}y(x)$ in a neighbourhood of their common point (Fig. 9.11). If we replace the curve ^{2}y by a straight line, then the straight line which has contact of the first order at least (at least two-point contact) with the curve is the tangent at this point.

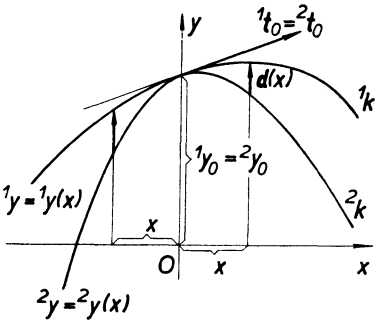


Fig. 9.11.

Definition 2. A curve ^{2}k of a given type (e.g. a circle) that has contact of highest possible order with the curve ^{1}k at their common point is said to be an *osculating curve* of the curve ^{1}k at the point considered. We say that the curve ^{2}k *osculates* the curve ^{1}k at this point.

REMARK 5. An osculating curve of the given curve is generally completely determined by the condition of osculation. In special cases, i.e. at some special points on the given curve (e.g. at the points of inflexion, at the vertices, etc.) the osculating curve

may have contact of an order that is higher than is the highest possible at an *ordinary* point.

Definition 3. The curve ${}^2k$ that has contact with the curve ${}^1k$ at their common point of an order higher than is generally the highest possible order, is said to be the *superosculating curve* of the curve ${}^1k$ at such point.

Example 1. In the general case a straight line and a curve may have, at a regular point of the given curve, contact of the first order at most (i.e. two-point contact). Of course, this osculating straight line is the tangent to the given curve. If a further condition regarding contact of a higher order is imposed, it relates only to the properties of the given curve and may be fulfilled, for example, at its points of inflexion. The *tangent at a point of inflexion* is also a superosculating straight line and its contact with the given curve is at least a three-point contact. We may proceed to add further conditions as far as the given curve possesses derivatives of higher order.

Example 2. Let us investigate the contact of the (plane) curve ${}^1k$,

$$ {}^1x = t\,,\ {}^1y = {}^1y(t)\,,\ {}^1z = 0 \ \text{(i.e. the curve } {}^1y = {}^1y(x)) \tag{5} $$

with the straight line

$$ {}^2x = t\,,\quad {}^2y = 0\,,\quad {}^2z = 0 \tag{6} $$

(i.e. with the x-axis). The necessary and sufficient condition that the curve (5) and the straight line (6) have contact of the first order at least ($q = 1$), at the point $t = 0$ (which is assumed to be their common point), is

$$ {}^1y(0) = {}^2y(0) = 0\,,\quad {}^1\dot{y}_0 = {}^2\dot{y}_0 = 0\,. \tag{7} $$

From this condition it follows that the straight line (6) is the *tangent* to the curve (5) at the contact point (0, 0). Therefore the function ${}^1y(x)$ may be expanded, in a sufficiently small neighbourhood of the point (0,0), into the power series

$$ {}^1y = \frac{x^2}{2!}\,{}^1\ddot{y}_0 + \frac{x^3}{3!}\,{}^1\dddot{y}_0 + \ldots + \frac{x^n}{n!}\left(\frac{\mathrm{d}^n\,{}^1y}{\mathrm{d}x^n}\right)_0 + \ldots \tag{8} $$

(see Fig. 9.12b for the case ${}^1\ddot{y}_0 \neq 0$, full line). The necessary and sufficient condition that the curve (5) and the tangent (6) have, at the point (0, 0), contact of the second order at least (see (4)), is that in addition to (7), at least ${}^1\ddot{y}_0 = 0$. From (8) it then follows that if ${}^1\ddot{y}_0 \neq 0$, the curve crosses the tangent ${}^2y = 0$ at the point (0,0) (Fig. 9.12a, full line). The point (0, 0) is called the *point of inflexion* of the curve (it is an *ordinary inflexion*, an *inflexion of the first order*). Similarly it can be shown that the vanishing of all derivatives of the function ${}^1y$ with respect to x up to the q-th derivative inclusive is a necessary and sufficient condition for contact at least of order q between the given curve and the tangent at the point (0, 0). If q is even, $q > 2$ and

$(d^{q+1}\,{}^1y/dx^{q+1})_0 \neq 0$, the curve crosses the tangent line at the point (0, 0) (Fig. 9.12a, the dashed line) and we speak of a *higher inflexion* (an *inflexion of higher order*, an inflexion of the order q). If q is odd, $q > 1$, the curve remains in a sufficiently small neighbourhood of the point (0, 0), on the same side of the tangent (Fig. 9.12b, the dashed line). In this case the contact point (0, 0) is called a *flat point.*

Example 3. Let us consider two (plane) curves ${}^1k$ and ${}^2k$,

$$ {}^1y = {}^1y(x), \quad {}^2y = {}^2y(x), \tag{9} $$

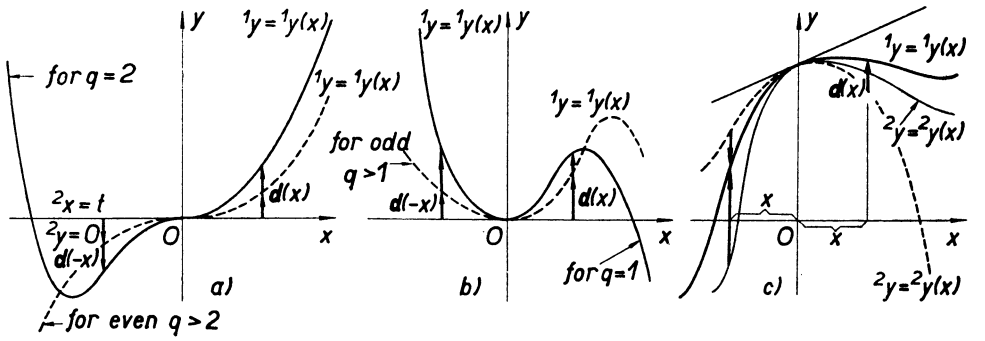


Fig. 9.12.

for which the relations

$$ {}^1y_0 = {}^2y_0, \quad {}^1y_0^{(p)} = {}^2y_0^{(p)} \quad (p = 1, 2, \ldots, q) $$

hold and where

$$ {}^1y_0^{(q+1)} \neq {}^2y_0^{(q+1)}. $$

In a sufficiently small neighbourhood of the point $x = 0$ the difference $d(x) = {}^1y(x) - {}^2y(x)$ may be represented by the expansion

$$ d(x) = \frac{x^{q+1}}{(q+1)!}({}^1y_0^{(q+1)} - {}^2y_0^{(q+1)}) + \frac{x^{q+2}}{(q+2)!}({}^1y_0^{(q+2)} - {}^2y_0^{(q+2)}) + \ldots. \tag{10} $$

If q is *even*, then $d(x)$ changes its sign at the point $x = 0$ (for sufficiently small x) and the curve ${}^2k$ *crosses* the curve ${}^1k$ at the point $x = 0$ (Fig. 9.12c, the curve ${}^2k$ is the dashed line). If q is odd, then $d(x)$ keeps its sign (for sufficiently small x) and the curve ${}^2k$ lies, in a neighbourhood of the point $x = 0$, *on the same side* of the curve ${}^1k$ (Fig. 9.12c, the curve ${}^2k$ is drawn as a narrow line).

Definition 4. A circle which passes through a point of a (space) curve (at which $k_1 \neq 0$) and has there contact of the second order at least (three-point contact) is called the *osculating circle* (the *circle of curvature*) of the curve at that point. The centre of

this circle is called the *centre of curvature* and its radius is the *radius of curvature* at that point.

REMARK 6. In general, contact of the second order exactly (a three-point contact), of a curve and a circle (or a plane) is contact of highest possible order. This means that, at a regular point of a curve, its osculating circle (or its osculating plane) is uniquely determined and we cannot construct a circle (or a plane) with contact of the third or higher order at that point. But there may be such points on the curve (at which $k_2 = 0$) that admit the existence of a circle (or a plane) with contact of a higher order.

Theorem 2. *The osculating circle at a point of a curve lies in the osculating plane of the curve at that point, its radius is equal to the radius of curvature* r_1 *of the curve and the radius vector of its centre is*

$$\bar{\mathbf{r}} = \mathbf{r} + r_1\mathbf{n} \quad (\mathbf{r} \text{ is the radius vector of the point of the curve}),$$

i.e. *the centre lies on the principal normal constructed at the point of the curve under consideration.*

REMARK 7. If two curves have the same osculating circle at their common point, they have contact of the second order at least (three-point contact) at that point, and conversely.

Theorem 3. *The orthogonal projection of a curve into the osculating plane (at a given point of the curve) is a plane curve which has a common osculating circle with the curve at the point mentioned.*

Theorem 4. *A curve* $y = y(x)$ *has at every point* (x, y) *for which* $\ddot{y} \neq 0$ (i.e., for example, the points of inflexion are excluded) *a unique osculating circle; its radius (the radius of curvature)* r_1 *and the coordinates* (m, n) *of its centre S are given by the expressions*

$$r_1 = \frac{(1 + \dot{y}^2)^{3/2}}{|\ddot{y}|}, \quad m = x - \dot{y}\frac{1 + \dot{y}^2}{\ddot{y}}, \quad n = y + \frac{1 + \dot{y}^2}{\ddot{y}}. \tag{11}$$

REMARK 8. If the osculating circle of a curve at its point has *exactly* three-point contact with this curve, then it crosses the curve at the contact point.

Theorem 5. *In the case of the curve* $F(x, y) = 0$, *the radius* r_1 *(the radius of curvature) and the coordinates* (m, n) *of the centre S of the osculating circle are given by the expressions*

$$r_1 = \frac{(F_x^2 + F_y^2)^{3/2}}{|J|}, \quad m = x - F_x\frac{F_x^2 + F_y^2}{J}, \quad n = y - F_y\frac{F_x^2 + F_y^2}{J}, \tag{12}$$

where

$$J = F_{xx}F_y^2 - 2F_{xy}F_xF_y + F_{yy}F_x^2 \neq 0. \tag{13}$$

(*It is assumed that the tangent at the point* (x, y) *of the curve is not parallel to the y-axis.*)

Theorem 6. *A curve* $x = \varphi(t)$, $y = \psi(t)$ *has at every point t, for which* $\dot{\varphi}\ddot{\psi} - \ddot{\varphi}\dot{\psi} \neq 0$, *a unique osculating circle; its radius (the radius of curvature)* r_1 *and the coordinates m, n of its centre S are given by the expressions*

$$r_1 = \frac{(\dot{\varphi}^2 + \dot{\psi}^2)^{3/2}}{|\dot{\varphi}\ddot{\psi} - \ddot{\varphi}\dot{\psi}|}, \quad m = \varphi - \dot{\psi}\frac{\dot{\varphi}^2 + \dot{\psi}^2}{\dot{\varphi}\ddot{\psi} - \ddot{\varphi}\dot{\psi}}, \quad n = \psi + \dot{\varphi}\frac{\dot{\varphi}^2 + \dot{\psi}^2}{\dot{\varphi}\ddot{\psi} - \ddot{\varphi}\dot{\psi}}. \tag{14}$$

REMARK 9. To find the *points of inflexion* of a curve $y = f(x)$ we solve the equation of the curve and the equation $\ddot{y} = 0$. Such points are really points of inflexion if the order of the highest non-vanishing derivative is odd ($y^{(p)} \neq 0$, $p > 2$, p odd). The inflexion points at which the tangent is perpendicular to the x-axis must be found separately. The coordinates of a point of inflexion of a curve $F(x, y) = 0$ may be found from the common solution of the equation of the curve and the equation

$$J \equiv F_{xx}F_y^2 - 2F_{xy}F_xF_y + F_{yy}F_x^2 = 0,$$

which is satisfied by the coordinates of the point of inflexion (this condition is only a necessary one). But the equation $J = 0$ is also satisfied by the *coordinates of all singular points* of the curve ($F_x = F_y = 0$). Thus, among the points obtained, there will be the points of inflexion and also the singular points.

The coordinates of a point of inflexion of a curve $x = \varphi(t)$, $y = \psi(t)$ may be determined from the equation $\dot{\varphi}\ddot{\psi} - \ddot{\varphi}\dot{\psi} = 0$, which is satisfied by the parameter of a point of inflexion (again, this is only a necessary condition because the last equation is satisfied by *all singular points* of the given curve).

Definition 5. If a point of a curve admits the existence of a circle which at this point has contact of the third order at least (four-point contact) with the curve, then such a circle is called the *superosculating circle* of the curve at that point.

Theorem 7. *At a point* (x, y) *of a curve* $y = y(x)$ *for which*

$$3\dot{y}\ddot{y}^2 - (1 + \dot{y}^2)\dddot{y} = 0 \tag{15}$$

the osculating circle has with the curve contact of the third order at least (four-point contact), i.e. *the circle is the superosculating circle.*

REMARK 10. The superosculating circle that has exactly four-point contact with a given plane curve at its point (x, y) lies, in a certain neighbourhood of the point of contact, on the same side of the curve.

Theorem 8. *The radius of a superosculating circle is a stationary value of the radius of curvature along a given curve (we assume that* $r_1(t)$ *has a sufficient number of derivatives).*

Definition 6. A point on a plane curve at which the radius of curvature attains its (relative) extreme value is called an *apex* of the curve; it is a point at which the plane curve has contact of an odd order with the superosculating circle.

Definition 7. The straight line through the centre of the osculating circle at a point of a (space) curve and perpendicular to the corresponding osculating plane is called the *polar line* associated with that point.

REMARK 11. The polar line associated with a point of a curve is parallel to the binormal at that point.

REMARK 12. A curve, for which the equation

$$\frac{r_1}{r_2} + (r_1' r_2)' = 0 \quad \left(r_1' = \frac{dr_1}{ds}, \quad r_2 \neq 0\right)$$

is satisfied at every of its points, lies entirely on a sphere and is called a *spherical curve.* The osculating circle at any point of such curve is the intersection of the sphere, on which the curve lies, with the osculating plane at that point.

9.6. Asymptotes. Singular Points of Plane Curves

Definition 1. Let a point T be given on a plane curve and let v be the distance of the point T from a straight line p. If $\lim v = 0$, when at least one of the coordinates of the point T tends to $\pm\infty$, then the straight line p is called an *asymptote* of the given curve.

There may be two kinds of asymptotes of the curve $y = f(x)$: If $\lim f(x) = \pm\infty$, when $x \to c$ (also only when $x \to c+$, or $x \to c-$, respectively), then the straight line $x = c$ (parallel to the y-axis) is an asymptote of the curve. For asymptotes that are not parallel to the y-axis, the following theorem holds:

Theorem 1. *If a curve $y = f(x)$ is defined in the interval $[a, +\infty)$ and the limits*

$$k = \lim_{x \to +\infty} \frac{f(x)}{x}, \tag{1}$$

$$q = \lim_{x \to +\infty} (f(x) - kx) \tag{2}$$

exist, then the curve has an asymptote; its equation is

$$y = kx + q\,. \tag{3}$$

REMARK 1. An entirely similar theorem may be formulated for the interval $(-\infty, a]$.

The tangent to a curve at infinity is an asymptote of this curve. But there are asymptotes of another kind (see Theorem 2):

Theorem 2. *If the equation of the given curve is of the form*

$$y = kx + q + \mu(x) \tag{4}$$

where $\lim_{x \to +\infty} \mu(x) = 0$, *then the curve has the asymptote* $y = kx + q$.

Theorem 3. *If, for a plane curve* $y = f(x)$,

$$\lim_{|x| \to +\infty} y = q \quad \text{or} \quad \lim_{|y| \to +\infty} x = c\,,$$

then the curve has the asymptote

$$y = q \quad \text{or} \quad x = c\,, \tag{5}$$

respectively.

Theorem 4. *If the equation of an algebraic curve can be put in the form*

$$x^n \varphi(y) + x^{n-1} \varphi_1(y) + x^{n-2} \varphi_2(y) + \ldots = 0\,,$$

or

$$y^m \psi(x) + y^{m-1} \psi_1(x) + y^{m-2} \psi_2(x) + \ldots = 0$$

(we suppose that the polynomials $\varphi(y)$, $\varphi_1(y)$, ..., *or* $\psi(x)$, $\psi_1(x)$, ..., *respectively, have no common factor) then the curve has asymptotes parallel to the x-axis or to the y-axis, given by*

$$\varphi(y) = 0 \quad \textit{or} \quad \psi(x) = 0\,, \quad \textit{respectively}\,. \tag{6}$$

Theorem 5. *If an algebraic curve of degree* $n \geqq 2$,

$$F(x, y) \equiv \varphi_n(x, y) + \varphi_{n-1}(x, y) + \ldots = 0\,,$$

where $\varphi_h(x, y)$ *denotes the sum of terms of degree h, has the asymptote* $y = kx + q$, *then k satisfies the equation*

$$\varphi_n(1, k) = 0 \tag{7}$$

and q the equation

$$q = -\frac{\varphi_{n-1}(1, k)}{\varphi_n'(1, k)} \quad \left(\varphi_n'(1, k) = \frac{\partial \varphi_n}{\partial y}(1, k) \neq 0\right), \tag{8}$$

respectively.

REMARK 2. Theorem 5 thus states the following: Substitute $y = kx + q$ into the equation of the given algebraic curve and equate the coefficient of the two highest powers of the variable x to zero. From the first condition we compute the values k_i

$(i = 1, 2, \ldots, n)$ and substitute these values into the second condition to determine the q_i corresponding to the particular k_i. In this way we find also the asymptotes parallel to the x-axis (but not those parallel to the y-axis). If $\varphi_n'(1, k) = \varphi_{n-1}(1, k) = 0$, then we find q from the equation

$$\tfrac{1}{2}q^2\,\varphi_n''(1, k) + q\,\varphi_{n-1}'(1, k) + \varphi_{n-2}(1, k) = 0\,.$$

Example 1. The curve

$$x^3 + y^3 - 3axy = 0$$

(folium of Descartes) is an algebraic curve. We substitute $y = kx + q$ and equate the coefficients of x^3 and x^2 to zero, so that

$$x^3 + k^3x^3 + 3k^2qx^2 + 3kq^2x + q^3 - 3akx^2 - 3aqx = 0$$

giving

$$1 + k^3 = 0 \quad \text{and} \quad 3k^2q - 3ak = 0\,.$$

The only real solution to the first condition is $k = -1$, which when substituted into the second condition gives

$$3q + 3a = 0\,.$$

Hence $q = -a$ and the asymptote is

$$y = -x - a\,.$$

Theorem 6. *There exists only one straight line passing through a regular point of a curve (it is the tangent) such that at this point, the line and the curve have two-point contact at least. For all other straight lines there is only one-point contact.*

Definition 2. A singular point (x_0, y_0) of the curve $F(x, y) = 0$ is called a *double point* of the curve if $F_x^0 = F_y^0 = 0$ and if, at the same time, not all partial derivatives of the second order of $F(x, y)$ vanish at this point.

Theorem 7. *There are at most two (real) straight lines passing through a double point (x_0, y_0) of a curve $F(x, y) = 0$ such that the point (x_0, y_0) is their point of intersection with the curve and has a multiplicity at least three.*

The slopes k of these straight lines satisfy the equation

$$F_{xx}^0 + 2F_{xy}^0 k + F_{yy}^0 k^2 = 0\,. \tag{9}$$

There are two or one or no such straight lines according as to whether

$$F_{xx}^0 F_{yy}^0 - (F_{xy}^0)^2 < 0\,, \quad or \quad = 0\,, \quad or \quad > 0\,, \tag{10}$$

respectively.

Definition 3. The straight lines that pass through a double point of a curve $F(x, y) = 0$ and the slopes of which satisfy equation (9) are called the *tangents to the curve, at its double point.*

Definition 4. A double point of a curve $F(x, y) = 0$ for which

$$\text{a)} \quad F_{xx}F_{yy} - F_{xy}^2 < 0\,,$$

is called a *double point with (two) distinct tangents* or a *node* (the curve has two branches through the singular point, each of them touching the corresponding tangent (Fig. 9.13a));

$$\text{b)} \quad F_{xx}F_{yy} - F_{xy}^2 > 0\,,$$

is called an *isolated point* (the curve has no (real) tangent at this point (Fig. 9.13c, point P));

$$\text{c)} \quad F_{xx}F_{yy} - F_{xy}^2 = 0\,,$$

is called a *cusp* (the curve has two branches which tend to the singular point from one side only and lie on opposite sides of the tangent (Fig. 9.13b)).

REMARK 3. In case c) the curve may have a *point of self-tangency* (a double cusp) (Fig. 9.13d).

REMARK 4. If at a point (x_0, y_0) of a curve $F(x, y) = 0$, in addition to $F_x^0 = F_y^0 = 0$, $F_{xx}^0 = F_{yy}^0 = F_{xy}^0 = 0$ and at least one of the derivatives of the third order of $F(x, y)$ at that point is non-zero, then the curve has a *triple point* (a singular point of multiplicity three) at that point etc. Besides the multiple points, singular points of other kind may also exist, e.g. the *end point* (on the curve $y = e^{1/x}$, see Fig. 9.14a) or the *angular point* (Fig. 9.14b).

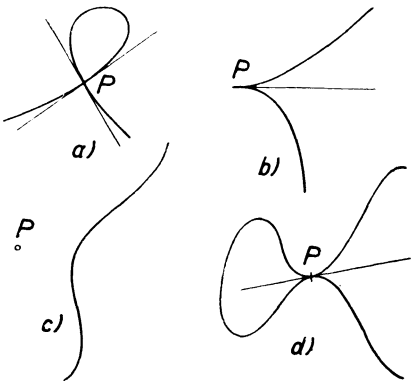


Fig. 9.13.

Theorem 8. *The coordinates* (X, Y) *of a current point on the tangents at the double point* $(0, 0)$ *or* (x_0, y_0) *of a given curve* $F(x, y) = 0$ *satisfy the equations*

$$F_{xx}^0 X^2 + 2F_{xy}^0 XY + F_{yy}^0 Y^2 = 0$$

(*the derivatives being computed at the point* $(0, 0)$) (11)

or

$$F^0_{xx}(X - x_0)^2 + 2F^0_{xy}(X - x_0)(Y - y_0) + F^0_{yy}(Y - y_0)^2 = 0$$
(*the derivatives being computed at the point* (x_0, y_0)). (12)

REMARK 5. A curve $F(x, y) = 0$ has a singular (multiple) point at points whose coordinates satisfy the equations

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

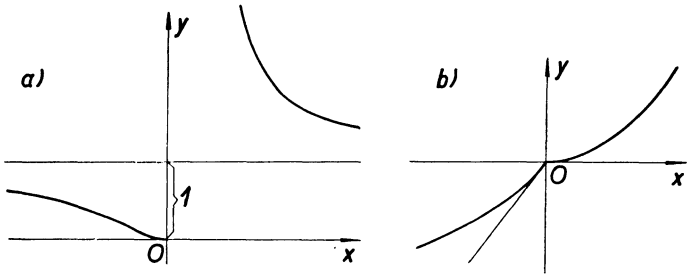


Fig. 9.14.

A deeper investigation of these points can be made by means of the analysis of higher derivatives. In the case of algebraic curves we may proceed as follows: If an algebraic curve passes through the origin and if the origin is a regular point, then we obtain the equation of the tangent at this point by equating to zero the sum of terms of the first order of the equation of the curve. If the lowest terms of the equation of an algebraic curve are of degree at least two then the origin is its singular point. The lowest degree of terms of the equation of the curve gives the multiplicity of the singular point (at the origin). If we equate to zero the sum of lowest terms of the equation of the curve we obtain an equation which represents the tangents at the singular point.

REMARK 6. A curve $x = \varphi(t)$, $y = \psi(t)$ may have a node at a point (x_0, y_0) when the relations $x_0 = \varphi(t_1) = \varphi(t_2)$, $y_0 = \psi(t_1) = \psi(t_2)$ hold for two different values t_1 and t_2 of the parameter. A cusp may occur if $\dot{\varphi}(t) = \dot{\psi}(t) = 0$ holds for a value of the parameter.

9.7. Envelopes of a One-parameter Family of Plane Curves

Definition 1. We say that the equation

$$F(x, y, c) = 0 \tag{1}$$

defines a *one-parameter family of curves* in a domain O of the xy-plane if

1. the function $F(x, y, c)$ is a continuous function of the variables x, y, c for $(x, y) \in O$ and c of an interval I,

2. for every $c \in I$, equation (1) defines a certain curve in the domain O (cf. § 9.2) in such a manner that to any two different values of $c \in I$ there correspond two different curves.

Example 1. The equation

$$y - cx^2 = 0 \tag{2}$$

defines a one-parameter family of parabolas passing through the origin (for $c = 0$ the parabola reduces to the straight line $y = 0$).

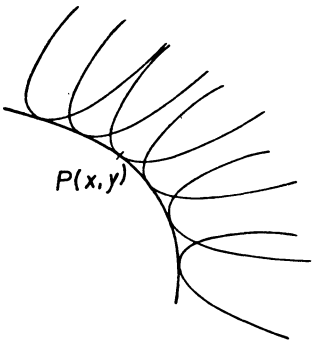


Fig. 9.15.

Definition 2. A curve is called (if such a curve acutally exists) the *envelope* of the family (1), if it touches every curve of the given family and, conversely, if every point $P(x, y)$ of the curve is the point of contact with a curve of the family (Fig. 9.15).

Theorem 1. *Let the function $F(x, y, c)$ have continuous partial derivatives*

$$\frac{\partial F}{\partial x}, \quad \frac{\partial F}{\partial y}, \quad \frac{\partial F}{\partial c}, \quad \frac{\partial^2 F}{\partial c\, \partial x}, \quad \frac{\partial^2 F}{\partial c\, \partial y}, \quad \frac{\partial^2 F}{\partial c^2} \tag{3}$$

in a neighbourhood of the point (x_0, y_0, c_0). At the point (x_0, y_0, c_0) let the following relations be fulfilled:

$$F(x_0, y_0, c_0) = 0, \quad \frac{\partial F}{\partial c}(x_0, y_0, c_0) = 0, \tag{4}$$

$$\frac{\partial^2 F}{\partial c^2} \neq 0, \quad \begin{vmatrix} \dfrac{\partial F}{\partial x}, & \dfrac{\partial F}{\partial y} \\ \dfrac{\partial^2 F}{\partial c\, \partial x}, & \dfrac{\partial^2 F}{\partial c\, \partial y} \end{vmatrix} \neq 0. \tag{5}$$

Then in a certain neighbourhood U of the point (x_0, y_0) and for c from a definite neighbourhood V of the point c_0, there exists an envelope of the family (1).

REMARK 1. We may obtain the equation of the envelope from the equations

$$F(x, y, c) = 0\,, \quad \frac{\partial F}{\partial c}(x, y, c) = 0 \tag{6}$$

by expressing x and y as functions of the variable c (this can be done because of the second condition (5)) or by expressing c as a function of the variables x, y (this can be done because of the first condition (5)) and substituting for c into the first equation (6),

$$F(x, y, c(x, y)) = 0 \tag{7}$$

(i.e. by "elimination" of the parameter c from equations (6)).

REMARK 2. The conditions (5) are sufficient but not necessary for the existence of the envelope in a neighbourhood of the point (x_0, y_0):

Example 2. The envelope of the family of curves (occasionally called parabolas of the fourth order)

$$F(x, y, c) \equiv y - (x - c)^4 = 0\,, \tag{8}$$

is evidently the x-axis; we easily obtain its equation $y = 0$ from equations (6),

$$y - (x - c)^4 = 0\,, \quad 4(x - c)^3 = 0\,. \tag{9}$$

However, at every point of this envelope, both expressions (5) are zero:

$$\frac{\partial^2 F}{\partial c^2} = -12(x-c)^2\,, \quad \begin{vmatrix} \dfrac{\partial F}{\partial x}, & \dfrac{\partial F}{\partial y} \\ \dfrac{\partial^2 F}{\partial c\,\partial x}, & \dfrac{\partial^2 F}{\partial c\,\partial y} \end{vmatrix} = \begin{vmatrix} -4(x-c)^3, & 1 \\ 12(x-c)^2, & 0 \end{vmatrix} = -12(x-c)^2$$

(in consequence of (8) and of the equation $y = 0$).

If conditions (5) are not fulfilled, equations (6) need not define the envelope:

Example 3. Let us consider the family of curves

$$F(x, y, c) \equiv (y - c)^2 - (x - c)^3 = 0\,. \tag{10}$$

(Fig. 9.16.) From (10) it follows that

$$\frac{\partial F}{\partial c} = -2(y - c) + 3(x - c)^2 = 0\,,$$

i.e.

$$y - c = \tfrac{3}{2}(x - c)^2\,. \tag{11}$$

Substituting (11) into (10) we obtain

$$\tfrac{9}{4}(x - c)^3 (x - c - \tfrac{4}{9}) = 0 .$$

Hence either $x - c = 0$ or $x - c - \frac{4}{9} = 0$.

1. $x - c = 0$. Then from (11) it follows that $y - c = 0$. Equations $x = c$, $y = c$ are parametric equations of the straight line $y = x$.

2. $x - c = \frac{4}{9}$. Then, by (11), $y - c = \frac{8}{27}$.

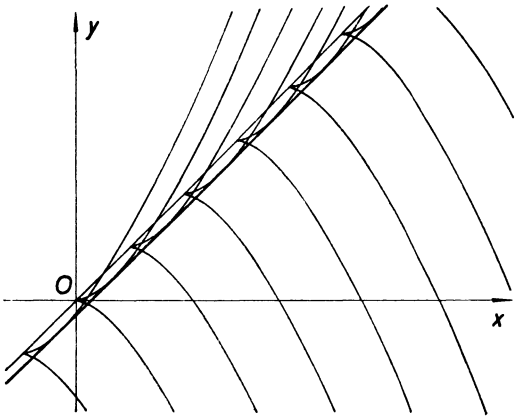


Fig. 9.16.

Equations

$$x = c + \tfrac{4}{9}, \quad y = c + \tfrac{8}{27} \tag{12}$$

are parametric equations of the straight line

$$y = x - \tfrac{4}{27} . \tag{13}$$

Equations (10) and (11) provide two straight lines, the straight line (13) and the straight line $y = x$. At every point of the former straight line the relation

$$\frac{\partial^2 F}{\partial c^2} \equiv 2 - 6(x - c) = 2 - \tfrac{8}{3} = -\tfrac{2}{3}$$

holds (using the first equation (12)) and

$$\begin{vmatrix} \dfrac{\partial F}{\partial x}, & \dfrac{\partial F}{\partial y} \\ \dfrac{\partial^2 F}{\partial c\, \partial x}, & \dfrac{\partial^2 F}{\partial c\, \partial y} \end{vmatrix} \equiv \begin{vmatrix} -3(x - c)^2, & 2(y - c) \\ 6(x - c), & -2 \end{vmatrix} = 6(x - c)[(x - c) - 2(y - c)] =$$

$$= \tfrac{8}{3}(\tfrac{4}{9} - \tfrac{16}{27}) = -\tfrac{32}{81}$$

(also using equations (12)). Thus, both expressions (5) are non-zero at all points of the straight line (13), which is the envelope of the family (10), since the other conditions of Theorem 1 are evidently fulfilled. The straight line $y = x$ (see Fig. 9.16) is the locus of singular points (the cusps) and evidently it is not an envelope of the curves (10). The reader may easily verify that the determinant (5) vanishes at every point of this straight line.

9.8. Parallel Curves, Gradient Curves, Evolutes and Involutes

Definition 1. Two curves are said to be *parallel curves* (so-called *pair of Bertrand curves*) if there exists a continuous one-to-one correspondence between their points such that both curves have the same principal normal at the corresponding points.

Theorem 1. *If two curves*

$$\mathbf{r} = \mathbf{r}(s), \quad \bar{\mathbf{r}} = \bar{\mathbf{r}}(s)$$

are parallel curves, then the relation

$$\bar{\mathbf{r}}(s) = \mathbf{r}(s) + c\,\mathbf{n}(s) \quad (c \text{ being a constant}) \tag{1}$$

holds.

REMARK 1. There *need not* exist a parallel curve to each *space* curve. The constant c in (1) determines the distance between corresponding points on the common principal normals. Every *plane* curve possesses an infinite number of parallel curves (for different values of the real constant c in (1)).

A (plane) curve k,

$$x = \varphi(t), \quad y = \psi(t)$$

has a parallel curve $\bar{k}$

$$\bar{x} = \varphi(t) + c\frac{\dot{\psi}}{\sqrt{(\dot{\varphi}^2 + \dot{\psi}^2)}}, \quad \bar{y} = \psi(t) - c\frac{\dot{\varphi}}{\sqrt{(\dot{\varphi}^2 + \dot{\psi}^2)}} \tag{2}$$

(c being an arbitrary real constant). To a given plane curve k we may construct a parallel curve $\bar{k}$ in this manner: We mark off a segment of constant length c on the normal at every point of the given curve, always on the same side of the curve, the starting point of the segment being the point on the given curve (Fig. 9.17). The end points of these segments form a plane curve $\bar{k}$ parallel to the given curve. If the parameter t can be eliminated from equations (2), we obtain the equation of a one-parameter family of curves (with the parameter c) parallel to the given curve. Any two curves of this family are mutually parallel. A plane curve $\bar{k}$ parallel to a given plane curve k is said to be an *equidistant curve* of the curve k (the curves k and $\bar{k}$ are said to be the *curves of equidistance* or *equidistant curves*). The equation of the equi-

distant curve to the curve $F(x, y) = 0$ (in the variables (X, Y)) may be obtained by the elimination of (x, y) from the equations

$$F(x, y) = 0, \quad Y - y = -\frac{dx}{dy}(X - x), \quad (X - x)^2 + (Y - y)^2 = c^2.$$

The radii of curvature at points on the common normal of two parallel curves differ by the constant length c (the centre of curvature is common). A circular helix is the only

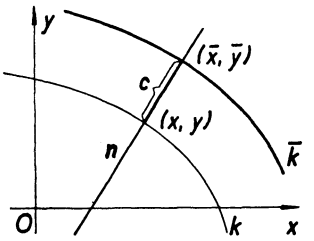


Fig. 9.17.

space curve that also has an infinite number of parallel curves, each of them being a circular helix (lying on a circular cylinder with the same axis as that of the given helix).

Definition 2. A *gradient curve with respect to a fixed direction* is a curve, the tangents to which make a constant angle ω with that fixed direction.

Theorem 2. *A necessary and sufficient condition that a curve be a gradient curve with respect to a given direction is that the relation*

$$k_2 \sin \omega - k_1 \cos \omega = 0 \tag{3a}$$

holds at every point of the curve (k_1 *and* k_2 *being the curvature and torsion respectively*).

REMARK 2. Gradient curves are sometimes called *cylindrical helices*. Their characteristic feature is that the equation

$$\frac{k_1}{k_2} = \tan \omega = k \quad (k = \text{const}). \tag{3b}$$

holds along the entire curve. The circular helix is a gradient curve with respect to its axis.

Definition 3. An orthogonal trajectory $\bar{k}$ of the tangents of a curve k is called an *involute* of the curve k. The curve k is called an *evolute* of $\bar{k}$.

REMARK 3. An *orthogonal trajectory* of the tangents of a curve is a curve that intersects the tangents to the given curve at right angles.

Theorem 3. *The equation of an involute* $\bar{k}$ *of the given evolute* $k \equiv \mathbf{r} = \mathbf{r}(s)$ *is*

$$\bar{\mathbf{r}} = \mathbf{r} - (s + c)\,\mathbf{t} \quad (c \text{ is a constant; Fig. 9.18}). \tag{4}$$

REMARK 4. If a curve k (the evolute) is a plane curve, then its involutes (for different values of the constant c) are also plane curves. Every two of these involutes have common normals in the tangents to the curve k (the evolute) and are parallel curves (Fig. 9.18).

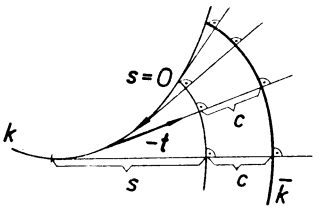


Fig. 9.18.

REMARK 5. To every curve there exists an infinite number of evolutes (forming a one-parameter family of curves) and of involutes (c in (4) is the parameter of the family of involutes).

Theorem 4. *The evolutes of a plane curve* $\bar{k}$ *are gradient curves (with respect to the perpendicular to the plane of the given curve). Among them, there is one plane evolute* k *formed by the centres of curvature of the given curve. The equation of this plane evolute is*

$$\mathbf{r} = \bar{\mathbf{r}} + \bar{r}_1 \bar{\mathbf{n}} \tag{5}$$

($\bar{r}_1$ *is the radius of curvature at the variable point of the given plane curve* $\bar{k}$).

REMARK 6. The *plane evolute* of a curve $x = \varphi(t)$, $y = \psi(t)$ is expressed parametrically by equations (9.5.14) in Theorem 9.5.6, p. 287, defining the coordinates (m, n) of the centre of curvature. By the elimination of t from these equations we obtain the equation of the evolute in the form $F(m, n) = 0$. If a curve is given by an equation $y = f(x)$ or $F(x, y) = 0$, we find the equation of its evolute by the elimination of the variables (x, y) from equations (9.5.11) (Theorem 9.5.4) and $y = f(x)$, or equations (9.5.12) and $F(x, y) = 0$, respectively. If a curve is given by an equation $y = f(x)$, we may retain the parametric expression for an evolute if we choose, for example, x as the parameter.

Example 1. For the parabola $y = x^2$, we have

$$\dot{y} = \frac{\mathrm{d}y}{\mathrm{d}x} = 2x\,, \quad \ddot{y} = \frac{\mathrm{d}^2 y}{\mathrm{d}x^2} = 2$$

and by (9.5.11)

$$m = x - 2x\,\frac{1 + 4x^2}{2} = -4x^3\,, \quad n = x^2 + \frac{1 + 4x^2}{2} = \tfrac{1}{2} + 3x^2\,.$$

From the first equation it follows that

$$x = -\sqrt[3]{\frac{m}{4}};$$

on substituting for x into the second equation, we obtain the equation of the evolute of the given parabola,

$$n = \tfrac{1}{2} + 3\sqrt[3]{\frac{m^2}{16}} \quad \text{or} \quad (n - \tfrac{1}{2})^3 = \tfrac{27}{16}m^2 .$$

Conversely the parabola $y = x^2$ is the involute of this curve.

REMARK 7. Eliminating m and n from the equations

$$m = x - \dot{y}\frac{1 + \dot{y}^2}{\ddot{y}}, \quad n = y + \frac{1 + \dot{y}^2}{\ddot{y}}, \quad F(m, n) = 0,$$

(see (9.5.11)) we obtain the *differential equation of the involutes* of the given curve $F(m, n) = 0$.

REMARK 8. The plane evolute of a given plane curve is the envelope of normals of this curve. The centre of curvature at a fixed point of a given plane curve is the limiting point of the point of intersection of the normal at that point and the normal at a point which tends to the given point along the curve. A normal of a given plane curve is a tangent to its plane evolute with the point of contact at the corresponding centre of curvature of the given plane curve.

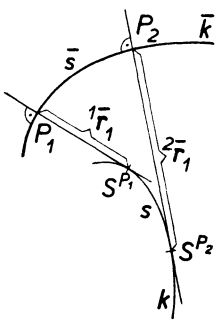


Fig. 9.19.

Theorem 5. *The length of the arc s on the plane evolute k of a given plane curve $\bar{k}$ is equal (in absolute value) to the difference in length of the radii of curvature of the curve $\bar{k}$ that lie on the normals touching the evolute at the end points of the arc s* (Fig. 9.19), i.e.

$$|s| = |^2\bar{r}_1 - {}^1\bar{r}_1| \tag{6}$$

($s = \overset{\frown}{S^{P_1}S^{P_2}}$ *is the arc of the evolute k,* S^{P_1} *and* S^{P_2} *are the centres of curvature*

corresponding to the points P_1 and P_2 on the given curve $\bar{k}$ and ${}^1\bar{r}_1$ and ${}^1\bar{r}_2$ are the corresponding radii of curvature).

REMARK 9. It is assumed in Theorem 5 that there are no points of inflexion or singular points on the arc $\bar{s} = \overset{\frown}{P_1P_2}$ of the curve $\bar{k}$ and that no inner point of this arc is a summit of the given curve $\bar{k}$.

REMARK 10. To every curve k it is possible to construct an infinite number of involutes as follows: we choose a fixed point on the curve k and then mark off on every tangent of the curve k a segment whose length is equal to the length of the arc of the curve k measured from the fixed point to the point of contact of the tangent; the starting point of the segment is the point of contact of the tangent, the end point of the segment is a point of the involute so constructed. The plane involutes of the same family have a single plane evolute in common. To a regular point (which is not a point of inflexion) of a plane curve with maximal or minimal curvature (i.e., to a summit of the curve) there corresponds a cusp of its plane evolute. The normal at a point of inflexion of a plane curve is an asymptote of its plane evolute. The plane evolute passes through the cusps of its plane involute having there its tangent perpendicular to the double tangent at the cusp of the involute (cf. Example 1). Using equation (6), we may determine the length of the arc of the evolute (or of any curve that may be considered as an evolute) if the radii of curvature of its involute are known.

9.9. Direction of the Tangent, Curvature and Asymptotes of Plane Curves in Polar Coordinates

For analytic representation of many plane curves, it is often convenient to use polar coordinates. In the system of polar coordinates a plane curve is represented by an equation expressing a relation between the polar coordinates (ϱ, φ):

$$F(\varrho, \varphi) = 0 \quad \text{or} \quad \varrho = f(\varphi).$$

The direction of a tangent to a curve expressed in polar coordinates is determined by the angle ϑ between the tangent and the radius vector of its point of contact; this angle ϑ is called the *direction angle of the tangent.*

Theorem 1. *The direction angle* ϑ (Fig. 9.20) *is determined by the relation*

$$\tan \vartheta = \varrho/\varrho' \quad (0 < \vartheta < \pi, \vartheta \neq \tfrac{1}{2}\pi, \varrho' = \mathrm{d}\varrho/\mathrm{d}\varphi \neq 0). \tag{1}$$

REMARK 1. The tangent to the curve at the point $P(\varrho, \varphi)$ makes an angle $\varphi + \vartheta$ with the polar axis o, the normal at the point P on the curve and the radius vector of the point P are inclined at an angle $\vartheta + \frac{1}{2}\pi$, while the same normal and the polar

axis o are inclined at an angle $\varphi + \vartheta + \frac{1}{2}\pi$. The distance v of the tangent t from the origin O is given by the relation

$$v = \frac{\varrho^2}{\sqrt{(\varrho^2 + \varrho'^2)}} \quad (\varrho \text{ being the radius vector of the point of contact}).$$

REMARK 2. In order to construct the tangent or the normal at the point P or to determine their mutual position, it may sometimes be convenient to find the (oriented) length of a segment on the tangent (the *length of the polar tangent* t) and on the

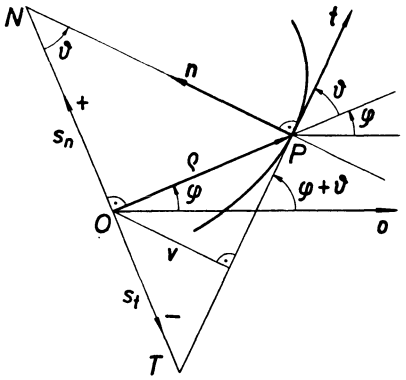


Fig. 9.20.

normal (*the length of the polar normal* n). These lengths are measured from the point of contact P to their respective points of intersection with the perpendicular to the radius vector of the point P through the origin O. We may also need the orthogonal projections of these two segments on the above-mentioned perpendicular, the so-called *polar subtangent* s_t and *subnormal* s_n. The relations

$$s_t = \varrho \tan \vartheta = \frac{\varrho^2}{\varrho'}, \quad s_n = \frac{\varrho}{\tan \vartheta} = \varrho' \tag{2}$$

hold for s_t and s_n (Fig. 9.20). (The segments s_t and s_n are taken either oriented as shown in Fig. 9.20, or in absolute values.) For the length t of the polar tangent and the length n of the polar normal the following relations hold:

$$t = \left|\frac{\varrho}{\varrho'}\right| s', \quad n = s' \quad \left(s' = \frac{\mathrm{d}s}{\mathrm{d}\varphi} = \sqrt{(\varrho^2 + \varrho'^2)}, \ \mathrm{d}s \text{ being the differential of the arc}\right). \tag{3}$$

Theorem 2. *The radius of the circle of curvature of the curve* $F(\varrho, \varphi) = 0$ *is given by the expression*

$$r_1 = \frac{(\varrho^2 + \varrho'^2)^{3/2}}{\varrho^2 + 2\varrho'^2 - \varrho\varrho''}. \tag{4}$$

The coordinates ϱ_0, φ_0 *of the centre of the circle of curvature satisfy the equations*

$$\varrho_0 \cos(\varphi_0 - \varphi) = \frac{\varrho(\varrho'^2 - \varrho\varrho'')}{\varrho^2 + 2\varrho'^2 - \varrho\varrho''},$$

and (5)

$$\varrho_0 \sin(\varphi_0 - \varphi) = \frac{\varrho'(\varrho^2 + \varrho'^2)}{\varrho^2 + 2\varrho'^2 - \varrho\varrho''}.$$

REMARK 3. If we eliminate the variables φ and ϱ from equations (5) and the equation $F(\varrho, \varphi) = 0$, we obtain the equation of the plane evolute of the given curve.

Definition 1. An *asymptote* of a curve $F(\varrho, \varphi) = 0$ is a straight line intersecting the polar axis at an angle, for which

$$\lim_{\varrho \to +\infty} \varphi = \alpha \tag{6}$$

and at a distance v from the origin for which

$$v = \lim_{\varphi \to \alpha} \left(-\frac{\varrho^2}{\varrho'} \right) \quad \text{(see (2))} \tag{7}$$

(on the assumption that the limits (6) and (7) exist).

REMARK 4. According to Definition 1, we find an asymptote (if it exists) of a curve as follows: First we determine its direction defined by the angle α, i.e. the angle of inclination to the polar axis and then its distance v from the origin. This distance is the limiting value of the polar subtangent (see (2)) for $\varphi \to \alpha$.

Example 1. For the hyperbolic spiral (reciprocal spiral)

$$\varrho = \frac{a}{\varphi} \quad (a > 0, \varphi > 0),$$

the relation $\varrho \to +\infty$ yields

$$\varphi \to 0, \quad \text{hence} \quad \alpha = 0.$$

Further,

$$v = \lim_{\varphi \to 0} \left(-\frac{\varrho^2}{\varrho'} \right) = \lim_{\varphi \to 0} \frac{a^2/\varphi^2}{a/\varphi^2} = a.$$

Thus the asymptote is a line parallel to, and at a distance a from the polar axis.

REMARK 5. Besides the asymptotes there may exist, with some plane curves, so-called *asymptotic points*. An asymptotic point of a plane curve is a point which does not lie on the curve but to which a point moving along the curve approaches to

within an arbitrarily small distance. (This occurs, for example, in the case of the clothoid; cf. § 4.8.) Of a similar meaning is the concept of an *asymptotic curve* (so-called *curvilinear asymptote*). For example, for the curve

$$\varrho = \frac{a\varphi}{\varphi - 1} \quad (a \neq 0 \text{ is a real constant})$$

we have $\lim \varrho = a$ when $|\varphi| \to +\infty$, thus the curve has an *asymptotic circle* of radius a and the curve considered approaches this circle asymptotically from outside and inside (for $\varphi \to +\infty$ and $\varphi \to -\infty$, respectively).

9.10. Supplementary Notes to Part A

a) To determine the *equations of the tangents drawn to the curve* $F(x, y) = 0$ *from an arbitrary point* (x_0, y_0), we find first the coordinates of their points of contact by solving the equations

$$\left.\begin{aligned} F(x, y) &= 0 \\ \text{and} \qquad (x_0 - x)\frac{\partial F}{\partial x} + (y_0 - y)\frac{\partial F}{\partial y} &= 0 \end{aligned}\right\} \tag{1}$$

and then substitute these coordinates into the equation of the tangent to the given curve.

b) To determine the *equations of the normals drawn to the curve* $F(x, y) = 0$ *from an arbitrary point* (x_0, y_0), we find those points at which the normals cut the given curve orthogonally (so-called *feet of the normals*), by solving the equations

$$F(x, y) = 0 \quad \text{and} \quad (x_0 - x)\frac{\partial F}{\partial y} - (y_0 - y)\frac{\partial F}{\partial x} = 0\,. \tag{2}$$

If a given algebraic curve is of degree n, then at most n^2 normals can be drawn from a given point to the curve.

Definition 1. The locus of the foot of the perpendicular from a given point (in the plane of the curve) on the tangent to a given curve is called the *pedal curve* of the given curve with respect to the given point (the *pole*).

c) We can find the *equation of the pedal curve* of the curve $F(x, y) = 0$ with respect to the pole (x_0, y_0) by elimination of the variables (x, y) from the equations

$$(X - x)\frac{\partial F}{\partial x} + (Y - y)\frac{\partial F}{\partial y} = 0\,, \quad (X - x_0)\frac{\partial F}{\partial y} - (Y - y_0)\frac{\partial F}{\partial x} = 0\,, \quad F(x, y) = 0\,.$$

The equation of the pedal curve will thus be expressed in the coordinates (X, Y).

d) *The angle between two curves* $y = f(x)$, $y = g(x)$ or $u(x, y) = 0$, $v(x, y) = 0$ at their common point is given by the equation

$$\tan \omega = \frac{\dfrac{dg}{dx} - \dfrac{df}{dx}}{1 + \dfrac{df}{dx}\dfrac{dg}{dx}}, \quad \text{or} \quad \tan \omega = \frac{\dfrac{\partial u}{\partial x}\dfrac{\partial v}{\partial y} - \dfrac{\partial u}{\partial y}\dfrac{\partial v}{\partial x}}{\dfrac{\partial u}{\partial x}\dfrac{\partial v}{\partial x} + \dfrac{\partial u}{\partial y}\dfrac{\partial v}{\partial y}},$$

respectively. (The derivatives are to be evaluated for the coordinates of the point of intersection of the curves.)

The curves $u(x, y) = 0$, $v(x, y) = 0$ *intersect at right angles* if

$$\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} = 0 .$$

e) The curves which intersect a one-parameter family $F(x, y, c) = 0$ of curves at right angles are called the *orthogonal trajectories* of this family and they form a one-parameter family of curves. Both families form an *orthogonal* net (provided exactly one curve of each family passes through each point). By eliminating the parameter c from the equations

$$\frac{\partial F}{\partial x} + \frac{\partial F}{\partial y}\frac{dy}{dx} = 0 , \quad F(x, y, c) = 0 ,$$

we obtain the differential equation $f(x, y, \dot{y}) = 0$ of the given family. By putting $-1/\dot{y}$ for $\dot{y}$ we obtain the differential equation

$$f\left(x, y, -\frac{1}{\dot{y}}\right) = 0$$

of the family of orthogonal trajectories (cf. § 17·6).

In polar coordinates, by eliminating the parameter c from the equations

$$\frac{\partial F}{\partial \varphi} + \frac{\partial F}{\partial \varrho}\frac{d\varrho}{d\varphi} = 0 , \quad F(\varrho, \varphi, c) = 0 ,$$

we obtain the differential equation $f(\varrho, \varphi, \varrho') = 0$ of the family of curves $F(\varrho, \varphi, c) = 0$. Writting $-\varrho^2/\varrho'$ for ϱ', we obtain the differential equation

$$f\left(\varrho, \varphi, -\frac{\varrho^2}{\varrho'}\right) = 0$$

of the orthogonal trajectories.

The curves which intersect a one-parameter family of curves at a constant angle $\omega \neq \frac{1}{2}\pi$ are called the *isogonal trajectories* of this family. If $f(x, y, \dot{y}) = 0$ is the

differential equation of the given family, then

$$f\left(x, y, \frac{\dot{y} - k}{1 + k\dot{y}}\right) = 0 \quad (k = \tan \omega)$$

is the differential equation of the isogonal trajectories. If, in polar coordinates, $f(\varrho, \varphi, \varrho') = 0$ is the differential equation of a family of curves, then

$$f\left(\varrho, \varphi, \frac{k\varrho^2 + \varrho\varrho'}{\varrho - k\varrho'}\right) = 0 \quad (k = \tan \omega)$$

is the differential equation of the isogonal trajectories.

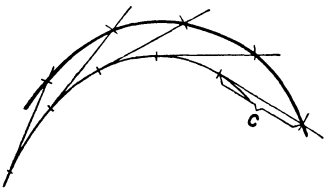


Fig. 9.21.

Definition 2. The locus of the end point of a segment of constant length c marked off on the tangent of a given curve, with the starting point at the point of contact, is called the *equitangential curve* of the given curve (Fig. 9.21).

f) We find the *equation of an equitangential curve* of the given curve $F(x, y) = 0$ by eliminating the variables (x, y) from the equations

$$F(x, y) = 0\,, \quad Y - y = \frac{\mathrm{d}y}{\mathrm{d}x}(X - x)\,, \quad (X - x)^2 + (Y - y)^2 = c^2\,.$$

The equation of the equitangential curve is then given in the coordinates (X, Y). The *equations of an equitangential curve* to the curve $x = \varphi(t)$, $y = \psi(t)$ are

$$X = \varphi + \frac{c\dot{\varphi}}{\sqrt{(\dot{\varphi}^2 + \dot{\psi}^2)}}\,, \quad Y = \psi + \frac{c\dot{\psi}}{\sqrt{(\dot{\varphi}^2 + \dot{\psi}^2)}}\,.$$

B. SURFACES

9.11. Definition and Equations of a Surface; Coordinates on a Surface

Definition 1. A *finite piecewise smooth surface*, defined *parametrically*, is a set of points (x, y, z) given by the equations

$$x = x(u, v)\,, \quad y = y(u, v)\,, \quad z = z(u, v)\,; \tag{1}$$

we suppose that the functions $x(u, v)$, $y(u, v)$, $z(u, v)$ are defined in a domain I which is a region of the type A (§ 14.1) containing, possibly, its boundary or a part of its boundary, and possess the following properties:

1. They are continuous and have piecewise continuous derivatives (Remark 12.1.8, p. 405) of the first order in I. (On the curves of discontinuity or at the boundary points, a derivative is taken to be the value of the corresponding continuous extension.)

2. The matrix

$$\mu = \begin{bmatrix} \frac{\partial x}{\partial u}, & \frac{\partial y}{\partial u}, & \frac{\partial z}{\partial u} \\ \frac{\partial x}{\partial v}, & \frac{\partial y}{\partial v}, & \frac{\partial z}{\partial v}, \end{bmatrix} \tag{2}$$

is everywhere – with the exception of at most a finite number of points – of rank $h = 2$ (i.e. at least one of its determinants of order two is non-vanishing).

REMARK 1. In Remark 9.2.1, we defined a closed curve and a smooth curve; in a somewhat similar fashion we can define a closed surface and a smooth surface. Briefly: a smooth surface has at every point a definite normal (§ 9.12) which varies continuously with the point on the surface. When a surface is considered in this chapter, a smooth surface or a piecewise smooth surface will be assumed. Furthermore, in the case of a surface, we often suppose (as the nature of the problem may require) that the functions $x(u, v)$, $y(u, v)$, $z(u, v)$ have derivatives of an order $r > 1$, without explicitly pointing out this condition. (The same remark applies to the surface (4).)

The arguments (u, v) in equations (1) are called the *parameters of the surface* and equations (1) are called the *parametric equations of the surface.* Every pair of numbers (u, v) from I is called a *point of the surface* (because to every pair of numbers (u, v) there corresponds a definite point (x, y, z) on the surface); such a point is said to be a *regular* (*general, ordinary*) *point of the surface* if, at this point, the functions (1) possess continuous partial derivatives and if the matrix (2) is of rank $h = 2$. Otherwise it is called a *singular point of the surface.* (However, see also Definition 9.12.2 and Remark 9.12.1.)

REMARK 2. If equations (1) are of the form

$$x = u\,, \quad y = v\,, \quad z = z(u, v)\,,$$

the equation

$$z = z(x, y) \quad \text{or} \quad z = f(x, y) \tag{3}$$

is called the *explicit equation of the surface.*

REMARK 3. A surface may be defined *implicitly* as a set of points satisfying an equation

$$F(x, y, z) = 0\,. \tag{4}$$

When we speak, in the following text, of a *surface given implicitly*, then we shall suppose that the function $F(x, y, z)$ is continuous and has continuous or piecewise continuous partial derivatives of the first order in the domain considered and that, with the exception of a finite number of points on the surface, at least one of them

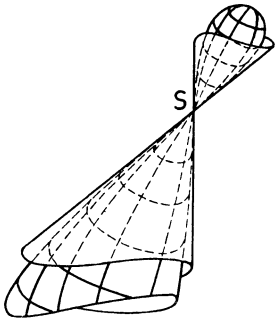


Fig. 9.22.

is non-zero. (In the case of piecewise continuous derivatives, the derivatives are to be understood in the sense of a continuous extension – cf. Definition 1.) If the relations

$$\frac{\partial F}{\partial x} = 0\,, \quad \frac{\partial F}{\partial y} = 0\,, \quad \frac{\partial F}{\partial z} = 0 \tag{5}$$

hold at a point (x, y, z) on the surface (4), then such a point is said to be a *singular point on the surface*. A point on the surface (4) at which at least one of the partial derivatives (5) is non-zero is called a *regular* (*general, ordinary*) *point of the surface* (4). (Cf., however, Remark 9.12.1 which may be applied in a similar form to equation (4).)

REMARK 4. A singular point on the surface (4) at which at least one partial derivative of the second order of the function F is non-zero is called a *conical point of the surface* (the point S in Fig. 9.22).

We shall suppose in the following text that the surface does not intersect itself at its regular points.

If we denote by $\boldsymbol{r}$ the radius vector of a point on the surface, the coordinates of which are, in the parametric representation (1), $x(u, v)$, $y(u, v)$, $z(u, v)$, we can use, instead of equations (1), the single symbolic equation

$$\boldsymbol{r} = \boldsymbol{r}(u, v) = \boldsymbol{i}x(u, v) + \boldsymbol{j}y(u, v) + \boldsymbol{k}z(u, v) \tag{6}$$

for the equation of the surface (the so-called *vector equation*).

Definition 2. The set of points (u_0, v) from I (where u_0 is fixed) for which the derivatives

$$\frac{\partial x(u_0, v)}{\partial v}, \quad \frac{\partial y(u_0, v)}{\partial v}, \quad \frac{\partial z(u_0, v)}{\partial v}$$

do not vanish simultaneously, is called the *parametric u-curve* (briefly *u-curve*) on the surface (1). The *parametric v-curve* (briefly the *v-curve*) is defined in a similar way.

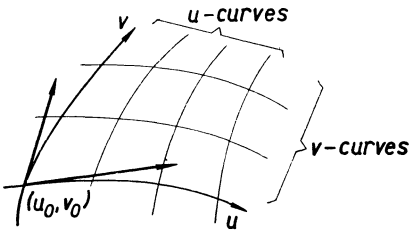


Fig. 9.23.

REMARK 5. If one of the parameters in equations (1) is kept constant while the other varies continuously, we obtain two one-parameter families of curves, one being formed by the u-curves and the other by the v-curves. These curves are called the *coordinate curves* (Fig. 9.23). Their equations on the surface are:

$$u = u_0 = \text{const.}, \quad v = t, \tag{7a}$$

(the u-curve; only the parameter v varies along the curve);

$$u = t, \quad v = v_0 = \text{const.}, \tag{7b}$$

(the v-curve; only the parameter u varies along the curve).

The parameters u and v constitute the so-called *curvilinear coordinates* (u, v) of a point on the surface (1) in such a domain where

1. no two curves from the same family intersect,

2. one and only one curve from each family passes through every point of the surface.

The parametric equations of a u-curve (with variable parameter v) or the parametric equations of a v-curve (with variable parameter u) — in space — are

$$x = x(u_0, v), \quad y = y(u_0, v), \quad z = z(u_0, v), \tag{8a}$$

or

$$x = x(u, v_0), \quad y = y(u, v_0), \quad z = z(u, v_0), \tag{8b}$$

respectively.

The *direction cosines of the tangent* of these coordinate curves are proportional to the numbers

$$\frac{\partial x(u_0, v)}{\partial v}, \quad \frac{\partial y(u_0, v)}{\partial v}, \quad \frac{\partial z(u_0, v)}{\partial v} \tag{9a}$$

or

$$\frac{\partial x(u, v_0)}{\partial u}, \quad \frac{\partial y(u, v_0)}{\partial u}, \quad \frac{\partial z(u, v_0)}{\partial u}, \tag{9b}$$

respectively.

Theorem 1. *The tangents of a u-curve and a v-curve constructed at their common regular point are different.* (See Fig. 9.23, the point (u_0, v_0).)

9.12. Curves on Surfaces, Tangent Planes and Normal Lines

Definition 1. The equations

$$u^* = u^*(u, v), \quad v^* = v^*(u, v) \tag{1}$$

express a (*regular*) *transformation of parameters* in a two-dimensional domain I^* from I, if the functions (1), defined in I^*, possess the following properties:

1. They are continuous and have continuous derivatives of order at least $m \geqq r$ (cf. Remark 9.11.1);

2. the determinant

$$\Delta = \begin{vmatrix} \dfrac{\partial u^*}{\partial u}, & \dfrac{\partial u^*}{\partial v} \\ \dfrac{\partial v^*}{\partial u}, & \dfrac{\partial v^*}{\partial v} \end{vmatrix} \tag{2}$$

is non-vanishing in I^*.

Definition 2. Let (u_0, v_0) be a singular point on a surface. Suppose there exist two functions $u^* = u^*(u, v)$, $v^* = v^*(u, v)$, continuous in a neighbourhood of the point (x_0, y_0) and having there continuous partial derivatives of the order $m \geqq r$ (see Definition 1) with $\Delta \neq 0$. Moreover, let the point (u_0, v_0) be not a singular point with respect to the new parameters. Then the point (u_0, v_0) is called an *unsubstantially singular point* (a *pole*) with respect to the parameters (u, v). Every singular point which is not a pole is called *substantially singular.*

REMARK 1. Thus, an unsubstantially singular point is a singular point of a surface only with respect to a certain system of coordinates on the surface. For example on

a sphere, if we choose the parallels of latitude for the u-curves, the meridians (parallels of longitude) for the v-curves, then, according to Remark 9.11.1, p. 344, the "north" and "south" poles are singular points of the sphere. But according to Definition 2 these points are unsubstantially singular (because they may be regular in some other system of coordinates.)

The notion of a general curve on a surface is introduced by the following definition:

Definition 3. The equations

$$u = u(t), \quad v = v(t), \quad t \in i, \tag{3}$$

express parametrically a *curve on a surface* (9.11.1) provided that functions (3), defined in an interval i, have the following properties in the interval i:

1. They are continuous and possess continuous first derivatives which do not vanish simultaneously;
2. the points (3) lie for all $t \in i$ in the domain I (see Definition 9.11.1);
3. the elements of the matrix (9.11.2) vanish simultaneously at a finite number of points at most.

Equations (3) are called the *parametric equations of a curve on the surface* (9.11.1).

REMARK 2. If equations (3) are of the form

$$u = t, \quad v = v(t),$$

then the equation

$$v = v(u) \tag{4}$$

is called the *explicit equation of a curve on the surface* (9.11.1). Equations (3) and (4) are the equations of a curve expressed in the coordinates on the surface. The equations

$$x = x(u(t), v(t)), \quad y = y(u(t), v(t)), \quad z = z(u(t), v(t)) \tag{5a}$$

or briefly

$$\mathbf{r} = \mathbf{r}(u(t), v(t)) \quad (\mathbf{r} \text{ being the radius vector of a point on the curve}) \tag{5b}$$

give the parametric representation of the same curve, which lies on the surface considered, in orthogonal cartesian coordinates (i.e. parametric representation of a space curve).

Definition 4. We also say that the equation

$$f(u, v) = 0 \tag{6}$$

expresses, in a definite neighbourhood of a certain (regular) point of a surface, a *curve on the surface defined implicitly* in I, if $f(u, v)$ is defined in I and has con-

tinuous first partial derivatives in I such that at least one of them is non-zero at every point. The equation (6) is called the *implicit equation of a curve on the surface.*

REMARK 3. The functional relation between the curvilinear coordinates defines (under the above-mentioned conditions) a curve on the surface (and conversely).

Theorem 1. *Let us consider all the curves that lie on a surface* $\mathbf{r} = \mathbf{r}(u, v)$ *and pass through a regular point* (u_0, v_0) *of this surface, and themselves have a regular point at this point. Then the tangent lines at* (u_0, v_0) *of all these curves lie in a plane* (Fig. 9.24).

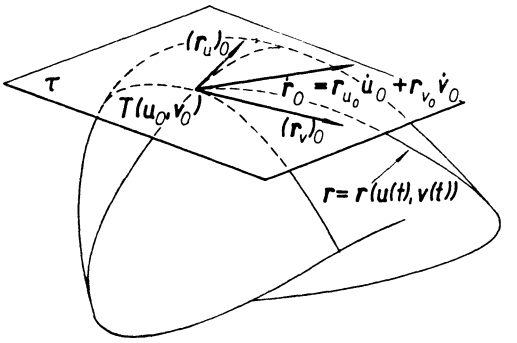


Fig. 9.24.

Definition 5. The plane mentioned in Theorem 1 is called the *tangent plane of the surface* $\mathbf{r} = \mathbf{r}(u, v)$, at its regular point (u_0, v_0). The point (u_0, v_0) is called the *point of contact* of this plane (the plane τ in Fig. 9.24).

REMARK 4. The tangential vector of a curve on a surface is the so-called *tangential vector of the surface* (shortly the *vector of the surface*) at the corresponding point of contact.

Theorem 2. *At a regular point* (u_0, v_0) *of a surface* $\mathbf{r} = \mathbf{r}(u, v)$ *there exists just one tangent plane and its equation in orthogonal cartesian coordinates* (X, Y, Z) *is*

$$[\mathbf{R} - \mathbf{r}_0, (\mathbf{r}_u)_0, (\mathbf{r}_v)_0] \equiv$$

$$\equiv \begin{vmatrix} X - x_0, & Y - y_0, & Z - z_0 \\ \left(\frac{\partial x}{\partial u}\right)_0, & \left(\frac{\partial y}{\partial u}\right)_0, & \left(\frac{\partial z}{\partial u}\right)_0 \\ \left(\frac{\partial x}{\partial v}\right)_0, & \left(\frac{\partial y}{\partial v}\right)_0, & \left(\frac{\partial z}{\partial v}\right)_0 \end{vmatrix} \equiv \begin{vmatrix} X - x_0, & Y - y_0, & Z - z_0 \\ (x_u)_0, & (y_u)_0, & (z_u)_0 \\ (x_v)_0, & (y_v)_0, & (z_v)_0 \end{vmatrix} = 0 \qquad (7)$$

(*where* $\mathbf{R}$ *is the vector of the variable point on the plane,* $(\mathbf{r}_u)_0 = (\partial \mathbf{r}/\partial u)_0$, $(\mathbf{r}_v)_0 = (\partial \mathbf{r}/\partial v)_0$).

REMARK 5. The linearly independent (i.e. non-collinear) vectors $\mathbf{r}_u = \partial\mathbf{r}/\partial u$ and $\mathbf{r}_v = \partial\mathbf{r}/\partial v$, with the starting point at a regular point (u, v) of a surface, form the tangential vectors of the corresponding v-curve and u-curve at this point, respectively. The tangential vector $\dot{\mathbf{r}}$ of a curve $\mathbf{r} = \mathbf{r}(u(t), v(t))$ on the surface passing through its regular point (u, v) is given by the linear combination of the vectors $\mathbf{r}_u$ and $\mathbf{r}_v$,

$$\dot{\mathbf{r}} = \mathbf{r}_u\dot{u} + \mathbf{r}_v\dot{v}\,, \quad \dot{u} = \frac{\mathrm{d}u}{\mathrm{d}t}\,, \quad \dot{v} = \frac{\mathrm{d}v}{\mathrm{d}t}\,.$$

The ratio $\dot{u}/\dot{v}$ defines, at the point (u, v) of the surface, the direction on the surface in which the curve $u = u(t)$, $v = v(t)$ on the surface passes through the point considered (Fig. 9.24). The tangent plane τ of the surface at its regular point (u, v) is determined by the vectors $\mathbf{r}_u$ and $\mathbf{r}_v$, with the starting point of both vectors at (u, v). A regular (or removably singular, p. 309) point of a surface is a point at which there exists just one tangent plane.

Theorem 3. *The tangent plane at a regular point (x, y, z) of a surface $z = z(x, y)$ has the equation*

$$(X - x)\,p + (Y - y)\,q - (Z - z) = 0 \tag{8}$$

(where $p = \partial z/\partial x$, $q = \partial z/\partial y$, and (X, Y, Z) are the coordinates of the variable point on the plane).

Theorem 4. *The tangent plane at a regular point (x, y, z) of a surface $F(x, y, z) = 0$ has the equation*

$$\frac{\partial F}{\partial x}(X - x) + \frac{\partial F}{\partial y}(Y - y) + \frac{\partial F}{\partial z}(Z - z) = 0 \tag{9}$$

((X, Y, Z) being the coordinates of the variable point on the plane).

Definition 6. The *normal* $\mathbf{n} = \mathbf{n}(u, v)$ *to a surface* $\mathbf{r} = \mathbf{r}(u, v)$, at a regular point (u, v), is the unit vector with its starting point at (u, v), perpendicular to the tangent plane at (u, v) and oriented so that

$$[\mathbf{n}\mathbf{r}_u\mathbf{r}_v] > 0\,,$$

where $[\mathbf{n}\mathbf{r}_u\mathbf{r}_v]$ is the mixed product (scalar triple product) of the vectors $\mathbf{n}$, $\mathbf{r}_u$, $\mathbf{r}_v$ (Fig. 9.25).

Theorem 5. *The unit vector* $\mathbf{n}$ *of the normal to the surface $x = x(u, v)$, $y = y(u, v)$, $z = z(u, v)$ is*

$$\mathbf{n} = \frac{\mathbf{r}_u \times \mathbf{r}_v}{\sqrt{(EG - F^2)}}\,, \tag{10}$$

where

$$E = \mathbf{r}_u \cdot \mathbf{r}_u = \left(\frac{\partial x}{\partial u}\right)^2 + \left(\frac{\partial y}{\partial u}\right)^2 + \left(\frac{\partial z}{\partial u}\right)^2,$$

$$F = \mathbf{r}_u \cdot \mathbf{r}_v = \frac{\partial x}{\partial u}\frac{\partial x}{\partial v} + \frac{\partial y}{\partial u}\frac{\partial y}{\partial v} + \frac{\partial z}{\partial u}\frac{\partial z}{\partial v}, \tag{11}$$

$$G = \mathbf{r}_v \cdot \mathbf{r}_v = \left(\frac{\partial x}{\partial v}\right)^2 + \left(\frac{\partial y}{\partial v}\right)^2 + \left(\frac{\partial z}{\partial v}\right)^2.$$

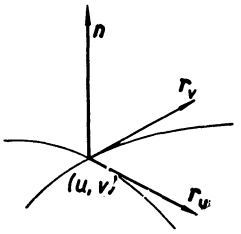


Fig. 9.25.

REMARK 6. At a regular point of a surface the relation $D^2 = EG - F^2 > 0$ always holds. The expression

$$D = \sqrt{(EG - F^2)} = |\mathbf{r}_u \times \mathbf{r}_v| = [\mathbf{n}\mathbf{r}_u\mathbf{r}_v] > 0 \tag{12}$$

is called the *discriminant of the surface.* We have

$$D^2 = \begin{vmatrix} E, & F \\ F, & G \end{vmatrix}. \tag{13}$$

The lengths of the tangential vectors $\mathbf{r}_u$ and $\mathbf{r}_v$ to a surface are

$$|\mathbf{r}_u| = \sqrt{E}\,, \quad |\mathbf{r}_v| = \sqrt{G}\,.$$

Theorem 6. *The direction cosines* (n_x, n_y, n_z) *of the normal* $\mathbf{n}$ *to a surface* $\mathbf{r} = \mathbf{r}(u, v)$, *are given by the expressions*

$$n_x = \frac{\begin{vmatrix} \frac{\partial y}{\partial u}, & \frac{\partial z}{\partial u} \\ \frac{\partial y}{\partial v}, & \frac{\partial z}{\partial v} \end{vmatrix}}{D}, \quad n_y = \frac{\begin{vmatrix} \frac{\partial z}{\partial u}, & \frac{\partial x}{\partial u} \\ \frac{\partial z}{\partial v}, & \frac{\partial x}{\partial v} \end{vmatrix}}{D}, \quad n_z = \frac{\begin{vmatrix} \frac{\partial x}{\partial u}, & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v}, & \frac{\partial y}{\partial v} \end{vmatrix}}{D}. \tag{14}$$

Theorem 7. *The direction cosines of the normal to a surface* $z = f(x, y)$ *are given by the expressions*

$$n_x = \frac{-p}{\sqrt{(p^2 + q^2 + 1)}}, \quad n_y = \frac{-q}{\sqrt{(p^2 + q^2 + 1)}}, \quad n_z = \frac{1}{\sqrt{(p^2 + q^2 + 1)}} \tag{15}$$

(*where* $p = \partial z/\partial x$, $q = \partial z/\partial y$).

Theorem 8. *The direction cosines of the normal to a surface* $F(x, y, z) = 0$ *are given by the expressions*

$$n_x = \frac{F_x}{J}, \quad n_y = \frac{F_y}{J}, \quad n_z = \frac{F_z}{J} \tag{16}$$

(*where* $F_x = \partial F/\partial x$, $F_y = \partial F/\partial y$, $F_z = \partial F/\partial z$ *and* $J = \sqrt{(F_x^2 + F_y^2 + F_z^2)}$).

The equations of a normal to the surface $F(x, y, z) = 0$ *at its regular point* (x, y, z) *are*

$$\frac{X - x}{F_x} = \frac{Y - y}{F_y} = \frac{Z - z}{F_z} \tag{17}$$

((X, Y, Z) *being the coordinates of the variable point on the normal*).

REMARK 7. The direction ratios of the normal to a surface $F(x, y, z) = 0$, or $z = f(x, y)$ are (F_x, F_y, F_z), or $(\partial f/\partial x, \partial f/\partial y, -1)$, respectively.

Example 1. If we choose the spherical coordinates u and v (polar coordinates in space) (Fig. 9.26) for the representation of a spherical surface of radius r with its centre at the origin of the cartesian coordinate system, then the equations

$$x = r \sin u \cos v, \quad y = r \sin u \sin v, \quad z = r \cos u \tag{18}$$
$$(0 \leqq u \leqq \pi, \quad 0 \leqq v < 2\pi)$$

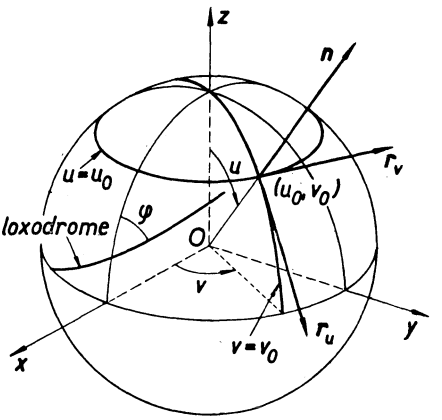


Fig. 9.26.

are the parametric equations of this spherical surface. The geometrical interpretation of the parameters u and v may be seen in Fig. 9.26. The net of the coordinate curves is formed by the parallel circles (u-curves) and the half-meridians (v-curves) of the spherical surface. For the vector $\mathbf{r}_u$ (the direction of the tangent to a meridian) and the vector $\mathbf{r}_v$ (the direction of the tangent to a parallel circle), at the point (u, v), we

obtain from (18) (for the sake of clarity we use here the notation $\boldsymbol{a} = \begin{cases} a_1 \\ a_2 \\ a_3 \end{cases}$ instead of $\boldsymbol{a} = (a_1, a_2, a_3)$)

$$\boldsymbol{r}_u(u, v) = \begin{cases} r \cos u \cos v , \\ r \cos u \sin v , \\ -r \sin u , \end{cases} \quad \boldsymbol{r}_v(u, v) = \begin{cases} -r \sin u \sin v , \\ r \sin u \cos v , \\ 0 . \end{cases}$$

For the direction ratios of the vector of the normal we obtain

$$\boldsymbol{r}_u \times \boldsymbol{r}_v = \begin{cases} r^2 \sin^2 u \cos v , \\ r^2 \sin^2 u \sin v , \\ r^2 \sin u \cos u , \end{cases} \quad \text{i.e.} \quad \boldsymbol{n} = \begin{cases} \sin u \cos v , \\ \sin u \sin v , \\ \cos u \end{cases}$$

and, using (18), we obtain the equation of the tangent plane at the point (u, v) in the form

$$X \sin u \cos v + Y \sin u \sin v + Z \cos u - r = 0 .$$

Example 2. Let the equations

$$x = u \cos v , \quad y = u \sin v , \quad z = f(v) \quad (u \in (-\infty, +\infty), \quad v \in (-\tfrac{1}{2}\pi, \tfrac{1}{2}\pi)) \tag{19}$$

represent a surface. By the elimination of u and v from (19) we obtain the explicit equation of the surface in the form

$$z = f\left(\tan^{-1} \frac{y}{x}\right) . \tag{20}$$

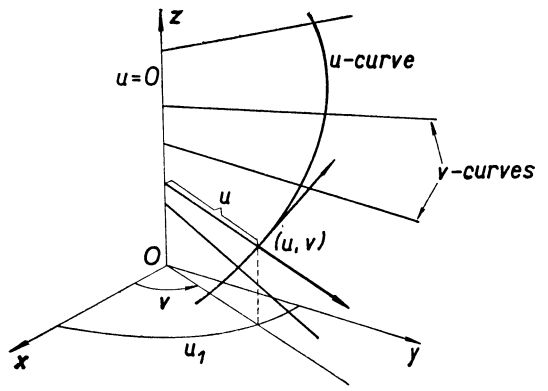


Fig. 9.27.

For $u = u_0 \neq 0$ the parametric u-curve is the intersection of the surface (20) and the circular cylinder $x^2 + y^2 = u_0^2$. The equation $u = 0$ denotes the z-axis. The v-curves are the sections of the surface (20) cut by the planes $y = x \tan v_0$ and therefore constitute the straight lines which intersect the z-axis at right angles (Fig. 9.27). This

surface (which contains a one-parameter family of straight lines) is a *ruled surface* (in this case a *skew surface* or a *scroll*, i.e. a non-developable surface), in fact a *right conoid*. The equation of the tangent plane at the point (u, v) is, by (7),

$$\begin{vmatrix} X - u\cos v, & Y - u\sin v, & Z - f(v) \\ \cos v, & \sin v, & 0 \\ -u\sin v, & u\cos v, & \mathrm{d}f/\mathrm{d}v \end{vmatrix} =$$

$$= (X\sin v - Y\cos v)\frac{\mathrm{d}f}{\mathrm{d}v} + [Z - f(v)]\,u = 0\,.$$

In particular, if we choose for the function $z = f(v)$, the function $z = cv$ ($c \neq 0$ being a real constant), we obtain a right conoid and screw surface, known as a *helicoid*; its equation is

$$z = c\tan^{-1}\frac{y}{x}$$

(the u-curves are formed by coaxial helices). At the point (x, y, z) we obtain, from (8), the equation of the tangent plane of this screw surface

$$cyX - cxY + (x^2 + y^2)\,Z = (x^2 + y^2)\,z$$

and, from (15), the equations of the normal line

$$\frac{X - x}{cy} = \frac{Y - y}{-cx} = \frac{Z - z}{x^2 + y^2}\,.$$

Example 3. Let

$$\bar{x} = \bar{x}(u)\,,\quad \bar{y} = \bar{y}(u)\,,\quad \bar{z} = \bar{z}(u)\,,\quad u \in I\,, \tag{21}$$

be the parametric equations of a space curve, all points of which are regular for $u \in I$. The equations

$$x = \bar{x}(u) + v\frac{\mathrm{d}\bar{x}}{\mathrm{d}u}\,,\quad y = \bar{y}(u) + v\frac{\mathrm{d}\bar{y}}{\mathrm{d}u}\,,\quad z = \bar{z}(u) + v\frac{\mathrm{d}\bar{z}}{\mathrm{d}u} \tag{22}$$

are the parametric equations of the *tangent surface* (or the tangent developable) of the space curve (21). The given curve is called the *cuspidal edge*, or *edge of regression* of this surface. The matrix μ (see (9.11.2)) is of rank $h < 2$ for $v = 0$. The parametric curves are on the one hand the curves $u = \text{const.}$, i.e. the tangents to the curve (21), and on the other $v = \text{const.}$, i.e. the curves which have the same distance v, measured along the tangent to the given curve, from the given curve (i.e. from the curve $v = 0$). Every point of the given curve is a singular point of the surface (22) (see Fig. 9.28). The surface (22) which contains a one-parameter family of straight lines (the u-curves)

is a ruled surface, but is *developable*. The equation of the tangent plane at a regular point (u, v) (for which $v \neq 0$) is

$$\left[\boldsymbol{R} - \bar{\boldsymbol{r}}, \frac{\mathrm{d}^2\bar{\boldsymbol{r}}}{\mathrm{d}u^2}, \frac{\mathrm{d}\bar{\boldsymbol{r}}}{\mathrm{d}u}\right] = 0 .$$

This equation (which is independent of v) is, however, an equation of the osculating plane of the curve $\bar{\boldsymbol{r}} = \bar{\boldsymbol{r}}(u)$ at its point (u). Any tangent plane of the given surface is a tangent plane at every point $v \neq 0$ to a fixed u-curve, i.e. to a definite straight

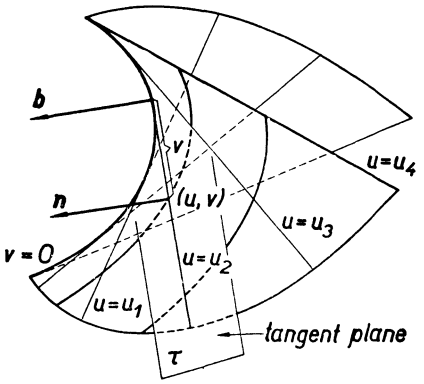


Fig. 9.28.

line on the surface (the so-called *generator*, or *ruling of the surface*). The normal $\boldsymbol{n}$ to the surface does not change along this straight line and is parallel to the binormal $\boldsymbol{b}$ to the given curve at the corresponding point of contact $(u, 0)$.

Example 4. According to (9), the surface

$$xyz = a^3 \quad (a \neq 0 \text{ being a real constant})$$

has at the point (x, y, z) the tangent plane with the equation

$$\frac{X}{x} + \frac{Y}{y} + \frac{Z}{z} = 3 .$$

9.13. Envelope of a One-parameter Family of Surfaces, Ruled Surfaces (Torses and Scrolls)

Definition 1. We say that an equation

$$F(x, y, z, c) = 0 \tag{1}$$

defines a *one-parameter family of surfaces* in a domain O of the three-dimensional space xyz if

1. the function $F(x, y, z, c)$ is a continuous function of the variables (x, y, z, c) for $(x, y, z) \in O$ and for c from an interval I,

2. for every $c \in I$, equation (1) defines in the domain O a certain surface (cf. § 9.11) and to every two different values of $c \in I$ there correspond two different surfaces in the domain O.

Definition 2. A surface is said to be the *envelope* of the family (1) (provided such a surface exists) if it touches every surface of the given family, and conversely, if its every point (x, y, z) is a point of contact with a surface of the family (1).

Theorem 1. *Let a point* $(x_0, y_0, z_0) \in O$ *and a number* $c_0 \in I$ *exist such that the equations*

$$F(x_0, y_0, z_0, c_0) = 0\,, \quad \frac{\partial F}{\partial c}(x_0, y_0, z_0, c_0) = 0 \tag{2}$$

hold.

Let the function $F(x, y, z, c)$ *have continuous partial derivatives*

$$\frac{\partial F}{\partial x}\,, \quad \frac{\partial F}{\partial y}\,, \quad \frac{\partial F}{\partial z}\,, \quad \frac{\partial F}{\partial c}\,, \quad \frac{\partial^2 F}{\partial c\,\partial x}\,, \quad \frac{\partial^2 F}{\partial c\,\partial y}\,, \quad \frac{\partial^2 F}{\partial c\,\partial z}\,, \quad \frac{\partial^2 F}{\partial c^2} \tag{3}$$

in a certain neighbourhood of the point (x_0, y_0, z_0, c_0). *Further at the point* (x_0, y_0, z_0, c_0), *let*

$$\frac{\partial^2 F}{\partial c^2} \neq 0 \tag{4}$$

and let the matrix

$$\begin{bmatrix} \dfrac{\partial F}{\partial x}\,, & \dfrac{\partial F}{\partial y}\,, & \dfrac{\partial F}{\partial z} \\[2ex] \dfrac{\partial^2 F}{\partial c\,\partial x}\,, & \dfrac{\partial^2 F}{\partial c\,\partial y}\,, & \dfrac{\partial^2 F}{\partial c\,\partial z} \end{bmatrix} \tag{5}$$

be of rank

$$h = 2\,. \tag{6}$$

Then, in a certain neighbourhood U *of the point* (x_0, y_0, z_0) *and for* c *from a certain neighbourhood* V *of the point* c_0, *there exists an envelope of the family* (1).

REMARK 1. The equation of the envelope may be obtained, for example, from the equations

$$F(x, y, z, c) = 0\,, \quad \frac{\partial F}{\partial c}(x, y, z, c) = 0 \tag{7}$$

by expressing c from the second equation as a function of (x, y, z) and substituting into the first equation; by eliminating c we obtain

$$F(x, y, z, c(x, y, z)) = 0\,. \tag{8}$$

For a fixed c, equations (7) give the so-called *characteristic curve*, or simply the *characteristic*, of the family (1), along which the envelope touches the corresponding surfaces of the given family.

REMARK 2. By considering (4) and (6), we can make a remark, similar to Remark 9.7.1, on the envelope of a given family of curves. In the next example, the reader may verify the fulfilment of the above-mentioned assumptions:

Example 1. The equation

$$x^2 + y^2 + (z - c)^2 = r^2(c) \tag{9}$$

is the equation of a one-parameter family of spheres, whose centres are on the z-axis and whose radii depend on the distance c between the centre of the sphere and the origin. Equation (9), together with the second equation (7),

$$-(z - c) = r(c)\frac{\mathrm{d}r(c)}{\mathrm{d}c} \tag{10}$$

express (c being fixed) the characteristic of a sphere from the family investigated. By using (9) and (10), the equation of this characteristic may be expressed in the form

$$x^2 + y^2 = r^2(c)\left[1 - \left(\frac{\mathrm{d}r(c)}{\mathrm{d}c}\right)^2\right]; \quad z = c - r(c)\frac{\mathrm{d}r(c)}{\mathrm{d}c}\,. \tag{11}$$

The first of equations (11) represents a circular cylinder with its axis along the z-axis; the second of equations (11) represents a plane which is perpendicular to the z-axis, so that the characteristic is a circle lying in this plane and having its centre on the z-axis. We obtain the equation of the envelope of the family (9) by eliminating the parameter c from equations (11) in the form

$$x^2 + y^2 = f(z)\,, \tag{12}$$

which is the equation of a surface of revolution with its axis of revolution in the z-axis. If, for example, $r^2(c) = 2c - 1$, equations (11) reduce, after rearrangement, to

$$x^2 + y^2 = 2(c - 1)\,, \quad z = c - 1\,.$$

The equation of the envelope of the family $x^2 + y^2 + (z - c)^2 = 2c - 1$ will be obtained by the elimination of the parameter c from the two preceding equations, in the form

$$x^2 + y^2 = 2z\,;$$

this is the equation of a paraboloid of revolution generated by the revolution of the parabola $2z = x^2$ about the z-axis.

REMARK 3. Surfaces of revolution are examples of envelopes of a one-parameter family of spheres. The characteristics are circles.

Definition 3. If the equations

$$F(x, y, z, c) = 0\,, \quad \frac{\partial F}{\partial c} = 0\,, \quad \frac{\partial^2 F}{\partial c^2} = 0 \quad \left(\frac{\partial^3 F}{\partial c^3} \neq 0 \text{ for all } c \text{ from } I\right) \tag{13}$$

determine a curve, then this curve is called the *edge of regression* of the given one-parameter family of surfaces.

REMARK 4. We also speak of an edge of regression in the case where equations (13) define a finite number of points.

Theorem 2. *If the edge of regression is a curve, then at each of its points* (x_0, y_0, z_0) *the characteristic corresponding to a parameter* c_0 *and the edge of regression are tangential to each other,* i.e. *every characteristic of the surface of the family is a tangent to the edge of regression* (*at the so-called focal point*).

Theorem 3. *If the edge of regression of a one-parameter family of planes is a space curve, then this family of planes is formed by the osculating planes of the edge of regression and the characteristics are the tangents to the edge of regression.*

Theorem 4. *The envelope of a one-parameter family of planes, whose edge of regression is a space curve, is the tangent surface of this space curve.*

REMARK 5. The envelopes of one-parameter families of planes are developable ruled surfaces. The straight lines of the surface (so-called *rulings* or *generators*) are the characteristics of the family (of planes). The tangent surfaces of space curves, all cones and cylinders belong to this group. The tangent surfaces of gradient curves (Definition 9.8.2) are called the *gradient surfaces.* All tangent planes of such a surface make a constant angle with a fixed direction.

REMARK 6. In Examples 9.12.2 and 9.12.3 two substantially different cases of ruled surfaces were shown. The ruled surfaces are of great importance in technical applications.

Definition 4. A *ruled surface* is a surface such that, through every point of it, there passes at least one straight line lying entirely on it.

REMARK 7. We may also represent a ruled surface as a locus of a straight line moving continuously in space, i.e. a ruled surface is a one-parameter family of straight lines where the corresponding parameter changes continuously through an interval.

Theorem 5. *The parametric equations of a ruled surface are*

$$\begin{aligned} x(u, v) &= \bar{x}(u) + v \,.\, a(u), \\ y(u, v) &= \bar{y}(u) + v \,.\, b(u), \\ z(u, v) &= \bar{z}(u) + v \,.\, c(u), \end{aligned} \quad \text{i.e.} \quad \mathbf{r}(u, v) = \bar{\mathbf{r}}(u) + v \,.\, \mathbf{p}(u), \tag{14}$$

where $\bar{x} = \bar{x}(u)$, $\bar{y} = \bar{y}(u)$, $\bar{z} = \bar{z}(u)$ *are the parametric equations of the so-called director curve, and* $a(u)$, $b(u)$, $c(u)$ *are continuous functions of the argument* u *and determine the direction ratios of the straight lines (so-called rulings or generators) on the surface.*

Theorem 6. *If on a ruled surface* (14), *it is not possible to determine a value of* u *for which the ruling is such that the tangent planes at the points along that ruling form a pencil of planes, the ruled surface is a cylinder, cone, or tangent surface of a (space) curve. The characteristic feature of these surfaces is that every tangent plane touches the surface along an entire ruling.*

Definition 5. A ruled surface of the type mentioned in Theorem 6 is called a *developable surface* or a *torse.* A ruled surface which possesses a property such that the tangent planes at the points along its rulings (with the possible exception of a finite number of rulings called *torsal lines*) form a pencil of planes, is called an *undevelopable ruled surface,* a *skew surface* or a *scroll.* The ruling of an undevelopable ruled surface is called the *torsal line* if it has the same tangent plane at all its regular points.

REMARK 8. Every ruling on a cone, cylinder or tangent surface of a space curve (i.e. on a developable ruled surface) is a torsal line. The skew conicoids (the hyperbolic paraboloid and the hyperboloid of one sheet) and the helicoid from Example 9.12.2 serve as examples of skew surfaces that have no torsal lines.

Theorem 7 (*Chasles' Theorem*). *The tangent planes of a skew surface along its rulings form pencils of planes (their axes are the rulings). There is a projectivity between the pencil of tangent planes and the range of their points of contact on a ruling,* i.e. *the cross-ratio of four tangent planes to a skew surface at points of a ruling is equal to the cross-ratio of the points of contact.*

Theorem 8. *A necessary and sufficient condition that a ruled surface* (14) *be developable is*

$$[\dot{\mathbf{r}}, \mathbf{p}, \dot{\mathbf{p}}] = 0, \tag{15}$$

where $\dot{\mathbf{r}} = \dfrac{\mathrm{d}\bar{\mathbf{r}}}{\mathrm{d}u}$, $\dot{\mathbf{p}} = \dfrac{\mathrm{d}\mathbf{p}}{\mathrm{d}u}$.

Theorem 9. *If a ruled surface is given by the equations*

$$x = a(t)\, z + m(t), \quad y = b(t)\, z + n(t)$$

(a, b, m *and* n *being functions of the same parameter* t*), then a necessary and*

sufficient condition that this surface be developable is

$$\dot{a}\dot{n} = \dot{b}\dot{m} \quad \left(\dot{a} = \frac{da}{dt}, \quad \dot{n} = \frac{dn}{dt}, \text{ etc.}\right).$$

Theorem 10. *The equation*

$$f_{xx}f_{yy} - f_{xy}^2 = 0 \quad \left(f_{xx} = \frac{\partial^2 f}{\partial x^2}, \quad f_{yy} = \frac{\partial^2 f}{\partial y^2}, \quad f_{xy} = \frac{\partial^2 f}{\partial x\, \partial y}\right) \tag{16}$$

is the differential equation of developable surfaces (for surfaces expressed by the equation $z = f(x, y)$).

9.14. First Fundamental Form of the Surface

Theorem 1. *The square of the differential of the arc of a curve $u = u(t)$, $v = v(t)$ on a surface $\mathbf{r} = \mathbf{r}(u, v)$ is given by the formula*

$$ds^2 = E\,du^2 + 2F\,du\,dv + G\,dv^2 \equiv I\,. \tag{1}$$

Definition 1. The quadratic differential form (1) is called the *first fundamental* (or *metric*) *form of a surface*, ds is called the *linear element* (the *element of the arc*) on the surface and E, F, G are called the *first fundamental coefficients* (see Theorem 9.12.5).

Theorem 2. *If φ is the angle between the curves $u = u(t)$, $v = v(t)$ and $\bar{u} = \bar{u}(\bar{t})$, $\bar{v} = \bar{v}(\bar{t})$ which lie on a surface $\mathbf{r} = \mathbf{r}(u, v)$* and *pass through any of its regular points, then*

$$\cos\varphi = \frac{E\dot{u}\dot{\bar{u}} + F(\dot{u}\dot{\bar{v}} + \dot{\bar{u}}\dot{v}) + G\dot{v}\dot{\bar{v}}}{\sqrt{(E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2)}\,.\,\sqrt{(E\dot{\bar{u}}^2 + 2F\dot{\bar{u}}\dot{\bar{v}} + G\dot{\bar{v}}^2)}}\,. \tag{2}$$

REMARK 1. For the parametric curves of a surface we have that $ds^2 = G\,dv^2$ (for $u = \text{const.}$) and $ds^2 = E\,du^2$ (for $v = \text{const.}$).

If the second curve of Theorem 2 is a parametric curve, then

$$\cos\varphi = \frac{E\dot{u} + F\dot{v}}{\sqrt{(E)}\,\sqrt{(E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2)}} \quad \text{or} \quad \cos\varphi = \frac{F\dot{u} + G\dot{v}}{\sqrt{(G)}\,\sqrt{(E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2)}} \tag{3}$$

for $\bar{v}(t) = \bar{v}_0$ or $\bar{u}(t) = \bar{u}_0$, respectively (the curve is oriented in the sense of the increasing parameter). If the curves of Theorem 2 intersect at right angles, then

$$(E\,du + F\,dv)\,d\bar{u} + (F\,du + G\,dv)\,d\bar{v} = 0\,.$$

Theorem 3. *For the angle made by a parametric v-curve with a parametric u-curve (in this order) the following relations hold:*

$$\cos\varphi = \frac{F}{\sqrt{(EG)}}, \quad \sin\varphi = \frac{D}{\sqrt{(EG)}} \quad (D = \sqrt{(EG - F^2)}). \tag{4}$$

Theorem 4. *A necessary and sufficient condition that the parametric curves intersect at right angles at every point of a surface, i.e., that the parametric net on a surface be orthogonal, is*

$$F(u, v) \equiv 0 \quad (\textit{identically}). \tag{5}$$

Example 1. For a *sphere* (Example 9.12.1, p. 314)

$$x = r\sin u\cos v, \quad y = r\sin u\sin v, \quad z = r\cos u$$

we obtain (using (9.12.11))

$$E = r^2, \quad F = 0, \quad G = r^2\sin^2 u.$$

Hence the metric form is

$$\mathrm{d}s^2 = r^2(\mathrm{d}u^2 + \sin^2 u\,\mathrm{d}v^2),$$

and this relation holds for the element of arc of every curve on the sphere. A curve which makes a constant angle φ $(0 < \varphi < \frac{1}{2}\pi)$ with the meridians of a sphere is called a *loxodrome* on the sphere (Fig. 9.26). Its equation in the coordinates u and v is

$$v + \tan\varphi \ln\tan\frac{u}{2} = c \quad (c \text{ being a real constant}).$$

The length of its arc s is given by

$$s = r\int\sqrt{\left[1 + \sin^2 u\left(\frac{\mathrm{d}v}{\mathrm{d}u}\right)^2\right]}\,\mathrm{d}u.$$

By differentiation of the equation of the loxodrome we obtain its differential equation

$$\frac{\mathrm{d}v}{\mathrm{d}u} = \frac{-\tan\varphi}{\sin u}$$

and for the length of arc s, e.g. between $u = \frac{1}{4}\pi$ and $u = \frac{1}{2}\pi$

$$s = r\int_{\pi/4}^{\pi/2}\sqrt{(1 + \tan^2\varphi)}\,\mathrm{d}u = r\sqrt{(1 + \tan^2\varphi)}\,.\,[u]_{\pi/4}^{\pi/2} = \frac{\pi r}{4}\sqrt{(1 + \tan^2\varphi)}.$$

Theorem 5. *The area of the parallelogram, two adjacent sides of which are tangent vectors of the surface, is* (Fig. 9.29)

$$\mathrm{d}P = D \begin{vmatrix} \mathrm{d}u, & \mathrm{d}v \\ \mathrm{d}\bar{u}, & \mathrm{d}\bar{v} \end{vmatrix} . \tag{6a}$$

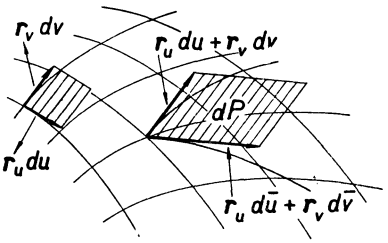


Fig. 9.29.

Definition 2. The expression (6a) for $\mathrm{d}P$ is called the *element of area* of the surface $\boldsymbol{r} = \boldsymbol{r}(u, v)$.

REMARK 2. If for the tangent vectors of Theorem 5 we choose the tangent vector of parametric curves (Fig. 9.29), then equation (6a) reduces to

$$\mathrm{d}P = D\,\mathrm{d}u\,\mathrm{d}v\,. \tag{6b}$$

The *area of a region of a surface* $x = x(u, v)$, $y = y(u, v)$, $z = z(u, v)$, over a domain O, may be found by evaluating the double integral

$$P = \iint_O \sqrt{(EG - F^2)}\,\mathrm{d}u\,\mathrm{d}v\,. \tag{7}$$

REMARK 3. For a surface $z = z(x, y)$ we obtain the first fundamental coefficients in the form

$$E = 1 + \left(\frac{\partial z}{\partial x}\right)^2, \quad F = \frac{\partial z}{\partial x}\,\frac{\partial z}{\partial y}, \quad G = 1 + \left(\frac{\partial z}{\partial y}\right)^2,$$

the *discriminant* $D = \sqrt{[1 + (\partial z/\partial x)^2 + (\partial z/\partial y)^2]}$ and the *element of area*

$$\mathrm{d}P = \sqrt{\left[1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2\right]}\,\mathrm{d}x\,\mathrm{d}y = \sqrt{(1 + p^2 + q^2)}\,\mathrm{d}x\,\mathrm{d}y\,.$$

For the *differential* $\mathrm{d}s$ *of the arc* of a curve on the surface $z = f(x, y)$, we obtain

$$\mathrm{d}s = \sqrt{[(1 + p^2)\,\mathrm{d}x^2 + 2pq\,\mathrm{d}x\,\mathrm{d}y + (1 + q^2)\,\mathrm{d}y^2]}\,.$$

9.15. Second Fundamental Form of the Surface, Shape of the Surface with Respect to its Tangent Plane

Theorem 1. *Along a curve* $u = u(s)$, $v = v(s)$ (s *being the arc of the curve*) *on the surface* $\mathbf{r} = \mathbf{r}(u, v)$ *the following equation holds*:

$$-\mathrm{d}\mathbf{r} \,.\, \mathrm{d}\mathbf{n} = L\,\mathrm{d}u^2 + 2M\,\mathrm{d}u\,\mathrm{d}v + N\,\mathrm{d}v^2 \equiv II\,, \tag{1}$$

where

$$L = -\mathbf{r}_u \,.\, \mathbf{n}_u\,, \quad 2M = -(\mathbf{r}_u \,.\, \mathbf{n}_v + \mathbf{r}_v \,.\, \mathbf{n}_u)\,, \quad N = -\mathbf{r}_v \,.\, \mathbf{n}_v \tag{2}$$

($\mathbf{n}_u = \partial\mathbf{n}/\partial u$, $\mathbf{n}_v = \partial\mathbf{n}/\partial v$, and $\mathbf{n}$ *is the unit normal vector of the surface*).

Definition 1. The form (1) is called the *second fundamental form of the surface*, while the coefficients L, M, N are called the *second fundamental coefficients of the surface.*

Theorem 2. *The following relations hold*:

$$L = \frac{[\mathbf{r}_{uu}\mathbf{r}_u\mathbf{r}_v]}{\sqrt{(EG - F^2)}}\,, \quad M = \frac{[\mathbf{r}_{uv}\mathbf{r}_u\mathbf{r}_v]}{\sqrt{(EG - F^2)}}\,, \quad N = \frac{[\mathbf{r}_{vv}\mathbf{r}_u\mathbf{r}_v]}{\sqrt{(EG - F^2)}}\,, \tag{3}$$

where

$$\mathbf{r}_{uu} = \frac{\partial^2\mathbf{r}}{\partial u^2}\,, \quad \mathbf{r}_{uv} = \frac{\partial^2\mathbf{r}}{\partial u\,\partial v}\,, \quad \mathbf{r}_{vv} = \frac{\partial^2\mathbf{r}}{\partial v^2}\,.$$

Theorem 3. *For a surface* $z = f(x, y)$ *we have*

$$L = \frac{r}{\sqrt{(1 + p^2 + q^2)}}\,, \quad M = \frac{s}{\sqrt{(1 + p^2 + q^2)}}\,, \quad N = \frac{t}{\sqrt{(1 + p^2 + q^2)}} \tag{4}$$

$$\left(r = \frac{\partial^2 z}{\partial x^2}\,, \quad s = \frac{\partial^2 z}{\partial x\,\partial y}\,, \quad t = \frac{\partial^2 z}{\partial y^2}\,, \quad p = \frac{\partial z}{\partial x}\,, \quad q = \frac{\partial z}{\partial y}\right).$$

Definition 2. A regular point of a surface at which

$$LN - M^2 > 0 \quad \text{or} \quad LN - M^2 = 0 \quad \text{or} \quad LN - M^2 < 0 \tag{5}$$

is called an *elliptic, parabolic* or *hyperbolic point of the surface*, respectively.

Theorem 4. *In a sufficiently small neighbourhood of a regular point P which is an elliptic, or hyperbolic point, the surface lies on one side, or on both sides of the tangent plane* τ *at P, respectively. The tangent plane* τ *at an elliptic, parabolic or hyperbolic point P of a surface cuts the surface in a curve for which the point P is a double point with imaginary conjugate, real coincident, or real distinct tangents, respectively* (Fig. 9.30).

REMARK 1. If a surface is given by the equation $z = f(x, y)$, then at an *elliptic point* $f_{xx}f_{yy} - f_{xy}^2 > 0$, at a *hyperbolic point* $f_{xx}f_{yy} - f_{xy}^2 < 0$, and at a *parabolic point* $f_{xx}f_{yy} - f_{xy}^2 = 0$. Developable surfaces have only parabolic points (with the exception of singular points).

Example 1. Every point of an elliptic paraboloid is elliptic; a hyperbolic paraboloid and a skew helicoid consist exclusively of hyperbolic points.

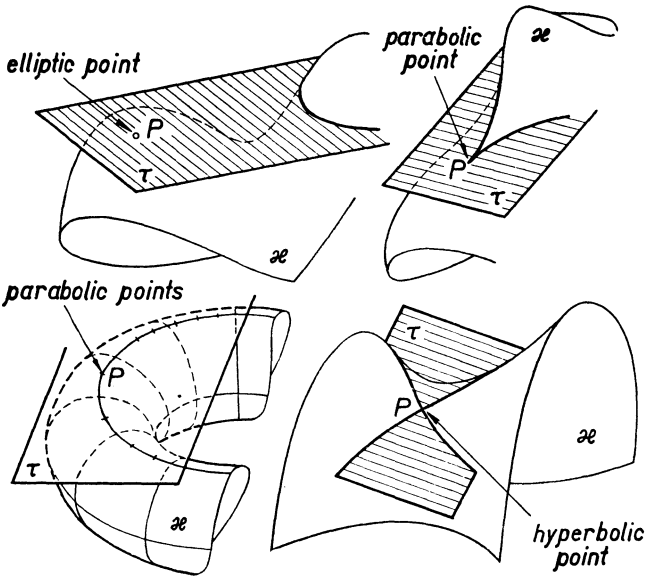


Fig. 9.30.

REMARK 2. The directions dv/du of the tangent vectors at a point of a surface which are tangent vectors of the section of the surface by the tangent plane at this point, are called the *asymptotic directions* and they satisfy the equation

$$L\,du^2 + 2M\,du\,dv + N\,dv^2 = 0\,. \tag{6}$$

If the equation of a surface is of the form $z = f(x, y)$, then equation (6) for the asymptotic directions reduces to

$$f_{xx} + 2f_{xy}k + f_{yy}k^2 = 0 \quad \left(k = \frac{dy}{dx}\right). \tag{7}$$

9.16. Curvature of a Surface

Theorem 1. *All curves on a surface which pass through a regular point P of the surface and have the same osculating plane at P have also the same curvature at P.*

REMARK 1. The radius of curvature of the curve of the section cut by a plane passing through a point P of the surface $\boldsymbol{r} = \boldsymbol{r}(u, v)$ is

$$r = \frac{E\,du^2 + 2F\,du\,dv + G\,dv^2}{|L\,du^2 + 2M\,du\,dv + N\,dv^2|}\cos\vartheta \quad (II \neq 0;\ \text{cf. Theorem 9.15.1}) \tag{1}$$

(ϑ being the angle between the plane of section and the normal to the surface at P; Fig. 9.31). Formula (1) also holds for space curves on the surface (in this case we

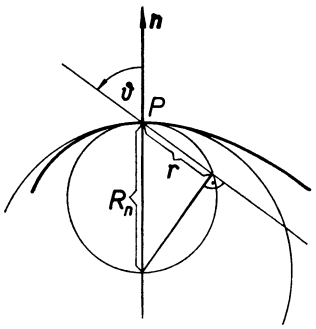


Fig. 9.31.

consider the corresponding osculating plane in place of the plane of section). A curve cut on a surface by a plane that contains a normal to the surface is called a *curve of normal section* (Fig. 9.31).

Theorem 2 (*Theorem of Meusnier*). *A curve of section passing through a regular point P on a surface, has at P a radius of curvature which is the orthogonal projection of the radius of curvature R_n of the curve of normal section* (*into the osculating plane of the first curve at P*), *while both curves of section have a common tangent line at P* (Fig. 9.31, schematic), i.e.

$$r = R_n \cos\vartheta\,. \tag{2}$$

REMARK 2. The circles of curvature of all the curves of section on a surface, through a regular point P and with the same tangent line t at P, lie on a sphere of radius R_n (the radius of curvature of the curve of normal section through the common tangent t) with its centre on the normal to the surface at P. Theorem 2 holds also for space curves on a surface.

REMARK 3. If, for example, we consider a regular point P of a surface of revolution (Fig. 9.32, schematic), then the centre of curvature S of the curve of normal section lying in the plane through the tangent line to the corresponding parallel circle at P is on the axis of the surface.

REMARK 4. The *normal curvature* $1/R$ in a given direction at a point P of a surface $\mathbf{r} = \mathbf{r}(u, v)$ is

$$\frac{1}{R} = \frac{L\,du^2 + 2M\,du\,dv + N\,dv^2}{E\,du^2 + 2F\,du\,dv + G\,dv^2} = \frac{\varepsilon}{R_n} \quad (\varepsilon = \pm 1); \tag{3a}$$

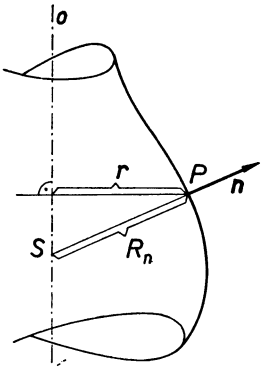


Fig. 9.32.

if a surface is given by an equation $z = f(x, y)$, then the following relation holds for the *radius of curvature of a curve of normal section* at a point of the surface:

$$\varepsilon R_n = \frac{\sqrt{(p^2 + q^2 + 1)}}{f_{xx}\cos^2\alpha + 2f_{xy}\cos\alpha\cos\beta + f_{yy}\cos^2\beta} \tag{3b}$$

(where α, β are the angles which the tangent of the curve of normal section at the point considered makes with the axes x and y).

If a regular point P of a surface $z = f(x, y)$ is the origin of the coordinate system and the normal to the surface at the point P is the z-axis (i.e. the tangent plane of the surface at P is the coordinate plane xy), then the expression for the curvature of a curve of normal section at P is of the form

$$\frac{\varepsilon}{R_n} = (f_{xx})_0\cos^2\varphi + 2(f_{xy})_0\sin\varphi\cos\varphi + (f_{yy})_0\sin^2\varphi \tag{4a}$$

or

$$\frac{\varepsilon}{R_n} = \tfrac{1}{2}(f_{xx} + f_{yy})_0 + \tfrac{1}{2}(f_{xx} - f_{yy})_0\cos 2\varphi + (f_{xy})_0\sin 2\varphi \tag{4b}$$

(φ being the angle between the tangent vector of the curve of normal section at P and the x-axis).

Theorem 3. *Among the curves of normal section at a point P of a surface there exist at least two curves in mutually perpendicular planes such that the normal*

curvature of one curve has a maximum value and that of the other curve a minimum value at P.

Definition 1. The curves of normal section of a surface for which the corresponding normal curvatures have extreme values are called the *principal curves of normal section*, and their radii of curvature R_1 and R_2 the *principal radii of curvature*, at the point considered on the surface.

Theorem 4 (*Euler's Theorem*). *The curvature* $1/R_n$ *of a curve of normal section at a regular point of a surface is given by the formula*

$$\frac{1}{R_n} = \frac{\cos^2 \delta}{\varepsilon R_1} + \frac{\sin^2 \delta}{\varepsilon R_2} \quad (\varepsilon = \pm 1) \tag{5}$$

(δ *being the angle between the plane of the curve of normal section and the plane of the first principal curve of normal section*).

REMARK 5. Let us introduce in the tangent plane at a point P of a surface, cartesian coordinates such that the x- or y-axis is in the tangent line of the first, or second, principal curve of normal section at P, respectively. If the point P is elliptic, hyperbolic, or parabolic (in this case, let the curvature of the second curve of normal section be equal to zero), let us construct, in the tangent plane at P, the ellipse, two hyperbolas, or two parallel straight lines given by the equations

$$\frac{x^2}{R_1} + \frac{y^2}{R_2} = 1\,, \quad \text{or} \quad \pm\frac{x^2}{R_1} \mp \frac{y^2}{R_2} = 1\,, \quad \text{or} \quad \frac{x^2}{R_1} = 1\,, \quad \text{respectively}, \tag{6}$$

(Fig. 9.33).

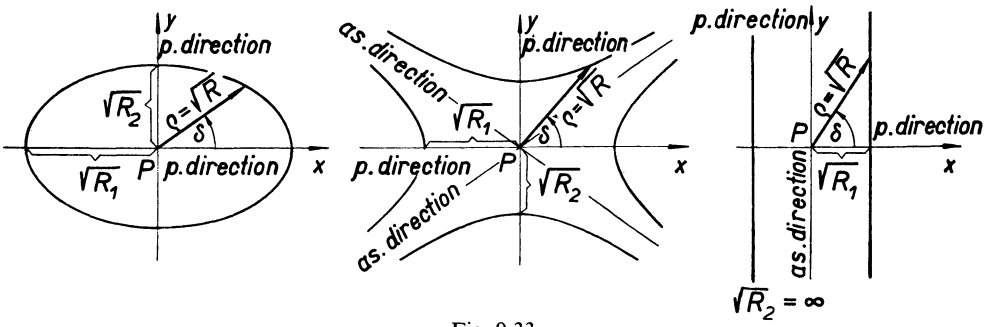


Fig. 9.33.

Then the length of radius vector ϱ of each point of the ellipse, hyperbolas, or the straight lines is the square root of the radius of curvature of the curve of normal section whose plane passes through the radius vector ϱ at the point P. The angle δ between ϱ and the first principal direction is the angle δ from equation (5).

Definition 2. The ellipse, two hyperbolas, or two parallel straight lines (6) in the tangent plane at a point of a surface is called the *indicatrix of Dupin.*

Definition 3. An elliptic point of a surface for which $R_1 = R_2$ is called an *umbilical point*, an *umbilic*, or a *circular point* of the surface.

REMARK 6. The indicatrix of Dupin at an umbilical point of a surface is a circle. Spheres are the only surfaces which have every point an umbilic.

Theorem 5. *The following relations hold*:

$$\pm \frac{1}{R_1} \cdot \frac{1}{R_2} = \frac{LN - M^2}{EG - F^2}, \quad \frac{1}{\varepsilon R_1} + \frac{1}{\varepsilon R_2} = \frac{EN - 2FM + GL}{EG - F^2} \quad (\varepsilon = \pm 1). \tag{7}$$

The principal radii R_1 and R_2 are the roots of the equation

$$(EG - F^2)\frac{1}{R^2} - (EN - 2FM + GL)\frac{1}{R} + (LN - M^2) = 0. \tag{8}$$

Definition 4. The product

$$K = \pm \frac{1}{R_1} \cdot \frac{1}{R_2} = \frac{LN - M^2}{EG - F^2} = \frac{W^2}{D^2} \qquad (LN - M^2 = W^2) \tag{9}$$

of the principal normal curvatures $1/\varepsilon R_1$, $1/\varepsilon R_2$ at a regular point of a surface is called the *total curvature* or the *Gaussian curvature* at the point of the surface.

The average

$$H = \frac{1}{2}\left(\frac{1}{\varepsilon R_1} + \frac{1}{\varepsilon R_2}\right) = \frac{EN - 2FM + GL}{2D^2} \quad (\varepsilon = \pm 1) \tag{10}$$

of the principal normal curvatures $1/\varepsilon R_1$, $1/\varepsilon R_2$ of a surface at a regular point is called the *mean curvature of the surface* at the point considered.

Theorem 6. *The average of the normal curvatures of two curves of normal section in mutually perpendicular planes at a point of a surface is constant, and equal to the mean curvature of the surface at that point.*

REMARK 7. At a regular point P of a surface $z = f(x, y)$, the Gaussian or mean curvature at P is given by the formula

$$K = \frac{f_{xx}f_{yy} - f_{xy}^2}{(f_x^2 + f_y^2 + 1)^2} \tag{11}$$

or

$$H = \frac{1}{2}\frac{(1 + f_y^2)f_{xx} - 2f_x f_y f_{xy} + (1 + f_x^2)f_{yy}}{(f_x^2 + f_y^2 + 1)^{3/2}}, \tag{12}$$

respectively.

REMARK 8. If P is an elliptic, parabolic or hyperbolic point of the surface, then $K > 0$ (i.e. $LN - M^2 > 0$), or $K = 0$ (i.e. $LN - M^2 = 0$ or at least one of the principal curvatures is zero), or $K < 0$ (i.e. $LN - M^2 < 0$), respectively. For developable surfaces (and for planes) $K = 0$ identically, and conversely. For so-called *minimal surfaces* the equation $H = 0$ holds identically.

9.17. Lines of Curvature

Definition 1. A curve on a surface whose tangent line at every point lies in the principal direction on the surface (Definition 9.16.1) is called a *line of curvature.*

Theorem 1. *The differential equation of the lines of curvature is of the form*

$$(LF - ME)\,du^2 + (LG - NE)\,du\,dv + (MG - NF)\,dv^2 = 0\,. \tag{1}$$

REMARK 1. Through a regular (not umbilical) point of a surface there pass two real lines of curvature intersecting at right angles. The lines of curvature constitute an *orthogonal conjugate net* on a surface.

Theorem 2. *A necessary and sufficient condition for a point of a surface* $\mathbf{r} = \mathbf{r}(u, v)$ *to be an umbilical point is*

$$L = \lambda E\,, \quad M = \lambda F\,, \quad N = \lambda G \quad (\lambda \neq 0)\,. \tag{2a}$$

If a surface is given by an equation $z = f(x, y)$, *then the equations*

$$\frac{1 + f_x^2}{f_{xx}} = \frac{f_x f_y}{f_{xy}} = \frac{1 + f_y^2}{f_{yy}} \tag{2b}$$

express the conditions for an umbilical point of the surface.

REMARK 2. For the radii of curvature of curves of normal section at an umbilical point of a surface $z = f(x, y)$ the following relation holds:

$$R_n = \frac{1 + f_x^2}{f_{xx}}\sqrt{(1 + f_x^2 + f_y^2)}\,. \tag{3}$$

Theorem 3. *A curve on a surface is a line of curvature of the surface if and only if the ruled surface of normals to the surface at points of the curve is a developable surface.*

REMARK 3. Every curve on spheres and planes is a line of curvature.

9.18. Asymptotic Curves

Definition 1. An *asymptotic curve* (or *asymptotic line*) is a curve on the surface, to which the tangent line, at every point of it, has asymptotic direction (Remark 9.15.2).

Theorem 1. *The differential equation of asymptotic curves is of the form*

$$L\,du^2 + 2M\,du\,dv + N\,dv^2 = 0\,. \tag{1}$$

REMARK 1. Only the plane has the property that every curve in it is an asymptotic curve.

Theorem 2. *Real asymptotic curves exist only on that part of a surface where all points are hyperbolic* $(K < 0)$ *or parabolic* $(K = 0)$. *Through every hyperbolic point of a surface there pass two real distinct asymptotic curves. Through every parabolic point of a surface there passes exactly one real asymptotic curve.*

Theorem 3. *If a curve of section of a surface by its tangent plane has a double point at the corresponding point of contact, then the tangent lines of the curve at the double point are in asymptotic directions.*

Theorem 4. *An asymptotic curve on a surface is a curve such that the osculating plane at every point of the curve coincides with the tangent plane of the surface at the same point.*

REMARK 2. If a surface is given by the equation $z = f(x, y)$, then the equation

$$f_{xx}\,dx^2 + 2f_{xy}\,dx\,dy + f_{yy}\,dy^2 = 0$$

is the differential equation of the orthogonal projection of asymptotic curves of the surface onto the coordinate plane xy.

The curves of normal section of a surface which touch an asymptotic curve have zero curvature at their common point. An asymptotic curve follows a direction of zero normal curvature on the surface. The tangent lines of asymptotic curves at every point of the surface are identical with the asymptotes of the corresponding indicatrix of Dupin. On a ruled surface one family of asymptotic curves is constituted by the generators of the surface.

9.19. Fundamental Equations of Weingarten, Gauss and Codazzi

The formulae of the following theorems give the relations between the vectors $\boldsymbol{n}_u$, $\boldsymbol{n}_v$ and the vectors $\boldsymbol{r}_u$, $\boldsymbol{r}_v$ ($\partial\boldsymbol{n}/\partial u = \boldsymbol{n}_u$, $\partial\boldsymbol{r}/\partial u = \boldsymbol{r}_u$, etc.) corresponding to the Frenet formulae of the theory of curves.

Theorem 1 (*The Weingarten Equations*). *The following relations exist between the vectors* $\mathbf{n}_u$, $\mathbf{n}_v$ *and* $\mathbf{r}_u$, $\mathbf{r}_v$:

$$\left.\begin{aligned} \mathbf{n}_u &= \frac{FM - GL}{D^2}\mathbf{r}_u + \frac{FL - EM}{D^2}\mathbf{r}_v\,; \quad \mathbf{r}_u = \frac{MF - NE}{W^2}\mathbf{n}_u + \frac{ME - LF}{W^2}\mathbf{n}_v\,; \\ \mathbf{n}_v &= \frac{FN - GM}{D^2}\mathbf{r}_u + \frac{FM - EN}{D^2}\mathbf{r}_v\,; \quad \mathbf{r}_v = \frac{MG - NF}{W^2}\mathbf{n}_u + \frac{MF - LG}{W^2}\mathbf{n}_v \end{aligned}\right\} \quad (1)$$

(*where* $W^2 = LN - M^2$).

Theorem 2 (*The Gauss Equations*). *The following relations exist between the vectors* $\mathbf{r}_{uu}$, $\mathbf{r}_{uv}$, $\mathbf{r}_{vv}$ *and* $\mathbf{r}_u$, $\mathbf{r}_v$, $\mathbf{n}$ (*where* $\partial^2\mathbf{r}/\partial u\,\partial v = \mathbf{r}_{uv}$, *etc.*):

$$\left.\begin{aligned} \mathbf{r}_{uu} &= \frac{GE_u - 2FF_u + FE_v}{2D^2}\mathbf{r}_u + \frac{-FE_u + 2EF_u - EE_v}{2D^2}\mathbf{r}_v + L\mathbf{n}\,, \\ \mathbf{r}_{uv} &= \frac{GE_v - FG_u}{2D^2}\mathbf{r}_u + \frac{EG_u - FE_v}{2D^2}\mathbf{r}_v + M\mathbf{n}\,, \\ \mathbf{r}_{vv} &= \frac{-FG_v + 2GF_v - GG_u}{2D^2}\mathbf{r}_u + \frac{EG_v - 2FF_v + FG_u}{2D^2}\mathbf{r}_v + N\mathbf{n} \end{aligned}\right\} \quad (2)$$

(*where* $E_u = \partial E/\partial u$, $F_v = \partial F/\partial v$, *etc.*).

The following relations express the mutual dependence of the six functions E, F, G; L, M, N which correspond to the same surface:

Theorem 3 (*The Codazzi or Mainardi Equations*):

$$\left.\begin{aligned} &(EG - 2FF + GE)(L_v - M_u) - (EN - 2FM + GL)(E_v - F_u) + \\ &\qquad + \begin{vmatrix} E, & E_u, & L \\ F, & F_u, & M \\ G, & G_u, & N \end{vmatrix} = 0\,, \\ &(EG - 2FF + GE)(M_v - N_u) - (EN - 2FM + GL)(F_v - G_u) + \\ &\qquad + \begin{vmatrix} E, & E_v, & L \\ F, & F_v, & M \\ G, & G_v, & N \end{vmatrix} = 0 \end{aligned}\right\} \quad (3)$$

(*where* $L_v = \partial L/\partial v$, $M_u = \partial M/\partial u$, *etc.*).

Theorem 4 (*The Gauss Theorem Egregium*).

$$K = -\frac{1}{4D^4}\begin{vmatrix} E, & E_u, & E_v \\ F, & F_u, & F_v \\ G, & G_u, & G_v \end{vmatrix} - \frac{1}{2D}\left(\frac{\partial}{\partial v}\frac{E_v - F_u}{D} - \frac{\partial}{\partial u}\frac{F_v - G_u}{D}\right). \quad (4)$$

9.20. Geodesic Curvature, Geodesic Curves and Gradient Curves on a Surface

Definition 1. The *geodesic curvature* k_g of a curve on a surface at a regular point P is defined by the relation

$$k_g = k_1 \sin \vartheta \tag{1}$$

(ϑ being the angle between the principal normal to the curve and the normal to the surface at P and k_1 the curvature of this curve at P).

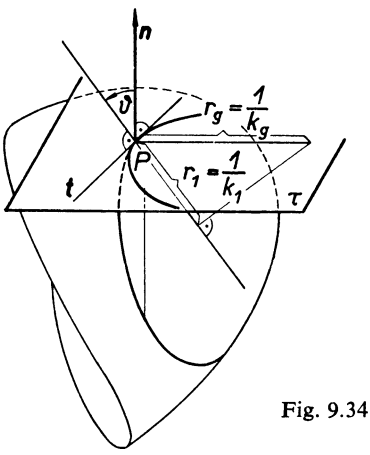


Fig. 9.34.

Theorem 1. *The geodesic curvature of a curve on a surface at a point P is equal to the curvature of the orthogonal projection of the curve onto the tangent plane of the surface at P* (Fig. 9.34).

If a curve $\mathbf{r}(s) = \mathbf{r}(u(s), v(s))$ on a surface $\mathbf{r} = \mathbf{r}(u, v)$ is given, then for its geodesic curvature k_g the following relation holds:

Theorem 2.

$$k_g = \frac{1}{r_g} = [\mathbf{r}_s \mathbf{r}_{ss} \mathbf{n}] . \tag{2}$$

REMARK 1. If a curve depends on a parameter t different from the arc s, then

$$k_g = \frac{1}{ds^3} [d\mathbf{r}, d^2\mathbf{r}, \mathbf{n}] \tag{3}$$

holds instead of (2).

Definition 2. A curve on a surface is called a *geodesic curve*, or simply a *geodesic* if at every point of the curve its osculating plane contains the corresponding normal to the surface.

Theorem 3. *A necessary and sufficient condition that a curve on a surface is a geodesic is that the geodesic curvature of this curve at every point is zero,* i.e. $k_g = 0$ *at every point of the curve.*

REMARK 2. The geodesics on a surface thus satisfy the condition (the differential equation)

$$[\boldsymbol{r}_s \boldsymbol{r}_{ss} \boldsymbol{n}] = 0 \quad \text{or} \quad [\mathrm{d}\boldsymbol{r}, \mathrm{d}^2\boldsymbol{r}, \boldsymbol{n}] = 0 . \tag{4}$$

A straight line on a surface is a geodesic on this surface.

Theorem 4. *On the assumption that the curvilinear coordinates on a surface constitute an orthogonal net, the differential equation of geodesics on the surface can be put in the form*

$$\sqrt{(EG)}\frac{\mathrm{d}\varphi}{\mathrm{d}s} + \tfrac{1}{2}G_u\frac{\mathrm{d}v}{\mathrm{d}u} - \tfrac{1}{2}E_v = 0 \tag{5}$$

($\mathrm{d}s^2 = E\,\mathrm{d}u^2 + G\,\mathrm{d}v^2$,

$\mathrm{d}\varphi = [(EG\,\mathrm{d}u\,\mathrm{d}v^2 - \frac{1}{2}G\,\mathrm{d}E\,\mathrm{d}u\,\mathrm{d}v + \frac{1}{2}E\,\mathrm{d}G\,\mathrm{d}u\,\mathrm{d}v)/\sqrt{(EG)}]\,\mathrm{d}s^2$, *$\varphi$ being the angle which the geodesic makes with the parametric v-curve at a point of the surface*).

Theorem 5. *From among all arcs on a given (properly chosen) part of a surface joining two definite points of the surface the shortest arc is on a geodesic.*

REMARK 3. All straight lines on a plane and on a ruled surface are geodesics. The helix is a geodesic on a cylinder, etc. A geodesic on a surface is determined by a point of the surface and its tangent line at that point. Through every regular point of a surface there passes a one-parameter family of geodesics. When a developable surface is developed upon a plane, the arc of every geodesic on the surface becomes a line segment in the plane.

Definition 3. The *gradient curves* (the *curves of greatest slope*) on a surface are the orthogonal trajectories of the level lines of the surface with respect to a given plane.

Theorem 6. *The equation*

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{f_y}{f_x} \tag{6}$$

is the differential equation of the orthogonal projection of the gradient curves of a surface $z = f(x, y)$ with respect to the coordinate plane xy, onto this plane.

10. SEQUENCES AND SERIES OF CONSTANT TERMS. INFINITE PRODUCTS

By KAREL REKTORYS

References: [1], [17], [26], [28], [30], [54], [59], [60], [74], [75], [84], [91], [92], [96], [99], [111], [119], [122], [137], [148], [158], [160], [164].

For sequences and series of variable terms see Chapters 15 and 16.

10.1. Sequences of Constant Terms

Definition 1. If to every natural number n we assign a number a_n (which may be real or complex) and order the numbers a_n according to their increasing suffixes, then we say that we have formed a *sequence*. We denote it by $a_1, a_2, a_3, \ldots, a_n, \ldots$ or briefly by $\{a_n\}$.

Example 1. The relation $a_n = (n-1)/n$ defines the sequence

$$0, \tfrac{1}{2}, \tfrac{2}{3}, \tfrac{3}{4}, \ldots .$$

Definition 2. We say that a sequence $\{a_n\}$ possesses a (*finite*) *limit* a (in other words: *tends* to a number a, *converges* with limit a, is said to converge to a limit a), if, to any arbitrary number $\varepsilon > 0$, there corresponds a number n_0 (depending in

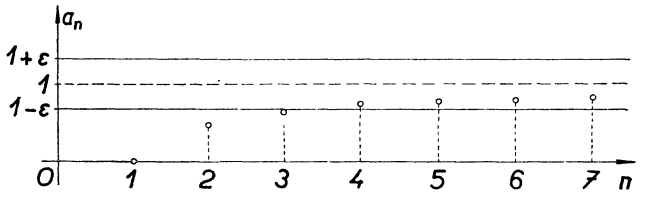


Fig. 10.1.

general on the choice of the number ε), such that $|a_n - a| < \varepsilon$ holds for every $n > n_0$ (Fig 10.1). We then write

$$\lim_{n\to\infty} a_n = a .$$

(When dealing with sequences, it is customary to write simply $n \to \infty$ rather than $n \to +\infty$.)

REMARK 1. Roughly speaking: A sequence $\{a_n\}$ possesses a limit a if the number a_n "approaches closer and closer" to the number a as the suffix n increases.

Definition 3. When a sequence possesses a (finite) limit, it is said to be *convergent.* Otherwise it is said to be *divergent.*

Theorem 1. *A sequence $\{a_n\}$ can have at most one limit.*

Definition 4. We say that $\{a_n\}$ *diverges* to $+\infty$ (has an *infinite limit* $+\infty$, is definitely *divergent* with the limit $+\infty$), if, for any arbitrary number K, there exists a number n_0 (depending on the choice of the number K), such that $a_n > K$ for every $n > n_0$. We then write

$$\lim_{n\to\infty} a_n = +\infty .$$

A corresponding meaning holds if we write

$$\lim_{n\to\infty} a_n = -\infty .$$

Example 2. The sequence given in Example 1 is convergent and the number 1 is its limit. For any given $\varepsilon > 0$, it is sufficient to choose as n_0 any number greater than $1/\varepsilon$ (Fig. 10.1). For, if $n_0 > 1/\varepsilon$ and $n > n_0$, then

$$|a_n - 1| = \left|\frac{n-1}{n} - 1\right| = \frac{1}{n} < \frac{1}{n_0} < \varepsilon .$$

Example 3. The sequence defined by the relation $a_n = 2^n$ (i.e. the sequence 2, 4, 8, ...) diverges to $+\infty$, i.e.

$$\lim_{n\to\infty} 2^n = +\infty .$$

Example 4. The sequence 0, 1, 0, 1, 0, 1, ... is divergent (its terms do not tend to any (single) number). As a rule, sequences of this type are said to *oscillate.*

Theorem 2. *A sequence $\{a_n\}$ is convergent if and only if it fulfils the following (Bolzano-Cauchy) condition: Given any (arbitrarily small) number $\varepsilon > 0$, there exists a number n_0 (depending on the choice of the number ε) such that the relation $|a_m - a_n| < \varepsilon$ is valid for every pair of numbers m, n, such that $m > n_0$, $n > n_0$.*

Theorem 3. *If $\{a_n\}$, $\{b_n\}$ are convergent sequences such that*

$$\lim_{n\to\infty} a_n = a , \quad \lim_{n\to\infty} b_n = b ,$$

then

$$\lim_{n\to\infty} (a_n \pm b_n) = a \pm b , \quad \lim_{n\to\infty} ka_n = ka \quad (k \text{ a constant}) , \quad \lim_{n\to\infty} |a_n| = |a| ,$$

$$\lim_{n\to\infty} a_n b_n = ab , \quad \lim_{n\to\infty} \frac{a_n}{b_n} = \frac{a}{b} \quad \text{for} \quad b \neq 0 .$$

(*Thus each of the sequences mentioned above is convergent. In the last case, we omit from the sequence* $\{a_n/b_n\}$ *the terms with suffixes for which* $b_n = 0$ *and of which there are a finite number, because* $b \neq 0$.)

Example 5. Find

$$\lim_{n\to\infty} \frac{2n^2 - n + 1}{7n^2 - 5n + 2}.$$

Because the sequences $\{a_n\}$ (in the numerator) and $\{b_n\}$ (in the denominator) are divergent, we cannot apply Theorem 3 directly. However, for every n

$$\frac{2n^2 - n + 1}{7n^2 - 5n + 2} = \frac{2 - \frac{1}{n} + \frac{1}{n^2}}{7 - \frac{5}{n} + \frac{2}{n^2}},$$

so that

$$\lim_{n\to\infty} \frac{2n^2 - n + 1}{7n^2 - 5n + 2} = \lim_{n\to\infty} \frac{2 - \frac{1}{n} + \frac{1}{n^2}}{7 - \frac{5}{n} + \frac{2}{n^2}} = \frac{2}{7},$$

because

$$\lim_{n\to\infty} \frac{1}{n} = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{1}{n^2} = 0.$$

REMARK 2. Theorem 3 can be generalized to cases where the sequences $\{a_n\}$, $\{b_n\}$ diverge to $+\infty$ or $-\infty$. If, for example,

$$\lim_{n\to\infty} a_n = +\infty \quad \text{and} \quad \lim_{n\to\infty} b_n = -\infty$$

then

$$\lim_{n\to\infty} (a_n - b_n) = +\infty \quad \text{and} \quad \lim_{n\to\infty} a_n b_n = -\infty.$$

But we cannot apply the theorem directly when the computation leads to so-called indeterminate expressions of the type $\infty - \infty$, $0 \cdot \infty$ or ∞/∞.

Theorem 4. *Let*

$$\lim_{n\to\infty} a_n = a, \quad \lim_{n\to\infty} b_n = a$$

and let $\{c_n\}$ *be a sequence such that* $a_n \leqq c_n \leqq b_n$ *holds for every* n. *Then*

$$\lim_{n\to\infty} c_n = a.$$

REMARK 3. In other words: If a sequence $\{c_n\}$ lies between two convergent sequences $\{a_n\}$ and $\{b_n\}$ which have the same limit, then $\{c_n\}$ is also convergent and has this same limit.

Theorem 5. *Let $\{a_n\}$ be a sequence of complex numbers and for every n let $a_n = \alpha_n + i\beta_n$, where α_n, β_n are real. Then $\{a_n\}$ is convergent if and only if both sequences $\{\alpha_n\}$ and $\{\beta_n\}$ are convergent. Moreover, the relation*

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \alpha_n + i \lim_{n\to\infty} \beta_n$$

holds.

Theorem 6. *Let $\lim_{n\to\infty} a_n = a$, $\lim_{n\to\infty} b_n = b$. If, for every n, $a_n \leqq b_n$, then $a \leqq b$.*

REMARK 4. If $a_n < b_n$ for every n, then we cannot conclude that $a < b$, but only that $a \leqq b$. For example:

$$\{a_n\} = \left\{\frac{1}{2n}\right\}, \quad \{b_n\} = \left\{\frac{1}{n}\right\}.$$

Then $a_n < b_n$ for every n, but $a = b = 0$.

Definition 5. Let us consider the sequence $\{a_n\}$. A sequence $\{a_{k_n}\}$, where k_n are positive integers, such that $k_1 < k_2 < k_3 < \ldots$, is called a *sub-sequence of the sequence* $\{a_n\}$.

Example 6. If, for the sequence $0, 1, 0, 1, 0, 1, \ldots$, we choose $k_1 = 2$, $k_2 = 4$, $k_3 = 6, \ldots$, we obtain the sub-sequence $1, 1, 1, \ldots$.

Example 7. Let us form a sub-sequence of the sequence $\{a_n\} = \{(n-1)/n\}$ in such a way that we take only every third term, i.e. $k_1 = 3$, $k_2 = 6$, $k_3 = 9, \ldots$ Then the resulting sequence is

$$\frac{2}{3}, \frac{5}{6}, \frac{8}{9}, \ldots .$$

Theorem 7. *If $\lim_{n\to\infty} a_n = a$, then $\lim_{n\to\infty} a_{k_n} = a$ also.*

REMARK 5. The converse of the Theorem is not true. If a sub-sequence is convergent then the original sequence need not be convergent. For example, the sequence $0, 1, 0, 1, 0, 1, \ldots$ from Example 6 is divergent but its sub-sequence $1, 1, 1, \ldots$ is convergent.

Definition 6. A sequence $\{a_n\}$ is said to be *bounded above*, or *bounded below*, or *bounded* if there exists a finite number K_1, or K_2, or M, respectively, such that

$$a_n < K_1, \quad \text{or} \quad a_n > K_2, \quad \text{or} \quad |a_n| < M, \quad \text{respectively}$$

for every n.

Definition 7. A number d is called a *point of accumulation* or *limiting point* of a sequence $\{a_n\}$ if an infinite number of terms of the given sequence lie in every arbitrarily small ε-neighbourhood of the point d (i.e. in the interval $(d - \varepsilon, d + \varepsilon)$). (The term cluster point is also used.)

Theorem 8 (*The Bolzano-Weierstrass Theorem*). *Every bounded sequence* $\{a_n\}$ *possesses at least one limiting point. There always exist* (*even when there are infinitely many limiting points*) *a greatest limiting point and a least limiting point; we denote them by*:

$$\limsup_{n\to\infty} a_n \quad \text{or} \quad \overline{\lim_{n\to\infty}}\, a_n\,, \quad \liminf_{n\to\infty} a_n \quad \text{or} \quad \underline{\lim_{n\to\infty}}\, a_n\,,$$

respectively.

REMARK 6 The expression

$$\overline{\lim_{n\to\infty}}\, a_n = +\infty\,,$$

which is occasionally to be found in the literature, stands for the assertion "However large a number K be chosen, the sequence $\{a_n\}$ always possesses infinitely many terms such that $a_n > K$." The expression

$$\underline{\lim_{n\to\infty}}\, a_n = -\infty$$

has a corresponding meaning. For example, for the sequence $a_n = (-n)^n$ the relations

$$\overline{\lim_{n\to\infty}}\, a_n = +\infty\,, \quad \underline{\lim_{n\to\infty}}\, a_n = -\infty$$

hold.

Theorem 9. *The sequence* $\{a_n\}$ *is convergent if, and only if, the numbers* $\overline{\lim_{n\to\infty}}\, a_n$ *and* $\underline{\lim_{n\to\infty}}\, a_n$ *are finite and*

$$\overline{\lim_{n\to\infty}}\, a_n = \underline{\lim_{n\to\infty}}\, a_n\,,$$

i.e. if the sequence $\{a_n\}$ *possesses one and only one limiting point* (*which is not infinite*).

Example 8. The sequence 0, 1, 0, 1, 0, 1, ... in Example 4 possesses two limiting points: the point 0 and the point 1. Thus

$$\overline{\lim_{n\to\infty}}\, a_n = 1\,, \quad \underline{\lim_{n\to\infty}}\, a_n = 0\,.$$

Then the fact that this sequence is not convergent follows from Theorem 9.

Definition 8. A sequence is said to be

$$\left.\begin{array}{l} \textit{strictly increasing} \\ \textit{strictly decreasing} \\ \textit{decreasing} \\ \textit{increasing} \end{array}\right\} \text{ if, for every } n, \quad \begin{array}{l} a_{n+1} > a_n \\ a_{n+1} < a_n \\ a_{n+1} \leqq a_n \\ a_{n+1} \geqq a_n \,. \end{array}$$

All such sequences will be referred to as *monotonic sequences,* the first two as *strictly monotonic sequences.*

Theorem 10. *Every increasing (and thus every strictly increasing) sequence which is bounded above possesses a limit (equal to the l.u.b. of all the values a_n). Similarly every decreasing (and thus every strictly decreasing) sequence which is bounded below is convergent (and its limit is equal to the g. l. b. of all the values a_n).*

Theorem 11. *The sequences*

$$a_n = \left(1 + \frac{1}{n}\right)^n, \quad b_n = \left(1 + \frac{1}{n}\right)^{n+1}$$

(the former being strictly increasing, and the latter strictly decreasing) possess the same limit, the number e *(so-called) which is the base of natural logarithms. Thus*

$$\mathrm{e} = \lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^n = \lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^{n+1} = 2{\cdot}718,281,828,459,0\ldots.$$

Theorem 12. *If $a > 0$, $\lim_{n\to\infty} b_n = b$ (b_n need not be rational*; see § 1.9), *then*

$$\lim_{n\to\infty} a^{b_n} = a^{\lim\limits_{n\to\infty} b_n} = a^b \,.$$

Example 9.

$$\lim_{n\to\infty} \sqrt[n]{a} = 1 \ (a > 0), \text{ because } \sqrt[n]{a} = a^{1/n} \text{ and } \lim_{n\to\infty} \frac{1}{n} = 0 \,.$$

Theorem 13. *The sequence*

$$a_1 \,,\ a_1 + d \,,\ \ldots,\ a_1 + nd \,,\ \ldots,$$

called an arithmetic progression, is divergent for every $d \neq 0$.

The sequence

$$a_1, a_1 q, a_1 q^2, \ldots, a_1 q^n, \ldots,$$

called a geometric progression, has limit 0 for $|q| < 1$ and limit a_1 for $q = +1$; if $a_1 \neq 0$, then for any other value of q this sequence is divergent.

REMARK 7. It is often convenient to replace the calculation of the limit of a sequence by the calculation of the limit of a suitable function as $x \to +\infty$. If $f(x)$ is a function such that $f(n) = a_n$ for every positive integer n, then from the existence of the limit $\lim_{x\to+\infty} f(x) = A$, it follows that $\lim_{n\to\infty} a_n = A$.

Example 10. To calculate

$$\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n$$

we make use of the function

$$f(x) = \left(1+\frac{a}{x}\right)^x.$$

We have

$$\left(1+\frac{a}{x}\right)^x = \left[e^{\ln(1+a/x)}\right]^x = e^{x\,\ln(1+a/x)},$$

while

$$\lim_{x\to+\infty}\left\{x\ln\left(1+\frac{a}{x}\right)\right\} = \lim_{x\to+\infty}\frac{\ln(1+a/x)}{1/x} =$$

$$= \lim_{z\to0+}\frac{\ln(1+az)}{z} = \lim_{z\to0+}\frac{a}{1+az} = a$$

(according to l'Hôpital's Rule, Theorem 11.8.1). Because the exponential function is continuous everywhere we have

$$\lim_{x\to+\infty} e^{x\,\ln(1+a/x)} = e^{\lim\limits_{x\to+\infty} x\,\ln(1+a/x)} = e^a.$$

Hence

$$\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = e^a.$$

Theorem 14. *Survey of Important Formulae and Limits*

1. $\lim_{n\to\infty}(a_n \pm b_n) = \lim_{n\to\infty} a_n \pm \lim_{n\to\infty} b_n$; $\lim_{n\to\infty} ka_n = k\lim_{n\to\infty} a_n$ (k a constant);

   $\lim_{n\to\infty}|a_n| = \left|\lim_{n\to\infty} a_n\right|$; $\lim_{n\to\infty} a_n b_n = \lim_{n\to\infty} a_n \cdot \lim_{n\to\infty} b_n$;

   $$\lim_{n\to\infty}\frac{a_n}{b_n} = \frac{\lim\limits_{n\to\infty} a_n}{\lim\limits_{n\to\infty} b_n} \quad (\lim_{n\to\infty} b_n \neq 0;\ \text{cf. Theorem 3}).$$

2. $a_n \leqq b_n \Rightarrow \lim_{n\to\infty} a_n \leqq \lim_{n\to\infty} b_n$ (Theorem 6).

3. $\lim_{n\to\infty} a_{k_n} = \lim_{n\to\infty} a_n$ (Theorem 7).

4. $\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = e^a$ *for every* a; *in particular* $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = e$.

5. $\lim_{n\to\infty} a^n = 0$ *for* $|a| < 1$.

6. $\lim_{n\to\infty}\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}-\ln n\right) = C = 0{\cdot}577,215,664,9\ldots$ (*Euler's Constant*).

7. $\lim_{n\to\infty}\sqrt[n]{n!} = +\infty$; $\lim_{n\to\infty}\sqrt[n]{\frac{1}{n!}} = 0$; $\lim_{n\to\infty}\frac{\sqrt[n]{n!}}{n} = \frac{1}{e}$; $\lim_{n\to\infty}\frac{n!}{n^n} = 0$.

8. $\lim_{n\to\infty}\frac{n!}{n^n e^{-n}\sqrt{n}} = \sqrt{(2\pi)}$ (*Stirling's Formula*).

9. $\lim_{n\to\infty}\left[\frac{2.4.6.\ldots.2n}{1.3.5.\ldots.(2n-1)}\right]^2.\frac{1}{2n} = \frac{\pi}{2}$ (*Wallis's Product*).

10. $\lim_{n\to\infty}\frac{1^k+2^k+\ldots+n^k}{n^{k+1}} = \frac{1}{k+1}$ (*k a positive integer*).

11. $\lim_{n\to\infty}\frac{1^2+3^2+5^2+\ldots+(2n-1)^2}{n^3} = \frac{4}{3}$.

12. *If* $a_n > 0$ *and* $\lim_{n\to\infty}\frac{a_{n+1}}{a_n} = a$ *then* $\lim_{n\to\infty}\sqrt[n]{a_n} = a$ *also.*

13. *If* $\lim_{n\to\infty} a_n = a$ *then* $\lim_{n\to\infty}\frac{a_1+a_2+\ldots+a_n}{n} = a$ *also.*

10.2. Infinite Series (of Constant Terms)

Definition 1. Let the sequence $\{a_n\}$ be given. The symbolical expression

$$a_1 + a_2 + a_3 + \ldots = \sum_{n=1}^{\infty} a_n \tag{1}$$

is called the (*infinite*) *series* corresponding to the given sequence.

The sum

$$s_n = a_1 + a_2 + \ldots + a_n$$

is called a *partial sum of the series* (1).

Definition 2. If the sequence of partial sums $s_1, s_2, s_3, \ldots$ is convergent (see definition 10.1.3) and possesses a (finite) limit s, we say that the series is *convergent* and *has the sum* s. If the sequence $\{s_n\}$ is divergent then we say that the series (1) is *divergent.*

Example 1. For the so-called *geometric* series $1 + q + q^2 + \ldots$ we have

$$s_n = \frac{1 - q^n}{1 - q} \quad (q \neq 1).$$

If $|q| < 1$ then

$$s = \lim_{n \to \infty} s_n = \frac{1}{1 - q},$$

so the series is convergent. If $q = 1$ then $s_n = n$, so the sequence of partial sums diverges to $+\infty$ (see Definition 10.1.4). (In this case we say that the *series has the sum* $+\infty$.) For $q = -1$, $s_1 = 1$, $s_2 = 0$, $s_3 = 1$, $s_4 = 0$, ... and, the sequence s_n having no limit, the series is divergent (we speak of an *oscillating series*).

Example 2. The *arithmetic* series

$$a_1 + (a_1 + d) + (a_1 + 2d) + \ldots$$

is divergent when at least one of the numbers a_1, d is not zero. For $d > 0$ its sum is $+\infty$, for $d < 0$ its sum is $-\infty$.

Example 3. The series

$$1 + \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{4} + \ldots$$

(the so-called *harmonic series*) is divergent (its sum is $+\infty$, see Examples 6 and 7.)

REMARK 1. The problem of deciding whether a given series is convergent or not is a very important one. When we know that a series is convergent we may (according to the definition) determine an approximation to its sum to any desired degree of accuracy by considering the finite sum of a sufficiently large number of its terms. It is for this reason that in the following text we are so concerned with tests for the convergence of infinite series. To determine exactly the sum of a series is generally a difficult problem, and theorems on the differentiation and integration of power series (§ 15.4) and on double series (see Remark 14) etc. provide an effective help.

Theorem 1. *For the series* $a_1 + a_2 + a_3 + \ldots$ *to be convergent it is necessary that*

$$\lim_{n \to \infty} a_n = 0.$$

REMARK 2. This condition is not, however, sufficient for the convergence of a given series as we can see in the case of the harmonic series in Example 3.

Theorem 2 (*The Bolzano-Cauchy Condition*). *A necessary and sufficient condition for the convergence of the series* $a_1 + a_2 + a_3 + \ldots$ *is that, for any given* $\varepsilon > 0$, *there exists a number* n_0 *such that for every* $n > n_0$ *and every positive integer* p *the relation*

$$|a_n + a_{n+1} + a_{n+2} + \ldots + a_{n+p}| < \varepsilon$$

holds.

Theorem 3. *If the series* $a_1 + a_2 + a_3 + \ldots$ *and* $b_1 + b_2 + b_3 + \ldots$ *are convergent with respective sums* s *and* t, *then the series whose n-th terms are respectively* $a_n \pm b_n$ *and* ka_n (k *constant*), *are also convergent and*

$$\sum_{n=1}^{\infty}(a_n \pm b_n) = s \pm t\,, \quad \sum_{n=1}^{\infty} ka_n = ks\,.$$

REMARK 3. For the multiplication of infinite series see Theorems 21, 22 and 24.

REMARK 4. If the series $\sum_{n=1}^{\infty}(a_n + b_n)$ is convergent then neither of the series $\sum_{n=1}^{\infty} a_n$, $\sum_{n=1}^{\infty} b_n$ need be convergent. For example, the series $(1 - 1) + (1 - 1) + (1 - 1) + \ldots$ is convergent as each of its terms is zero but neither the series $1 + 1 + 1 + \ldots$ nor the series $-1 - 1 - 1 - \ldots$ is convergent. The series without parentheses is also divergent.

Theorem 4. *If the series*

$$|a_1| + |a_2| + |a_3| + \ldots \tag{2}$$

converges, then the series

$$a_1 + a_2 + a_3 + \ldots \tag{3}$$

also converges.

Definition 3. If the series (2) converges then the series (3) is said to be *absolutely convergent*. If the series (3) converges but the series (2) does not, then the series (3) is said to be *conditionally convergent*.

Example 4. The series whose n-th term is

$$a_n = \frac{(-1)^{n+1}}{n} \tag{4}$$

(that is the series

$$1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} + \ldots),$$

converges (see Example 9), but the corresponding series of absolute values (2), that is

$$1 + \tfrac{1}{2} + \tfrac{1}{3} + \ldots$$

is the harmonic series (see Example 3) which is divergent. Hence the series

$$1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} + \ldots \tag{5}$$

is conditionally convergent.

Theorem 5. *If* (1) *is absolutely convergent then every series that arises from the series* (1) *by a rearrangement of its terms* (*i.e. by a change of the order of its terms*) *is also absolutely convergent and possesses the same sum.*

Theorem 6. *If* (1) *is conditionally convergent then we can obtain from* (1), *by a suitable rearrangement of its terms, a series converging to any given value, or even one which diverges.*

Theorem 7. *If we remove from an infinite series a finite number of its terms we change in general the sum of the series, but alter nothing regarding its convergence or divergence.*

REMARK 5. If we omit in a given series the terms with value zero (finite or infinite in number) then we change neither the convergence nor the sum of the series. *We can thus suppose in the following exposition that the series considered have no zero terms,* so that instead of series with non-negative terms we need consider only *series with positive terms.* This is important in particular in the case of d'Alembert's Ratio Test (Theorem 13) where the term a_n occurs as a divisor.

Theorem 8. *A series of positive terms converges if and only if the sequence* s_n *of its partial sums* (see Definition 1) *is bounded above.*

REMARK 6. When we say that the terms a_n of a given series have some property for *almost every* n, then we mean that they have that property for all n with the exception of a finite number of terms at most. For example, if the relation $a_n < a_{n+1}$ holds for almost every n, it means that this inequality fails for a finite number of indices n at most.

Theorem 9 (*Comparison Test*). *Let*

$$a_1 + a_2 + a_3 + \ldots, \tag{6}$$

$$b_1 + b_2 + b_3 + \ldots \tag{7}$$

be two series of positive (Remark 5) *terms and suppose* $a_n \leqq b_n$ *for almost every* n (Remark 6). *Then, from the convergence of the series* (7) *follows the convergence of the series* (6), *while from the divergence of the series* (6) *follows the divergence of the series* (7).

(A series such as (7) is called a *majorant* of the series (6).)

Theorem 10. *Let* (6), (7) *be two series of positive terms and* $K \neq 0$ *be a finite number. If the limit*

$$\lim_{n\to\infty} \frac{a_n}{b_n} = K$$

exists then both the series (6), (7) *are simultaneously either convergent or divergent.*

Theorem 11. *If the series of positive terms* $a_1 + a_2 + a_3 + \ldots$ *is convergent and* $c_1, c_2, c_3, \ldots$ *are positive numbers, for which* $c_n < A$ *for almost every* n (A *being a constant*) *then the series*

$$c_1a_1 + c_2a_2 + c_3a_3 + \ldots$$

is also convergent.

Theorem 12. *If* $a_1 + a_2 + a_3 + \ldots$, $b_1 + b_2 + b_3 + \ldots$ *are two series of positive terms, the latter being convergent, and if*

$$\frac{a_{n+1}}{a_n} \leqq \frac{b_{n+1}}{b_n}$$

for almost every n (Remark 6), *then the former of these series is also convergent.*

Theorem 13. *Let* $a_1 + a_2 + a_3 + \ldots$ *be a series of positive terms. If*

$\lim_{n\to\infty} \sqrt[n]{a_n} = l$ and $l < 1$	then the series is	convergent	(Cauchy's Root Test).
$\lim_{n\to\infty} \sqrt[n]{a_n} = l$ and $l > 1$		divergent	
$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = k$ and $k < 1$		convergent	(d'Alembert's Ratio Test).
$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = k$ and $k > 1$		divergent	
$\lim_{n\to\infty} n\left(\frac{a_n}{a_{n+1}} - 1\right) = m$ and $m > 1$		convergent	(Raabe's Test).
$\lim_{n\to\infty} n\left(\frac{a_n}{a_{n+1}} - 1\right) = m$ and $m < 1$		divergent	

REMARK 7. Some statements of Theorem 13 are valid under more general assumptions. In particular, for a given series to be convergent, it is sufficient that one of the relations

$$\overline{\lim_{n\to\infty}} \sqrt[n]{a_n} < 1, \quad \overline{\lim_{n\to\infty}} \frac{a_{n+1}}{a_n} < 1$$

(Theorem 10.1.8) holds. On the other hand, if $\sqrt[n]{a_n} \geqq 1$ or $a_{n+1}/a_n \geqq 1$ for almost every n, then the series is divergent.

REMARK 8. The tests of Cauchy and d'Alembert are inconclusive when $l = 1$ and $k = 1$, respectively. For Raabe's test we have the result: "If the relation

$$n\left(\frac{a_n}{a_{n+1}} - 1\right) \leqq 1$$

holds for almost every n, then the series diverges". Using Cauchy's test we can often conclude that a series converges, even when d'Alembert's test fails (but not vice versa). If Cauchy's test fails, it is often possible to reach a conclusion with the help of Raabe's test:

Example 5. For the series $\sum_{n=1}^{\infty} \frac{1}{n^2}$ neither d'Alembert's nor Cauchy's Test gives a conclusion. Using Raabe's test we have

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

so the series is convergent.

Example 6. According to Remark 8 we can conclude that the harmonic series $\sum_{n=1}^{\infty} \frac{1}{n}$ is divergent, because $n\left(\frac{a_n}{a_{n+1}} - 1\right) = n\left(\frac{n+1}{n} - 1\right) = 1$ for every n.

Theorem 14 (*Integral Test*). *Let $f(x)$ be a non-negative decreasing function defined in an interval $[a, \infty)$ $(a > 0)$ such that $f(n) = a_n$ for every positive integer n. Then the series $a_1 + a_2 + a_3 + \ldots$ and the integral $\int_a^{\infty} f(x)\,dx$ both converge or both diverge.*

Example 7. Let us find for which $\alpha > 0$ the series

$$1 + \frac{1}{2^\alpha} + \frac{1}{3^\alpha} + \frac{1}{4^\alpha} + \ldots \tag{8}$$

converges (for $\alpha \leqq 0$ it is evidently divergent). In Theorem 14 let us choose $f(x) = 1/x^\alpha$. The integral

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

is convergent if and only if $\alpha > 1$ (see Theorem 13.8.9, p. 529). Thus, by Theorem 14 the series (8) is convergent for $\alpha > 1$ and divergent for $\alpha \leqq 1$. (In particular, from this result it follows that the harmonic series diverges, since $\alpha = 1$ in this case.)

Theorem 15 (*Cauchy's Theorem*). *Let* $a_1 + a_2 + a_3 + \ldots$ *be a series of positive decreasing terms* $(a_n \geqq a_{n+1} \geqq \ldots > 0)$ *and* c *be a positive integer such that* $c > 1$. *Then the given series and the series*

$$ca_c + c^2 a_{c^2} + c^3 a_{c^3} + \ldots$$

both converge or both diverge.

Example 8. Let us find for which $\alpha \geqq 0$ the series

$$\frac{1}{2 \ln^\alpha 2} + \frac{1}{3 \ln^\alpha 3} + \frac{1}{4 \ln^\alpha 4} + \ldots = \sum_{n=2}^{\infty} \frac{1}{n \ln^\alpha n} \tag{9}$$

converges. (For $\alpha < 0$ the series is evidently divergent; it is sufficient to compare it with the harmonic series.) The conditions of Theorem 15 are fulfilled, so let us choose $c = 2$. By Theorem 15 our series converges if and only if the series

$$2a_2 + 4a_4 + 8a_8 + \ldots$$

with the general term

$$2^k a_{2^k} = \frac{2^k}{2^k \ln^\alpha 2^k} = \frac{1}{(k \ln 2)^\alpha} = \frac{1}{\ln^\alpha 2} \cdot \frac{1}{k^\alpha}$$

converges. According to Example 7 the series with the general term $1/k^\alpha$ converges if and only if $\alpha > 1$. Thus the series (9) converges for $\alpha > 1$ and diverges for $\alpha \leqq 1$.

REMARK 9. The criteria (tests) expressed in Theorems 13–15 are valid for series of positive (Remark 5) terms. For other series they are useful because we can prove the convergence of the series (3) on the basis of Theorem 4 by proving the convergence of the series (2) for which our tests are applicable. From the divergence of the series (2) the divergence of the series (3) does not follow. Let us remark, however, that from the relations

$$\lim \frac{|a_{n+1}|}{|a_n|} > 1 \quad \text{or} \quad \lim \sqrt[n]{|a_n|} > 1$$

follows not only the divergence of the series (2) but also the divergence of the series (3).

Theorems on differentiation and integration of power series provide further effective means for investigating the convergence (and also for finding the sum) of many series (§ 15.4).

REMARK 10. The series $\sum_{n=1}^{\infty} a_n$ of complex terms $a_n = \alpha_n + i\beta_n$ (α_n, β_n real) is convergent if and only if both series $\sum_{n=1}^{\infty} \alpha_n$, $\sum_{n=1}^{\infty} \beta_n$ converge (and then

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} \alpha_n + i \sum_{n=1}^{\infty} \beta_n).$$

To prove the convergence of a series of complex terms we can often with advantage use Theorem 4 on absolute convergence and so transform the problem to the investigation of the convergence of a series with positive terms.

Definition 4. A series

$$a_1 - a_2 + a_3 - a_4 \ldots, \quad a_n \geqq 0 \tag{10}$$

(with alternating positive and negative terms) is called an *alternating series.*

Theorem 16 (*Leibniz's Rule*). *If, for the series* (10), *the relations*

$$a_1 \geqq a_2 \geqq a_3 \geqq \ldots \quad \text{and} \quad \lim a_n = 0$$

hold, then the series (10) *is convergent. The absolute value of the difference* $s - s_n$ *between the sum* s *of this series and the partial sum* s_n *is less than or equal to the number* a_{n+1} *and this difference has the same sign as the* $(n+1)$*-th term in the series* (10).

Example 9. By Theorem 16 the alternating series

$$1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} + \ldots$$

is convergent. If we consider, for instance, the first eight terms, then $0 < s - s_8 < \frac{1}{9}$.

Theorem 17 (*Dirichlet's Test*). *Let two sequences* $\{a_n\}$, $\{b_n\}$ *be given, where* $\{b_n\}$ *is a monotonic sequence* (Definition 10.1.8) *such that* $\lim_{n\to\infty} b_n = 0$. *If* $\left|\sum_{k=1}^{n} a_k\right| \leqq K$ *for every* n (K *constant*), *then the series* $\sum_{k=1}^{\infty} a_k b_k$ *is convergent.*

Theorem 18 (*Abel's Test*). *Let two sequences* $\{a_n\}$, $\{b_n\}$ *be given, where* $\{b_n\}$ *is bounded and monotonic. If the series* $\sum_{k=1}^{\infty} a_k$ *converges, then the series* $\sum_{k=1}^{\infty} a_k b_k$ *is also convergent.*

Example 10. Let $b_1 \geqq b_2 \geqq b_3 \geqq \ldots \geqq 0$ and $\lim_{n\to\infty} b_n = 0$. Then the series

$$b_1 \sin x + b_2 \sin 2x + b_3 \sin 3x + \ldots \tag{11}$$

is convergent for every x: If $x = 2k\pi$ (k integral), the sum of the series is zero. For $x \neq 2k\pi$

$$a_1 + a_2 + \ldots + a_n = \sin x + \sin 2x + \ldots + \sin nx = \frac{\sin \frac{1}{2}nx}{\sin \frac{1}{2}x} \sin \frac{n+1}{2} x$$

(see Theorem 25), so

$$\left|\sum_{k=1}^{n} a_k\right| \leqq \left|\frac{1}{\sin \frac{1}{2}x}\right|$$

for every n, and the convergence of the series (11) for every fixed $x \neq 2k\pi$ follows from Theorem 17.

Definition 5. Let us consider the "two-dimensional" array of (real or complex) numbers

$$\begin{array}{l} a_{11}, a_{12}, a_{13}, a_{14}, \dots, \\ a_{21}, a_{22}, a_{23}, a_{24}, \dots, \\ a_{31}, a_{32}, a_{33}, a_{34}, \dots, \\ \dots\dots\dots\dots\dots \end{array} \tag{12}$$

The expression

$$\lim_{\substack{m\to\infty \\ n\to\infty}} a_{mn} = a$$

means: "For an arbitrarily chosen $\varepsilon > 0$ there exists a number K such that the relation

$$|a_{mn} - a| < \varepsilon$$

holds whenever both $m > K$ and $n > K$."

Definition 6. Let us write

$$s_{mn} = \sum_{i=1}^{m} \sum_{k=1}^{n} a_{ik}.$$

If the limit

$$\lim_{\substack{m\to\infty \\ n\to\infty}} s_{mn} = s$$

exists, then we write

$$\sum_{i,k=1}^{\infty} a_{ik} = s \tag{13}$$

and say that the so-called *double series* $\sum_{i,k=1}^{\infty} a_{ik}$ formed from the array (12), is *convergent* and *has the sum s.*

REMARK 11. Let us form from the array (12) an arbitrary series such that it includes every term from the array (12) exactly once. In particular, let us form the series

$$a_{11} + a_{12} + a_{21} + a_{13} + a_{22} + a_{31} + \dots \tag{14}$$

(we note down successively from the array (12) the terms that have the same sum of indices). If any one of these series converges *absolutely*, in particular the series (14), then all such series converge absolutely and possess (Theorem 5) the same sum s. It is then possible to sum the terms of the array (12) in any order. In particular it

follows that $\sum_{i,k=1}^{\infty} a_{ik} = s$ [see (13)]. We speak of the *absolute convergence of the double series* (13). Theorem 19 provides a simple criterion for absolute convergence.

REMARK 12. We meet double series in particular when dealing with Fourier series in two-dimensional domains. For example, the equation

$$\sum_{m,n=1}^{\infty} A_{mn} \sin mx \sin ny = f(x, y) \tag{15}$$

(A_{mn} are the Fourier coefficients of the function $f(x, y)$) means, according to Definition 5: For every arbitrarily small $\varepsilon > 0$ there exists a number K such, that for every pair of numbers M, N, for which both the relations $M > K$, $N > K$ hold, the relation

$$\left| \sum_{m=1}^{M} \sum_{n=1}^{N} A_{mn} \sin mx \sin ny - f(x, y) \right| < \varepsilon \tag{16}$$

holds. If we know (from the theorems on Fourier series and from the properties of the function $f(x, y)$) that the series (15) converges absolutely to $f(x, y)$ at the point (x, y), we can sum the terms of that series in an arbitrary order.

Theorem 19. *Let the series* $\sum_{k=1}^{\infty} |a_{ik}|$ *be convergent for every* i. *Let us write* $\sum_{k=1}^{\infty} |a_{ik}| = \sigma_i$. *If the series* $\sum_{i=1}^{\infty} \sigma_i$ *converges, then the double series* (13) *is absolutely convergent.*

Theorem 20. *Let the double series* (13) *be absolutely convergent. Then the series* $\sum_{k=1}^{\infty} a_{ik} = s_i$ *are absolutely convergent for every* i. *The series* $\sum_{i=1}^{\infty} s_i$ *is also absolutely convergent and its sum* s *is equal to the sum of the double series* (13).

REMARK 13. The summing in theorems 19 and 20 is done first by rows, then by columns. The theorems, however, retain their validity if the summing is done first by columns and then by rows.

REMARK 14. The fact that an absolutely convergent double series can be summed in an arbitrary order, may be used to *improve the rapidity of convergence* of simple series. In particular, the series

$$s = \frac{x}{1 - x} + \frac{x^2}{1 - x^2} + \frac{x^3}{1 - x^3} + \dots \tag{17}$$

is absolutely convergent for every x for which $|x| < 1$, as follows, for example, from d'Alembert's Test (Theorem 13). But this series converges slowly for $|x|$ in the vici-

nity of 1. By means of the so-called *Clausen's transformation* this series changes into the series

$$s = x\frac{1+x}{1-x} + x^4\frac{1+x^2}{1-x^2} + x^9\frac{1+x^3}{1-x^3} + x^{16}\frac{1+x^4}{1-x^4} + \ldots, \tag{18}$$

which converges much faster. Namely, the i-th term of the series (17) is the sum of the series formed by members of the i-th row of the array

$$\begin{array}{l} x,\ x^2, x^3, x^4,\ \ldots, \\ x^2, x^4, x^6, x^8,\ \ldots, \\ x^3, x^6, x^9, x^{12}, \ldots, \\ \ldots\ldots\ldots\ldots \end{array} \tag{19}$$

Each of these series converges absolutely for $|x| < 1$. The series corresponding to the series $\sum_{i=1}^{\infty} \sigma_i$ from Theorem 19 is here the series

$$\frac{|x|}{1-|x|} + \frac{|x|^2}{1-|x|^2} + \frac{|x|^3}{1-|x|^3} + \ldots$$

which is convergent (e.g. according to d'Alembert's Test) for $|x| < 1$. Thus, the double series corresponding to the array (19) is absolutely convergent by Theorem 19. We form the sum in the following way:

$$s = U_1 + U_2 + U_3 + \ldots$$

where U_1 is the sum of the elements from the first row and the first column in (19), U_2 is the sum of the remaining elements of the second row and the second column and so on. It follows that

$$U_1 = x + 2x^2 + 2x^3 + 2x^4 + \ldots = x + 2x^2\frac{1}{1-x} = x\frac{1+x}{1-x},$$

$$U_2 = x^4 + 2x^6 + 2x^8 + 2x^{10} + \ldots = x^4 + 2x^6\frac{1}{1-x^2} = x^4\frac{1+x^2}{1-x^2}.$$

In this way, we proceed to the series (18).

REMARK 15. By analogy with the definition of double series we can define *triple series*, etc.

Theorem 21 (*Multiplication or Product of Series*). *Let the series* $\sum_{n=1}^{\infty} a_n$ *and* $\sum_{n=1}^{\infty} b_n$ *be absolutely convergent and have the sums* s *and* t, *respectively. Then the double series*

$$\sum_{i,k=1}^{\infty} a_i b_k$$

is absolutely convergent and has the sum s . t (i.e. we can multiply the series as we would multiply polynomials and then sum in an arbitrary order).

Theorem 22 (*Cauchy Product of Series*). *Let* $\sum_{n=1}^{\infty} a_n$, $\sum_{n=1}^{\infty} b_n$ *be convergent series (with sums s, t, respectively), one of which at least is absolutely convergent. Then the series*

$$\sum_{n=1}^{\infty} U_n, \text{ where } U_n = a_1 b_n + a_2 b_{n-1} + \ldots + a_{n-1} b_2 + a_n b_1$$

is convergent and has the sum s . t.

Theorem 23. *Let us write in the customary way* $s_n = a_1 + a_2 + \ldots + a_n$. *Let* $\sum_{n=1}^{\infty} a_n$ *be convergent, that is* $\lim_{n\to\infty} s_n = s$. *Then also*

$$\lim_{n\to\infty} S_n = \lim_{n\to\infty} \frac{s_1 + s_2 + \ldots + s_n}{n} = s .$$

REMARK 16. The converse of Theorem 23 does not hold. We say that a series, for which the sequence S_n converges, is *summable by arithmetic means of the first order* [or *Cesàro summable*; we write: *summable* (C, 1)]. This method assigns a "sum" to certain divergent series. For some definitions of the summability and particularly for the applicability of divergent series to asymptotic expansions (representations) see § 15.7. We shall mention one application here:

Theorem 24. *If* $a_1 + a_2 + a_3 + \ldots$ *and* $b_1 + b_2 + b_3 + \ldots$ *are two convergent series with respective sums s and t then the series*

$$a_1 b_1 + (a_1 b_2 + a_2 b_1) + (a_1 b_3 + a_2 b_2 + a_3 b_1) + \ldots$$

is summable by arithmetic means (Remark 16) *to the sum s . t (i.e.* $\lim_{n\to\infty} S_n = s . t$).

Theorem 25 (*Survey of Important Formulae*).

1. $\sum_{n=1}^{\infty} (a_n \pm b_n) = \sum_{n=1}^{\infty} a_n \pm \sum_{n=1}^{\infty} b_n$, $\sum_{n=1}^{\infty} k a_n = k \sum_{n=1}^{\infty} a_n$ (k a constant) (Theorem 3).

2. $$a_n > 0, \quad \lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \left\{ \begin{array}{l} k < 1 \\ \\ k > 1 \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} \sum_{n=1}^{\infty} a_n \text{ is convergent} \\ \\ \sum_{n=1}^{\infty} a_n \text{ is divergent} \end{array} \right\} \text{(Theorem 13);}$$

$$a_n > 0, \quad \lim_{n\to\infty} \sqrt{a_n} = \left\{ \begin{array}{l} l < 1 \\ \\ l > 1 \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} \sum_{n=1}^{\infty} a_n \text{ is convergent} \\ \\ \sum_{n=1}^{\infty} a_n \text{ is divergent} \end{array} \right\} \text{(Theorem 13);}$$

$$a_n > 0\,,\quad \lim_{n\to\infty} n\left(\frac{a_n}{a_{n+1}} - 1\right) = \left\{\begin{array}{l} m > 1 \\ \\ m < 1 \end{array}\right\} \Rightarrow \left\{\begin{array}{l} \sum_{n=1}^{\infty} a_n \text{ is convergent} \\ \\ \sum_{n=1}^{\infty} a_n \text{ is divergent} \end{array}\right\} \text{(Theorem 13).}$$

3. *The arithmetic series* $a_1 + a_2 + a_3 + \ldots$, *where* $a_n = a_1 + (n-1)\,d$, *is convergent if and only if* $a_1 = 0$ *and* $d = 0$. *The partial sum of the first n terms is*

$$a_1 + a_2 + \ldots + a_n = \tfrac{1}{2}n(a_1 + a_n) = \tfrac{1}{2}n[2a_1 + (n-1)\,d]\,.$$

4. *The geometric series*

$$a + aq + aq^2 + \ldots \quad (a \neq 0)$$

converges if and only if $|q| < 1$. *The sum of the series is*

$$s = \frac{a}{1-q}\,.$$

The partial sum is:

$$s_n = a\,\frac{q^n - 1}{q - 1} \quad (q \neq 1)\,.$$

5. $1 + 2 + 3 + \ldots + n = \tfrac{1}{2}n(n+1)$;

$1^2 + 2^2 + 3^2 + \ldots + n^2 = \tfrac{1}{6}n(n+1)(2n+1)$;

$1^3 + 2^3 + 3^3 + \ldots + n^3 = \tfrac{1}{4}n^2(n+1)^2$;

$1^4 + 2^4 + 3^4 + \ldots + n^4 = \tfrac{1}{30}n(n+1)(2n+1)(3n^2 + 3n - 1)$.

6. $\dfrac{1}{1^2} + \dfrac{1}{2^2} + \dfrac{1}{3^2} + \ldots = \dfrac{\pi^2}{6}$;

$\dfrac{1}{1^4} + \dfrac{1}{2^4} + \dfrac{1}{3^4} + \ldots = \dfrac{\pi^4}{90}$;

$\dfrac{1}{1^6} + \dfrac{1}{2^6} + \dfrac{1}{3^6} + \ldots = \dfrac{\pi^6}{945}$;

$\dfrac{1}{1^2} + \dfrac{1}{3^2} + \dfrac{1}{5^2} + \ldots = \dfrac{\pi^2}{8}$.

7. $$\left.\begin{array}{l} \sin x + \sin 2x + \ldots + \sin nx = \dfrac{\sin \frac{1}{2}nx}{\sin \frac{1}{2}x} \sin \dfrac{1}{2}(n+1)\,x\,, \\ \\ \cos x + \cos 2x + \ldots + \cos nx = \dfrac{\sin \frac{1}{2}nx}{\sin \frac{1}{2}x} \cos \dfrac{1}{2}(n+1)\,x\,, \end{array}\right\} \begin{array}{l} x \neq 2k\pi\,, \\ k \text{ an integer.} \end{array}$$

8. *Let* $\frac{1}{p} + \frac{1}{q} = 1$. *The following results then hold:*

(a) *If* $a_n \geqq 0$, $b_n \geqq 0$, $p > 1$, *then*

$$\sum_{n=1}^{\infty} a_n b_n \leqq \left(\sum_{n=1}^{\infty} a_n^p\right)^{1/p} \cdot \left(\sum_{n=1}^{\infty} b_n^q\right)^{1/q} \quad \textit{(Hölder's Inequality)}$$

(b) *If* $a_n \geqq 0$, $b_n \geqq 0$, $p \geqq 1$, *then*

$$\left[\sum_{n=1}^{\infty} (a_n + b_n)^p\right]^{1/p} \leqq \left(\sum_{n=1}^{\infty} a_n^p\right)^{1/p} + \left(\sum_{n=1}^{\infty} b_n^p\right)^{1/p} \quad \textit{(Minkowski's Inequality)}$$

(c) *From* (a) *for* $p = 2$ $(a_n \geqq 0, b_n \geqq 0)$ *it follows that*

$$\left(\sum_{n=1}^{\infty} a_n b_n\right)^2 \leqq \sum_{n=1}^{\infty} a_n^2 \cdot \sum_{n=1}^{\infty} b_n^2 \quad \textit{(Schwarz or Schwarz-Cauchy Inequality)}$$

(so, if the series $\sum_{n=1}^{\infty} a_n^2$, $\sum_{n=1}^{\infty} b_n^2$ *converge, then the series* $\sum_{n=1}^{\infty} a_n b_n$ *also converges).*

Theorem 26 (*The Sums of Some Series*).

1. $1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \ldots + \frac{1}{n!} + \ldots = \mathrm{e}$.

2. $1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \ldots + (-1)^n \frac{1}{n!} + \ldots = \frac{1}{\mathrm{e}}$.

3. $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots + (-1)^{n+1} \frac{1}{n} + \ldots = \ln 2$.

4. $1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots + \frac{1}{2^n} + \ldots = 2$.

5. $1 - \frac{1}{2} + \frac{1}{4} - \frac{1}{8} + \ldots + (-1)^n \frac{1}{2^n} + \ldots = \frac{2}{3}$.

6. $1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots + (-1)^{n+1} \frac{1}{2n-1} + \ldots = \frac{\pi}{4}$.

7. $\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \ldots + \frac{1}{n(n+1)} + \ldots = 1$.

8. $\frac{1}{1 \cdot 3} + \frac{1}{3 \cdot 5} + \frac{1}{5 \cdot 7} + \ldots + \frac{1}{(2n-1)(2n+1)} + \ldots = \frac{1}{2}$.

9. $\frac{1}{1 \cdot 3} + \frac{1}{2 \cdot 4} + \frac{1}{3 \cdot 5} + \ldots + \frac{1}{n(n+2)} + \ldots = \frac{3}{4}$.

10. $\frac{1}{3 \cdot 5} + \frac{1}{7 \cdot 9} + \frac{1}{11 \cdot 13} + \ldots + \frac{1}{(4n-1)(4n+1)} + \ldots = \frac{1}{2} - \frac{\pi}{8}$.

11. $\frac{1}{1 \cdot 2 \cdot 3} + \frac{1}{2 \cdot 3 \cdot 4} + \ldots + \frac{1}{n(n+1)(n+2)} + \ldots = \frac{1}{4}$.

12. $\frac{1}{1 \cdot 2 \ldots . l} + \frac{1}{2 \cdot 3 \ldots . (l+1)} + \ldots + \frac{1}{n \ldots . (n+l-1)} + \ldots =$

$= \frac{1}{(l-1)(l-1)!}$.

REMARK 17. For series with variable terms see Chap. 15 where the application of power series to the numerical summation of series is dealt with. For the summation of infinite series by means of integral transformations see the article "The Summation of Infinite Series by means of Integral Transformations" by D. MAYER and J. NEČAS, *Aplikace matematiky* 1 (1956), No. 3, pp. 165–185. (In Czech, English summary.)

10.3. Infinite Products

Definition 1. Suppose we are given the sequence of (real or complex) numbers $p_1, p_2, p_3, \ldots$. Let us define

$$P_n = p_1 \cdot p_2 \cdot p_3 \cdot \ldots \cdot p_n . \tag{1}$$

The symbol

$$\prod_{n=1}^{\infty} p_n = p_1 \cdot p_2 \cdot p_3 \cdot \ldots \tag{2}$$

is called an *infinite product*. If $\lim_{n \to \infty} P_n$ exists, then this limit is called the *value of the infinite product* (2).

Definition 2. We say that the product (2) is *convergent* if, either (i) the limit $\lim_{n \to \infty} P_n$, finite and *different from zero*, exists, or (ii) in the product (2) there is only a finite number of factors equal to zero and their omission leads again to a finite limit *different from zero*. (In the latter case the infinite product has the value 0 according to Definition 1.)

In every other case we shall say that (2) is *divergent*.

Example 1. The product

$$\tfrac{1}{1} \cdot \tfrac{1}{2} \cdot \tfrac{1}{3} \cdot \dots$$

is divergent because

$$P_n = \frac{1}{n!} \quad \text{and} \quad \lim_{n \to \infty} P_n = 0 .$$

REMARK 1. The reader's attention is drawn to the fact that there is no uniformity in mathematical literature regarding the definition of the convergence of an infinite product.

REMARK 2. In applications, the investigation of infinite products frequently relates to cases where the factors p_n are of the form $1 + a_n$.

Theorem 1. *Let the product*

$$(1 + |a_1|)(1 + |a_2|)(1 + |a_3|) \dots \tag{3}$$

be convergent. Then the product

$$(1 + a_1)(1 + a_2)(1 + a_3) \dots \tag{4}$$

is also convergent.

Definition 3. If the product (3) converges then the product (4) is said to be *absolutely convergent.*

Theorem 2. *If the series $a_1 + a_2 + a_3 + \dots$ is absolutely convergent* (see Definition 10.2.3) *then the product* (4) *is absolutely convergent (and conversely) and its value does not depend on the ordering of its factors. (We say in this case that its factors may be rearranged.)*

REMARK 3. If the series $a_1 + a_2 + a_3 + \dots$ is only conditionally convergent (Definition 10.2.3) then the product (4) need not be convergent.

Theorem 3. *For every x (real or complex) the relations*

$$\sin x = x\left(1 - \frac{x^2}{\pi^2 \cdot 1^2}\right)\left(1 - \frac{x^2}{\pi^2 \cdot 2^2}\right)\left(1 - \frac{x^2}{\pi^2 \cdot 3^2}\right) \dots = x \prod_{n=1}^{\infty} \left(1 - \frac{x^2}{n^2\pi^2}\right),$$

$$\cos x = \left(1 - \frac{2^2x^2}{\pi^2 \cdot 1^2}\right)\left(1 - \frac{2^2x^2}{\pi^2 \cdot 3^2}\right)\left(1 - \frac{2^2x^2}{\pi^2 \cdot 5^2}\right) \dots = \prod_{n=1}^{\infty} \left(1 - \frac{2^2x^2}{\pi^2(2n-1)^2}\right)$$

hold. Also

$$\frac{\pi}{2} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \dots \quad (\textit{Wallis's Product}).$$

REMARK 4. For the expression of the Γ function as an infinite product see § 13.11.

11. DIFFERENTIAL CALCULUS OF FUNCTIONS OF A REAL VARIABLE

By KAREL REKTORYS

References: [1], [4], [15], [17], [25], [26], [30], [31], [40], [52], [54], [59], [60], [68], [74], [91], [96], [109], [111], [119], [121], [122], [123], [139], [142], [145], [148], [158], [160], [175].

11.1. The Concept of a Function. Composite Functions. Inverse Functions

Notation: x is a real number. Instead of "the number x" we often say "the point x".

The *closed interval* $[a, b]$ is the set of all x, for which $a \leqq x \leqq b$;
the *open interval* (a, b) is the set of all x, for which $a < x < b$;
the *semi-closed (semi-open) interval* $[a, b)$ is the set of all x, for which $a \leqq x < b$;
the *semi-closed (semi-open) interval* $(a, b]$ is the set of all x, for which $a < x \leqq b$.

The interval $(a, +\infty)$	(briefly (a, ∞))	is the set of all x, for which $x > a$;
the interval $[a, +\infty)$	(briefly $[a, \infty)$)	is the set of all x, for which $x \geqq a$;
the interval $(-\infty, a)$		is the set of all x, for which $x < a$;
the interval $(-\infty, a]$		is the set of all x, for which $x \leqq a$;
the interval $(-\infty, +\infty)$	(briefly $(-\infty, \infty)$)	is the set of all real numbers x.

We shall write the interval I in speaking of an interval without special reference to its end-points. The notation $x \in M$ means: x is an element belonging to the set M. For example, $x \in [a, b]$ means that x is in the interval $[a, b]$.

Definition 1. We say that a *real function* is defined on a set M of real numbers, if a rule (relation) is given by virtue of which to each number $x \in M$ there corresponds *exactly one* real number y. The number x is called the *independent variable (argument)*, y is called the *dependent variable*. The set M is called the *domain of definition of the function*.

A function is generally denoted by the letters $f, g, \ldots$. The value y of the function corresponding to an arbitrary point $x \in M$ is denoted by $f(x)$, $g(x)$, etc. Instead of "the function f" we often say "the function $f(x)$" or "the function $y = f(x)$".

Example 1. The area y of a square is a function of the length x of its side, $y = x^2$. The domain of definition is the interval $(0, +\infty)$, since the length of the side of the square is always expressed by a positive number.

REMARK 1. The domain of definition of a function is most often an interval, e.g. the interval $[-1, 1]$ for the function $y = \arcsin x$, etc. But every "reasonable" function need not have an interval as its domain of definition. For example, the function $y = \tan x$ is defined in the interval $(-\infty, +\infty)$ from which the points $\pm\frac{1}{2}\pi$, $\pm\frac{3}{2}\pi$, $\pm\frac{5}{2}\pi$, ... are excluded.

REMARK 2. The relationship defining the function need not be given by an equation (i.e. by an analytic formula, from which the value of the dependent variable y can be calculated for a given value of the independent variable x) as was the case in Example 1. Frequently, in applications, the correspondence between the independent variable x and the dependent variable y is established by a graph, expecially in cases where the values of the independent and dependent variables can be read off the graph with adequate accuracy. When we perform different types of measurements we compile a table of measured values. From this table we often try to obtain values of a function for the whole domain of definition (e.g., by interpolation). The function may be given also as the limit of a sequence of functions, as the sum of a series of functions, etc. Often a relationship between x and y is given (most frequently by an equation) from which it is necessary to determine the single-valued correspondence between the dependent variable y and the independent variable x (Fig. 11.1). (E.g. $x^2 + y^2 = 25$; in the neighbourhood of the point $(3, 4)$ the function $y = f(x)$ will be given by the relation $y = \sqrt{(25 - x^2)}$. There we say that the *explicit* function $y = \sqrt{(25 - x^2)}$ is given in a neighbourhood of the point $(3, 4)$ by the *implicit*

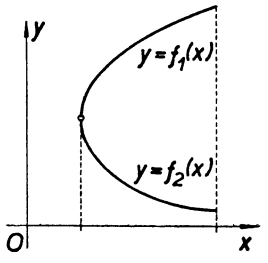


Fig. 11.1. The curve in this figure is the graph of two functions, $y = f_1(x)$ and $y = f_2(x)$; it is not the graphical representation of a single-valued function.

equation $x^2 + y^2 = 25$.) See § 12.9 for more detailed treatment. *Until further notice the concept of a function means the (single-valued) function as defined in Definition* 1.

REMARK 3. If to each real number x from a certain domain M there corresponds a complex number $y = y_1 + iy_2$ we say that $y = f(x) = f_1(x) + if_2(x)$ is a *complex function of the real variable* x. The study of these functions is reduced to that of

the real functions $f_1(x)$ and $f_2(x)$ (e.g. the derivative is defined by the relation $f'(x) = f_1'(x) + if_2'(x)$, etc.) so that in the following we shall deal only with *real* functions of a real variable. (For functions where the independent variable is complex see Chap. 20.)

REMARK 4. If the functional relationship is given by an analytical formula then we are interested in those x for which the formula has a sense (in the domain of real numbers). The set of those x is then accepted as the domain of definition of the given function. For example, the domain of definition of the function given by the formula $y = \sqrt{(4 - x^2)}$ is taken to be the interval $[-2, 2]$ (for $|x| > 2$, $\sqrt{(4 - x^2)}$ is no longer a real number).

Definition 2. By the *graph of the function* $y = f(x)$ is understood the set of all points (x, y) in the plane xy (with a cartesian coordinate system $(O; x, y)$) such that $x \in M$, $y = f(x)$. Instead of "the graph of the function $y = f(x)$" we often say "the curve $y = f(x)$". The coordinate x is called the *first coordinate* (the *abscissa*, *x-coordinate*), the coordinate y is called the *second coordinate* (the *ordinate*, *y-coordinate*).

Example 2. Graphs of the trigonometric functions are given on p. 72.

Definition 3. Let the function $y = f(x)$ be defined in the interval M_1. We say that the function $y = f(x)$ maps the interval M_1 *into* (*in*) the interval M_2 if for every $x \in M_1$ it follows that $y \in M_2$.

If in addition it is possible to find, for every $y \in M_2$, at least one $x \in M_1$ such that $y = f(x)$ then we say that the function $y = f(x)$ maps the interval M_1 *onto* (*on*) the interval M_2.

Example 3. The interval $[0, 7\pi]$ is mapped by the function $y = \sin x$ onto the interval $[-1, 1]$, because for every $x \in [0, 7\pi]$, $y = f(x) \in [-1, 1]$ and to any number $y \in [-1, 1]$ we can find even several $x \in [0, 7\pi]$ such that $y = \sin x$. We can say as well that the interval $[0, 7\pi]$ is mapped *into* any interval that includes the interval $[-1, 1]$, e.g. into the interval $[-5, 10]$.

Definition 4 (*of a Composite Function*). Let the interval M_1 be mapped by the function $z = f(x)$ into the interval M_2. Let $y = g(z)$ be a function defined in the domain M_2. The function $y = g(f(x))$ is said to be a *composite function* of the functions $z = f(x)$ and $y = g(z)$.

REMARK 5. Thus, $y = g(f(x))$ has the following meaning (Fig. 11.2): When we choose any number $x \in M_1$ then by means of the relation $z = f(x)$ we can evaluate z. To this number z we can then by means of the relation $y = g(z)$ evaluate y. Thus y is determined uniquely by the choice of the number $x \in M_1$, so that finally $y = g(f(x))$ is a function of the variable x only. (On how to use composite functions for finding derivatives see Theorem 11.5.5, p. 382.)

Example 4. The function $y = \sqrt{(1 - \frac{1}{2}\sin^2 x)}$ may be "decomposed" into the functions $y = \sqrt{z}$, $z = 1 - \frac{1}{2}\sin^2 x$. As the interval M_1 we can take the interval $(-\infty, +\infty)$ because the function $z = 1 - \frac{1}{2}\sin^2 x$ maps the interval $(-\infty, +\infty)$ onto the interval $[\frac{1}{2}, 1]$ and consequently also into the interval $[0, +\infty)$, which is the domain of definition of the function $y = \sqrt{z}$.

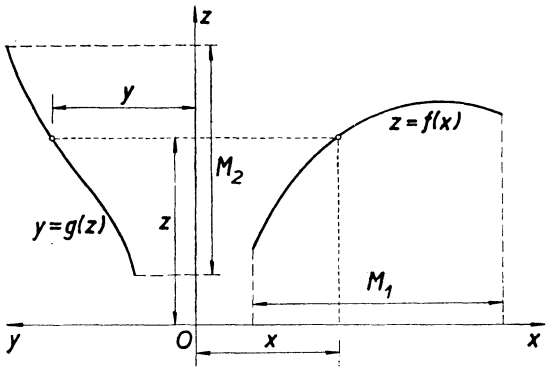


Fig. 11.2.

REMARK 6. In general, we define composite functions (and also mappings) for sets other than intervals. It is then necessary to replace in definitions 3 and 4 the word "interval" by the word "set".

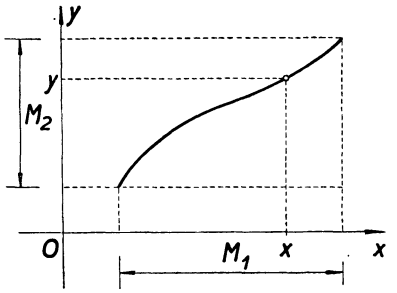


Fig. 11.3.

Definition 5. Let the interval M_1 be mapped by the function $y = f(x)$ onto the interval M_2 (see definition 3) with a *one-to-one correspondence* (Fig. 11.3), which means that not only to every $x \in M_1$ there corresponds exactly one $y \in M_2$, but also to every $y \in M_2$ there corresponds exactly one $x \in M_1$ such that $y = f(x)$. Because to every $y \in M_2$ there corresponds just one $x \in M_1$, a function is defined on the interval M_2 which we denote by $x = \varphi(y)$. This function is called the *inverse function* of the function $y = f(x)$. Conversely, the function $y = f(x)$ is the inverse function of the function $x = \varphi(y)$.

REMARK 7. The one-to-one correspondence is ensured, for example, if $y = f(x)$ is strictly increasing or decreasing in M_1 (or if the function $x = \varphi(y)$ is strictly increasing or decreasing in M_2) (see Fig. 11.3). This case is very common.

REMARK 8. It is possible to add a remark to the definition of inverse function similar to Remark 6.

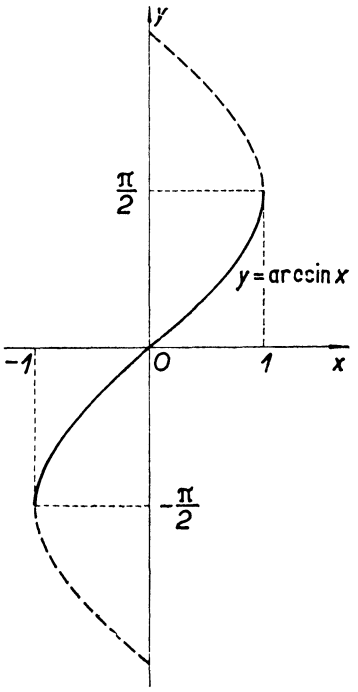


Fig. 11.4.

Example 5. The function $x = \sin y$ is a strictly increasing function of the variable y in the interval $[-\frac{1}{2}\pi, \frac{1}{2}\pi]$. The corresponding interval of the variable x is $[-1, 1]$ (see Fig. 11.4). The inverse function of the function $x = \sin y$ is called *arcussinus x* and is denoted by $y = \arcsin x$ (or $y = \sin^{-1} x$). Thus this function is defined in the interval $[-1, 1]$ (of the independent variable x). (For details see § 2.11.)

Example 6. The interval $(-\infty, +\infty)$ of the independent variable x is mapped by the function $y = x^2$ onto the interval $[0, +\infty)$ of the dependent variable y, but the correspondence between x and y is not one-to-one, because, for example, $f(2) = f(-2) = 4$. But the interval $[0, +\infty)$ of the variable x is mapped by the given function onto the interval $[0, +\infty)$ in a one-to-one correspondence. The corresponding inverse function is then $x = +\sqrt{y}$ as we easily derive from the equation $y = x^2$, and is defined in the interval $0 \leqq y < +\infty$.

REMARK 9. We draw attention to the fact that, according to Definition 5, the function $y = \arcsin x$ $(-1 \leqq x \leqq 1)$ is the inverse of the function $x = \sin y$

$(-\frac{1}{2}\pi \leqq y \leqq \frac{1}{2}\pi)$ (Example 5). Both equations have the same meaning and both functions have the same *graphical representation* (in the chosen coordinate system xy). When we interchange the variables x and y in one of them (e.g. in Example 5 we write $y = \sin x$ instead of $x = \sin y$) then the graphs of the functions will be symmetrical with respect to the straight line $y = x$ (i.e. the straight line bisecting the angle between the positive x-axis and the positive y-axis). (See Fig. 11.5.)

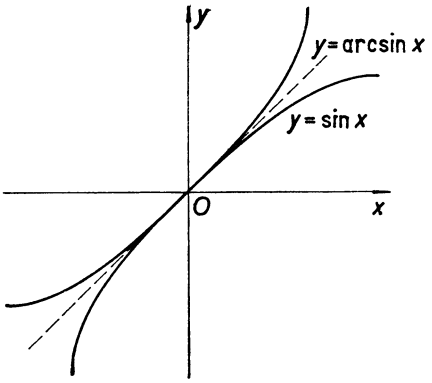


Fig. 11.5.

11.2. Elementary Functions. Algebraic Functions, Transcendental Functions. Even and Odd Functions. Bounded Functions

A function $y = f(x)$ is called *algebraic* (in a domain M) if it satisfies identically an equation

$$F(x, y) = 0 \tag{1}$$

where $F(x, y)$ is a polynomial in the variables x, y. For example, the function $y = \sqrt{(1 - x^2)}$, $x \in [-1, 1]$, is algebraic since it satisfies in the interval mentioned the equation $1 - x^2 - y^2 \equiv 0$ and $1 - x^2 - y^2$ is a polynomial in the variables x, y. Functions which are not algebraic are called *transcendental functions.*

REMARK 1. Algebraic functions include, first, *polynomials* (or *rational integral functions*) and *fractional rational functions* (or briefly *rational functions*), i.e. functions of the form

$$y = \frac{a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0}{b_m x^m + b_{m-1}x^{m-1} + \ldots + b_0},$$

where m, n are non-negative integers. Additional examples are the functions

$y = \sqrt{x}$ $(x \geqq 0)$, $y = \sqrt[3]{(1-x^2)}$ $(-1 \leqq x \leqq 1)$, $y = \sqrt{(1+x^2)}$ $(-\infty < x < +\infty)$,

etc. (For more details on polynomials see § 1.14.)

Transcendental functions include the general power $y = x^n$ ($x > 0$, n irrational), trigonometric, hyperbolic and exponential functions and their inverse functions. All these functions are called *elementary transcendental functions*. Further transcendental functions are defined by means of differential equations and integrals (so-called *higher transcendental functions*; e.g. $g(x) = \int_0^x e^{-t^2}\,dt$).

REMARK 2. For the trigonometric and hyperbolic functions and their inverse functions in more detail, see Chap. 2.

In Definition 1.9.2, p. 51, the meaning of the symbol a^b ($a > 0$) for b irrational is explained. If b is a constant and a a variable, we get the *general power* with standard notation: $y = x^n$ ($x > 0$, n any real number). The function $y = x^n$ is continuous and strictly increasing or strictly decreasing or constant in the interval $(0, \infty)$ according as $n > 0$, or $n < 0$, or $n = 0$, respectively. For certain n it is possible to extend the domain of definition of the function $y = x^n$ to values of x other than $x > 0$. For example, the relation $y = x^2$ has sense for all x (the domain of definition is $(-\infty, +\infty)$).

If the exponent b in the expression a^b changes its value and the base a remains constant we get the *exponential function* $y = a^x$ ($a > 0$); its domain of definition is $(-\infty, +\infty)$. For the graph of the function $y = a^x$ see Fig. 11.6. The function is

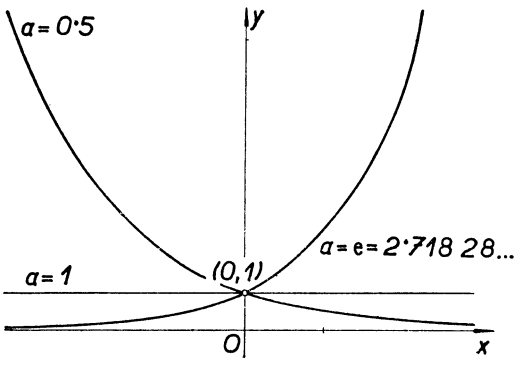


Fig. 11.6.

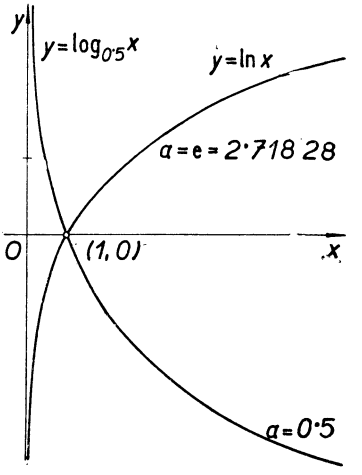


Fig. 11.7.

positive for all x, it is increasing for $a > 1$, constant for $a = 1$ ($y \equiv 1$) and decreasing for $0 < a < 1$. For $a = e = 2{\cdot}718, 281, 828, 459, 0\ldots$ (Theorem 10.1.11) we obtain the important function $y = e^x$ (frequently used notation: $y = \exp(x)$).

If $a > 0$, $a \neq 1$, the function $x = a^y$ maps the interval $-\infty < y < +\infty$ onto the interval $0 < x < +\infty$ in a one-to-one correspondence and so we can define (see Definition 11.1.5) the inverse function of the exponential function, called *logarithm of x to the base a*. Notation: $y = \log_a x$. For the graphical representation see Fig. 11.7.

The function is defined for $x > 0$ and is strictly increasing for $a > 1$ and strictly decreasing for $0 < a < 1$. For $a = e$ we write $y = \ln x$ (in the literature the notation $\lg x$, $\log x$ is also used) and this function is called the *natural logarithm of x*. This function is the inverse function of the function $x = e^y$.

REMARK 3. On the differentiation of the elementary functions see § 11.5. All current rules known from elementary mathematics, e.g. $a^{x_1} \cdot a^{x_2} = a^{x_1+x_2}$, $\log_a (x_1 \cdot x_2) = \log_a x_1 + \log_a x_2$ $(x_1 > 0, x_2 > 0)$, etc., are valid for general powers and for the exponential and logarithmic functions (for details see §§ 1.9 and 1.10). Further we have

$$\log_a x = \log_b x \cdot \log_a b ,$$

in particular

$$\log_{10} x = M \log_e x = M \ln x \approx 0{\cdot}434\,294 \ln x ,$$

$$\log_e x = \ln x = \frac{1}{M} \log_{10} x \approx 2{\cdot}302\,585 \log_{10} x .$$

The number M is called the *conversion modulus from natural to common logarithms*, the number $1/M$ is the *conversion modulus from common to natural logarithms*.

Definition 1. We call a function $y = f(x)$ an *even function* if $f(-x) = f(x)$, an *odd function* if $f(-x) = -f(x)$ for every x from the domain of definition of the function $f(x)$.

Example 1. The functions $y = x^2$, $y = x^4$, $y = \cos x$ are even in the interval $(-\infty, +\infty)$ (because $(-x)^2 = x^2$, $(-x)^4 = x^4$, $\cos(-x) = \cos x$). The functions $y = x$, $y = x^3$, $y = \sin x$ are odd in the interval $(-\infty, +\infty)$, because, for example, $\sin(-x) = -\sin x$ for every x.

Definition 2. The function $f(x)$ is called *bounded above (below) in the domain M*, if there exists a constant K (k) such that for all $x \in M$, $f(x) < K$ $(f(x) > k)$. If the function $f(x)$ is both bounded above and bounded below in the domain M, we say simply that $f(x)$ is *bounded in the domain M*.

11.3. Continuity. Types of Discontinuity. Functions of Bounded Variation

Definition 1. By a *δ-neighbourhood of a point a* we mean a set of all points x such that their distance from the point a is smaller than δ (or: such that they lie in the interval $(a - \delta, a + \delta)$ or for which $|x - a| < \delta$; we often denote such a neighbourhood by $U_\delta(a)$).

Definition 2 (*Cauchy's Definition of Continuity*). We say that $f(x)$ is *continuous at the point a* if, an *arbitrary* number $\varepsilon > 0$ being chosen, another number $\delta > 0$

exists (depending in general on the choice of the number ε), such that for *every* x in the δ-neighbourhood of the point a the relation

$$|f(x) - f(a)| < \varepsilon \tag{1}$$

holds.

REMARK 1. Roughly speaking: $f(x)$ is continuous at the point a, if $f(x)$ differs from $f(a)$ by a small enough quantity when x is sufficiently near to the point a (Fig. 11.8). Or also (writing $\Delta y = f(x + \Delta x) - f(x)$) if $\Delta y \to 0$ when $\Delta x \to 0$.

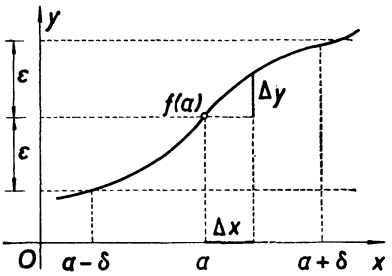


Fig. 11.8.

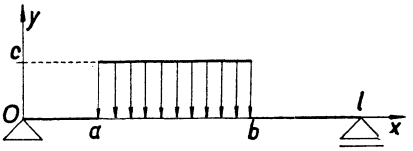


Fig. 11.9.

REMARK 2. It follows from the definition that: Should $f(x)$ be continuous at a point a, it must be defined in a certain neighbourhood of the point a (and so also at the point a itself).

REMARK 3. It is possible to define the continuity of a function $f(x)$ at the point a by means of sequences of x_n tending to the point a (the so-called *Heine definition*).

Definition 3. We say that $f(x)$ is *continuous from (on) the right at the point* a if, an arbitrary number $\varepsilon > 0$ being chosen, another number $\delta > 0$ exists, such that for every $x \geqq a$ in the δ-neighbourhood of the point a the relation

$$|f(x) - f(a)| < \varepsilon$$

holds. Analogous is the definition of *left-hand continuity*. (In these definitions we consider "the right-hand" or "the left-hand" δ-neighbourhood of the point a.)

Example 1. The function $y = \sqrt{(4 - x^2)}$ is continuous from the right (according to Theorems 4 and 5 and Remark 6) at the point $x = -2$ (to the left of the point -2 it is not defined at all). Similarly it is continuous from the left at the point $x = 2$.

REMARK 4. The function in Fig. 11. 9 illustrating the loading of a bar is continuous from the right at the point a, and continuous from the left at the point b, if we define the function at the points a, b by the value c. If we assign to our function the value 0 at these two points it will be continuous from the left at the point a (and discontinuous from the right) and continuous from the right at the point b. (Of course we cannot

assign to our function both values 0 and c at the point a; this would be in contradiction to the definition of a *single-valued function*).

Theorem 1. *The function $f(x)$ is continuous at the point a if and only if it is continuous from the right and continuous from the left at this point.*

Theorem 2. *The function $f(x)$ is continuous at the point a if and only if it is defined at the point a and has a limit at the point a* (see Definition 11.4.1) *equal to the number $f(a)$.*

Definition 4. We say that $f(x)$ is *continuous in a domain M*, if it is continuous at every point of this domain. It is continuous in $[a, b]$ if it is continuous in (a, b) and continuous from the right at a and from the left at b.

Theorem 3. *If $f(x)$ has a derivative at the point a* (Definition 11.5.1), *then $f(x)$ is continuous at a.*

REMARK 5. The converse theorem is not valid (cf. Remark 11.5.3, p. 379).

Theorem 4. *If $f(x)$ and $g(x)$ are continuous at a, then the functions $k \cdot f(x)$ (k being a constant), $f(x) + g(x), f(x) - g(x), f(x) \cdot g(x)$ are also continuous at a; if $g(a) \neq 0$ then also $f(x)/g(x)$ is continuous at a.* A similar theorem is true on the continuity from the right and from the left.

Theorem 5. *A composite function composed of continuous functions is also continuous.* Precisely: *If $f(x)$ is continuous at a and $g(z)$ continuous at the corresponding point $z_0 = f(a)$, then the function $y = g(f(x))$* (see Definition 11.1.4) *is continuous at the point a.*

REMARK 6. On the basis of Theorems 4 and 5 it is easy to show that the great majority of functions we meet with in applications are continuous functions. Especially all polynomials and rational functions with non-vanishing denominators are continuous functions. The function $y = \sqrt{x}$ is continuous for $x \geqq 0$. Further, trigonometric functions are continuous (with the exception of points in the neighbourhood of which they are not bounded, e.g. the function tan x is not continuous at the point $\frac{1}{2}\pi$), inverse trigonometric functions, exponential functions, logarithmic functions and functions generated from them by addition, subtraction, multiplication and division (with non-vanishing divisors) as well as their composite functions are continuous.

Definition 5 (*Points of Discontinuity*). We say that the point a is a *point of discontinuity of the first kind* for the function $f(x)$ (the point a_1 in Fig. 11.10) if there exist a finite right-hand limit and a finite left-hand limit of the function $f(x)$ at the point a (denoted by the symbols $f(a + 0)$ or $f(a - 0)$, respectively, see Remark 11.4.2, p. 371) and if $f(a + 0) \neq f(a - 0)$. We call the number $f(a + 0) - f(a - 0)$ *the jump of the function $f(x)$ at the point a.*

If at least one of the one-sided limits does not exist then we call the point a a *point of discontinuity of the second kind* of the function $f(x)$ (the point a_3 in Fig. 11.10).

If the finite limit $\lim_{x \to a} f(x) = A$ exists but either the function $f(x)$ is not defined at the point a or $f(a) \neq A$, then we say that $f(x)$ has a *removable discontinuity at the point* a (the point a_2 in Fig. 11.10).

Example 2. The function $f(x) = \sin(1/x)$ has a discontinuity of the second kind at the point $x = 0$. The function represented in Fig. 11.9 has a discontinuity of the first kind at the points a and b. The function

$$g(x) = \frac{\sin x}{x}$$

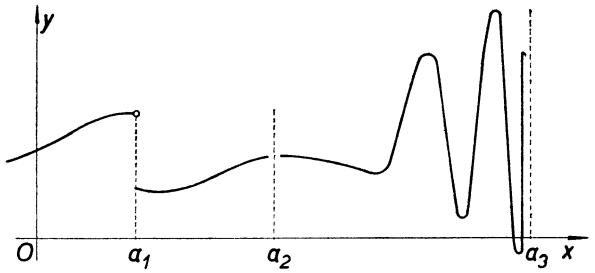


Fig. 11.10.

has a removable discontinuity at the point $x = 0$ because it is not defined at the point $x = 0$ but

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

(cf. Theorem 11.4.9, p. 377, formula 2).

Definition 6. A function $f(x)$ defined in the interval $[a, b]$ is called *sectionally* or *piecewise continuous in the interval* $[a, b]$ if it is continuous in $[a, b]$ except at a finite number of points of discontinuity of the first kind.

Example 3. The function illustrated in Fig. 11.9 is sectionally continuous in the interval $[0, l]$.

Theorem 6. *A function* $f(x)$ *continuous in* $[a, b]$ *takes on a greatest and a least value in* $[a, b]$. Precisely: *There exist at least one point* $x_1 \in [a, b]$ *such that* $f(x_1) \geqq f(x)$ *for all* $x \in [a, b]$ *and at least one point* $x_2 \in [a, b]$ *such that* $f(x_2) \leqq \leqq f(x)$ *for all* $x \in [a, b]$.

REMARK 7. There may be several such points. For example, the function $y = \sin x$ attains in the interval $[-2\pi, 2\pi]$ its maximum value at the points $-\frac{3}{2}\pi$ and $\frac{1}{2}\pi$, its minimum value at the points $-\frac{1}{2}\pi$ and $\frac{3}{2}\pi$.

Theorem 7. *Let* $f(x)$ *be continuous in* $[a, b]$, $f(a) \neq f(b)$, *let* c *be any number between* $f(a)$ *and* $f(b)$ (*i.e. either* $f(a) < c < f(b)$ *or* $f(a) > c > f(b)$). *Then there exists at least one point* $x_0 \in (a, b)$ *such that* $f(x_0) = c$.

REMARK 8. Thus a function continuous in $[a, b]$ assumes every value between $f(a)$ and $f(b)$. Especially, if $f(x)$ is continuous in $[a, b]$ and $f(a) \cdot f(b) < 0$, then $f(x)$ has at least one zero in (a, b).

Theorem 8. *A function $f(x)$, continuous in the interval $[a, b]$, is uniformly continuous in this interval, that is, to any arbitrary $\varepsilon > 0$ there exists $\delta > 0$ (depending only on the choice of the number ε) such that for every two points x_1, x_2 from the interval $[a, b]$, the distance of which is smaller than δ, the relation*

$$|f(x_1) - f(x_2)| < \varepsilon$$

holds.

REMARK 9. Theorems 6 and 8 do not hold in an *open* interval as we can easily verify for the function $1/x$ in the interval $(0, 1)$.

Theorem 9. (*Weierstrass's Theorem*). *It is possible to approximate uniformly in $[a, b]$ with an arbitrary accuracy every function continuous in $[a, b]$ by means of a sequence of polynomials, that is, to every $\varepsilon > 0$ there exists a polynomial $P_n(x)$ such that*

$$|f(x) - P_n(x)| < \varepsilon \quad \text{for all} \quad x \in [a, b].$$

Definition 7. Let $f(x)$ be defined in $[a, b]$. Let us divide this interval into subintervals by means of the points $a = x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n = b$ and let us form the sum

$$V = \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})| .$$

V is a non-negative number depending on the choice of the points of division $x_1, x_2, \ldots, x_{n-1}$. If we choose all possible n and all possible divisions of the interval $[a, b]$, then the numbers V form a set of non-negative numbers. Its lowest upper bound (Definition 1.3.3) is called the *variation* (or more precisely the *total variation*) *of the function $f(x)$ in the interval $[a, b]$*. We denote this by

$$\overset{b}{\underset{a}{V}}(f) .$$

If $\overset{b}{\underset{a}{V}}(f)$ is a finite number, then $f(x)$ is said to be *of bounded variation in $[a, b]$*.

Example 4.

$$\overset{2\pi}{\underset{0}{V}}(\cos x) = 4$$

(it is sufficient to divide the interval $[0, 2\pi]$ into the intervals $[0, \pi]$, $[\pi, 2\pi]$ in which $\cos x$ is decreasing or increasing, respectively.

Theorem 10. *If $f(x)$ has a bounded derivative in $[a, b]$ or if $f(x)$ is monotonic in $[a, b]$ or if $f(x)$ is continuous and attains a finite number of maximum and minimum values in $[a, b]$, then $f(x)$ is of bounded variation in $[a, b]$.*

Theorem 11. *$f(x)$ is a function of bounded variation if and only if it can be expressed as the difference of two non-decreasing functions.*

11.4. Limit. Infinite Limits. Evaluation of Limits. Some Important Limits. Symbols $O(g(x))$, $o(g(x))$

Definition 1. We say that $f(x)$ has the *limit A at the point a* (in more detail *the finite limit A*) if to any arbitrary $\varepsilon > 0$ there exists a $\delta > 0$ (depending in general

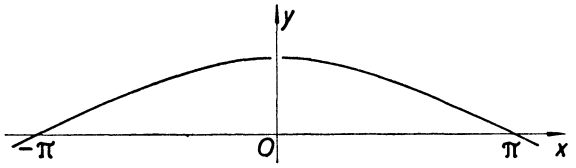


Fig. 11.11.

on the choice of the number ε) such that for all x from the δ-neighbourhood of the point a, *from which we exclude the point a* (i.e. for all x, for which $0 < |x - a| < \delta$) the relation

$$|f(x) - A| < \varepsilon$$

holds.

We write

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

REMARK 1. Roughly speaking: $f(x)$ has the limit A at the point a, if $f(x)$ differs from the number A by as little as we please when x is sufficiently near to the point a.

Example 1. The function

$$y = \frac{\sin x}{x}$$

(Fig. 11.11) is not defined at the point $a = 0$ (therefore it cannot be continuous at that point, cf. Remark 11.3.2) but it has a limit at that point equal to 1 (see Theorem 9). This example shows why we exclude the point a (here the point zero) from our consideration, because our function need not be defined at this point at all.

REMARK 2. The *limit from the right* or the *limit from the left* is defined quite analogously as in Definition 1 (we take into consideration only those x that lie in the "right" or "left" neighbourhood of the point a while the point a is excluded). We

write

$$\lim_{x \to a+} f(x) = A \quad \text{or} \quad \lim_{x \to a-} f(x) = B$$

or often

$$f(a + 0) = A \quad \text{or} \quad f(a - 0) = B .$$

Example 2. For the function represented in Fig. 11.9 we see that $f(a + 0) = c$, $f(a - 0) = 0$.

Theorem 1. *The function $f(x)$ has a limit at the point a if and only if it has both limits from the right and from the left at this point and if these two limits are equal.*

Theorem 2. *The function $f(x)$ is continuous at the point a if and only if it is defined at this point and if*

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

REMARK 3. If the function $f(x)$ is continuous at the point a, we compute the limit of $f(x)$ at that point very easily by putting $x = a$ in the formula of $f(x)$. For example, the function $y = \sin x$ is continuous at the point $a = \frac{1}{3}\pi$, hence

$$\lim_{x \to \pi/3} \sin x = \sin \frac{\pi}{3} = \frac{\sqrt{3}}{2} .$$

Theorem 3. *The function $f(x)$ has a (finite) limit at the point a if and only if the Bolzano–Cauchy condition is satisfied: To any arbitrary $\varepsilon > 0$ there exists a $\delta > 0$ such that for every pair of numbers x_1, x_2, $0 < |x_1 - a| < \delta$, $0 < |x_2 - a| < \delta$ the relation $|f(x_1) - f(x_2)| < \varepsilon$ holds.*

Theorem 4. *If $f(x)$ has the limit A and $g(x)$ the limit B both at the point a, then the functions $k \,.\, f(x)$ ($k = const$), $f(x) \pm g(x)$, $f(x) \,.\, g(x)$, $f(x)/g(x)$ (if $B \neq 0$) have limits at the point a and the relations*

$$\lim_{x \to a} k \,.\, f(x) = kA , \quad \lim_{x \to a} [f(x) \pm g(x)] = A \pm B$$

$$\lim_{x \to a} f(x)\, g(x) = AB , \quad \lim_{x \to a} \frac{f(x)}{g(x)} = \frac{A}{B}$$

hold. A similar theorem holds for the limit from the right or from the left.

REMARK 4. This theorem facilitates the practical computation of limits of many functions; cf. the similar Remark 11.3.6.

Theorem 5 (*Limit of Composite Functions*). *Let*

$$\lim_{x \to a} f(x) = A , \quad \lim_{z \to A} g(z) = B$$

and let $\delta > 0$ *exist such that for all* x, *for which*

$$0 < |x - a| < \delta, \text{ the relation} f(x) \neq A \text{ holds}. \tag{1}$$

Then

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

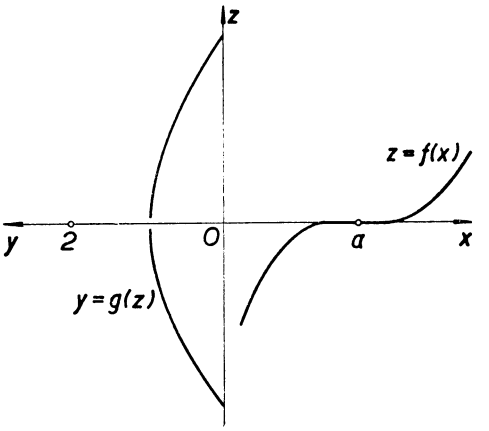


Fig. 11.12.

REMARK 5. If condition (1) is not satisfied, an incorrect result may be obtained as we can see from Fig. 11.12 where $f(x) = 0$ in some neighbourhood of the point a. Because here $A = f(a) = 0$, $g(0) = 2$, we have

$$\lim_{x \to a} g(f(x)) = 2$$

whereas

$$B = \lim_{z \to 0} g(z) = 1 .$$

Definition 2 (*Infinite Limit*). We say that $f(x)$ has the *infinite limit* $+\infty$ *at the point* a (we write $\lim_{x \to a} f(x) = +\infty$) if to any (arbitrarily great) number $K > 0$ there exists a $\delta > 0$ such that for all x from the δ-neighbourhood of the point a (except the point a itself) the relation $f(x) > K$ holds.

REMARK 6. We can give similar definitions of the *infinite limit* $-\infty$ and of the *infinite limit from the right* or *from the left*. For example, the infinite limit $-\infty$ from the right is defined in this way:

Definition 3. We say that the function $f(x)$ has the *infinite limit* $-\infty$ *from the right at the point* a if to any (arbitrarily great) number $L > 0$ there exists a $\delta > 0$ such that for all x from the interval $(a, a + \delta)$ the relation

$$f(x) < -L$$

holds.

Theorem 6. *If $f(x)$ is continuous from the right at the point a and $f(a) > 0$, and if for a certain $\delta > 0$ $g(x) > 0$ in the interval $(a, a + \delta)$ and*

$$\lim_{x \to a+} g(x) = 0 ,$$

then

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

REMARK 7. Similar theorems hold (with the corresponding sign) for various combinations $f(a) \gtrless 0$, $g(x) \gtrless 0$, and also for the limit from the left and for the limit.

Example 3.

$$\lim_{x \to 0+} \frac{1}{x} = +\infty , \quad \lim_{x \to 0-} \frac{1}{x} = -\infty .$$

For $b > 0$,

$$\lim_{x \to 0+} \frac{b}{x} = +\infty , \quad \lim_{x \to 0-} \frac{b}{x} = -\infty ;$$

for $b < 0$,

$$\lim_{x \to 0+} \frac{b}{x} = -\infty , \quad \lim_{x \to 0-} \frac{b}{x} = +\infty .$$

Example 4.

$$\lim_{x \to 2+} \frac{2x^2 - 5x + 1}{x^2 - x - 2} = \lim_{x \to 2+} \frac{\dfrac{2x^2 - 5x + 1}{x + 1}}{x - 2} = -\infty ,$$

because $x^2 - x - 2 = (x - 2)(x + 1)$, and if we write

$$f(x) = \frac{2x^2 - 5x + 1}{x + 1} , \quad g(x) = x - 2,$$

then

$$f(2) = -\tfrac{1}{3} < 0 , \quad g(x) > 0 \quad \text{for} \quad x > 2 \quad \text{and} \quad \lim_{x \to 2+} g(x) = 0 .$$

The limit from the left at the point $x = 2$ is $+\infty$ because $g(x) < 0$ for $x < 2$.

REMARK 8. If the numerator of the quotient $f(x)/g(x)$ also vanishes or if both $f(x)$ and $g(x)$ become infinite as $x \to a$ (an expression of the type $0/0$ or ∞/∞), then, in many cases, we can conclude whether the quotient has a limit, by means of l'Hospital's Rule (Theorems 11.8.1, 11.8.2).

If, however, $f(x)$ is bounded in the neighbourhood of the point a (i.e. $|f(x)| < M$) and

$$\lim_{x \to a} g(x) = +\infty \quad \text{or} \quad \lim_{x \to a} g(x) = -\infty ,$$

then

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

(and similarly for the limit from the right or from the left).

Example 5.

$$\lim_{x \to 0+} \frac{\sin \dfrac{1 + x^2}{x}}{e^{1/x}} = 0$$

because

$$\left| \sin \frac{1 + x^2}{x} \right| \leqq 1$$

for all $x \neq 0$ and

$$\lim_{x \to 0+} e^{1/x} = +\infty .$$

Definition 4 (*Limits at the Points at Infinity*). We say that $f(x)$ has the *limit* A *at the point at infinity* $+\infty$ (or briefly, *at the point* $+\infty$) if to an arbitrary $\varepsilon > 0$ there exists an x_0 such that for all $x > x_0$

$$|f(x) - A| < \varepsilon .$$

We write

$$\lim_{x \to +\infty} f(x) = A .$$

REMARK 9. The definition of the limit at the point at infinity $-\infty$ is analogous. The definitions of the *infinite* limit at the points at infinity $+\infty$ or $-\infty$ are analogous. For example:

Definition 5. We say that the function $f(x)$ has the *infinite limit* $+\infty$ *at the point at infinity* $-\infty$ if to an arbitrary $K > 0$ there exists such a number x_0 that for all $x < x_0$

$$f(x) > K .$$

We write

$$\lim_{x \to -\infty} f(x) = +\infty .$$

Example 6.

$$\lim_{x \to +\infty} \frac{1}{x} = 0 , \quad \lim_{x \to +\infty} x = +\infty , \quad \lim_{x \to -\infty} x^3 = -\infty ,$$

$$\lim_{x \to +\infty} \frac{x^2 - 2x + 3}{2x - 1} = \lim_{x \to +\infty} x \frac{1 - 2/x + 3/x^2}{2 - 1/x} = +\infty$$

(cf. Remark 12).

REMARK 10. We often write ∞ only instead of $+\infty$.

REMARK 11. L'Hospital's rule may also be applied to find limits of the form $0/0$, ∞/∞ at the points at infinity (§ 11.8).

REMARK 12. For computation with finite limits at points at infinity we apply Theorem 4. For computation with infinite limits as $x \to a$ or $x \to +\infty$ or $x \to -\infty$ we may apply Theorem 4 unless the result is of an "indeterminate form" ($0 \cdot \infty$, $\infty - \infty$, etc.). For example,

$$\lim_{x \to 0+} \frac{\sin x}{x^2} = \lim_{x \to 0+} \frac{\sin x}{x} \cdot \frac{1}{x} = 1 \cdot (+\infty) = +\infty .$$

We naturally obtain the same result if we apply Theorem 11.8.1. (Cf. also Example 6.)

Theorem 7. *If $f(x) \leqq g(x)$, for all x in some neighbourhood of the point a (with the possible exception of the point a itself), then $\lim_{x \to a} f(x) \leqq \lim_{x \to a} g(x)$ if both these limits exist. A similar theorem holds for the limit from the right or from the left. (If $a = +\infty$ or $a = -\infty$, then we take into consideration all x greater (or smaller) then a certain number x_0 (instead of a "neighbourhood of the point a").)*

REMARK 13. If $f(x) < g(x)$ for all x in a neighbourhood of the point a, then the relation $\lim_{x \to a} f(x) < \lim_{x \to a} g(x)$ need not hold, the sign of equality may also be valid. For example, in a sufficiently small neighbourhood of the origin $|x| > x^2$ for all x ($x \neq 0$), but $\lim_{x \to 0} |x| = \lim_{x \to 0} x^2 = 0$.

Theorem 8. *If $f(x) \leqq g(x) \leqq h(x)$ for all x in a neighbourhood of the point a (with the possible exception of the point a itself) and if the limits $\lim_{x \to a} f(x) = A$, $\lim_{x \to a} h(x) = A$ exist, then $\lim_{x \to a} g(x)$ exists also and is equal to A.*

REMARK 14. Theorem 8 has a simple geometrical interpretation: If the graph of $g(x)$ lies between the graphs of $f(x)$ and $h(x)$ and if both $f(x)$ and $h(x)$ tend to the same value as $x \to a$, then $g(x)$ also tends to the same value.

Definition 6. We say that $f(x)$ is *of the order $O(g(x))$ in the neighbourhood of the point a if the expression $|f(x)/g(x)|$* is bounded for all $x \neq a$ from a certain neighbourhood of the point a. If

$$\lim_{x \to a} \left| \frac{f(x)}{g(x)} \right| = 0 ,$$

we say that $f(x)$ is *of the order $o(g(x))$* or *of a smaller order than $g(x)$, in the neighbourhood of the point a.*

REMARK 15. We naturally demand that the expression $|f(x)/g(x)|$ has sense in the neighbourhood of the point a ($x \neq a$), i.e. that $g(x) \neq 0$.

REMARK 16. One can see from the definition that if $f(x) = o(g(x))$, then also $f(x) = O(g(x))$; however, in general, the converse is not true.

REMARK 17. By the point a in Definition 6 we may understand also the point at infinity.

Example 7. $\sin x = O(x)$ in the neighbourhood of the point $x = 0$ because

$$\lim_{x\to 0}\left|\frac{\sin x}{x}\right| = 1;$$

$x^3 = o(e^x)$ for $x \to +\infty$ because $\lim\limits_{x\to+\infty}\dfrac{x^3}{e^x} = 0.$

Theorem 9 (*Some Important Limits*).

1. $\lim\limits_{x\to+\infty}\left(1+\dfrac{a}{x}\right)^x = e^a$, $\lim\limits_{x\to-\infty}\left(1+\dfrac{a}{x}\right)^x = e^a$,
2. $\lim\limits_{x\to 0}\dfrac{\sin x}{x} = 1$, $\lim\limits_{x\to 0}\dfrac{\tan x}{x} = 1$,
3. $\lim\limits_{x\to+\infty} a^x = +\infty \quad (a > 1)$, $\lim\limits_{x\to-\infty} a^x = 0 \quad (a > 1)$,
4. $\lim\limits_{x\to+\infty} a^x = 0 \ (0 < a < 1)$, $\lim\limits_{x\to-\infty} a^x = +\infty \ (0 < a < 1)$,
5. $\lim\limits_{x\to 0}\dfrac{e^x - 1}{x} = 1$, $\lim\limits_{x\to 0}\dfrac{a^x - 1}{x} = \ln a \ (a > 0)$,
6. $\lim\limits_{x\to+\infty}\dfrac{x^n}{e^{kx}} = 0 \quad (k > 0, n \textit{ arbitrary})$,
7. $\lim\limits_{x\to+\infty}\dfrac{(\ln x)^n}{x^\alpha} = 0 \quad (\alpha > 0, n \textit{ arbitrary})$,

   $\lim\limits_{x\to 0+} x^\alpha(-\ln x)^n = 0 \quad (\alpha > 0, n \textit{ arbitrary})$.

Especially $\lim\limits_{x\to 0+} x \ln x = 0.$

11.5. Derivative. Formulae for Computing Derivatives. Derivatives of Composite and Inverse Functions

Definition 1. If there exists the (finite) limit

$$\lim_{h\to 0}\frac{f(a+h) - f(a)}{h}, \tag{1}$$

we say that $f(x)$ has a *derivative at the point* a. The corresponding limit is denoted by $f'(a)$.

REMARK 1. The geometrical representation of the number $f'(a)$ is, as we can see from Fig. 11.13, the slope of the tangent to the curve, given by the equation $y = f(x)$, at the point a (because the tangent at the point a is the limiting position of the chord for $h \to 0$). In dynamics, if x stands for time, y for the length of path traversed by a particle up to the instant x, $y = f(x)$ for the equation of motion, then the meaning of the derivative is the limit of the average velocity, i.e. the instantaneous velocity v at the instant a.

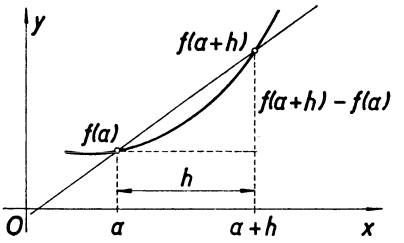


Fig. 11.13.

Definition 2. If there exists in (1) only the limit from the right or from the left, then we say that $f(x)$ has a *right-hand* or *left-hand derivative at the point* a. We write $f'_+(a)$ or $f'_-(a)$.

Theorem 1. *A necessary and sufficient condition that* $f(x)$ *has a derivative at the point* a *is that it has a right-hand and left-hand derivative at the point* a *and* $f'_+(a) = f'_-(a)$.

Definition 3. If (1) is an infinite limit (Definition 11.4.2) or an infinite limit from the right (or from the left), we say that $f(x)$ has an *infinite derivative at the point* a or an *infinite right-hand* (or *left-hand*) *derivative*. (One of the possible geometrical interpretations: The tangent to the curve $y = f(x)$ is vertical.)

REMARK 2. *When we say that* $f(x)$ *has a derivative at the point* a, *we shall always mean a finite derivative.*

Definition 4. If $f(x)$ has a derivative at every point $x \in (a, b)$, we say that $f(x)$ is *differentiable in the interval* (a, b) or that it has a *derivative in the interval* (a, b). Current notations for derivatives:

$$f'(x),\ y'(x),\ \frac{df}{dx},\ \frac{dy}{dx},\ \frac{d}{dx}f(x),\ \frac{d}{dx}y(x),\ y',\ f',\ [f(x)]'.$$

If $f(x)$ is differentiable in (a, b) and if it also has a right-hand derivative at a and a left-hand derivative at b, we say that $f(x)$ is *differentiable in* $[a, b]$ or that $f(x)$ has a *derivative in* $[a, b]$.

Definition 5. A function $y = f(x)$ that has a continuous derivative in $[a, b]$ is called a *smooth function in* $[a, b]$. The curve $y = f(x)$, its graphical representation, is also said to be *smooth*.

Definition 6. If there exists the (finite) limit

$$\lim_{h \to 0} \frac{f'(a + h) - f'(a)}{h},$$

we say that $f(x)$ has a *second derivative at a*. We write it as $f''(a)$. Similarly we define *higher derivatives*. We write them as $f'''(a)$, $f^{(4)}(a)$, $f^{(5)}(a)$, etc.

Theorem 2. *If $f(x)$ has a derivative at the point a, then it is continuous at this point.*

REMARK 3. The converse is not true as we can see when $f(x) = |x|$ (Fig. 11.14), which is continuous at the origin but has no derivative there (it has there the right-hand derivative equal to $+1$ and the left-hand derivative equal to -1).

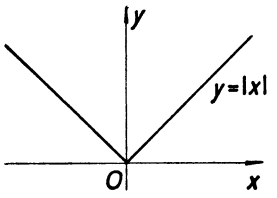


Fig. 11.14.

Theorem 3 (*Fundamental Formulae*). *We denote a derivative by a dash. Unless the contrary is stated, all the formulae are valid for all x and every value of the constants referred to.*

1. $(x^n)' = nx^{n-1}$ ($x > 0$, *n arbitrary*). It is possible to enlarge the domain of validity for the variable x for some n. For example, $(x^3)' = 3x^2$ holds for all x; $(x^{1/3})' = \frac{1}{3}x^{-2/3}$ holds for all $x \neq 0$.

2. If $f(x) = \text{const.}$, *then* $f'(x) = 0$.

3. $(a^x)' = a^x \ln a \quad (a > 0)$, $\quad (e^x)' = e^x$.

4. $(\log_a x)' = \dfrac{1}{x \ln a} \quad (a > 0, a \neq 1, x > 0)$, $\quad (\ln x)' = \dfrac{1}{x} \quad (x > 0)$,

$$[\ln(kx)]' = \frac{1}{x} \quad (k \textit{ arbitrary but such that } kx > 0).$$

5. $(\sin x)' = \cos x$, $\quad (\cos x)' = -\sin x$.

6. $(\tan x)' = \dfrac{1}{\cos^2 x} \quad (x \neq \pm\frac{1}{2}\pi, \pm\frac{3}{2}\pi, \ldots)$,

$$(\cot x)' = -\frac{1}{\sin^2 x} \quad (x \neq 0, \pm\pi, \pm 2\pi, \ldots).$$

7. $(\arcsin x)' = \dfrac{1}{\sqrt{(1 - x^2)}} \quad (|x| < 1),$

$$(\arccos x)' = -\frac{1}{\sqrt{(1 - x^2)}} \quad (|x| < 1).$$

8. $(\arctan x)' = \dfrac{1}{1 + x^2}, \quad (\operatorname{arccot} x)' = -\dfrac{1}{1 + x^2}.$

9. $(\sinh x)' = \cosh x, \quad (\cosh x)' = \sinh x.$

10. $(\tanh x)' = \dfrac{1}{\cosh^2 x}, \quad (\coth x)' = -\dfrac{1}{\sinh^2 x} \quad (x \neq 0).$

Some other common formulae:

11. $(\sin x)^{(n)} = \sin\left(x + \dfrac{n\pi}{2}\right), \quad (\cos x)^{(n)} = \cos\left(x + \dfrac{n\pi}{2}\right).$

12. $(x^m)^{(n)} = m(m-1)\ldots(m-n+1)\,x^{m-n}$ ($x > 0$, *n a positive integer, m arbitrary*), *in particular* $(x^n)^{(n)} = n!$.

13. $(a^x)^{(n)} = a^x(\ln a)^n \quad (a > 0), \quad (e^x)^{(n)} = e^x,$

$$(\ln x)^{(n)} = (-1)^{n-1}\frac{(n-1)!}{x^n} \quad (x > 0).$$

14. $(\sin ax)' = a\cos ax, \quad (\cos ax)' = -a\sin ax.$

15. $$\left(\arcsin\frac{x}{a}\right)' = \frac{1}{\sqrt{(a^2 - x^2)}} \quad (a > 0, |x| < a),$$

$$= -\frac{1}{\sqrt{(a^2 - x^2)}} \quad (a < 0, |x| < |a|),$$

$$\left(\arccos\frac{x}{a}\right)' = -\frac{1}{\sqrt{(a^2 - x^2)}} \quad (a > 0, |x| < a),$$

$$= \frac{1}{\sqrt{(a^2 - x^2)}} \quad (a < 0, |x| < |a|).$$

16. $\left(\arctan\dfrac{x}{a}\right)' = \dfrac{a}{a^2 + x^2}, \quad \left(\operatorname{arccot}\dfrac{x}{a}\right)' = -\dfrac{a}{a^2 + x^2} \quad (a \neq 0).$

17. $(\ln |x + \sqrt{(x^2 + a)}|)' = \dfrac{1}{\sqrt{(x^2 + a)}}$ $(a \neq 0, x^2 + a > 0)$.

18. $[\sin(ax + b)]' = a \cos(ax + b)$, $[\cos(ax + b)]' = -a \sin(ax + b)$.

19. $(e^{ax+b})' = ae^{ax+b}$.

20. $[\ln(ax + b)]' = \dfrac{a}{ax + b}$ $(ax + b > 0)$.

21. $(x^x)' = x^x(1 + \ln x)$ $(x > 0)$.

22. $[f(x)^{g(x)}]' = [e^{g(x)\ln f(x)}]' = f(x)^{g(x)}\left[g'(x)\ln f(x) + \dfrac{g(x)f'(x)}{f(x)}\right]$ $(f(x) > 0)$.

23. $(\operatorname{arsinh} x)' = \dfrac{1}{\sqrt{(1 + x^2)}}$, $(\operatorname{arcosh} x)' = \dfrac{1}{\sqrt{(x^2 - 1)}}$ $(x > 1)$,

$(\operatorname{artanh} x)' = \dfrac{1}{1 - x^2}$ $(-1 < x < 1)$, $(\operatorname{arcoth} x)' = -\dfrac{1}{x^2 - 1}$ $(|x| > 1)$.

24. $\left(\operatorname{arsinh} \dfrac{x}{a}\right)' = \dfrac{1}{\sqrt{(a^2 + x^2)}}$ $(a > 0)$,

$= -\dfrac{1}{\sqrt{(a^2 + x^2)}}$ $(a < 0)$,

$\left(\operatorname{arcosh} \dfrac{x}{a}\right)' = \dfrac{1}{\sqrt{(x^2 - a^2)}}$ $(x > a > 0)$,

$= -\dfrac{1}{\sqrt{(x^2 - a^2)}}$ $(x < a < 0)$.

Theorem 4. *If the functions $f(x)$, $g(x)$ have a derivative at the point a, then also the functions $k \cdot f(x)$ (k = const.), $f(x) \pm g(x)$, $f(x)\,g(x)$ and, if $g(a) \neq 0$, $f(x)/g(x)$ each have a derivative at the point a. Moreover* (briefly written),

$$(kf)' = kf', \quad (f \pm g)' = f' \pm g', \quad (fg)' = f'g + fg', \quad \left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}.$$

Example 1. Applying the rule for differentiation of a quotient we find that for all $x \neq 0$ the function

$$y = \frac{\sin x}{x}$$

has the derivative

$$y' = \frac{x \cos x - \sin x}{x^2}.$$

Theorem 5 (*Differentiation of a Composite Function or the Chain Rule*). *Let* $y = g(f(x))$ *be a composite function where* $y = g(z)$, $z = f(x)$ (Definition 11.1.4). *If the function* $z = f(x)$ *has a derivative with respect to* x *at the point* a *and if the function* $y = g(z)$ *has a derivative with respect to* z *at the point* $z_0 = f(a)$, *then* $y = g(f(x))$ *has a derivative with respect to* x, *at the point* a, *equal to* $g'(z_0) \cdot f'(a)$; *in short notation*

$$\frac{dy}{dx} = \frac{dy}{dz} \cdot \frac{dz}{dx}. \tag{2}$$

REMARK 4. We therefore compute the derivative of the function $g(f(x))$ with respect to x at the point a by multiplying the derivative of the function $y = g(z)$ (with respect to z) at the point $z_0 = f(a)$ by the derivative of the function $z = f(x)$ (with respect to x) at the point a. A more exact form of equation (2) is

$$\frac{dy}{dx} = \frac{dg}{dz} \cdot \frac{df}{dx}. \tag{3}$$

Example 2. We express the function $y = \sin^5 x$ as a composite function as follows: $y = z^5$, $z = \sin x$. From (2), or (3) we have

$$y' = \frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx} = 5z^4 \cos x = 5 \sin^4 x \cos x .$$

Example 3. If $y = \sin(x^5)$, we choose $y = \sin z$, $z = x^5$; hence

$$y' = \frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx} = \cos z \,.\, 5x^4 = 5x^4 \cos(x^5) .$$

Theorem 6 (*Derivative of an Inverse Function*). *If* $y = f(x)$ *is the inverse function of the function* $x = g(y)$ (Definition 11.1.5) *and if* $g(y)$ *has a non-zero derivative* (*with respect to* y) *at the point* y_0, *then the function* $y = f(x)$ *has a derivative* (*with respect to* x) *at the corresponding point* $x_0 = g(y_0)$ *and the relation*

$$\frac{dy}{dx} = \frac{1}{\dfrac{dx}{dy}} \quad \left(\textit{or, more exactly: } \frac{df}{dx} = \frac{1}{\dfrac{dg}{dy}}\right)$$

holds.

Example 4.

$$y = \arcsin x \quad (\text{i.e. } y = \sin^{-1} x), \quad -1 \leqq x \leqq 1 ,$$

is the inverse function of $x = \sin y$ $\left(-\frac{1}{2}\pi \leqq y \leqq \frac{1}{2}\pi\right)$ (Example 11.1.5, p. 363). Then

$$\frac{dx}{dy} = \cos y \neq 0 \quad \text{for} \quad y \neq \pm\tfrac{1}{2}\pi ,$$

hence

$$\frac{dy}{dx} = \frac{1}{\cos y} = \frac{1}{+\sqrt{(1 - \sin^2 y)}} = \frac{1}{\sqrt{(1 - x^2)}} \quad (-1 < x < 1).$$

(We take the positive root because $\cos y > 0$ when $-\frac{1}{2}\pi < y < \frac{1}{2}\pi$.)

Theorem 7. *The derivatives of a function given parametrically by equations* $x = \varphi(t)$, $y = \psi(t)$:

$$y' = \frac{dy}{dx} = \frac{\psi'(t)}{\varphi'(t)}, \quad y'' = \frac{d}{dx}(y') = \frac{d}{dt}(y')\frac{dt}{dx} = \frac{\psi''(t)\,\varphi'(t) - \psi'(t)\,\varphi''(t)}{[\varphi'(t)]^3},$$

$$y''' = \frac{d^3y}{dx^3} = \frac{\varphi'^2\psi''' - \varphi'\psi'\varphi''' - 3\varphi'\varphi''\psi'' + 3\varphi''^2\psi'}{\varphi'^5} \quad (\varphi'(t) \neq 0).$$

Example 5. For the ellipse $x = a\cos t$, $y = b\sin t$ we have

$$y' = \frac{b\cos t}{-a\sin t} = -\frac{b}{a}\cot t = -\frac{b^2}{a^2}\frac{x}{y}, \quad t \neq k\pi \quad \text{(where } k \text{ is an integer)}.$$

REMARK 5. Similarly, for polar coordinates

$$x = \varrho\cos\varphi, \quad y = \varrho\sin\varphi,$$

the relations

$$dx = d\varrho\cos\varphi - \varrho\sin\varphi\,d\varphi, \quad dy = d\varrho\sin\varphi + \varrho\cos\varphi\,d\varphi$$

hold; hence

$$\frac{dy}{dx} = \frac{\sin\varphi\dfrac{d\varrho}{d\varphi} + \varrho\cos\varphi}{\cos\varphi\dfrac{d\varrho}{d\varphi} - \varrho\sin\varphi},$$

Some formulae that find frequent application:

Theorem 8. *For the derivative of a product of functions* $u_1(x), u_2(x), \ldots, u_n(x)$ *the following rule holds:*

$$(u_1u_2u_3\ldots u_n)' = u_1'u_2u_3\ldots u_n + u_1u_2'u_3\ldots u_n + \ldots + u_1u_2u_3\ldots u_n' =$$

$$= u_1u_2u_3\ldots u_n\left(\frac{u_1'}{u_1} + \frac{u_2'}{u_2} + \frac{u_3'}{u_3} + \ldots + \frac{u_n'}{u_n}\right)$$

(*we apply the last form if* $u_1 \neq 0, u_2 \neq 0, \ldots$). *In another form:*

$$\frac{(u_1u_2u_3 \ldots u_n)'}{u_1u_2u_3 \ldots u_n} = \frac{u_1'}{u_1} + \frac{u_2'}{u_2} + \frac{u_3'}{u_3} + \ldots + \frac{u_n'}{u_n}.$$

Theorem 9 (*Leibniz's Rule*).

$$(uv)^{(n)} = u^{(n)}v^{(0)} + \binom{n}{1}u^{(n-1)}v^{(1)} + \binom{n}{2}u^{(n-2)}v^{(2)} + \ldots + u^{(0)}v^{(n)}.$$

The upper indices in brackets stand for the order of the derivative; $u^{(0)} = u$, $v^{(0)} = v$; $\binom{n}{k}$ *are the binomial coefficients* (§ 1.12, p. 18).

Example 6. For the second derivative of a product of functions $u(x) \cdot v(x)$ the relation

$$(uv)'' = u''v + 2u'v' + uv''$$

holds. For example

$$(x^3 \sin x)'' = 6x \sin x + 6x^2 \cos x - x^3 \sin x.$$

REMARK 6. If the given function $y = h(x)$ is positive and if we can easily differentiate the function $\ln h(x)$, then we frequently use so-called *logarithmic differentiation*. Then $y = h(x) = e^{\ln h(x)}$, $y' = e^{\ln h(x)}[\ln h(x)]' = h(x) \cdot [\ln h(x)]'$.

Example 7. $y = x^x$ $(x > 0)$, $\ln x^x = x \ln x$, $[x \ln x]' = \ln x + x/x = \ln x + 1$, so $y' = x^x(\ln x + 1)$.

More generally:

$$y = f(x)^{g(x)} \quad (f(x) > 0), \quad \ln [f(x)^{g(x)}] = g(x) \ln f(x),$$

$$[g(x) \ln f(x)]' = g'(x) \ln f(x) + g(x)\frac{f'(x)}{f(x)},$$

so

$$y' = f(x)^{g(x)}\left[g'(x) \ln f(x) + g(x)\frac{f'(x)}{f(x)}\right].$$

11.6. Differential. Differences

Definition 1. We say that $f(x)$ is *differentiable* or *has a differential at the point a* if we can express its increment $\Delta f = f(a + h) - f(a)$ in the form

$$\Delta f = f(a + h) - f(a) = Ah + h\,\tau(h) \tag{1}$$

where A is a constant and

$$\lim_{h \to 0} \tau(h) = 0 . \tag{2}$$

Theorem 1. *The function $f(x)$ has a differential at the point a if and only if $f(x)$ has a derivative at the point a. The constant A in (1) is equal to $f'(a)$,* i.e.

$$f(a + h) - f(a) = f'(a)\, h + h\, \tau(h) . \tag{3}$$

Definition 2. The expression $f'(a)\, h$ is called the *differential of the function $f(x)$ at the point a*. We denote it by $\mathrm{d}f(a)$. At a general point we write $\mathrm{d}f(x)$ or $\mathrm{d}y$.

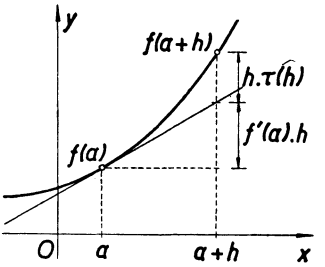


Fig. 11.15.

The geometrical interpretation (Fig. 11.15): If we replace the increment $\Delta f = f(a + h) - f(a)$ by the differential $f'(a)\, h$, then it means that we take only the increment on the tangent $y = f(a) + f'(a)\,(x - a)$ instead of the increment of the function $y = f(x)$. According to (3), the "error" $\Delta f - \mathrm{d}f$ is equal to the function $h\, \tau(h)$ of the variable h; if $h \to 0$, then $\mathrm{d}f \to 0$ and $h\, \tau(h) \to 0$. The condition (2) says that, if $f'(a) \neq 0$, then $h\, \tau(h)$ tends to zero at a "higher order" than $f'(a)\, h$. That is, the smaller is h, the smaller *relative* error we commit in replacing Δf by the differential $\mathrm{d}f$.

At a general point x, $\mathrm{d}f(x) = f'(x)\, h$. For the function $y = x$ we have $\mathrm{d}y = \mathrm{d}x = h$. This justifies the notation of differentials

$$\mathrm{d}f(x) = f'(x)\, \mathrm{d}x \quad \text{or} \quad \mathrm{d}y = f'(x)\, \mathrm{d}x$$

used almost exclusively nowadays.

We call the reader's attention to the fact that here $\mathrm{d}x$ is by no means an "infinitesimally small quantity", but it can assume any value. Sufficient accuracy when replacing the increment of a function by its differential is, of course, secured by (2) only for sufficiently small $\mathrm{d}x$.

Example 1. For the function $y = x^3$, $\mathrm{d}y = 3x^2\, \mathrm{d}x$; $\Delta y = (x + \mathrm{d}x)^3 - x^3 = 3x^2\, \mathrm{d}x + 3x\, \mathrm{d}x^2 + \mathrm{d}x^3$, hence $\tau(\mathrm{d}x) = 3x\, \mathrm{d}x + \mathrm{d}x^2$. (Cf. the notation in Remark 4.) Obviously for every x,

$$\lim_{\mathrm{d}x \to 0} \tau(\mathrm{d}x) = 0 .$$

Further,

$$\mathrm{d}y = 2{\cdot}7\,, \quad \Delta y = 2{\cdot}791 \quad \text{when} \quad x = 3\,, \quad \mathrm{d}x = 0{\cdot}1\,.$$
$$\mathrm{d}y = 0{\cdot}027\,, \quad \Delta y = 0{\cdot}027,009,001 \quad \text{when} \quad x = 3\,, \quad \mathrm{d}x = 0{\cdot}001\,.$$

REMARK 1. In technical subjects we often speak about an *increment of the function* instead of a *differential of the function.*

Taylor's formula gives a better approximation to the increment of a function than the differential does (§ 11.10, p. 396).

REMARK 2. If we replace the increment of a function by its differential, then it follows for the estimation of the error $R = [f(a + h) - f(a)] - f'(a)\,h$ by Taylor's formula that

$$|R| = \frac{|f''(a + \vartheta h)|}{2!}\,h^2\,, \quad 0 < \vartheta < 1\,.$$

REMARK 3. The differential is often used for the approximate determination of the error committed in computing the value of a quantity from the value of another quantity measured with some error. For example, if the radius of a sphere is measured as $a = 4$ cm and if we know that the error of that measurement is 0·1 mm at most, then the maximum error in the determination of the volume $V = \frac{4}{3}\pi x^3 = \frac{4}{3}\pi\,.\,4^3\ \mathrm{cm}^3$ is given approximately by the differential $V'\,.\,h = 4\pi x^2 h = 4\pi\,.\,4^2\,.\,0{\cdot}01\ \mathrm{cm}^3$ (about 0·75 per cent).

Definitior Let $f(x)$ have a second derivative which is continuous at the point x. The *sec* *ifferential of the function* $f(x)$ *at the point* x is the expression

$$\mathrm{d}^2 f(x) = f''(x)\,\mathrm{d}x^2\,.$$

gously we define *differentials of higher orders,* $\mathrm{d}^{(n)} f(x) = f^{(n)}(x)\,\mathrm{d}x^n$; it is sup- d that $f(x)$ has a continuous n-th derivative at the point x.

The second (n-th) differential is obtained formally as the differential of the first (or $(n-1)$-th) differential for the same constant h:

$$\mathrm{d}^2 f(x) = \mathrm{d}[h\,f'(x)] = hf''(x)\,.\,h = h^2 f''(x) = f''(x)\,\mathrm{d}x^2\,.$$

REMARK 4. The notation $\mathrm{d}x^n$ is usual for $(\mathrm{d}x)^n$, thus it does not stand for $\mathrm{d}(x^n)$. Similarly Δx^n is used instead of $(\Delta x)^n$. Thus Δx^n is not $\Delta(x^n)$.

Definition 4. The *first difference* $\Delta f(a)$ *of the function* $f(x)$ *at the point* a is defined as

$$\Delta f(a) = f(a + \Delta x) - f(a)\,.$$

The *second difference* is the difference of the first difference, $\Delta^2 f(a) = \Delta[\Delta f(a)] =$ $= [f(a + \Delta x + \Delta x) - f(a + \Delta x)] - [f(a + \Delta x) - f(a)] = f(a + 2\Delta x) - 2f(a + \Delta x) +$

$+ f(a)$. Generally, the *n-th difference* is the difference of the $(n-1)$-th difference,

$$\Delta^n f(a) = f(a + n\Delta x) - \binom{n}{1} f[a + (n-1)\Delta x] + \binom{n}{2} f[a + (n-2)\Delta x] - \\ - \ldots + (-1)^n f(a).$$

Theorem 2. *If $f(x)$ has the n-th derivative which is continuous at the point x, then*

$$f^{(n)}(x) = \lim_{\Delta x \to 0} \frac{\Delta^n f(x)}{\Delta x^n}.$$

For more detailed treatment of differences see Chap. 32.

11.7. General Theorems on Derivatives. Rolle's Theorem. Mean-Value Theorem

Theorem 1 (*Rolle's Theorem*). *If a function $f(x)$ is continuous in $[a, b]$, has a derivative (finite or infinite) in (a, b), and $f(a) = f(b)$, then there exists at least one point $c \in (a, b)$ such that $f'(c) = 0$* (so that tangent to the graph at c is horizontal).

Theorem 2 (*Mean-Value Theorem or Lagrange's Theorem*). *If a function $f(x)$ is continuous in $[a, b]$ and has a derivative (finite or infinite) in (a, b), then there exists at least one point $c \in (a, b)$ such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a} \quad \text{or} \quad f(b) - f(a) = (b - a) f'(c).$$

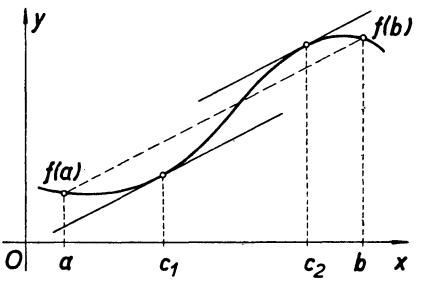


Fig. 11.16.

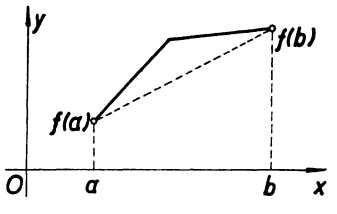


Fig. 11.17.

REMARK 1. Geometrical interpretation: There exists at least one point c (in Fig. 11.16 even two such points, c_1 and c_2) such that the tangent at c is parallel to the straight line joining the points $(a, f(a))$, $(b, f(b))$. For $f(a) = f(b)$ we have Rolle's Theorem. The theorem does not hold if $f(x)$ has a right-hand and left-hand derivative at an interior point, but these are not equal, as we can see in Fig. 11.17.

Theorem 3 (*The Generalized Mean-Value Theorem*). *If the functions* $f(x)$ *and* $g(x)$ *are continuous in* $[a, b]$ *and have derivatives in* (a, b) *(an infinite derivative of* $f(x)$ *is admitted),* $g'(x) \neq 0$ *in* (a, b), *then there exists at least one point* $c \in (a, b)$ *such that*

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(c)}{g'(c)}.$$

REMARK 2. For $g(x) = x$ we have the preceding Mean-Value theorem.

Theorem 4. *If* $f(x)$ *is continuous from the right at the point* a, *has a finite or infinite derivative when* $a < x < a + \delta$ $(\delta > 0)$ *and if the finite or infinite limit* $\lim_{x \to a+} f'(x)$ *exists, then there exists a derivative from the right of* $f(x)$ *at the point* a *which is equal to this limit.*

An analogous theorem holds for a left-hand neighbourhood of the point a.

11.8. The Computation of Certain Limits by means of l'Hospital's Rule

Theorem 1 (*Computation of Limits of the Form* $0/0$). *If*

$$\lim_{x \to a} f(x) = 0, \quad \lim_{x \to a} g(x) = 0$$

and

$$\lim_{x \to a} \frac{f'(x)}{g'(x)}$$

exists (finite or infinite), then

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

also exists and

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

A similar theorem is valid for the limit from the right or from the left (i.e. for $x \to a+$ or $x \to a-$) and for the limits at the points at infinity (for $x \to +\infty$ or $x \to -\infty$).

REMARK 1. Frequently it is necessary, if $f'(x)/g'(x)$ is again of "indeterminate form" (i.e. if again $\lim_{x \to a} f'(x) = 0$ and $\lim_{x \to a} g'(x) = 0$), to repeat the application of l'Hospital's rule.

Example 1.

$$\lim_{x \to 0} \frac{1 - \cos x}{x^2} = \lim_{x \to 0} \frac{\sin x}{2x} = \lim_{x \to 0} \frac{\cos x}{2} = \frac{1}{2}.$$

REMARK 2. L'Hospital's Rule cannot be used if one of the functions tends to zero when $x \to a$ while the other does not.

REMARK 3. Note that $f(x)/g(x)$ is not differentiated as a quotient; the functions in the numerator and the denominator are differentiated separately.

Theorem 2 (*Computation of Limits of the Form* ∞/∞). *Let* $\lim_{x\to a} |g(x)| = +\infty$ (*we do not suppose anything about* $\lim_{x\to a} f(x)$, *not even the existence of that limit*). *Then, if*

$$\lim_{x\to a} \frac{f'(x)}{g'(x)}$$

exists (*finite or infinite*) *so does*

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

and

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

Theorem 2 holds analogously for $x \to a+$, $x \to a-$, $x \to +\infty$, $x \to -\infty$, respectively.

Example 2.

$$\lim_{x\to+\infty} \frac{x^3 + 5x - 2}{x^2 - 1} = \lim_{x\to+\infty} \frac{3x^2 + 5}{2x} = \lim_{x\to+\infty} \frac{6x}{2} = +\infty.$$

Example 3. For $a > 1$, n a positive integer,

$$\lim_{x\to+\infty} \frac{a^x}{x^n} = \lim_{x\to+\infty} \frac{a^x \ln a}{nx^{n-1}} = \ldots = \lim_{x\to+\infty} \frac{a^x (\ln a)^n}{n!} = +\infty.$$

In particular,

$$\lim_{x\to+\infty} \frac{e^x}{x^n} = +\infty.$$

This fact is expressed by the statement that the exponential function increases faster than any power of x, when $x \to +\infty$.

REMARK 4. The computation of limits of the form $0 \cdot \infty$, $\infty - \infty$ is frequently reduced to the preceding forms:

Example 4.

$$\lim_{x\to0+} x \ln x = \lim_{x\to0+} \frac{\ln x}{\frac{1}{x}} = \lim_{x\to0+} \frac{\frac{1}{x}}{-\frac{1}{x^2}} = \lim_{x\to0+} (-x) = 0.$$

Example 5.

$$\lim_{x\to 0+}\left(\frac{1}{\sin x}-\frac{1}{x}\right)=\lim_{x\to 0+}\frac{x-\sin x}{x\sin x}=\lim_{x\to 0+}\frac{1-\cos x}{\sin x+x\cos x}=$$

$$=\lim_{x\to 0+}\frac{\sin x}{2\cos x-x\sin x}=0\,.$$

We try to reduce "indeterminate expressions" of other forms to the preceding forms also:

Example 6.

$$\lim_{x\to 0+}x^x=\lim_{x\to 0+}e^{x\ln x}=\exp\Big(\lim_{x\to 0+}x\ln x\Big)=e^0=1\,.$$

Here we have applied the result of example 4, the relation $x=e^{\ln x}$ (following from the definition of $\ln x$) and the continuity of the exponential function,

$$\lim_{z\to z_0}e^z=\exp\Big(\lim_{z\to z_0}z\Big)\,.$$

11.9. Investigation of a Function. Graphical Representation. Monotonic Functions. Concavity. Convexity. Points of Inflection. Maxima and Minima

In this paragraph the briefer term "the function" instead of "the graph of the function" is frequently used.

Definition 1. We say that $f(x)$ is *strictly increasing at the point a* if, in a certain neighbourhood of the point a,

$$\begin{aligned} f(x)&>f(a) \quad \text{when} \quad x>a\,, \\ f(x)&<f(a) \quad \text{when} \quad x<a \end{aligned} \tag{1}$$

(the points x_1, x_4 in Fig. 11.18). Analogously we define a function *strictly decreasing at the point a.*

If in (1) $f(x)\geqq f(a)$ when $x>a$ and $f(x)\leqq f(a)$ when $x<a$, we say that $f(x)$ is *increasing at the point a*; analogously we speak of a function being *decreasing at a.* For example the function $f(x)=$ const. is both increasing and decreasing at a.

Theorem 1. *If $f'(a)>0$, then $f(x)$ is strictly increasing at the point a; if $f'(a)<0$, then $f(x)$ is strictly decreasing at the point a.*

REMARK 1. If $f(x)$ is strictly increasing at every point of an interval I, we say that it is *strictly increasing in I.* The following definition is equivalent: $f(x)$ is called

strictly increasing in I if for every pair of points x_1, x_2 of this interval, satisfying $x_1 < x_2$, the relation

$$f(x_1) < f(x_2) \tag{2}$$

holds. *Increasing* (with the sign $\leqq$ in (2)), *strictly decreasing*, and *decreasing functions in I* are defined analogously. All such functions are called *monotonic in I*. Strictly increasing or strictly decreasing functions are called *strictly monotonic in I*. *If $f'(x) > 0$ ($f'(x) < 0$) in I, then $f(x)$ is strictly increasing (decreasing) in I.*

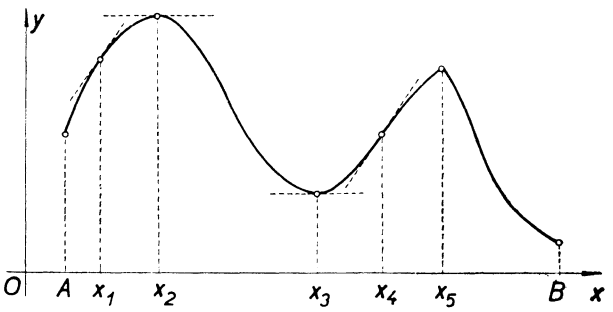


Fig. 11.18.

Definition 2. If, in a certain neighbourhood U of the point a, the graph of the function $f(x)$ lies below the tangent (or on the tangent) drawn at the point $(a, f(a))$ (i.e. if the relation $f(x) \leqq f(a) + (x - a)f'(a)$ holds in U; the points x_1, x_2 in Fig. 11.18), we say that $f(x)$ is *concave at the point a*. If the graph lies above the tangent (or on the tangent), i.e. the relation $f(x) \geqq f(a) + (x - a)f'(a)$ holds in U, we say that $f(x)$ is *convex at the point a* (the point x_3 in Fig. 11.18).

REMARK 2. If in these cases the graph of the function coincides in a certain neighbourhood U of the point a with the tangent *only* at the point of contact (this is the case we meet most frequently in applications), the function is called *strictly concave* or *strictly convex at a*.

Theorem 2. *If $f''(a) > 0$, then $f(x)$ is strictly convex at a; if $f''(a) < 0$, then $f(x)$ is strictly concave at a.*

REMARK 3. If $f(x)$ is convex at every point of an interval I, we say that it is *convex in the interval I*. Analogously a function is defined to be *strictly convex* or *concave* or *strictly concave in an interval I*.

If $f''(x) > 0$ (or $f''(x) < 0$) everywhere in I, then $f(x)$ is strictly convex (or strictly concave) in I.

Definition 3. If the graph of a function crosses, at the point $x = a$, its tangent at this point, we say that $f(x)$ has a *point of inflection at the point a* (the point x_4 in Fig. 11.18).

Theorem 3. *If* $f''(a) = 0$, $f'''(a) \neq 0$, *then* $f(x)$ *has a point of inflection at* a. *Further if* $f'''(a) < 0$, *the graph of the function crosses its tangent from above* (the point x_4 in Fig. 11.18); *if* $f'''(a) > 0$, *the graph crosses its tangent from below. If* $f''(a) \neq 0$, *then* $f(x)$ *has no point of inflection at* a.

Definition 4. If

$$f(x) \leqq f(a) \quad (\text{or} \quad f(x) \geqq f(a)) \tag{3}$$

in some neighbourhood U of the point a, we say that $f(x)$ has a *relative maximum* (or a *relative minimum*) *at the point* a. If the sign of equality in (3) holds in U at the point a only (this is the case we meet most frequently in applications) we speak of a *strict relative maximum* or *s. r. minimum* (the points x_2, x_3, x_5 in Fig. 11.18).

Theorem 4. *If* $f'(a) = 0, f''(a) > 0$, *then* $f(x)$ *has a strict relative minimum at* a; *if* $f'(a) = 0, f''(a) < 0$, *then* $f(x)$ *has a strict relative maximum at* a. *If* $f'(a) \neq 0$, *then* $f(x)$ *has no relative extremum at* a.

Definition 5. If $f(x) \geqq f(a)$ for all x from a given interval (more generally from a domain) M, we say that $f(x)$ has a *minimum on* M *at the point* $a \in M$ (the point B in Fig. 11.18). A *maximum* on M is defined analogously (the point x_2 in Fig. 11.18). (Both these extremes are often called *total* or *absolute.*)

REMARK 4. As we can observe in Fig. 11.18, absolute extremes on M need not always be at the points of the relative extremes (the point B in Fig. 11.18); it is necessary to investigate also the values of the function at the boundary points of the domain M. In the same figure, we can see that the relative extremes occur not only at the points where $f'(a) = 0$ (Theorem 4) but they may occur also at points at which $f(x)$ has no derivative at all (the point x_5 in Fig. 11.18).

REMARK 5. Even if a function $f(x)$ has a sufficient number of higher derivatives at the point a, Theorems 1–4 may prove to be ineffective when several of the first derivatives vanish. Then the following theorems may be useful:

Theorem 5. *Let* $f''(a) = f'''(a) = \ldots = f^{(n-1)}(a) = 0$, $f^{(n)}(a) \neq 0$ $(n \geqq 2)$. *Then*

$f(x)$ *is strictly convex at* a *when* $f^{(n)}(a) > 0$, n *even*,
$f(x)$ *is strictly concave at* a *when* $f^{(n)}(a) < 0$, n *even*,
$f(x)$ *has a point of inflection at* a (*and crosses the tangent from below*) *when* $f^{(n)}(a) > 0$, n *odd*,
$f(x)$ *has a point of inflection at* a (*and crosses the tangent from above*) *when* $f^{(n)}(a) < 0$, n *odd*.

Theorem 6. *Let* $f'(a) = f''(a) = \ldots = f^{(n-1)}(a) = 0$, $f^{(n)}(a) \neq 0$ $(n \geqq 1)$. *Then*
$f(x)$ *has a strict relative minimum at* a *when* $f^{(n)}(a) > 0$, n *even*,
$f(x)$ *has a strict relative maximum at* a *when* $f^{(n)}(a) < 0$, n *even*,
$f(x)$ *is strictly increasing at* a *when* $f^{(n)}(a) > 0$, n *odd*,
$f(x)$ *is strictly decreasing at* a *when* $f^{(n)}(a) < 0$, n *odd*.

REMARK 6. Putting $n = 2, 3$ in Theorem 5, we obtain Theorems 2, 3 respectively. Putting $n = 1, 2$ in Theorem 6, we obtain Theorems 1, 4 respectively.

REMARK 7. If $f'(a) = 0$ (the zeros of the derivative are generally known as "*stationary*" points of the function $f(x)$) and if the computation of the second derivative or of higher derivatives is not complicated, then we decide easily whether there is an extremum at the point a (and its type) by Theorem 4 or 6. If the computation of derivatives is rather lengthy, we may use the following theorem:

Theorem 7. *Let $f'(a) = 0$ and $f'(x) > 0$ when $x < a$, $f'(x) < 0$ when $x > a$ in a certain neighbourhood U of the point a.* (We say briefly that the derivative is *changing its sign from positive to negative.*) *Then $f(x)$ has a strict relative maximum at the point a.*

If $f'(a) = 0$ and $f'(x) < 0$ when $x < a$, $f'(x) > 0$ when $x > 0$ in U, then $f(x)$ has a strict relative minimum at the point a.

If $f'(a) = 0$ and $f'(x) > 0$ (or $f'(x) < 0$) when $x < a$ as well as when $x > a$ in U, then $f(x)$ is strictly increasing (strictly decreasing) at the point a.

Similarly we have:

Theorem 8. *If $f''(a) = 0$ and if, in a certain neighbourhood U of the point a, $f''(x) > 0$ when $x < a$ and $f''(x) < 0$ when $x > a$, then $f(x)$ has a point of inflection at the point a and the graph of $f(x)$ crosses the tangent from above. If $f''(a) = 0$ and $f''(x) < 0$ when $x < a$, $f''(x) > 0$ when $x > a$ in U, then $f(x)$ has a point of inflection at the point a and the graph crosses the tangent from below.*

Example 1 (The *Investigation of a Function*). Using the theorems of this paragraph, let us investigate the characteristic features of the function

$$f(x) = x + \frac{4}{x} \tag{4}$$

and plot its graph approximately.

The functional relation (4) is defined for every $x \neq 0$. Also the domain of definition is (Remark 11.1.4, p. 361) the interval $(-\infty, +\infty)$ from which the point $x = 0$ is excluded. The function (4) has derivatives of all orders. Especially

$$f'(x) = 1 - \frac{4}{x^2}, \tag{5}$$

$$f''(x) = \frac{8}{x^3}. \tag{6}$$

From (5) it follows that $f'(x) = 0$ for $x_1 = 2, x_2 = -2$. By (6) $f''(x_1) > 0, f''(x_2) < 0$, hence by Theorem 4, $f(x)$ has a strict relative minimum at the point x_1 and a strict relative maximum at the point x_2. We easily compute $f(2) = 4$, $f(-2) = -4$ (the points A, B in Fig. 11.19). For $x \neq \pm 2$, $f'(x) \neq 0$ and thus the given function has

no other relative extremes. For $|x| > 2$ we have by (5) $f'(x) > 0$ and so (Remark 1) the function $f(x)$ is strictly increasing in the intervals $(2, +\infty)$ and $(-\infty, -2)$. In the intervals $(0, 2)$, $(-2, 0)$ we have $f'(x) < 0$ by (5) and $f(x)$ is strictly decreasing there (Fig. 11.19).

By (6), $f''(x) > 0$ for $x > 0$, hence $f(x)$ is strictly convex in the interval $(0, +\infty)$ (Remark 3). In the interval $(-\infty, 0)$, $f''(x) < 0$ and thus $f(x)$ is strictly concave (Fig. 11.19).

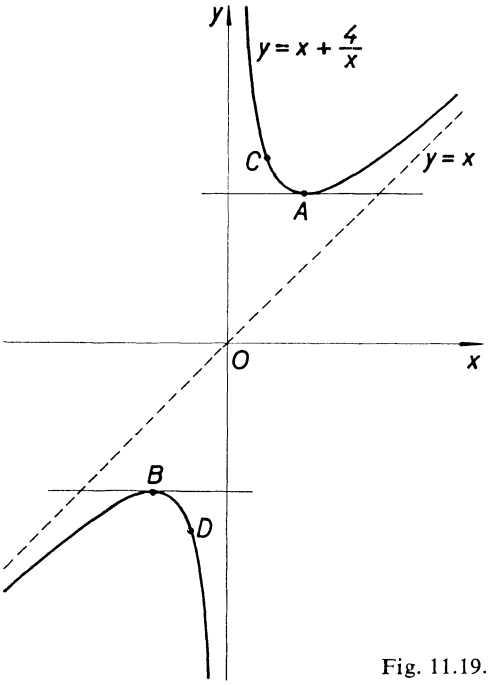


Fig. 11.19.

Because $f''(x) \neq 0$ everywhere in the domain of definition M, $f(x)$ has nowhere a point of inflection (Theorem 3).

For an approximate plotting of the graph it is useful to determine the asymptotes of the graph. Evidently,

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

so that the straight line $x = 0$ is a vertical asymptote of the graph (§ 9.6, p. 288). Further, there exist finite limits for $x \to +\infty$ and $x \to -\infty$,

$$k = \lim \frac{f(x)}{x} = \lim \frac{x + \dfrac{4}{x}}{x} = \lim \left(1 + \frac{4}{x^2}\right) = 1\,,$$

$$q = \lim [f(x) - kx] = \lim \left(x + \frac{4}{x} - x \right) = \lim \frac{4}{x} = 0 ,$$

so that the straight line

$$y = x$$

is a (non-vertical) asymptote of the graph (Theorem 9.6.1, p. 326).

To facilitate the drawing of the graph, we observe that the investigated function is odd (Definition 11.2.1, p. 404),

$$f(-x) = -x + \frac{4}{-x} = -\left(x + \frac{4}{x} \right) = -f(x) ,$$

thus its graph is symmetric with respect to the origin. (Moreover, we could have used this property earlier for the investigation of the given function for $x > 0$ only: If $f(x)$ has a strict relative minimum at $x = 2$, then it has a strict relative maximum at $x = -2$, etc.) For the construction of the graph it is advantageous, further, to work out the values of the function itself at some points where their computation is easy, e.g. at the points $x = 1$, $x = -1$ (the points C, D in Fig. 11.19).

Thus the investigation of the properties of the function $f(x)$ and the preparation for the approximate drawing of its graph are finished. Briefly, we say that we have performed the investigation of the given function.

Example 2. Of all rectangles of perimeter 20 cm, to find that with the greatest area.

Denoting the lengths of the sides by x, y, the area is $P = xy$ with the condition $2x + 2y = 20$, hence $y = 10 - x$, and $P = x(10 - x) = 10x - x^2$. We look for that value of $x \in [0, 10]$ for which P assumes its maximum value. Evidently we have to find the *relative* maximum, since $P(0) = P(10) = 0$. Because the function $P(x)$ has a derivative everywhere it may have a maximum only at the point where $P' = 0$ (Theorem 4). From the equation $P'(x) = 0$ or $10 - 2x = 0$ we obtain $x = 5$. Indeed, there is a maximum at the point $x = 5$ (by Theorem 4), because $P''(5) = -2 < 0$. (Thus the square, the length of side of which is 5 cm, has a maximum area of all rectangles of the given perimeter 20 cm.)

Example 3. Of all right circular cylinders of given volume V, to find that with the least surface S.

We have

$$V = \pi r^2 h , \tag{7}$$

$$S = 2\pi rh + 2\pi r^2 . \tag{8}$$

If we choose for instance r as independent variable, then we can express h using (7) by

$$h = \frac{V}{\pi r^2} , \tag{9}$$

for V is fixed. Putting (9) in (8),

$$S = \frac{2V}{r} + 2\pi r^2 , \tag{10}$$

which gives S as a function of the single variable r. We will find the minimum of this function for $r \in (0, +\infty)$. If we put $S' = 0$, i.e.

$$-\frac{2V}{r^2} + 4\pi r = 0 , \tag{11}$$

we obtain

$$r_{\min} = \sqrt[3]{\frac{V}{2\pi}} .$$

For this value of r the surface S really attains its minimum value in the interval $(0, +\infty)$ because for $r > r_{\min}$, $S' > 0$ and for $0 < r < r_{\min}$, $S' < 0$ (as follows from the left-hand side of equation (11)). For $r_{\min}$ we obtain from (9)

$$h_{\min} = \sqrt[3]{\frac{4V}{\pi}} = 2r_{\min} .$$

Thus, the resulting cylinder is such that its height is equal to the diameter of its base.

Example 4. Discuss the behaviour of the function

$$f(x) = x^3 e^{-x}$$

at the point $x = 0$. By easy computation we obtain $f(0) = 0$, $f'(0) = 0$, $f''(0) = 0$, $f'''(0) > 0$. By Theorem 6, $f(x)$ is increasing at the point $x = 0$; by Theorem 5, it has a point of inflection there and crosses its tangent from below (its behaviour in the neighbourhood of that point is represented in Fig. 11.20).

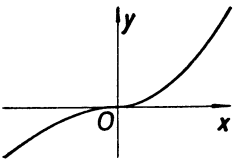


Fig. 11.20.

11.10. Taylor's Theorem

Theorem 1 (*Taylor's Theorem*). *Let $f(x)$ have continuous derivatives up to the n-th order inclusive in $[a, a + h]$ (or in $[a + h, a]$ if h is negative) and a continuous derivative of the $(n + 1)$-th order in $(a, a + h)$ (or in $(a + h, a)$). Then*

$$f(a + h) = f(a) + \frac{f'(a)}{1!} h + \frac{f''(a)}{2!} h^2 + \ldots + \frac{f^{(n)}(a)}{n!} h^n + R_{n+1} \tag{1}$$

where the expression for the remainder R_{n+1} may be put in one of these forms:

$$R_{n+1} = \frac{f^{(n+1)}(a + \vartheta h)}{(n+1)!} h^{n+1} \quad (0 < \vartheta < 1) \qquad (\textit{Lagrange form}), \quad (2)$$

$$R_{n+1} = \frac{f^{(n+1)}(a + \eta h)}{n!} (1 - \eta)^n h^{n+1} \quad (0 < \eta < 1) \quad (\textit{Cauchy form}), \quad (3)$$

$$R_{n+1} = \int_a^{a+h} f^{(n+1)}(t) \frac{(a + h - t)^n}{n!} \, dt \qquad (\textit{integral form}). \quad (4)$$

REMARK 1. If we write $h = x - a$, we obtain the frequently used form

$$f(x) = f(a) + \frac{f'(a)}{1!}(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \ldots + \frac{f^{(n)}(a)}{n!}(x - a)^n + R_{n+1} \quad (5)$$

where

$$R_{n+1} = \frac{f^{(n+1)}[a + \vartheta(x - a)]}{(n+1)!}(x - a)^{n+1} \quad (0 < \vartheta < 1) \qquad (6)$$

or

$$R_{n+1} = \frac{f^{(n+1)}[a + \eta(x - a)]}{n!}(1 - \eta)^n (x - a)^{n+1} \quad (0 < \eta < 1), \qquad (7)$$

or

$$R_{n+1} = \int_a^x f^{(n+1)}(t) \frac{(x - t)^n}{n!} \, dt. \qquad (8)$$

REMARK 2. In particular, if $a = 0$, then

$$f(x) = f(0) + \frac{f'(0)}{1!} x + \frac{f''(0)}{2!} x^2 + \ldots + \frac{f^{(n)}(0)}{n!} x^n + R_{n+1}, \qquad (9)$$

where it is sufficient to put $a = 0$ in (6), (7), (8), respectively, to obtain the expression for the remainder R_{n+1}. Formula (9) is called *Maclaurin's formula.*

REMARK 3. We use Taylor's or Maclaurin's Theorem in order to compute the values of a function or to express approximately a given function by means of a polynomial in the neighbourhood of the point a (or zero). Because we know the values of several functions at the origin and can then easily compute the values of the derivatives at this point, formula (9) finds the most frequent application; note, that x (or h in (1)) may be negative.

REMARK 4. If we consider only the first two terms on the right of (1), then we obtain (except for the remainder R_2) the replacing of the difference $f(a + h) - f(a)$ by the differential $hf'(a)$. For $n = 0$ we obtain from (1) the Mean-Value Theorem.

Example 1. Let us try to approximate the function $y = \sin x$ in the neighbourhood of the origin by means of a polynomial of degree 4 and let us estimate the error.

We apply (9); $f(x) = \sin x, f(0) = 0, f'(0) = 1, f''(0) = 0, f'''(0) = -1, f^{(4)}(0) = 0$, hence

$$\sin x = x - \frac{x^3}{3!} + R_5 . \tag{10}$$

According to formula 11, p. 380, $(\sin x)^{(5)} = \sin(x + \frac{5}{2}\pi) = \cos x$. According to (6), for $a = 0$

$$R_5 = \frac{\cos \vartheta x}{5!} x^5 \quad (0 < \vartheta < 1) . \tag{11}$$

Let us consider the interval $[-\frac{1}{10}, \frac{1}{10}]$. Because $|\cos \vartheta x| \leqq 1$, we shall have for all x in this interval $|R_5| \leqq (\frac{1}{10})^5/5! = 1/12{,}000{,}000$.

Example 2. Let us compute the approximate value of $\sin 3°$.

In radians

$$x = \frac{2\pi}{360} . 3 = \frac{\pi}{60} \doteq 0{\cdot}052{,}359{,}878 .$$

Further

$$\frac{1}{3!}\left(\frac{\pi}{60}\right)^3 \doteq 0{\cdot}000{,}023{,}925 ,$$

thus by (10)

$$\sin 3° \approx \frac{\pi}{60} - \frac{1}{3!}\left(\frac{\pi}{60}\right)^3 \doteq 0{\cdot}052{,}335{,}953$$

with an error (by (11)) of less than

$$\frac{1}{5!}\left(\frac{\pi}{60}\right)^5 \doteq 3{\cdot}3 \times 10^{-9} .$$

Theorem 2. *Let $f(x)$ have derivatives of all orders in $[a, x]$ (or in $[x, a]$ if $x < a$). Then a necessary and sufficient condition that the series*

$$f(a) + \frac{f'(a)}{1!}(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \ldots$$

converges and has the sum $f(x)$ is that $\lim_{n \to \infty} R_{n+1} = 0$ (for the specified x).

11.11. Approximate Expressions. Computation with Small Numbers

REMARK 1. If we take into consideration only the first few terms of the right-hand side of equation (11.10.1), p. 396, we obtain an approximating formula for the

evaluation of the number $f(a + h)$. If need be, we can use the formulae (11.10.2)–(11.10.4) or (11.10.6)–(11.10.8) for the eventual estimation of the error. Some frequently used approximations (we denote the approximations by $\approx$, ε is a relatively small number (in absolute value) not necessarily positive) are:

Theorem 1.

1. $(1 \pm \varepsilon)^n \approx 1 \pm n\varepsilon$, $(1 \pm \varepsilon)^2 \approx 1 \pm 2\varepsilon$, $\sqrt{(1 \pm \varepsilon)} \approx 1 \pm \frac{1}{2}\varepsilon$,

$$\frac{1}{1 \pm \varepsilon} \approx 1 \mp \varepsilon, \quad \frac{1}{(1 \pm \varepsilon)^2} \approx 1 \mp 2\varepsilon, \quad \frac{1}{\sqrt{(1 \pm \varepsilon)}} \approx 1 \mp \tfrac{1}{2}\varepsilon.$$

2. $a^\varepsilon \approx 1 + \varepsilon \ln a$, $e^\varepsilon \approx 1 + \varepsilon$.

3. $(1 \pm \varepsilon)(1 \pm \delta)(1 \pm \eta) \approx 1 \pm \varepsilon \pm \delta \pm \eta$,

$$\frac{(1 \pm \varepsilon)(1 \pm \delta)}{(1 \pm \eta)(1 \pm \varkappa)} \approx 1 \pm \varepsilon \pm \delta \mp \eta \mp \varkappa.$$

4. *For positive numbers $p \approx q$ the relation $\sqrt{(pq)} \approx \frac{1}{2}(p + q)$ holds.*

5. $\sin \varepsilon \approx \varepsilon$, $\cos \varepsilon \approx 1$, $\tan \varepsilon \approx \varepsilon$,

$\sin(x + \varepsilon) \approx \sin x + \varepsilon \cos x$, $\cos(x + \varepsilon) \approx \cos x - \varepsilon \sin x$,

$$\tan(x + \varepsilon) = \tan x + \frac{\varepsilon}{\cos^2 x},$$

$$e^\varepsilon \approx 1 + \varepsilon, \quad \ln(1 + \varepsilon) \approx \varepsilon, \quad \ln(x + \varepsilon) \approx \ln x + \frac{\varepsilon}{x},$$

$$\ln \frac{x + \varepsilon}{x - \varepsilon} \approx \frac{2\varepsilon}{x}.$$

} *first approximation*,

6. $$\sin \varepsilon \approx \varepsilon - \frac{\varepsilon^3}{3!}, \quad \cos \varepsilon \approx 1 - \frac{\varepsilon^2}{2!}, \quad \tan \varepsilon \approx \varepsilon + \frac{\varepsilon^3}{3},$$

$$e^\varepsilon \approx 1 + \varepsilon + \frac{\varepsilon^2}{2!} + \frac{\varepsilon^3}{3!},$$

$$\ln(1 + \varepsilon) \approx \varepsilon - \tfrac{1}{2}\varepsilon^2 + \tfrac{1}{3}\varepsilon^3,$$

$$\ln(x + \varepsilon) \approx \ln x + \frac{\varepsilon}{x} - \frac{1}{2}\frac{\varepsilon^2}{x^2} + \frac{1}{3}\frac{\varepsilon^3}{x^3},$$

$$\ln \frac{x + \varepsilon}{x - \varepsilon} \approx 2\left(\frac{\varepsilon}{x} + \frac{\varepsilon^3}{3x^3}\right).$$

} *higher approximations*.

(For higher approximations we have considered the first four terms of the right-hand side of equation (11.10.1)).

Example 1. Let us compute $\sqrt{9{,}986}$.

Using the relation $\sqrt{(1-\varepsilon)} \approx 1 - \frac{1}{2}\varepsilon$, we have:

$$\sqrt{9{,}986} = \sqrt{(10{,}000 - 14)} = 100\sqrt{\left(1 - \frac{14}{10{,}000}\right)} \approx 100\left(1 - \frac{7}{10{,}000}\right) = 99{\cdot}93\,.$$

Estimation of error: If we take $f(x) = \sqrt{(1-x)}$, then by (11.10.2) ($a = 0$, $h = 0{\cdot}001{,}4$, $0 < \vartheta < 1$)

$$|R_2| = \left|\frac{[\sqrt{(1-x)}]''_{x=0.001,4\vartheta}}{2!}\, 0{\cdot}001{,}4^2\right| = \frac{0{\cdot}001{,}4^2}{4\,.\,2!\,[\sqrt{(1-x)}]^3_{x=0.001,4\vartheta}} < \frac{1}{4\,.\,10^6}\,.$$

Because the whole expression $\sqrt{(1 - 0{\cdot}001{,}4)}$ is multiplied by the number 100, we obtain the result $\sqrt{9{,}986} \doteq 99{\cdot}93$ with an error less than $10^{-4}/4$.

11.12. Survey of Some Important Formulae from Chapter 11

(Cf. also Theorems 11.4.9, 11.5.3, 11.5.8, 11.5.9.)

1. $\lim\limits_{x\to a}[f(x) \pm g(x)] = \lim\limits_{x\to a} f(x) \pm \lim\limits_{x\to a} g(x)$, $\lim\limits_{x\to a} k f(x) = k \lim\limits_{x\to a} f(x)$,

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

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{\lim\limits_{x\to a} f(x)}{\lim\limits_{x\to a} g(x)} \quad (\lim_{x\to a} g(x) \neq 0) \quad \text{(Theorem 11.4.4)}\,.$$

2. $\lim\limits_{x\to 0+} \frac{1}{x} = +\infty$, $\lim\limits_{x\to 0-} \frac{1}{x} = -\infty$,

$$\lim_{x\to 0+} \frac{a}{x} = \begin{cases} +\infty & \text{when } a > 0 \\ -\infty & \text{when } a < 0\,, \end{cases} \quad \lim_{x\to 0-} \frac{a}{x} = \begin{cases} -\infty & \text{when } a > 0\,, \\ +\infty & \text{when } a < 0\,. \end{cases}$$

3. $\lim\limits_{x\to a} \frac{f(x)}{g(x)} = \lim\limits_{x\to a} \frac{f'(x)}{g'(x)}$, *if at the same time* $\lim\limits_{x\to a} f(x) = \lim\limits_{x\to a} g(x) = 0$,

or if $\lim\limits_{x\to a} |g(x)| = +\infty$ (Theorems 11.8.1, 11.8.2).

4. $[f(x) \pm g(x)]' = f'(x) \pm g'(x)$, $[f(x)\, g(x)]' = f'(x)\, g(x) + f(x)\, g'(x)$,

$$\left[\frac{f(x)}{g(x)}\right]' = \frac{f'(x)\, g(x) - f(x)\, g'(x)}{g^2(x)} \quad \text{(Theorem 11.5.4)}\,,$$

$\frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx}$ (*differentiation of composite functions*, Theorem 11.5.5).

5. $\frac{dy}{dx} = \frac{1}{dx/dy}$ (*differentiation of inverse functions*, Theorem 11.5.6).

6. $(uv)^{(n)} = u^{(n)}v^{(0)} + \binom{n}{1}u^{(n-1)}v^{(1)} + \binom{n}{2}u^{(n-2)}v^{(2)} + \ldots + u^{(0)}v^{(n)}$.

7. $\frac{(u_1u_2 \ldots u_n)'}{u_1u_2 \ldots u_n} = \frac{u_1'}{u_1} + \frac{u_2'}{u_2} + \ldots + \frac{u_n'}{u_n}$, $u_1u_2 \ldots u_n \neq 0$.

8. $[f(x)^{g(x)}]' = f(x)^{g(x)}\left[g'(x)\ln f(x) + g(x)\frac{f'(x)}{f(x)}\right]$ ($f(x) > 0$; Example 11.5.7).

9. $df(x) = f'(x)\,dx$ (§ 11.6).

10. $f'(a) = 0$, $f''(a) > 0 \Rightarrow f(x)$ *has a strict relative minimum at the point* a,
$f'(a) = 0$, $f''(a) < 0 \Rightarrow f(x)$ *has a strict relative maximum at the point* a
(Theorem 11.9.4).

11. $f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_{n+1}$,

where e.g.

$$R_{n+1} = \frac{f^{(n+1)}[a + \vartheta(x-a)]}{(n+1)!}(x-a)^{n+1} \quad (0 < \vartheta < 1)$$

(*Taylor's formula*, Theorem 11.10.1, Remark 11.10.1).

12. $e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} + \frac{x^{n+1}}{(n+1)!}e^{\vartheta x}$, $0 < \vartheta < 1$;

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots + (-1)^{n-1}\frac{x^{2n-1}}{(2n-1)!} + (-1)^n\frac{x^{2n+1}}{(2n+1)!}\cos \vartheta x, \quad 0 < \vartheta < 1;$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \ldots + (-1)^n\frac{x^{2n}}{(2n)!} + (-1)^{n+1}\frac{x^{2n+2}}{(2n+2)!}\cos \vartheta x, \quad 0 < \vartheta < 1.$$

All three formulae hold for every x.

13. $1 + x > 0$, $x \neq 0 \Rightarrow (1+x)^n > 1 + nx$, n *any positive integer greater than* 1.

12. FUNCTIONS OF TWO OR MORE VARIABLES

By KAREL REKTORYS

References: [4], [17], [26], [31], [52], [54], [57], [59], [68], [80], [91], [96], [111], [112], [119], [122], [123], [134], [142], [148], [158], [160], [169].

12.1. Functions of Several Variables. Composite Functions. Limit, Continuity

Definition 1. Let us consider a set M of points (x, y) in the xy-plane (this set is most often a *region*; see Remark 1). We say that a *real function of real variables* x, y *is defined in* (*on* or *over*) *the set* M if a rule is given according to which *exactly one* real number z is assigned to each point (x, y) of M. The set M is called the *domain of definition of the function*. Similarly a function of n variables

$$z = f(x_1, x_2, \ldots, x_n)$$

may be defined.

Usually we denote functions by letters f, g, etc. At an arbitrary point $(x, y) \in M$ or $(x_1, x_2, \ldots, x_n) \in M$ we then write

$$z = f(x, y) \quad \text{or} \quad z = g(x_1, x_2, \ldots, x_n),$$

etc. Cf. also Remark to Definition 11.1.1, p. 359.

REMARK 1. The domain of definition is most often a *region* or a *closed region* (that is, a region with its boundary included). Both these concepts are defined in § 22.1. (Thus, for example, the interior of a circle, the interior of an ellipse, the whole xy-plane, etc., are regions. An example of a closed region is a circle with its circumference included, the so-called *closed circle*.)

Example 1. The function $z = x^2 y$ is defined in the whole plane. The function

$$z = \sqrt{(1 - x^2 - y^2)}$$

is defined at all points for which

$$1 - x^2 - y^2 \geqq 0 \quad \text{or} \quad x^2 + y^2 \leqq 1 .$$

For its domain of definition the closed region $x^2 + y^2 \leqq 1$ may be taken, i.e. the closed circle with its centre at the origin and radius equal to 1.

Similarly the function

$$z = \sqrt{(1 - x_1^2 - x_2^2 - \ldots - x_n^2)}$$

is defined in a "closed n-dimensional sphere"

$$x_1^2 + x_2^2 + \ldots + x_n^2 \leqq 1 .$$

REMARK 2. The geometrical interpretation of a function $z = f(x, y)$ (in so far as the function is a "reasonable" one) is a surface in three-dimensional space. Functions of more than three variables can no longer be represented in such a simple way.

REMARK 3. The function

$$z = h(f(x, y), g(x, y)) \tag{1}$$

is called a *composite function*, composed of the functions

$$u = f(x, y), \quad v = g(x, y), \tag{2}$$

$$z = h(u, v). \tag{3}$$

The functions (2) are defined in the set M, the function (3) in the set N and it is required that for every point $(x, y) \in M$ the relation $(u, v) \in N$ be satisfied. For a given point $(x, y) \in M$ we can then compute by (2) the values of u and v and by (3) the corresponding value of z.

Example 2. The function $z = (1 + x^2 + y^2)^{\sin xy}$ may be considered as a composite function by means of the relations $z = u^v$, $u = 1 + x^2 + y^2$, $v = \sin xy$. For the set M the whole plane xy may be taken because the function $z = u^v$ is defined at all points (u, v), where $u > 0$, and the function $u = 1 + x^2 + y^2$ is positive for all x, y.

Example 3. The function $z = \sqrt{(1 - x^2 \sin x)}$ may be considered as a function composed of the functions $z = \sqrt{(1 - uv)}$, $u = x^2$, $v = \sin x$. (This function is a function of only one variable x.)

Definition 2. The *distance between two points* (x_1, y_1), (x_2, y_2) is, by definition, the number

$$d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$$

(we use cartesian coordinates unless otherwise stated). Similarly, the distance between the points $(x_1, x_2, \ldots, x_n)$, $(y_1, y_2, \ldots, y_n)$ is defined as the number

$$d = \sqrt{[(y_1 - x_1)^2 + (y_2 - x_2)^2 + \ldots + (y_n - x_n)^2]}\,.$$

Definition 3. A set of all points, the distance of which from a point P is smaller than δ, is called a *δ-neighbourhood of the point P.* (Note that the point P itself belongs to the δ-neighbourhood of the point P.) It is often denoted by $U_\delta(P)$.

REMARK 4. In the xy-plane a δ-neighbourhood of a point P is formed by all points that lie inside the circle with centre at the point P and radius δ. In three-dimensional space a δ-neighbourhood is the interior of a sphere, etc.

Definition 4. We say that a function $z = f(x, y)$ has a *limit A at a point $P(x_0, y_0)$* if to every (arbitrarily small) number $\varepsilon > 0$, there exists a number $\delta > 0$ (depending, in general, on the choice of the number ε) such that for all points $(x, y) \neq P$ in the δ-neighbourhood of the point P the relation

$$|f(x, y) - A| < \varepsilon$$

holds.

The notations

$$\lim_{(x,y)\to P} f(x, y) = A\,, \qquad \lim_{(x,y)\to(x_0,y_0)} f(x, y) = A\,, \qquad \lim_{\substack{x\to x_0 \\ y\to y_0}} f(x, y) = A$$

are used.

The limit of a function of several variables is defined similarly.

REMARK 5. The intuitive meaning of Definition 4: $f(x, y)$ has a limit A at a point P, if $f(x, y)$ is sufficiently close to the value A for all points (x, y) that are sufficiently close to the point P, except possibly P itself.

Definition 5. We say that $f(x, y)$ is *continuous at a point* (x_0, y_0) if it is defined at this point and if, corresponding to an arbitrary $\varepsilon > 0$, there exists a $\delta > 0$, such that for all points (x, y) in the δ-neighbourhood of the point (x_0, y_0) the relation $|f(x, y) - f(x_0, y_0)| < \varepsilon$ holds. The continuity of a function of several variables is defined similarly.

REMARK 6. It follows from the definition of continuity that a function $f(x, y)$ continuous at the point (x_0, y_0) is defined in a definite neighbourhood of the point (x_0, y_0) (this point included).

Theorem 1. *Let a function $f(x, y)$ be defined at a point (x_0, y_0). Then it is continuous at that point if, and only if,* $\lim_{(x,y)\to(x_0,y_0)} f(x, y) = f(x_0, y_0)$. *Similarly for functions of several variables.*

Definition 6. If a function is continuous at every point of a region (more generally: of a set) M we say that it is *continuous in* (or *on*) M.

REMARK 7. When we say that $f(x, y)$ is continuous in a closed region $\bar{O}$, we mean that it is continuous in O and that on the boundary it is *continuous with regard to the points from* $\bar{O}$ (i.e. if the point (x_0, y_0) lies on the boundary, then we consider in Definition 5 only those points from the δ-neighbourhood of the point (x_0, y_0) that belong to $\bar{O}$). The same remark holds for functions of several variables.

Example 4. The function $z = \sqrt{(1 - x^2 - y^2)}$ is continuous in the closed region given by $x^2 + y^2 \leqq 1$, i.e. in a closed circle.

REMARK 8. We often have to deal with the case where $f(x, y)$ is defined and continuous in the region O and at the same time can be defined on the boundary h of this region in such a manner that the extended function is continuous in $\bar{O}$. Then we say that $f(x, y)$ is *continuously extensible on the boundary* h.

In a similar way we define continuous extensibility on the boundary for functions of several variables.

We also often come across the case where the given function is continuous in a closed region $\bar{O}$ and has in $\bar{O}$ continuous partial derivatives of the first order. It is to be understood that these derivatives are continuous in O and continuously extensible on the boundary. Similarly, we may speak of the continuity of higher derivatives in $\bar{O}$.

When we say that a function $f(x, y)$ is piecewise continuous in a region O of the type A (Definition 14.1.2) we shall mean that it is possible to divide the region O by means of a finite number of simple finite piecewise smooth curves (Definition 14.1.1) into a (finite) number of regions O_n of the type A so that $f(x, y)$ is continuous in every region O_n and continuously extensible on its boundary. (For example, the function considered in Example 14.1.2 is piecewise continuous in O.)

Similarly we define a piecewise continuous function in a three-dimensional region of the type A.

A function is said to be *piecewise smooth in* O when the function and its partial derivatives of the first order are piecewise continuous in O.

Further definitions and theorems (similar to those given in § 11.3, 11.4 for functions of one variable) can be formulated for functions of two or more variables.

Of most frequent application are the following theorems:

Theorem 2. *If* $f(x, y)$ *and* $g(x, y)$ *possess the limits* A *and* B, *respectively, at the point* (x_0, y_0), *then the functions* $k \cdot f(x, y)$ $(k = \text{const.})$, $f(x, y) \pm g(x, y)$, $f(x, y) \cdot g(x, y)$ *and* (*if* $B \neq 0$) $f(x, y)/g(x, y)$ *also possess a limit at the point* (x_0, y_0) *and the relations*

$$\lim_{(x,y)\to(x_0,y_0)} kf(x, y) = kA \ (k \text{ a constant}), \quad \lim_{(x,y)\to(x_0,y_0)} [f(x, y) \pm g(x, y)] = A \pm B,$$

$$\lim_{(x,y)\to(x_0,y_0)} f(x, y)\, g(x, y) = AB, \quad \lim_{(x,y)\to(x_0,y_0)} \frac{f(x, y)}{g(x, y)} = \frac{A}{B}.$$

hold.

A similar theorem holds in the n-dimensional case. A similar statement holds also for continuity.

Theorem 3. *A continuous function of continuous functions is itself continuous. In more detail (for functions of two variables): If* $u = f(x, y)$, $v = g(x, y)$ *are continuous at the point* (x_0, y_0) *and if* $z = h(u, v)$ *is continuous at the corresponding point* (u_0, v_0), *then* $z = h(f(x, y), g(x, y))$ *is continuous (considered as a function of the variables* x, y*) at the point* (x_0, y_0).

REMARK 9. On the basis of the last two theorems we may decide on the continuity of many functions which we come across in applications. In particular all polynomials in x and y, all rational functions (provided the denominator is non-zero), all functions composed of continuous functions (e.g., the function $z = xy \sin^2 x$), etc., are continuous.

Theorem 4. *If* $f(x, y)$ *is continuous in a region* O *and if* (x_1, y_1), (x_2, y_2) *are any two points in this region, then* $f(x, y)$ *takes on in* O *every value between* $f(x_1, y_1)$ *and* $f(x_2, y_2)$.

Theorem 5. *A function that is continuous in a bounded closed region* $\bar{O}$ *takes on a greatest value at least at one point* $(x_0, y_0) \in \bar{O}$ *(that is:* $f(x_0, y_0) \geqq f(x, y)$ *for all points* $(x, y) \in \bar{O}$*) and a least value at least at one point* $(x_1, y_1) \in \bar{O}$.

Theorem 6. *A function* $f(x, y)$ *that is continuous in a bounded closed region* $\bar{O}$ *is uniformly continuous there. This means: To an arbitrary* $\varepsilon > 0$ *there exists* $\delta > 0$ *depending only on the choice of the number* ε *(thus the same for the whole closed region* $\bar{O}$*) such that*

$$|f(x_2, y_2) - f(x_1, y_1)| < \varepsilon$$

for every pair of points $(x_1, y_1) \in \bar{O}$, $(x_2, y_2) \in \bar{O}$, *the distance between which is smaller than* δ.

REMARK 10. Theorems similar to Theorems 4, 5, 6 hold for functions of several variables.

12.2. Partial Derivatives. Change of Order of Differentiation

Definition 1. We say that a function $z = f(x, y)$ has a *partial derivative with respect to* x at the point (x_0, y_0) if the (finite) limit

$$\lim_{h \to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h}$$

exists.

The following notations are used:

$$\frac{\partial f}{\partial x}(x_0, y_0), \quad \frac{\partial z}{\partial x}(x_0, y_0), \quad f_x(x_0, y_0), \quad f'_x(x_0, y_0).$$

Similarly

$$\frac{\partial f}{\partial y}(x_0, y_0) = \frac{\partial z}{\partial y}(x_0, y_0) = f_y(x_0, y_0) = f'_y(x_0, y_0) = \lim_{k \to 0} \frac{f(x_0, y_0 + k) - f(x_0, y_0)}{k}.$$

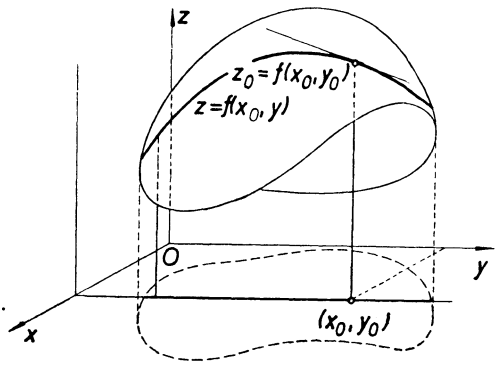


Fig. 12.1.

REMARK 1. The geometrical interpretation of the partial derivative is given in Fig. 12.1. The section of the surface $z = f(x, y)$ by the plane $x = x_0$ (a plane parallel to the coordinate plane $x = 0$) is the curve $z = f(x_0, y)$ (z depends only on y if x_0 is constant); $\partial f / \partial y(x_0, y_0)$ is the slope of the tangent to this curve at the point (x_0, y_0, z_0) indicated in Fig. 12.1.

REMARK 2. *Partial derivatives of functions of several variables* are defined in a similar manner. For example,

$$\frac{\partial f}{\partial x_2}(x_1, x_2, \ldots, x_n) = \lim_{h_2 \to 0} \frac{f(x_1, x_2 + h_2, x_3, \ldots, x_n) - f(x_1, x_2, x_3, \ldots, x_n)}{h_2}.$$

REMARK 3. To compute a partial derivative, we differentiate the given function regarding it as a function of the *single* variable, with respect to which the derivative is required. The other variables are treated as though they were constants.

Example 1.

$$z = x^3 y, \quad \frac{\partial z}{\partial x} = 3x^2 y$$

(during this differentiation y is kept constant);

$$z = \sin(xy), \quad \frac{\partial z}{\partial y} = x \cos(xy)$$

(during this differentiation x is constant).

REMARK 4. $\partial f/\partial x$, $\partial f/\partial y$ are again functions of x, y. *Derivatives of the second order* are defined by the relations:

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right), \quad \frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right), \quad \frac{\partial^2 f}{\partial x\,\partial y} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right), \quad \frac{\partial^2 f}{\partial y\,\partial x} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right).$$

The last two derivatives are called *mixed.*

The derivatives of second order of functions of several variables are defined similarly and so are those of higher orders. The order of differentiation is indicated by the order of the symbols ∂x, ∂y in the denominator. For example, $\partial^3 f/\partial x^2\,\partial y$ means that the function z has been differentiated first with respect to y, then with respect to x and then again with respect to x.

Example 2. Find $\partial^3 z/\partial y\,\partial x^2$, if $z = y^2 \sin x$. We have

$$\frac{\partial z}{\partial x} = y^2 \cos x\,, \quad \frac{\partial^2 z}{\partial x^2} = -y^2 \sin x\,, \quad \frac{\partial^3 z}{\partial y\,\partial x^2} = -2y \sin x\,.$$

Theorem 1. *Change of Order of Differentiation (Interchangeability of Mixed Derivatives). If*

$$\frac{\partial^2 f}{\partial x\,\partial y}\,, \quad \frac{\partial^2 f}{\partial y\,\partial x}$$

are both continuous at the point (x_0, y_0), *then they are equal at this point, i.e. the relation*

$$\frac{\partial^2 f}{\partial x\,\partial y} = \frac{\partial^2 f}{\partial y\,\partial x}$$

holds.

REMARK 5. Under similar assumptions theorems on the interchangeability of mixed derivatives of higher orders or on the interchangeability of mixed partial derivatives of functions of several variables also hold. If, for instance, $\partial^4 f/\partial y^2\,\partial z\,\partial x$ and $\partial^4 f/\partial x\,\partial y^2\,\partial z$ are both continuous at the point (x_0, y_0, z_0), then they are equal at that point. If the equality holds at every point of a given domain then, of course, the corresponding derivatives are equal in the whole domain.

Example 3. Let us consider the function $z = y^2 \sin x$ from Example 2. Then

$$\frac{\partial z}{\partial y} = 2y \sin x\,, \quad \frac{\partial^2 z}{\partial x\,\partial y} = 2y \cos x\,, \quad \frac{\partial^3 z}{\partial x^2\,\partial y} = -2y \sin x\,,$$

hence

$$\frac{\partial^3 z}{\partial x^2\,\partial y} = \frac{\partial^3 z}{\partial y\,\partial x^2}\,.$$

12.3. Total Differential

Definition 1. The function $z = f(x, y)$ is said to be *differentiable at the point* (x_0, y_0) when its increment $\Delta z = f(x_0 + h, y_0 + k) - f(x_0, y_0)$ can be expressed, in a certain neighbourhood of the point (x_0, y_0), in the form

$$\Delta z \equiv f(x_0 + h, y_0 + k) - f(x_0, y_0) = Ah + Bk + \varrho\tau(h, k), \tag{1}$$

where A, B are constants, $\varrho = \sqrt{(h^2 + k^2)}$ and

$$\lim_{\substack{h \to 0 \\ k \to 0}} \tau(h, k) = 0. \tag{2}$$

Note that in general τ contains h, k, x_0, y_0; x_0, y_0 are treated as constants. (See Definition 11.6.1 and Remark to Definition 11.6.2, p. 385.)

Theorem 1. *If $f(x, y)$ is differentiable at the point (x_0, y_0), then it possesses partial derivatives at (x_0, y_0) and the relations*

$$A = \frac{\partial f}{\partial x}(x_0, y_0), \quad B = \frac{\partial f}{\partial y}(x_0, y_0)$$

hold.

REMARK 1. If we pass to the customary notation $h = \mathrm{d}x$, $k = \mathrm{d}y$, we have

$$\Delta z = \frac{\partial f}{\partial x}\,\mathrm{d}x + \frac{\partial f}{\partial y}\,\mathrm{d}y + \varrho\tau(\mathrm{d}x, \mathrm{d}y). \tag{3}$$

Definition 2. If $f(x, y)$ is differentiable at (x_0, y_0), then the expression

$$\mathrm{d}z = \frac{\partial f}{\partial x}\,\mathrm{d}x + \frac{\partial f}{\partial y}\,\mathrm{d}y$$

is called the *total differential* of the function $z = f(x, y)$.

Theorem 2. *If $f(x, y)$ is differentiable at the point (x_0, y_0), then it is continuous at that point.*

REMARK 2. If $f(x, y)$ has partial derivatives of the first order at the point (x_0, y_0) then it need not be continuous at that point as a function of both variables x, y (it is continuous only as a function of the variable x on the straight line parallel to the x-axis drawn through the point (x_0, y_0) and as a function of the variable y on the straight line parallel to y-axis). This can be shown, for example, by the function

$$z(x, y) = \frac{xy}{x^2 + y^2} \quad \text{for} \quad (x, y) \neq (0, 0), \quad z(0, 0) = 0$$

which possesses derivatives of the first order at the origin (equal to zero) but is not continuous at this point. This example shows (see Theorem 2) that a function which possesses partial derivatives of the first order at (x_0, y_0) need not be differentiable at that point. The following theorem, however, holds:

Theorem 3. *If*

$$\frac{\partial f}{\partial x}, \quad \frac{\partial f}{\partial y}$$

are continuous at (x_0, y_0), *then* $f(x, y)$ *is differentiable at* (x_0, y_0) (*and therefore also continuous*).

Theorem 4. *If* $f(x, y)$ *is differentiable at* (x_0, y_0), *then the surface* $z = f(x, y)$ *possesses a tangent plane at the point* (x_0, y_0, z_0) (*where* $z_0 = f(x_0, y_0)$). *Its equation is*

$$z - z_0 = \left(\frac{\partial f}{\partial x}\right)_0 (x - x_0) + \left(\frac{\partial f}{\partial y}\right)_0 (y - y_0),$$

where

$$\left(\frac{\partial f}{\partial x}\right)_0 = \frac{\partial f}{\partial x}(x_0, y_0), \quad \left(\frac{\partial f}{\partial y}\right)_0 = \frac{\partial f}{\partial y}(x_0, y_0).$$

REMARK 3. In the same way as we approximate, in the case of a function of *one* variable, the increment of the function by its differential (geometrically: we substitute the increment on the tangent for the increment of the function) so here we approximate the increment of the function by its total differential (geometrically: we substitute the increment on the tangent plane for the increment of the function).

Example 1. $f(x, y) = x^3 + 4y^3$. Let us find the approximate value of $f(1{\cdot}11; 0{\cdot}58)$.

First, we have $f(1; 0{\cdot}5) = 1{\cdot}5$. If we compute the total differential for $\mathrm{d}x = 0{\cdot}11$, $\mathrm{d}y = 0{\cdot}08$ at this point (i.e. at the point $x_0 = 1$, $y_0 = 0{\cdot}5$), we obtain: $\mathrm{d}z = 3x^2\,\mathrm{d}x + 12y^2\,\mathrm{d}y = 3 \,.\, 0{\cdot}11 + 12 \,.\, 0{\cdot}5^2 \,.\, 0{\cdot}08 = 0{\cdot}57$. Hence $f(1{\cdot}11; 0{\cdot}58) \approx f(1; 0{\cdot}5) + \mathrm{d}z = 1{\cdot}5 + 0{\cdot}57 = 2{\cdot}07$. (The exact result is $f(1{\cdot}11; 0{\cdot}58) = 2{\cdot}148\,079$.)

Theorem 5. *If the functions*

$$P(x, y), \quad Q(x, y), \quad \frac{\partial P}{\partial y}(x, y), \quad \frac{\partial Q}{\partial x}(x, y)$$

are continuous in a simply-connected region O, *then a necessary and sufficient condition that the expression*

$$P(x, y)\,\mathrm{d}x + Q(x, y)\,\mathrm{d}y$$

be the total differential of a function $f(x, y)$ *in* O *is that the relation*

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$$

holds in O.

The conditions for a similar expression

$$P(x, y, z)\,\mathrm{d}x + Q(x, y, z)\,\mathrm{d}y + R(x, y, z)\,\mathrm{d}z$$

to be the total differential of a function $f(x, y, z)$ *are*

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}, \quad \frac{\partial P}{\partial z} = \frac{\partial R}{\partial x}, \quad \frac{\partial Q}{\partial z} = \frac{\partial R}{\partial y} \quad \textit{(simultaneously) in } O\,.$$

(For a more detailed treatment see § 14.7.)

Definition 3. Let $z = f(x, y)$ have a total differential in a neighbourhood of a point (x_0, y_0) and let the partial derivatives

$$\frac{\partial f}{\partial x}(x, y), \quad \frac{\partial f}{\partial y}(x, y)$$

have a total differential at the point (x_0, y_0). Then we say that $f(x, y)$ has a *total differential of the second order* (briefly a *second differential*). By this differential we understand the expression

$$\mathrm{d}^2 z = h^2 \frac{\partial^2 f}{\partial x^2}(x_0, y_0) + 2hk \frac{\partial^2 f}{\partial x\,\partial y}(x_0, y_0) + k^2 \frac{\partial^2 f}{\partial y^2}(x_0, y_0)\,.$$

Instead of h, k we often write $\mathrm{d}x$, $\mathrm{d}y$.

REMARK 4. Formally, we obtain the second differential as the differential of the first differential at the point (x_0, y_0) regarding h and k as constants (cf. Definition 11.6.3):

$$\mathrm{d}^2 z = \mathrm{d}\left(\frac{\partial f}{\partial x} h + \frac{\partial f}{\partial y} k\right) = h\,\mathrm{d}\left(\frac{\partial f}{\partial x}\right) + k\,\mathrm{d}\left(\frac{\partial f}{\partial y}\right) =$$

$$= h\left(h \frac{\partial^2 f}{\partial x^2} + k \frac{\partial^2 f}{\partial y\,\partial x}\right) + k\left(h \frac{\partial^2 f}{\partial x\,\partial y} + k \frac{\partial^2 f}{\partial y^2}\right) =$$

$$= h^2 \frac{\partial^2 f}{\partial x^2} + 2hk \frac{\partial^2 f}{\partial x\,\partial y} + k^2 \frac{\partial^2 f}{\partial y^2} = \left(h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y}\right)^2 f\,.$$

We have used here the symbolic operator notation which is convenient especially in the case of higher differentials.

Similarly we can also define higher differentials:

Definition 4. Let $f(x, y)$ and all its partial derivatives up to the $(n-2)$-th order have a total differential in a neighbourhood of the point (x_0, y_0). Let the partial derivatives of the $(n-1)$-th order have a total differential at the point (x_0, y_0). Then we say that the function $f(x, y)$ has *at the point* (x_0, y_0) *a total differential of the n-th order* (briefly an *n-th differential*) and by this differential we mean the expression

$$\mathrm{d}^n z = \left(h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y}\right)^n f =$$

$$= h^n \frac{\partial^n f}{\partial x^n} + \binom{n}{1} h^{n-1} k \frac{\partial^n f}{\partial x^{n-1} \partial y} + \ldots + \binom{n}{n-1} h k^{n-1} \frac{\partial^n f}{\partial x \partial y^{n-1}} + k^n \frac{\partial^n f}{\partial y^n}.$$

REMARK 5. In a similar way we define a *differentiable function of n variables.* Its m-th total differential is given by the formula

$$\mathrm{d}^m z = \left(h_1 \frac{\partial}{\partial x_1} + h_2 \frac{\partial}{\partial x_2} + \ldots + h_n \frac{\partial}{\partial x_n}\right)^m f.$$

REMARK 6. In contradistinction to the total differential we often speak of the *partial differential of the function* $z = f(x, y)$ *with respect to* x or y:

$$\mathrm{d}_x z = \frac{\partial f}{\partial x} \mathrm{d}x \quad \text{or} \quad \mathrm{d}_y z = \frac{\partial f}{\partial y} \mathrm{d}y.$$

12.4. Differentiation of Composite Functions

Theorem 1. *Let the functions* $u = f(x, y)$, $v = g(x, y)$ *be differentiable at the point* (x_0, y_0). (*In order that the functions be differentiable it is sufficient by Theorem* 12.3.3 *that they have continuous partial derivatives at that point.*) *Let the function* $z = h(u, v)$ *be differentiable at the corresponding point* (u_0, v_0) (where $u_0 = f(x_0, y_0)$, $v_0 = g(x_0, y_0)$). *Then the composite function* (see Remark 12.1.3)

$$z = h(f(x, y), g(x, y))$$

is a differentiable function (*as a function of the variables* x, y) *at the point* (x_0, y_0) *and*

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial x}, \quad \frac{\partial z}{\partial y} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial y}. \tag{1}$$

In somewhat more precise notation:

$$\frac{\partial z}{\partial x} = \frac{\partial h}{\partial u}\frac{\partial f}{\partial x} + \frac{\partial h}{\partial v}\frac{\partial g}{\partial x}, \quad \frac{\partial z}{\partial y} = \frac{\partial h}{\partial u}\frac{\partial f}{\partial y} + \frac{\partial h}{\partial v}\frac{\partial g}{\partial y}, \tag{2}$$

where the derivatives of the functions f, g are computed at the point (x_0, y_0) and those of the function h at the point (u_0, v_0).

Under similar assumptions regarding the functions

$$z = h(u_1, u_2, \ldots, u_n), \quad u_1 = f_1(x_1, x_2, \ldots, x_m), \quad \ldots, \quad u_n = f_n(x_1, x_2, \ldots, x_m)$$

the relation

$$\frac{\partial z}{\partial x_k} = \frac{\partial z}{\partial u_1}\frac{\partial u_1}{\partial x_k} + \frac{\partial z}{\partial u_2}\frac{\partial u_2}{\partial x_k} + \ldots + \frac{\partial z}{\partial u_n}\frac{\partial u_n}{\partial x_k}, \quad (k = 1, 2, \ldots, m)$$

holds.

Example 1. $z = (y \sin x)^{e^{x^2y}}$ $(y \sin x > 0)$. Let us put

$$z = u^v, \quad u = y \sin x, \quad v = e^{x^2y}.$$

By (1) or (2)

$$\frac{\partial z}{\partial x} = vu^{v-1}\, y \cos x + u^v \ln u\, 2xy\, e^{x^2y}$$

$$= u^v \left(\frac{e^{x^2y}}{u}\, y \cos x + 2xy\, e^{x^2y} \ln u\right)$$

$$= (y \sin x)^{e^{x^2y}} \cdot e^{x^2y} \cdot [\cot x + 2xy \ln (y \sin x)].$$

REMARK 1. When computing the *second* derivatives it is necessary to bear in mind that the functions $\partial h/\partial u$ and $\partial h/\partial v$ are functions of u, v and thus *they are to be differentiated with respect to x or y as composite functions* (according to (1), (2)). For example, using (2):

$$\frac{\partial^2 z}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial z}{\partial x}\right) = \frac{\partial}{\partial x}\left[\frac{\partial h}{\partial u}\frac{\partial f}{\partial x} + \frac{\partial h}{\partial v}\frac{\partial g}{\partial x}\right] =$$

$$= \frac{\partial}{\partial x}\left(\frac{\partial h}{\partial u}\right)\frac{\partial f}{\partial x} + \frac{\partial h}{\partial u}\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) + \frac{\partial}{\partial x}\left(\frac{\partial h}{\partial v}\right)\frac{\partial g}{\partial x} + \frac{\partial h}{\partial v}\frac{\partial}{\partial x}\left(\frac{\partial g}{\partial x}\right) =$$

$$= \left[\frac{\partial}{\partial u}\left(\frac{\partial h}{\partial u}\right)\frac{\partial f}{\partial x} + \frac{\partial}{\partial v}\left(\frac{\partial h}{\partial u}\right)\frac{\partial g}{\partial x}\right]\frac{\partial f}{\partial x} + \frac{\partial h}{\partial u}\frac{\partial^2 f}{\partial x^2} +$$

$$+ \left[\frac{\partial}{\partial u}\left(\frac{\partial h}{\partial v}\right)\frac{\partial f}{\partial x} + \frac{\partial}{\partial v}\left(\frac{\partial h}{\partial v}\right)\frac{\partial g}{\partial x}\right]\frac{\partial g}{\partial x} + \frac{\partial h}{\partial v}\frac{\partial^2 g}{\partial x^2} =$$

$$= \frac{\partial^2 h}{\partial u^2}\left(\frac{\partial f}{\partial x}\right)^2 + 2\frac{\partial^2 h}{\partial u\, \partial v}\frac{\partial f}{\partial x}\frac{\partial g}{\partial x} + \frac{\partial^2 h}{\partial v^2}\left(\frac{\partial g}{\partial x}\right)^2 + \frac{\partial h}{\partial u}\frac{\partial^2 f}{\partial x^2} + \frac{\partial h}{\partial v}\frac{\partial^2 g}{\partial x^2}.$$

(We have already applied the interchangeability of the order of differentiation, $\partial^2 h/\partial u\, \partial v = \partial^2 h/\partial v\, \partial u$.) Similarly,

$$\frac{\partial^2 z}{\partial y^2} = \frac{\partial^2 h}{\partial u^2}\left(\frac{\partial f}{\partial y}\right)^2 + 2\frac{\partial^2 h}{\partial u\,\partial v}\frac{\partial f}{\partial y}\frac{\partial g}{\partial y} + \frac{\partial^2 h}{\partial v^2}\left(\frac{\partial g}{\partial y}\right)^2 + \frac{\partial h}{\partial u}\frac{\partial^2 f}{\partial y^2} + \frac{\partial h}{\partial v}\frac{\partial^2 g}{\partial y^2},$$

$$\frac{\partial^2 z}{\partial x\,\partial y} = \frac{\partial^2 h}{\partial u^2}\frac{\partial f}{\partial x}\frac{\partial f}{\partial y} + \frac{\partial^2 h}{\partial u\,\partial v}\left(\frac{\partial f}{\partial x}\frac{\partial g}{\partial y} + \frac{\partial f}{\partial y}\frac{\partial g}{\partial x}\right) +$$

$$+ \frac{\partial^2 h}{\partial v^2}\frac{\partial g}{\partial x}\frac{\partial g}{\partial y} + \frac{\partial h}{\partial u}\frac{\partial^2 f}{\partial x\,\partial y} + \frac{\partial h}{\partial v}\frac{\partial^2 g}{\partial x\,\partial y}.$$

REMARK 2. We often come across the case where the composite function is given as follows:

$$z = h(x, y, u, v), \quad u = f(x, y), \quad v = g(x, y). \tag{3}$$

Here the function h contains the variable x partly directly, partly through the functions u, v so that

$$\frac{\partial z}{\partial x} = \frac{\partial h}{\partial x} + \frac{\partial h}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial h}{\partial v}\frac{\partial v}{\partial x}, \quad \frac{\partial z}{\partial y} = \frac{\partial h}{\partial y} + \frac{\partial h}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial h}{\partial v}\frac{\partial v}{\partial y}. \tag{4}$$

(In this case we cannot apply notation (1) and write $\partial z/\partial x$ instead of $\partial h/\partial x$.)

Example 2. $z = xu^2 + 2y^2 v$, $u = y \sin x$, $v = x \ln y$. By (4)

$$\frac{\partial z}{\partial x} = u^2 + 2xuy\cos x + 2y^2 \ln y = y^2 \sin^2 x + 2xy^2 \sin x \cos x + 2y^2 \ln y.$$

REMARK 3. The preceding differentiation could also be carried out directly after putting the expressions $y \sin x$ and $x \ln y$ for u and v into the formula $z = xu^2 + 2y^2 v$. But, if one (or more) of the functions (3) is given implicitly, then the substitution cannot (at least in the general case) be carried out. Then the application of formulae for differentiation of composite functions is necessary (cf. § 12.9).

12.5. Taylor's Theorem, the Mean-Value Theorem. Differentiation in a Given Direction

Theorem 1 (*Taylor's Theorem*). *Let the function $z = f(x, y)$ possess total differentials* (§ 12.3) *up to order $n + 1$ at every point of the closed segment u joining the points (x_0, y_0), $(x_0 + h, y_0 + k)$. Then*

$$f(x_0 + h, y_0 + k) = f(x_0, y_0) + \frac{\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right) f_{x_0, y_0}}{1!} + \ldots +$$

$$+ \frac{\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^n f_{x_0, y_0}}{n!} + R_{n+1}, \tag{1}$$

where the most often applied form of the remainder is Lagrange's form, namely

$$R_{n+1} = \frac{\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^{n+1} f_{c,d}}{(n+1)!}.$$

Here, the point (c, d) *is an interior point of the above-mentioned segment* u. *The suffixes* x_0, y_0 *or* c, d *denote the point at which the derivatives are to be computed.*

REMARK 1. For $n = 0$, the relation (1) reduces to the *mean-value theorem:*

$$f(x_0 + h, y_0 + k) - f(x_0, y_0) = h\frac{\partial f}{\partial x}(c, d) + k\frac{\partial f}{\partial y}(c, d).$$

REMARK 2. The generalization of *Taylor's Theorem* (or of the mean-value theorem) *to functions of several variables* is immediate:

$$f(x_1 + h_1, x_2 + h_2, \ldots, x_n + h_n) = f(x_1, x_2, \ldots, x_n) +$$

$$+ \frac{\left(h_1\frac{\partial}{\partial x_1} + \ldots + h_n\frac{\partial}{\partial x_n}\right) f_{x_1, x_2, \ldots, x_n}}{1!} + \ldots +$$

$$+ \frac{\left(h_1\frac{\partial}{\partial x_1} + \ldots + h_n\frac{\partial}{\partial x_n}\right)^m f_{x_1, x_2, \ldots, x_n}}{m!} + R_{m+1},$$

where

$$R_{m+1} = \frac{\left(h_1\frac{\partial}{\partial x_1} + \ldots + h_n\frac{\partial}{\partial x_n}\right)^{m+1} f_{c_1, c_2, \ldots, c_n}}{(m+1)!},$$

$c_k = x_k + \vartheta h_k$ and $0 < \vartheta < 1$. (For $n = 1$ we obtain Taylor's Theorem for functions of one variable, § 11.10.)

Definition 1. Let $\cos\alpha$, $\cos\beta$ be the direction-cosines of the oriented segment u joining the points (x_0, y_0), $(x_0 + h, y_0 + k)$. If $s = \sqrt{(h^2 + k^2)}$ denotes the length of this segment, then $h = s\cos\alpha$, $k = s\cos\beta$ (Fig. 12.2). The limit

$$\lim_{s \to 0} \frac{f(x_0 + s\cos\alpha, y_0 + s\cos\beta) - f(x_0, y_0)}{s} \tag{2}$$

(if it exists) is called the *derivative of the function* $f(x, y)$ *in the direction* $(\cos \alpha, \cos \beta)$ *at the point* (x_0, y_0) and is denoted by $\mathrm{d}f/\mathrm{d}s$.

Theorem 2. *If* $f(x, y)$ *is differentiable at* (x_0, y_0), *then*

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x}(x_0, y_0) \cos \alpha + \frac{\partial f}{\partial y}(x_0, y_0) \sin \alpha .$$

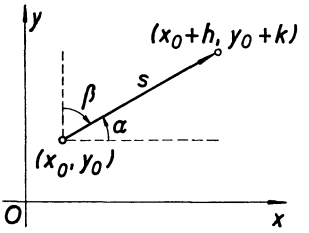


Fig. 12.2.

12.6. Euler's Theorem on Homogeneous Functions

Definition 1. The function $z = f(x, y)$ is said to be *homogeneous of degree n in a region O* if the relation

$$f(tx, ty) = t^n f(x, y)$$

holds identically for every point $(x, y) \in O$ and for every t from a certain neighbourhood of the point $t = 1$ (depending, in general, on the point (x, y)).

Example 1. The function

$$z = x^2 + y^2$$

is a homogeneous function of the second degree in the whole plane because (for every t)

$$(tx)^2 + (ty)^2 = t^2(x^2 + y^2) .$$

The function

$$z = \frac{1}{\sqrt{(x - y)}} \quad (x > y)$$

is, as we see in a similar way, a homogeneous function of degree $n = -\frac{1}{2}$ in the half-plane $x > y$.

Theorem 1 (*Euler's Theorem on Homogeneous Functions*). *If a function* $z = f(x, y)$, *homogeneous of degree n in a region O, has a total differential in O, then*

the relation

$$x\frac{\partial f}{\partial x} + y\frac{\partial f}{\partial y} = nf(x, y) \tag{1}$$

holds in O.

Example 2. The function $z = x^2 + y^2$ satisfies the relation

$$x\frac{\partial f}{\partial x} + y\frac{\partial f}{\partial y} = x \cdot 2x + y \cdot 2y = 2(x^2 + y^2).$$

REMARK 1. Under similar assumptions the relation

$$x_1\frac{\partial f}{\partial x_1} + \ldots + x_n\frac{\partial f}{\partial x_n} = mf(x_1, x_2, \ldots, x_n)$$

holds for a homogeneous function of degree m of n variables.

REMARK 2. The converse of Theorem 1 is also true: If a function $f(x, y)$ possesses a total differential in the region O and if equation (1) is satisfied everywhere in O, then $f(x, y)$ is a homogeneous function of degree n in O. A similar assertion holds for functions of several variables.

12.7. Regular Mappings. Functional Determinants

Definition 1. Let us consider m functions

$$\begin{aligned} y_1 &= f_1(x_1, x_2, \ldots, x_n), \\ y_2 &= f_2(x_1, x_2, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots\cdots \\ y_m &= f_m(x_1, x_2, \ldots, x_n) \end{aligned} \tag{1}$$

defined in an n-dimensional domain M. By system (1), to every point $X(x_1, x_2, \ldots, x_n$ of M, there corresponds a certain point $Y(y_1, y_2, \ldots, y_m)$ of the m-dimensional space E_m. This correspondence is called a *mapping*, or *transformation* (*of M into* E_m). The point X is called the *original* (or the *model* or the *inverse image*), the point Y is the *image* (or the *transform*). (On the simplest case $m = n = 1$ see Definition 11.1.3.)

Example 1. The parametric representation of a surface

$$x = f_1(u, v), \quad y = f_2(u, v), \quad z = f_3(u, v)$$

is a mapping, where the points $X(u, v)$ of the plane uv are the originals and the points $Y(x, y, z)$ of the space E_3 are the images.

Definition 2. The mapping

$$\begin{aligned} y_1 &= f_1(x_1, x_2, \ldots, x_n), \\ y_2 &= f_2(x_1, x_2, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots\cdots \\ y_n &= f_n(x_1, x_2, \ldots, x_n) \end{aligned} \tag{2}$$

(where the number of functions is the same as the number of variables) is said to be *regular in a region* M if each of the functions $y_1, y_2, \ldots, y_n$ possesses continuous partial derivatives of the first order in M and if the determinant

$$\begin{vmatrix} \dfrac{\partial f_1}{\partial x_1}, & \dfrac{\partial f_1}{\partial x_2}, & \ldots, & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1}, & \dfrac{\partial f_2}{\partial x_2}, & \ldots, & \dfrac{\partial f_2}{\partial x_n} \\ \cdots\cdots\cdots\cdots\cdots \\ \dfrac{\partial f_n}{\partial x_1}, & \dfrac{\partial f_n}{\partial x_2}, & \ldots, & \dfrac{\partial f_n}{\partial x_n} \end{vmatrix} \tag{3}$$

is different from zero in M.

REMARK 1. The determinant (3) is called the *functional determinant* of the given mapping or the *Jacobian*. We denote it by

$$\frac{\partial(y_1, y_2, \ldots, y_n)}{\partial(x_1, x_2, \ldots, x_n)}.$$

The Jacobian is continuous and non-zero in M; hence its sign does not change in M.

Definition 3. The mapping (1) is called *continuous at the point* $A(x_1, x_2, \ldots, x_n)$ if all the functions $f_1, f_2, \ldots, f_n$ are continuous at the point A.

REMARK 2. Evidently, every regular mapping is continuous but the converse is not true in general.

Definition 4. Let M and E_m have the same meaning as in Definition 1 and let N be a subregion of M; we do not exclude the possibility that $N = M$. Denote by Q the set of all points Y in E_m which correspond to all points $X \in N$ according to equations (1). We then say that the set N is transformed (or mapped) by equations (1) *onto* the set Q. If this correspondence has the further property that to each $Y \in Q$ there corresponds a unique original $X \in N$, then we say the *correspondence* is *one-to-one*. The so obtained mapping from Q to N (denoted by $Q \to N$), where thus Y is the original and X is the image, is called the *inverse mapping* to the original mapping $N \to Q$.

Theorem 1. *If a mapping is regular in M, then it is a one-to-one mapping in a sufficiently small neighbourhood of every interior point $X_0 \in M$.*

REMARK 3. Thus, this means that in a neighbourhood of every interior point $X_0 \in M$ there exists an inverse mapping, i.e. we can compute $x_1, x_2, \ldots, x_n$ from the system (2) as functions of the variables $y_1, y_2, \ldots, y_n$. The theorem has, however, a *local* character, i.e. a mapping that is regular in a region need not be a one-to-one mapping in the whole region.

Example 2. The Jacobian of the mapping $x = \varrho \cos \varphi$, $y = \varrho \sin \varphi$ (where ϱ and φ are the polar coordinates of the point (x, y)), namely

$$\frac{\partial(x, y)}{\partial(\varrho, \varphi)} = \begin{vmatrix} \frac{\partial x}{\partial \varrho}, & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial \varrho}, & \frac{\partial y}{\partial \varphi} \end{vmatrix} = \begin{vmatrix} \cos \varphi, & -\varrho \sin \varphi \\ \sin \varphi, & \varrho \cos \varphi \end{vmatrix} = \varrho ,$$

is non-vanishing when $\varrho > 0$ (ϱ is always non-negative) but, evidently, x and y are the same, for instance, when $\varrho = 1$, $\varphi = \frac{1}{2}\pi$ and $\varrho = 1$, $\varphi = \frac{5}{2}\pi$. If we restrict the values of φ, say, to the interval $0 \leqq \varphi < 2\pi$, then the mapping will be one-to-one for $\varrho > 0$. (But not for $\varrho = 0$ because, for example, for $\varrho = 0$, $\varphi = \frac{1}{2}\pi$ and $\varrho = 0$, $\varphi = \pi$ we obtain the same values $x = 0$, $y = 0$.)

REMARK 4. If $n = 1$ in (2), then we obtain the mapping $y = f(x)$. Its Jacobian is $f'(x)$. If, in a certain interval, $f'(x) \neq 0$, then the function $y = f(x)$ is strictly increasing or strictly decreasing in the whole interval and the inverse mapping exists in the whole interval.

Theorem 2 (*Theorem on the Preservation of the Region*). *The image of a region by a one-to-one regular mapping is again a region.*

Theorem 3. *If the mapping* (2) *is regular in a neighbourhood of the point $X_0 \in M$, then the inverse mapping* (see Remark 3) *is also regular in a neighbourhood of the corresponding point Y_0 and the values of the corresponding Jacobians are reciprocal,*

$$\frac{\partial(y_1, y_2, \ldots, y_n)}{\partial(x_1, x_2, \ldots, x_n)} = \frac{1}{\dfrac{\partial(x_1, x_2, \ldots, x_n)}{\partial(y_1, y_2, \ldots, y_n)}} .$$

Example 3. Let us consider the mapping

$$x = \varrho \cos \varphi , \quad y = \varrho \sin \varphi \tag{4}$$

(Example 2). Let us choose a point (x_0, y_0), for example so that $x_0 > 0$, $y_0 > 0$. Squaring equations (4) and adding them we obtain $x^2 + y^2 = \varrho^2$. Dividing them we have $y/x = \tan \varphi$. So we obtain the inverse mapping (in a neighbourhood of the chosen point)

$$\varrho = \sqrt{(x^2 + y^2)}, \quad \varphi = \arctan \frac{y}{x}. \tag{5}$$

Its Jacobian is

$$\frac{\partial(\varrho, \varphi)}{\partial(x, y)} = \begin{vmatrix} \dfrac{x}{\sqrt{(x^2 + y^2)}}, & \dfrac{y}{\sqrt{(x^2 + y^2)}} \\ -\dfrac{y}{x^2} \dfrac{1}{1 + y^2/x^2}, & \dfrac{1}{x} \dfrac{1}{1 + y^2/x^2} \end{vmatrix} =$$

$$= \frac{1}{\sqrt{(x^2 + y^2)(1 + y^2/x^2)}} \begin{vmatrix} x, & y \\ -y/x^2, & 1/x \end{vmatrix} = \frac{1}{\sqrt{(x^2 + y^2)}} = \frac{1}{\varrho}$$

in accordance with Theorem 3 (the value of the Jacobian of the mapping (4) is ϱ (Example 2)).

Theorem 4. *The mapping which is the resultant mapping obtained by combining two regular mappings is again a regular mapping. Its Jacobian is equal to the product of the Jacobians of the individual mappings:*

$$\frac{\partial(y_1, y_2, \ldots, y_n)}{\partial(x_1, x_2, \ldots, x_n)} = \frac{\partial(y_1, y_2, \ldots, y_n)}{\partial(u_1, u_2, \ldots, u_n)} \cdot \frac{\partial(u_1, u_2, \ldots, u_n)}{\partial(x_1, x_2, \ldots, x_n)}.$$

12.8. Dependence of Functions

Let us consider m functions defined in an n-dimensional region O,

$$\begin{aligned} y_1 &= f_1(x_1, x_2, \ldots, x_n), \\ y_2 &= f_2(x_1, x_2, \ldots, x_n), \\ &\ldots\ldots\ldots\ldots\ldots \\ y_m &= f_m(x_1, x_2, \ldots, x_n), \end{aligned} \tag{1}$$

having continuous partial derivatives of the first order in O.

Theorem 1. *Let the so-called Jacobian matrix of the functions* (1),

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1}, & \frac{\partial f_1}{\partial x_2}, & \dots, & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1}, & \frac{\partial f_2}{\partial x_2}, & \dots, & \frac{\partial f_2}{\partial x_n} \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ \frac{\partial f_m}{\partial x_1}, & \frac{\partial f_m}{\partial x_2}, & \dots, & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \tag{2}$$

be of rank h $(0 < h < m)$ *in a neighbourhood* U *of the point* $A(a_1, a_2, \dots, a_n) \in O$. *Thus, at the point* A (*and, hence, owing to the continuity of the functions* $\partial f_i/\partial x_k$ *also in a neighbourhood of the point* A) *at least one minor of order* h *is non-zero; let it be, for example, the minor*

$$\frac{\partial(y_1, y_2, \dots, y_h)}{\partial(x_1, x_2, \dots, x_h)},$$

Let us denote by $B(b_1, b_2, \dots, b_h)$ *the point with the coordinates*

$$b_k = f_k(a_1, a_2, \dots, a_n), \quad k = 1, 2, \dots, h.$$

Then there exist functions

$$F_1(y_1, y_2, \dots, y_h), \; F_2(y_1, y_2, \dots, y_h), \; \dots, \; F_{m-h}(y_1, y_2, \dots, y_h), \tag{3}$$

with continuous partial derivatives of the first order in a sufficiently small neighbourhood of the point B, *such that at every point* $X(x_1, x_2, \dots, x_n)$ *from a sufficiently small neighbourhood* Ω *of the point* A *the equations*

$$\begin{aligned} y_{h+1} &= F_1(y_1, y_2, \dots, y_h), \\ y_{h+2} &= F_2(y_1, y_2, \dots, y_h), \\ &\dots\dots\dots\dots\dots\dots \\ y_m &= F_{m-h}(y_1, y_2, \dots, y_h) \end{aligned} \tag{4}$$

hold.

REMARK 1. The functions (1) are said in this case (i.e. when some of them can be expressed as functions of the others) to be *dependent in* Ω. In the opposite case we say that the functions (1) are *independent* in Ω. This case occurs when $h = m$. In particular the functions defining a regular mapping (Definition 12.7.2) are independent.

REMARK 2. If the matrix (2) is of rank h in the whole domain O, we cannot, in general, affirm that the conclusion of Theorem 1 holds in the whole domain O;

Theorem 1 has a *local* character (therefore we often speak of *local dependency* of functions). It may, of course, happen that equations (4) hold in the whole domain O.

Example 1. The rank of the Jacobian of the system of functions

$$u_1 = x^2 + y^2 + z^2, \quad u_2 = x + y + z, \quad u_3 = xy + xz + yz \tag{5}$$

is less than 3 because the corresponding determinant

$$\begin{vmatrix} 2x, & 2y, & 2z \\ 1, & 1, & 1 \\ y + z, & x + z, & x + y \end{vmatrix}$$

is zero in the whole three-dimensional space xyz (as may be easily verified). Hence the functions (5) are locally dependent. It is easy to find that $u_3 = \frac{1}{2}(u_2^2 - u_1)$. Therefore, the functions (5) are dependent in the whole three-dimensional space.

Example 2. The functions $y_1 = \sin x$, $y_2 = \cos x$ are dependent, for instance, in the interval $[0, \pi]$ because the relation $y_1 = \sqrt{(1 - y_2^2)}$ holds for every x from this interval (see, however, Example 4 and Remark 4).

Definition 1. We say that the function $y_r(x_1, x_2, \ldots, x_n)$ is a *linear combination of the functions* $y_1(x_1, x_2, \ldots, x_n), \ldots, y_{r-1}(x_1, x_2, \ldots, x_n)$ *in the region* O if it is possible to find constants $c_1, c_2, \ldots, c_{r-1}$ such that

$$y_r = c_1 y_1 + c_2 y_2 + \ldots + c_{r-1} y_{r-1} \text{ identically in } O .$$

Definition 2. The functions (1) are said to be *linearly dependent in* O if at least one of them can be expressed as a linear combination of the others. In the opposite case the functions (1) are said to be *linearly independent in* O.

REMARK 3. The following definition is equivalent to Definition 2:

Definition 3. The functions (1) are said to be *linearly dependent in* O if m constants $c_1, c_2, \ldots, c_m$, *at least one of which is different from zero,* exist such that

$$c_1 y_1 + c_2 y_2 + \ldots + c_m y_m = 0 \text{ identically in } O . \tag{6}$$

If the identity (6) holds only when all c_k in (6) are equal to zero, the functions (1) are said to be *linearly independent in* O.

Example 3. The functions 1, x, x^2 are linearly independent in every interval I because (as is well known from algebra) the equation

$$c_1 + c_2 x + c_3 x^2 \equiv 0$$

holds in I only when $c_1 = c_2 = c_3 = 0$. The functions

$$f_1(x) = \sin^2 x, \quad f_2(x) = 4, \quad f_3(x) = \cos^2 x$$

are linearly dependent in every interval because

$$4 \sin^2 x - 4 + 4 \cos^2 x \equiv 0 .$$

Theorem 2. *Let the functions* (1) *be square-integrable in O* (§ 16.1) *and let their number be n. Let us denote* (cf. Definition 16.1.2 and Remark 16.1.7) *by* (f_i, f_k) *the scalar product of functions* $f_i(x_1, x_2, \ldots, x_n), f_k(x_1, x_2, \ldots, x_n)$, i.e.

$$(f_i, f_k) = \int_O f_i(x_1, x_2, \ldots, x_n) f_k(x_1, x_2, \ldots, x_n) \, dx_1 \, dx_2 \ldots dx_n . \tag{7}$$

Then the necessary and sufficient condition for the functions (1) *to be linearly dependent in O is the vanishing of the so-called Gram determinant,*

$$G = \begin{vmatrix} (f_1, f_1), & (f_1, f_2), & \ldots, & (f_1, f_n) \\ (f_2, f_1), & (f_2, f_2), & \ldots, & (f_2, f_n) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ (f_n, f_1), & (f_n, f_2), & \ldots, & (f_n, f_n) \end{vmatrix} = 0 . \tag{8}$$

Example 4. Let us apply (8) to the investigation of the linear dependence of the functions $f_1(x) = \sin x, f_2(x) = \cos x$ in the interval $[0, \pi]$. We have

$$(f_1, f_1) = \int_0^\pi \sin^2 x \, dx = \tfrac{1}{2}\pi , \quad (f_2, f_2) = \int_0^\pi \cos^2 x \, dx = \tfrac{1}{2}\pi ,$$

$$(f_1, f_2) = (f_2, f_1) = \int_0^\pi \sin x \cos x = 0 .$$

So

$$G = \begin{vmatrix} \frac{1}{2}\pi, & 0 \\ 0, & \frac{1}{2}\pi \end{vmatrix} = \tfrac{1}{4}\pi^2 \neq 0$$

and the functions under consideration are linearly independent in O. (Note that, by Example 2, the functions are dependent in the sense defined in Remark 1.)

REMARK 4. If the functions (1) are *linearly* dependent, then they are naturally dependent in the sense of Remark 1. If they are independent in accordance with Remark 1, then they are also *linearly* independent.

For a simple criterion (the so-called *Wronski determinant* or *Wronskian*) in the investigation of the linear independence of functions that are integrals of a linear differential equation see §17.11.

12.9. Theorem on Implicit Functions. Equations $f(x, y) = 0, f(x, y, z) = 0$

Definition 1. Let an equation $f(x, y) = 0$ be given. We say that the function $y = \varphi(x)$ is a *solution of this equation in the domain M* if the relation $f(x, \varphi(x)) = 0$

holds identically in M. The functions defined in this sense by the equation $f(x, y) = 0$ are said to be *given implicitly*, or are briefly called *implicit functions.*

Example 1. Consider $2 \ln x - x^2 + e^y - y = 0$. In the interval $(0, +\infty)$ the function $y = 2 \ln x$ is a solution of this equation.

Theorem 1. *Let us consider an equation*

$$f(x, y) = 0 \tag{1}$$

and a point (x_0, y_0) *such that* $f(x_0, y_0) = 0$. *Let* $f(x, y)$ *have continuous partial derivatives of the first order in a neighbourhood of this point and let*

$$\frac{\partial f}{\partial y}(x_0, y_0) \neq 0 .$$

Then, in a certain neighbourhood of the point x_0, *there exists a unique continuous solution* $y = \varphi(x)$ *of equation* (1) *that satisfies the condition* $\varphi(x_0) = y_0$. *The function* $y = \varphi(x)$ *has a continuous derivative* $y' = \varphi'(x)$ *in a neighbourhood of the point* x_0. *This derivative may be computed from the equation*

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} y' = 0 \tag{2}$$

or

$$y' = -\frac{\partial f / \partial x}{\partial f / \partial y} . \tag{3}$$

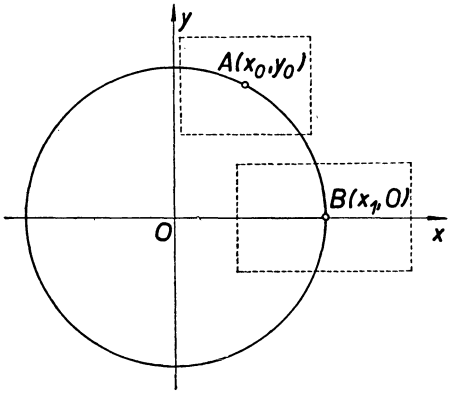


Fig. 12.3.

REMARK 1. *Geometrical interpretation*: If $f(x, y)$ is a "reasonable" function, then by equation (1) a curve in the plane xy is given; e.g. the equation

$$x^2 + y^2 - 25 = 0$$

represents a circle (Fig. 12.3). If we choose a point (x_0, y_0) on the curve, then the question arises if it is possible to express *all* points of this curve in a neighbourhood

of the point (x_0, y_0) by an explicit single valued function $y = \varphi(x)$. It can be seen from Fig. 12.3 that, for example, in the neighbourhood of the point A it certainly is possible (indeed, here this function can be found directly because y may be computed from the given equation, $y = +\sqrt{(25 - x^2)}$), but this is not so at the point B: However small we choose the neighbourhood of the point B, two values of y will always correspond to a single value of $x \in (x_1 - \delta, x_1)$ (if *all* points of the circle in

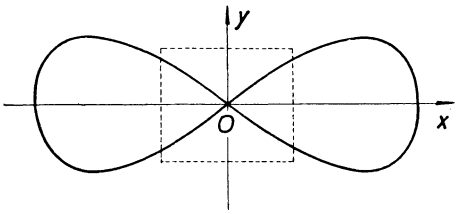


Fig. 12.4.

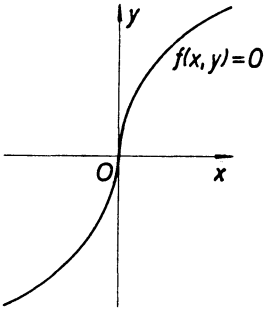


Fig. 12.5. $x - y^3 = 0$; at the point (0, 0) we have $\partial f/\partial y = 0$.

the neighbourhood of the point B are to be considered), and not a single value as is required by the definition of a function (Definition 11.1.1). We note that

$$\frac{\partial f}{\partial y} = 2y = 0$$

at the point $B(x_1, 0)$, so that even in this simple case we can see the importance of the condition $\frac{\partial f}{\partial y}(x_0, y_0) \neq 0$.

Likewise it can be seen that, in the case of the lemniscate $(x^2 + y^2)^2 + 2c^2(y^2 - x^2) = 0$ (Fig. 12.4), it is not possible to express *all* its points by a unique function $y = \varphi(x)$ in any (non-zero) neighbourhood of the origin – however small this neighbourhood is chosen. Here also we note that $\frac{\partial f}{\partial y}(0, 0) = 0$.

The condition $\frac{\partial f}{\partial y}(x_0, y_0) \neq 0$ is not, however, a necessary condition – as can be seen from Fig. 12.5.

REMARK 2. Since $\frac{\partial f}{\partial y}$ is continuous in a neighbourhood of (x_0, y_0), and since $\frac{\partial f}{\partial y} \neq 0$ at (x_0, y_0), it is possible to compute y' from (2) in a certain neighbourhood of the point (x_0, y_0). If $\frac{\partial f}{\partial y}(x_0, y_0) = 0$ and simultaneously $\frac{\partial f}{\partial x}(x_0, y_0) \neq 0$, then

this means geometrically that the tangent to the curve $f(x, y) = 0$ at the point (x_0, y_0) is parallel to the y-axis.

REMARK 3. Equation (2) results from (1) formally by differentiating equation (1) with respect to x as a composite function of a single variable x (i.e. when considering y as a function of x, $y = \varphi(x)$). Similarly, by differentiating equation (2) with respect to x (e.g. under the condition that $f(x, y)$ has continuous partial derivatives of the second order), treating the left-hand side of (2) as a composite function of the single variable x, we obtain an equation from which we can compute y'', namely

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y\, \partial x} y' + \left(\frac{\partial^2 f}{\partial x\, \partial y} + \frac{\partial^2 f}{\partial y^2} y' \right) y' + \frac{\partial f}{\partial y} y'' = 0 .$$

Or (making use of the interchangeability of mixed derivatives),

$$\frac{\partial^2 f}{\partial x^2} + 2 \frac{\partial^2 f}{\partial x\, \partial y} y' + \frac{\partial^2 f}{\partial y^2} y'^2 + \frac{\partial f}{\partial y} y'' = 0 . \tag{4}$$

By further differentiation of this equation we obtain an equation for y''', etc.

Example 2. Let us compute y' and y'' for the function given implicitly by the equation

$$x^2 + y^2 - 25 = 0 \tag{5}$$

(a circle) at the point (3, 4).

The chosen point does satisfy equation (5) (i.e. lies on the circle (5)) and

$$\frac{\partial f}{\partial y} = 2y = 8 \neq 0 .$$

Hence the condition $\partial f/\partial y \neq 0$ is fulfilled. From (5) we have (because $\partial f/\partial x = 2x$)

$$2x + 2yy' = 0 \tag{6}$$

or (when we put $x = 3$, $y = 4$) $y' = -\frac{3}{4}$. Further differentiation of equation (6) (divided by 2) gives

$$1 + y'y' + yy'' = 0 .$$

We substitute $y = 4$ and for y' the computed value $y' = -\frac{3}{4}$:

$$1 + (\tfrac{3}{4})^2 + 4y'' = 0 ,$$

$$y'' = -\tfrac{25}{64} .$$

Hence, the function $y = \varphi(x)$ given by equation (5) and such that $\varphi(3) = 4$, has the following values of derivatives for $x = 3$: $y' = -\frac{3}{4}$, $y'' = -\frac{25}{64}$.

REMARK 4. The foregoing example is a very simple one and was used here only as an illustration; y' and y'' could have been computed directly by differentiating the equation $y = \sqrt{(25 - x^2)}$. But for implicit equations, where $y = \varphi(x)$ cannot be computed on the basis of current elementary functions (for example, in the case of the equation $e^y \sin x + x^2y^2 - x \ln y + 1 = 0$), the application of the theorem considered is essential.

Theorem 2. *Let us consider the equation*

$$f(x, y, z) = 0\,. \tag{7}$$

Let

$$f(x_0, y_0, z_0) = 0 \tag{8}$$

and let $f(x, y, z)$ have continuous partial derivatives of the first order in a neighbourhood of the point (x_0, y_0, z_0). Further, let

$$\frac{\partial f}{\partial z}(x_0, y_0, z_0) \neq 0\,. \tag{9}$$

Then, in a certain neighbourhood of the point (x_0, y_0), there exists a unique continuous solution

$$z = \varphi(x, y) \tag{10}$$

of equation (7) that satisfies the condition $\varphi(x_0, y_0) = z_0$. The function $z = \varphi(x, y)$ has continuous partial derivatives of the first order in a neighbourhood of the point (x_0, y_0) and these derivatives can be determined from the relations

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial x} = 0\,, \quad \frac{\partial f}{\partial y} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial y} = 0\,. \tag{11}$$

REMARK 5. Equations (11) result formally from equation (7) if we differentiate this equation partially with respect to x (or with respect to y) considering z as a function of x and y. In the same way as was shown in Remark 3 we can obtain higher derivatives by differentiation of (11). For example, if we differentiate the first equation (11) with respect to x, we obtain an equation for $\partial^2 z/\partial x^2$:

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial z\,\partial x}\frac{\partial z}{\partial x} + \left(\frac{\partial^2 f}{\partial x\,\partial z} + \frac{\partial^2 f}{\partial z^2}\frac{\partial z}{\partial x}\right)\frac{\partial z}{\partial x} + \frac{\partial f}{\partial z}\frac{\partial^2 z}{\partial x^2} = 0\,,$$

or, after rearrangement,

$$\frac{\partial^2 f}{\partial x^2} + 2\frac{\partial^2 f}{\partial x\,\partial z}\frac{\partial z}{\partial x} + \frac{\partial^2 f}{\partial z^2}\left(\frac{\partial z}{\partial x}\right)^2 + \frac{\partial f}{\partial z}\frac{\partial^2 z}{\partial x^2} = 0\,.$$

Example 3. Let us find the equation of the tangent plane of the ellipsoid

$$\frac{x^2}{12} + \frac{y^2}{27} + \frac{z^2}{3} - 1 = 0 \tag{12}$$

at the point (2, 3, 1).

This point does satisfy equation (12) and at this point

$$\frac{\partial f}{\partial z} = \frac{2z}{3} = \frac{2}{3} \neq 0 .$$

Condition (9) is thus satisfied. The equation of the tangent plane is (Theorem 12.3.4, p. 410)

$$z - z_0 = \left(\frac{\partial f}{\partial x}\right)_0 (x - x_0) + \left(\frac{\partial f}{\partial y}\right)_0 (y - y_0) .$$

By (11)

$$\frac{2x}{12} + \frac{2z}{3}\frac{\partial z}{\partial x} = 0 , \quad \frac{2y}{27} + \frac{2z}{3}\frac{\partial z}{\partial y} = 0 .$$

If we insert the values $x_0 = 2$, $y_0 = 3$, $z_0 = 1$, we get

$$\left(\frac{\partial f}{\partial x}\right)_0 = \frac{\partial z}{\partial x}(2, 3) = -\frac{1}{2} , \quad \left(\frac{\partial f}{\partial y}\right)_0 = \frac{\partial z}{\partial y}(2, 3) = -\frac{1}{3} .$$

Thus, the equation of the tangent plane is

$$z - 1 = -\tfrac{1}{2}(x - 2) - \tfrac{1}{3}(y - 3)$$

or

$$3x + 2y + 6z - 18 = 0 .$$

Example 4. By the equations

$$uz - 2e^{vz} = 0 , \quad u = x^2 + y^2 , \quad v^2 - xy \ln v - 1 = 0 \tag{13}$$

z is given as a composite function of the variables x, y (by means of the functions u, v). We wish to find $\partial z/\partial x$ at the point $x = 0$, $y = e$ (the corresponding values of u, v, z are $u = e^2$, $v = 1$, $z = 2$). According to the rule for differentiation of composite functions (Theorem 12.4.1) we obtain

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial x} . \tag{14}$$

Let us denote $uz - 2e^{vz}$ by $f(u, v, z)$. Then

$$\frac{\partial f}{\partial z} = u - 2v\, e^{vz} , \quad \frac{\partial f}{\partial z}(e^2, 1, 2) = -e^2 \neq 0 .$$

Thus, according to (11),

$$\frac{\partial f}{\partial u} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial u} = 0\,, \quad z + (u - 2v\,\mathrm{e}^{vz})\frac{\partial z}{\partial u} = 0\,, \quad 2 - \mathrm{e}^2\frac{\partial z}{\partial u} = 0$$

and hence

$$\frac{\partial z}{\partial u} = \frac{2}{\mathrm{e}^2}\,.$$

Similarly

$$\frac{\partial f}{\partial v} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial v} = 0\,, \quad -2z\,\mathrm{e}^{vz} + (u - 2v\,\mathrm{e}^{vz})\frac{\partial z}{\partial v} = 0\,,$$

$$-4\mathrm{e}^2 - \mathrm{e}^2\frac{\partial z}{\partial v} = 0\,, \quad \frac{\partial z}{\partial v} = -4\,.$$

Further (after substitution of the values $x = 0$, $y = \mathrm{e}$, $v = 1$)

$$\frac{\partial u}{\partial x} = 2x = 0\,.$$

If

$$g(x, y, v) = v^2 - xy\ln v - 1\,, \quad \text{then} \quad \frac{\partial g}{\partial v} = 2v - \frac{xy}{v} = 2 \neq 0\,.$$

Hence

$$\frac{\partial g}{\partial x} + \frac{\partial g}{\partial v}\frac{\partial v}{\partial x} = 0\,, \quad -y\ln v + \left(2v - \frac{xy}{v}\right)\frac{\partial v}{\partial x} = 0\,, \quad 0 + 2\frac{\partial v}{\partial x} = 0\,, \quad \frac{\partial v}{\partial x} = 0\,.$$

If we insert all partial results into (14) we obtain:

$$\frac{\partial z}{\partial x}(0, \mathrm{e}) = \frac{2}{\mathrm{e}^2}\,.\,0 - 4\,.\,0 = 0\,.$$

REMARK 6. Instead of using equation (14) we may compute $\partial z/\partial x$ by differentiation of the first equation of (13) with respect to x taking into consideration the fact that u, v and z contain x:

$$\frac{\partial u}{\partial x}z + u\frac{\partial z}{\partial x} - 2\mathrm{e}^{vz}\left(\frac{\partial v}{\partial x}z + v\frac{\partial z}{\partial x}\right) = 0$$

($\partial u/\partial x$ and $\partial v/\partial x$ have naturally to be found from the second and third equations of (13)). We arrive at the same result. The advantage of this method is that it can be

applied without difficulty even in the case where the function f contains explicitly not only u and v but also x and y, when equation (14) cannot be applied (cf. Remark 12.4.2).

REMARK 7. The problem becomes somewhat complicated when the second and third equations of (13) do not contain u and v separately but when u and v are given (as functions of x and y) by the equations

$$g(x, y, u, v) = 0\,, \quad h(x, y, u, v) = 0\,. \tag{15}$$

The required derivatives $\partial u/\partial x$, $\partial v/\partial x$ are found by solving the equations derived from (15) by differentiating with respect to x:

$$\frac{\partial g}{\partial x} + \frac{\partial g}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial g}{\partial v}\frac{\partial v}{\partial x} = 0\,, \quad \frac{\partial h}{\partial x} + \frac{\partial h}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial h}{\partial v}\frac{\partial v}{\partial x} = 0$$

(see § 12.10).

12.10. Theorem on Implicit Functions. General Case

Definition 1. Let us have a system of equations

$$\begin{aligned} &F_1(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m) = 0\,,\\ &F_2(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m) = 0\,,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &F_m(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m) = 0\,. \end{aligned} \tag{1}$$

By a solution of system (1) (in a region O) we understand the functions

$$\begin{aligned} y_1 &= f_1(x_1, x_2, \ldots, x_n)\,,\\ y_2 &= f_2(x_1, x_2, \ldots, x_n)\,,\\ &\ldots\ldots\ldots\ldots\ldots\\ y_m &= f_m(x_1, x_2, \ldots, x_n) \end{aligned} \tag{2}$$

such that if we substitute f_1 for y_1, etc., into (1), all the equations of the system become identities in $x_1, x_2, \ldots, x_n$ in the region O. The functions (2) are said to be *given implicitly by equations* (1). Briefly they are *implicit functions.*

Theorem 1. *Let*

1. *the point* $(x_1^0, x_2^0, \ldots, x_n^0, y_1^0, y_2^0, \ldots, y_m^0)$ *satisfy all the equations of system* (1),

2. *the functions* (1) *possess continuous partial derivatives of the first order with respect to all variables in a neighbourhood of this point,*

3. *the determinant*

$$\frac{\partial(F_1, F_2, \ldots, F_m)}{\partial(y_1, y_2, \ldots, y_m)} = \begin{vmatrix} \frac{\partial F_1}{\partial y_1}, & \frac{\partial F_1}{\partial y_2}, \ldots, & \frac{\partial F_1}{\partial y_m} \\ \frac{\partial F_2}{\partial y_1}, & \frac{\partial F_2}{\partial y_2}, \ldots, & \frac{\partial F_2}{\partial y_m} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \frac{\partial F_m}{\partial y_1}, & \frac{\partial F_m}{\partial y_2}, \ldots, & \frac{\partial F_m}{\partial y_m} \end{vmatrix} \tag{3}$$

be non-zero at the point $(x_1^0, x_2^0, \ldots, x_n^0, y_1^0, y_2^0, \ldots, y_m^0)$.

Then, in a certain neighbourhood of the point $(x_1^0, x_2^0, \ldots, x_n^0)$, *there exists a unique system of continuous functions* (2) *which is the solution of equations* (1) *in this neighbourhood and for which the equations*

$$\begin{aligned} f_1(x_1^0, x_2^0, \ldots, x_n^0) &= y_1^0, \\ f_2(x_1^0, x_2^0, \ldots, x_n^0) &= y_2^0, \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ f_m(x_1^0, x_2^0, \ldots, x_n^0) &= y_m^0 \end{aligned}$$

hold simultaneously. Each of the functions (2) *possesses continuous partial derivatives of the first order with respect to all variables* $x_1, x_2, \ldots, x_n$ *in the neighbourhood of the point considered.*

REMARK 1. The formulae for the computation of derivatives are rather cumbersome, expecially those for the computation of derivatives of higher orders (existence of these derivatives is established if, for example, all the functions (1) possess continuous derivatives of the order considered in a neighbourhood of the point $(x_1^0, x_2^0, \ldots, x_n^0, y_1^0, y_2^0, \ldots, y_m^0)$). The computation of derivatives is carried out practically by differentiating the equations (1) with respect to the corresponding variable x_k and considering simultaneously $y_1, y_2, \ldots, y_m$ as functions of the variables $x_1, x_2, \ldots, x_n$. For example, to compute the derivatives with respect to x_2 we have:

$$\begin{aligned} &\frac{\partial F_1}{\partial x_2} + \frac{\partial F_1}{\partial y_1}\frac{\partial y_1}{\partial x_2} + \frac{\partial F_1}{\partial y_2}\frac{\partial y_2}{\partial x_2} + \ldots + \frac{\partial F_1}{\partial y_m}\frac{\partial y_m}{\partial x_2} = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\frac{\partial F_m}{\partial x_2} + \frac{\partial F_m}{\partial y_1}\frac{\partial y_1}{\partial x_2} + \frac{\partial F_m}{\partial y_2}\frac{\partial y_2}{\partial x_2} + \ldots + \frac{\partial F_m}{\partial y_m}\frac{\partial y_m}{\partial x_2} = 0. \end{aligned} \tag{4}$$

This system of equations for $\partial y_1/\partial x_2, \partial y_2/\partial x_2, \ldots, \partial y_m/\partial x_2$ is uniquely solvable because the determinant (3) is non-zero.

Example 1. Let y_1 and y_2 be given implicitly by the equations

$$x_1 e^{y_2} + y_1 \ln x_2 - e = 0, \quad x_1 y_1 + x_2 e^{y_2} - (2 + e) = 0. \tag{5}$$

Let us compute $\partial y_1/\partial x_1$, $\partial y_2/\partial x_1$ at the point $x_1 = 1$, $x_2 = 1$.

From equations (5) it follows (at the point $x_1 = 1$, $x_2 = 1$) that $y_2 = 1$, $y_1 = 2$. (In the general case, when there are more possibilities, it is necessary to state in advance for what values y_1, y_2 the problem is to be solved.) We differentiate equations (5) with respect to x_1 considering (see Remark 1) y_1 and y_2 dependent on x_1:

$$e^{y_2} + x_1 e^{y_2} \frac{\partial y_2}{\partial x_1} + \frac{\partial y_1}{\partial x_1} \ln x_2 = 0, \quad y_1 + x_1 \frac{\partial y_1}{\partial x_1} + x_2 e^{y_2} \frac{\partial y_2}{\partial x_1} = 0. \tag{6}$$

When the numerical values of x_1, x_2, y_1, y_2 are substituted into (6), we obtain

$$0 \cdot \frac{\partial y_1}{\partial x_1} + e \frac{\partial y_2}{\partial x_1} + e = 0,$$

$$\frac{\partial y_1}{\partial x_1} + e \frac{\partial y_2}{\partial x_1} + 2 = 0,$$

hence

$$\frac{\partial y_2}{\partial x_1}(1, 1) = -1, \quad \frac{\partial y_1}{\partial x_1}(1, 1) = e - 2.$$

12.11. Introduction of New Variables. Transformations of Differential Equations and Differential Expressions (Especially into Polar, Spherical and Cylindrical Coordinates)

In the course of the solution of differential equations and in many other problems it may be convenient to introduce new variables. In this section we shall consider the most frequently occurring cases of introducing new variables, independent as well as dependent, and shall show how to express the given differential expressions with the help of the new variables.

REMARK 1. Throughout this section we suppose that the functions considered possess all the necessary (continuous) derivatives and that the correspondence between the original and the new variables is one-to-one.

(a) CASE OF ONE VARIABLE

(α) *Introduction of a new Independent Variable.* By introducing a new variable $x = \varphi(t)$, the function $y = f(x)$ becomes a composite function of the variable t.

We have to find the relation between $y' = \mathrm{d}y/\mathrm{d}x$ and $\dot{y} = \mathrm{d}y/\mathrm{d}t$ and, similarly, between higher derivatives.

By Theorem 11.5.5 (assuming $\dot{x} = \mathrm{d}x/\mathrm{d}t \neq 0$ and applying Theorem 11.5.6) we get

$$y' = \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}t} \cdot \frac{\mathrm{d}t}{\mathrm{d}x} = \dot{y} \cdot \frac{1}{\mathrm{d}x/\mathrm{d}t} = \dot{y} \cdot \frac{1}{\dot{x}}, \tag{1}$$

$$y'' = \frac{\mathrm{d}y'}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}\left(\dot{y} \cdot \frac{1}{\dot{x}}\right) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{y} \cdot \frac{1}{\dot{x}}\right)\frac{\mathrm{d}t}{\mathrm{d}x} = \left(\ddot{y}\,\frac{1}{\dot{x}} - \dot{y}\,\frac{\ddot{x}}{\dot{x}^2}\right)\frac{1}{\dot{x}}, \tag{2}$$

$$y''' = \frac{\mathrm{d}y''}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}\left[\ddot{y}\,\frac{1}{\dot{x}^2} - \dot{y}\,\frac{\ddot{x}}{\dot{x}^3}\right] = \frac{\mathrm{d}}{\mathrm{d}t}\left[\ddot{y}\,\frac{1}{\dot{x}^2} - \dot{y}\,\frac{\ddot{x}}{\dot{x}^3}\right]\frac{\mathrm{d}t}{\mathrm{d}x} =$$

$$= \left\{\dddot{y}\,\frac{1}{\dot{x}^2} - 3\ddot{y}\,\frac{\ddot{x}}{\dot{x}^3} - \dot{y}\,\frac{\dddot{x}\dot{x}^3 - 3\ddot{x}\dot{x}^2\ddot{x}}{\dot{x}^6}\right\}\frac{1}{\dot{x}} = \dddot{y}\,\frac{1}{\dot{x}^3} - 3\ddot{y}\,\frac{\ddot{x}}{\dot{x}^4} - \dot{y}\,\frac{\dddot{x}\dot{x} - 3\ddot{x}^2}{\dot{x}^5}. \tag{3}$$

To compute higher derivatives, we proceed in a similar way. Derivatives with respect to t expressed in terms of derivatives with respect to x may be obtained from equations (1), (2), (3), or directly:

$$\dot{y} = \frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\mathrm{d}y}{\mathrm{d}x} \cdot \frac{\mathrm{d}x}{\mathrm{d}t} = y' \cdot \dot{x}, \tag{4}$$

$$\ddot{y} = \frac{\mathrm{d}^2y}{\mathrm{d}t^2} = \frac{\mathrm{d}}{\mathrm{d}t}(y' \cdot \dot{x}) = \dot{x}\,\frac{\mathrm{d}}{\mathrm{d}t}(y') + y' \cdot \frac{\mathrm{d}}{\mathrm{d}t}(\dot{x}) =$$

$$= \dot{x}\,\frac{\mathrm{d}}{\mathrm{d}x}(y') \cdot \frac{\mathrm{d}x}{\mathrm{d}t} + y' \cdot \ddot{x} = \dot{x}^2y'' + \ddot{x}y', \quad \text{etc.} \tag{5}$$

Example 1. Let us transform the left-hand side of the differential equation

$$x^2y'' + 4xy' - 2y = 0 \tag{6}$$

by introducing a new independent variable t by the relation $x = \mathrm{e}^t$ $(x > 0)$.

Evidently, the correspondence is one-to-one for all t and the function e^t possesses derivatives of all orders. We may proceed according to (1), (2) or differentiate directly:

$$y' = \dot{y} \cdot \frac{1}{\dot{x}} = \dot{y} \cdot \frac{1}{\mathrm{e}^t} = \dot{y}\mathrm{e}^{-t}, \quad y'' = \frac{\mathrm{d}}{\mathrm{d}t}(\dot{y}\mathrm{e}^{-t})\frac{\mathrm{d}t}{\mathrm{d}x} =$$

$$= (\ddot{y}\mathrm{e}^{-t} - \dot{y}\mathrm{e}^{-t})\,\mathrm{e}^{-t} = \ddot{y}\mathrm{e}^{-2t} - \dot{y}\mathrm{e}^{-2t}.$$

After substituting into (6) we obtain

$$\mathrm{e}^{2t}(\ddot{y}\mathrm{e}^{-2t} - \dot{y}\mathrm{e}^{-2t}) + 4\mathrm{e}^t \cdot \dot{y}\mathrm{e}^{-t} - 2y = 0$$

or

$$\ddot{y} + 3\dot{y} - 2y = 0$$

(see § 17.13, Euler's differential equation).

(β) *Introduction of a New Dependent Variable.* Instead of the dependent variable y we shall introduce a new dependent variable z by the relation $y = \varphi(z)$ or $z = \psi(y)$. In the same way as in (4), (5) we get

$$y' = \frac{dy}{dx} = \frac{dy}{dz} \cdot \frac{dz}{dx} = \frac{dy}{dz} \cdot z' , \tag{7}$$

$$y'' = \frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dz} z'\right) = z' \frac{d}{dz}\left(\frac{dy}{dz}\right)\frac{dz}{dx} + \frac{dy}{dz}\frac{d}{dx}(z') =$$

$$= \frac{d^2y}{dz^2} z'^2 + \frac{dy}{dz} z'' , \quad \text{etc} . \tag{8}$$

Similarly for the relation $z = \psi(y)$:

$$z' = \frac{dz}{dx} = \frac{dz}{dy} y' , \quad z'' = \frac{d^2z}{dx^2} = \frac{d^2z}{dy^2} y'^2 + \frac{dz}{dy} y'' , \quad \text{etc} . \tag{9}$$

Example 2. If we introduce into the differential equation

$$y' e^x \cos y - x \sin y = \ln x \tag{10}$$

a new dependent variable by the relation $\sin y = z$, then $\cos y \, . \, y' = z'$. Substituting into (10) we obtain the linear differential equation

$$e^x z' - xz = \ln x \tag{11}$$

for the new unknown function $z(x)$.

REMARK 2. In the majority of cases we need not apply the general transformation formulae because simpler methods are available as we have seen in Examples 1 and 2. A very simple treatment is also possible in the case of so-called *homogeneous differential equations of the first order*, where (see § 17.3) we introduce instead of $y(x)$ a new dependent variable $z(x)$ by means of the relation $y = xz$. Then $y' = z + xz'$.

REMARK 3. It is relatively seldom that we introduce simultaneously both a new independent variable and a dependent variable. As a rule, the transformation may be carried out successively according to α) and β).

(b) CASE OF TWO OR MORE VARIABLES

If we introduce new independent variables

$$x = x(u, v) , \quad y = y(u, v) , \tag{12}$$

then the function $z = f(x, y)$ becomes a composite function of the variables u, v, $z = f(x(u, v), y(u, v))$. By Theorem 12.4.1 we have

$$\frac{\partial z}{\partial u} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial u}, \quad \frac{\partial z}{\partial v} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial v}. \tag{13}$$

Higher derivatives are computed similarly (for a more detailed treatment see Remark 12.4.1 where second derivatives are computed directly; it is necessary, however, in the formulae of that Remark to interchange (x, y) and (u, v) because in the present case $z = z(x, y)$, $x = x(u, v)$, $y = y(u, v)$, which is different from the notation of Theorem 12.4.1).

Solving (13) for $\partial z/\partial x$, $\partial z/\partial y$ we obtain (if $\partial(x, y)/\partial(u, v) \neq 0$; Remark 12.7.1) the derivatives $\partial z/\partial x$, $\partial z/\partial y$ expressed in terms of the derivatives $\partial z/\partial u$, $\partial z/\partial v$.

Example 3. Let us transform the differential equation

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = 0 \tag{14}$$

by introducing the polar coordinates

$$x = \varrho\cos\varphi, \qquad y = \varrho\sin\varphi \quad (\varrho > 0,\ 0 \leqq \varphi < 2\pi). \tag{15}$$

By (13) (writing $u = \varrho$, $v = \varphi$) we have

$$\frac{\partial z}{\partial \varrho} = \frac{\partial z}{\partial x}\cos\varphi + \frac{\partial z}{\partial y}\sin\varphi, \quad \frac{\partial z}{\partial \varphi} = \frac{\partial z}{\partial x}(-\varrho\sin\varphi) + \frac{\partial z}{\partial y}\varrho\cos\varphi. \tag{16}$$

Solving (16) for $\partial z/\partial x$, $\partial z/\partial y$ we obtain

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial \varrho}\cos\varphi - \frac{\partial z}{\partial \varphi}\frac{1}{\varrho}\sin\varphi, \quad \frac{\partial z}{\partial y} = \frac{\partial z}{\partial \varrho}\sin\varphi + \frac{\partial z}{\partial \varphi}\frac{1}{\varrho}\cos\varphi. \tag{17}$$

If we carry out these operations with the function $\partial z/\partial x$ instead of the function z, we obtain by (17)

$$\frac{\partial}{\partial x}\left(\frac{\partial z}{\partial x}\right) = \frac{\partial\left(\frac{\partial z}{\partial x}\right)}{\partial \varrho}\cos\varphi - \frac{\partial\left(\frac{\partial z}{\partial x}\right)}{\partial \varphi}\frac{1}{\varrho}\sin\varphi. \tag{18}$$

Substituting the right-hand side of the first equation (17) into the right-hand side of (18) we obtain

$$\begin{aligned}\frac{\partial^2 z}{\partial x^2} &= \cos\varphi\frac{\partial}{\partial \varrho}\left[\frac{\partial z}{\partial \varrho}\cos\varphi - \frac{\partial z}{\partial \varphi}\frac{1}{\varrho}\sin\varphi\right] - \frac{1}{\varrho}\sin\varphi\frac{\partial}{\partial \varphi}\left[\frac{\partial z}{\partial \varrho}\cos\varphi - \frac{\partial z}{\partial \varphi}\frac{1}{\varrho}\sin\varphi\right] = \\ &= \frac{\partial^2 z}{\partial \varrho^2}\cos^2\varphi - \frac{\partial^2 z}{\partial \varrho\,\partial \varphi}\frac{2}{\varrho}\cos\varphi\sin\varphi + \frac{\partial^2 z}{\partial \varphi^2}\frac{1}{\varrho^2}\sin^2\varphi + \\ &\quad + \frac{1}{\varrho}\frac{\partial z}{\partial \varrho}\sin^2\varphi + \frac{\partial z}{\partial \varphi}\frac{2}{\varrho^2}\sin\varphi\cos\varphi.\end{aligned} \tag{19}$$

Similarly, by application of the second equation (17) we get

$$\frac{\partial^2 z}{\partial y^2} = \frac{\partial^2 z}{\partial \varrho^2} \sin^2 \varphi + \frac{\partial^2 z}{\partial \rho \, \partial \varphi} \frac{2}{\rho} \sin \varphi \cos \varphi + \frac{\partial^2 z}{\partial \varphi^2} \frac{1}{\varrho^2} \cos^2 \varphi +$$

$$+ \frac{1}{\varrho} \frac{\partial z}{\partial \varrho} \cos^2 \varphi - \frac{\partial z}{\partial \varphi} \frac{2}{\varrho^2} \sin \varphi \cos \varphi . \tag{20}$$

Substituting (19), (20) into (14) we obtain the transformed equation

$$\frac{\partial^2 z}{\partial \rho^2} + \frac{1}{\rho^2} \frac{\partial^2 z}{\partial \varphi^2} + \frac{1}{\rho} \frac{\partial z}{\partial \rho} = 0 . \tag{21}$$

REMARK 4. Without making use of the above formulae we can — by a rather more laborious process — reach the same result directly, starting out from the transformation inverse to (15),

$$\varrho = \sqrt{(x^2 + y^2)}, \quad \varphi = \arctan \frac{y}{x} \quad (+ \text{ const.}) \tag{22}$$

and applying the equations

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial \varrho} \frac{\partial \varrho}{\partial x} + \frac{\partial z}{\partial \varphi} \frac{\partial \varphi}{\partial x}, \quad \frac{\partial z}{\partial y} = \frac{\partial z}{\partial \varrho} \frac{\partial \varrho}{\partial y} + \frac{\partial z}{\partial \varphi} \frac{\partial \varphi}{\partial y} .$$

The constant in the second equation (22) depends on the quadrant in which the point (x, y) lies. It is irrelevant when differentiating.

Another method, making use of the simplicity of equation (14) and the substitution (15), is the following: Differentiating the first equation (16) with respect to ϱ, we obtain

$$\frac{\partial^2 z}{\partial \rho^2} = \left(\frac{\partial^2 z}{\partial x^2} \cos \varphi + \frac{\partial^2 z}{\partial x \, \partial y} \sin \varphi \right) \cos \varphi + \left(\frac{\partial^2 z}{\partial y \, \partial x} \cos \varphi + \frac{\partial^2 z}{\partial y^2} \sin \varphi \right) \sin \varphi ,$$

and differentiating the second equation with respect to φ we have

$$\frac{\partial^2 z}{\partial \varphi^2} = \left(- \frac{\partial^2 z}{\partial x^2} \varrho \sin \varphi + \frac{\partial^2 z}{\partial x \, \partial y} \varrho \cos \varphi \right) (-\varrho \sin \varphi) - \frac{\partial z}{\partial x} \varrho \cos \varphi +$$

$$+ \left(- \frac{\partial^2 z}{\partial y \, \partial x} \varrho \sin \varphi + \frac{\partial^2 z}{\partial y^2} \varrho \cos \varphi \right) \varrho \cos \varphi - \frac{\partial z}{\partial y} \varrho \sin \varphi .$$

Then

$$\frac{\partial^2 z}{\partial \varrho^2} + \frac{1}{\varrho^2} \frac{\partial^2 z}{\partial \varphi^2} + \frac{1}{\varrho} \frac{\partial z}{\partial \varrho} = \frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} .$$

REMARK 5. The method of transformation in the case of several variables is quite similar. For example, the equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0 \tag{23}$$

is converted (i) by introducing cylindrical coordinates

$$x = \varrho \cos\varphi, \quad y = \varrho \sin\varphi, \quad z = z \tag{24}$$

into the equation

$$\frac{\partial^2 u}{\partial \varrho^2} + \frac{1}{\varrho^2}\frac{\partial^2 u}{\partial \varphi^2} + \frac{1}{\rho}\frac{\partial u}{\partial \rho} + \frac{\partial^2 u}{\partial z^2} = 0, \tag{25}$$

and (ii) by introducing spherical coordinates

$$x = r \sin\vartheta \cos\varphi, \quad y = r \sin\vartheta \sin\varphi, \quad z = r\cos\vartheta \tag{26}$$

into the equation

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \vartheta^2} + \frac{1}{r^2 \sin^2\vartheta}\frac{\partial^2 u}{\partial \varphi^2} + \frac{2}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\cot\vartheta\frac{\partial u}{\partial \vartheta} = 0. \tag{27}$$

The expression

$$\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2$$

is converted by the transformation (15) into the expression

$$\left(\frac{\partial z}{\partial \varrho}\right)^2 + \frac{1}{\rho^2}\left(\frac{\partial z}{\partial \varphi}\right)^2.$$

The expression

$$\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial u}{\partial z}\right)^2$$

is converted by the transformation (24) into the expression

$$\left(\frac{\partial u}{\partial \rho}\right)^2 + \frac{1}{\rho^2}\left(\frac{\partial u}{\partial \varphi}\right)^2 + \left(\frac{\partial u}{\partial z}\right)^2,$$

and by the transformation (26) into the expression

$$\left(\frac{\partial u}{\partial r}\right)^2 + \frac{1}{r^2 \sin^2\vartheta}\left(\frac{\partial u}{\partial \varphi}\right)^2 + \frac{1}{r^2}\left(\frac{\partial u}{\partial \vartheta}\right)^2.$$

REMARK 6. The transformation of a dependent variable may be carried out in the same way as in (β). If, for example, $z = \varphi(t)$, then

$$\frac{\partial z}{\partial x} = \frac{\mathrm{d}z}{\mathrm{d}t}\frac{\partial t}{\partial x}, \quad \frac{\partial z}{\partial y} = \frac{\mathrm{d}z}{\mathrm{d}t}\frac{\partial t}{\partial y},$$

whence we obtain $\partial t/\partial x$ and $\partial t/\partial y$. Further,

$$\frac{\partial^2 z}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\mathrm{d}z}{\mathrm{d}t}\frac{\partial t}{\partial x}\right) = \left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathrm{d}z}{\mathrm{d}t}\right)\frac{\partial t}{\partial x}\right]\frac{\partial t}{\partial x} + \frac{\mathrm{d}z}{\mathrm{d}t}\frac{\partial^2 t}{\partial x^2} = \frac{\mathrm{d}^2 z}{\mathrm{d}t^2}\left(\frac{\partial t}{\partial x}\right)^2 + \frac{\mathrm{d}z}{\mathrm{d}t}\frac{\partial^2 t}{\partial x^2}, \quad \text{etc.}$$

12.12. Extremes of Functions of Several Variables. Constrained Extremes. Lagrange's Method of Undetermined Coefficients. Extremes of Implicit Functions

Definition 1. The function $z = f(x, y)$ is said to have a *relative (local) maximum at the point* (x_0, y_0) if there exists a neighbourhood U of the point (x_0, y_0) such that the relation

$$f(x, y) \leqq f(x_0, y_0) \tag{1}$$

holds at every point of this neighbourhood. A *relative minimum* is defined similarly and so are *relative extremes* for functions of several variables.

If, in the neighbourhood U of the point (x_0, y_0), the equality in (1) occurs *only* at the point (x_0, y_0) we speak of a *strict relative maximum.* (Similarly for a minimum and for functions of several variables.)

Definition 2. The function $z = f(x, y)$ is said to have at the point (x_0, y_0) a *maximum in the region (set)* M (a so-called *absolute maximum*) if the relation $f(x, y) \leqq f(x_0, y_0)$ holds at every point (x, y) of M.

REMARK 1. An (*absolute*) *minimum on* M is defined similarly. The definition of absolute extremes for functions of several variables is also similar.

Theorem 1. *If the function* $z = f(x, y)$ *possesses partial derivatives of the first order in the region* O, *then it may attain a relative extreme only at a point* (x_0, y_0) *for which*

$$\frac{\partial f}{\partial x}(x_0, y_0) = 0, \quad \frac{\partial f}{\partial y}(x_0, y_0) = 0. \tag{2}$$

If, moreover, $z = f(x, y)$ *possesses a second total differential at this point (a sufficient condition for this is the continuity of the second partial derivatives) and if*

$$AC - B^2 > 0, \tag{3}$$

where

$$A = \frac{\partial^2 f}{\partial x^2}(x_0, y_0), \quad B = \frac{\partial^2 f}{\partial x\, \partial y}(x_0, y_0), \quad C = \frac{\partial^2 f}{\partial y^2}(x_0, y_0),$$

then there is in fact a relative extreme at the point (x_0, y_0), *namely a strict relative maximum when* $A < 0$, *a strict relative minimum when* $A > 0$. *If*

$$AC - B^2 < 0, \tag{4}$$

then there is no extreme at the point (x_0, y_0). *If this expression vanishes at the point* (x_0, y_0), *then there may be, but need not be, a relative extreme at the point* (x_0, y_0) *and further investigation is necessary.*

REMARK 2. As a rule, this further investigation is done by considering straight lines $y - y_0 = k(x - x_0)$ drawn through the point (x_0, y_0) and investigating the given function only on these lines as a function of the single variable x. (But if the given function has, for example, a strict relative maximum on every straight line going through the point (x_0, y_0), then generally it does not follow that it has a strict relative maximum at this point by Definition 1 when considered as a function of the variables x, y.) In addition: If $AC - B^2 = 0$ at the point (x_0, y_0) and $AC - B^2 > 0$, or $AC - B^2 < 0$ holds in a certain neighbourhood of the point (x_0, y_0) from which the point (x_0, y_0) is excluded, then $f(x, y)$ has, or has not a strict relative extreme at that point, respectively. (Here $A = \frac{\partial^2 f}{\partial x^2}(x, y)$, etc.)

Theorem 2. *If* $z = f(x_1, x_2, \ldots, x_n)$ *possesses partial derivatives of the first order in the region* O, *then necessary conditions for the existence of a relative extreme at a point* $P(x_1^0, x_2^0, \ldots, x_n^0) \in O$ *are*

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

at that point.

Sufficient conditions: Let, moreover, $f(x_1, x_2, \ldots, x_n)$ have continuous derivatives of the second order in a neighbourhood of the point P. Denote

$$\frac{\partial^2 f}{\partial x_i\, \partial x_k}(x_1^0, x_2^0, \ldots, x_n^0) = A_{ik},^* \quad i = 1, 2, \ldots, n, \quad k = 1, 2, \ldots, n;$$

form the matrix

$$\mathbf{A} = \begin{bmatrix} A_{11}, & A_{12}, & \ldots, & A_{1n} \\ A_{21}, & A_{22}, & \ldots, & A_{2n} \\ \ldots\ldots\ldots\ldots\ldots & & & \\ A_{n1}, & A_{n2}, & \ldots, & A_{nn} \end{bmatrix}.$$

* $A_{ik} = A_{ki}$ under the assumptions considered.

If this matrix is positive definite, i.e. if all its principal minors

$$A_{11}, \quad \begin{vmatrix} A_{11}, & A_{12} \\ A_{21}, & A_{22} \end{vmatrix}, \quad \ldots,$$

(p. 68) are positive, then $f(x_1, x_2, \ldots, x_n)$ has a strict relative minimum at the point $(x_1^0, x_2^0, \ldots, x_n^0)$.

If $\boldsymbol{A}$ is negative definite, i.e. if its principal minors of odd and even order are negative and positive, respectively, then $f(x_1, x_2, \ldots, x_n)$ has a strict relative maximum at the point $(x_1^0, x_2^0, \ldots, x_n^0)$.

REMARK 3. The solution of equations (2), and even more so that of equations (5), is difficult in the general case. (In the case of equations (2) we have to find – geometrically speaking – the points of intersection of the curves (2).) In applications, however, there are usually no difficulties in solving those equations. See also Example 31.5.1.

Example 1. Let us find the dimensions of a rectangular water-tank of volume 32 m^3 with minimal area of its base and vertical walls.

If we denote the dimensions of the base by x and y and the height of the tank by z, then we have for the volume

$$V = xyz \tag{6}$$

and for the area of the walls and base

$$P = xy + 2xz + 2yz\,. \tag{7}$$

P is a function of x, y and z but, in fact, it depends only upon two variables, because x, y and z are subject to the relation (6) where V is a given constant, so that (after substituting $z = V/xy$) we have

$$P = xy + \frac{2V}{y} + \frac{2V}{x}\,. \tag{8}$$

Equations (2) now give

$$\frac{\partial P}{\partial x} = 0\,, \quad \frac{\partial P}{\partial y} = 0 \text{ or } y - \frac{2V}{x^2} = 0\,, \quad x - \frac{2V}{y^2} = 0\,. \tag{9}$$

From (9) it follows that

$$x^2 y = 2V, \quad xy^2 = 2V; \tag{10}$$

by dividing the first of these equations by the second we get $x/y = 1$ or $x = y$. If we put $y = x$ in the first equation (10) we obtain $x = \sqrt[3]{(2V)}$ and then $y = \sqrt[3]{(2V)}$. From (6) it then follows that $z = \sqrt[3]{(V/4)}$. Substituting the numerical values we obtain $x = y = 4$ m, $z = 2$ m.

We have really found a (strict) relative extreme because

$$\frac{\partial^2 P}{\partial x^2} = \frac{4V}{x^3} = \frac{4V}{2V} = 2\,, \quad \frac{\partial^2 P}{\partial y^2} = \frac{4V}{y^3} = 2\,, \quad \frac{\partial^2 P}{\partial x\,\partial y} = 1$$

so that $AC - B^2 = 4 - 1 = 3 > 0$ in accordance with (3); also, $\partial^2 P/\partial x^2 = 2$ so that a minimum is assured. It is easy to verify that this relative minimum is a minimum in the whole region $x > 0$, $y > 0$.

REMARK 4. The *extremes* of a given function *in a domain* M can be expected at the points, where (2) or (5) are fulfilled, or where these derivatives do not exist, or, finally, on the boundary of the domain M. The last problem leads to the so-called *constrained extremes* (or *extremes with subsidiary conditions*).

Definition 3. We say that the function $z = f(x, y)$ has a *relative maximum* (*or minimum*) *on the curve* $g(x, y) = 0$ *at the point* (x_0, y_0) for which $g(x_0, y_0) = 0$ (so-called *relative constrained extreme*) if the relation $f(x_0, y_0) \geqq f(x, y)$ (or $f(x_0, y_0) \leqq f(x, y)$) holds at every point (x, y) on the curve $g(x, y) = 0$ in a certain neighbourhood of the point (x_0, y_0). If in this neighbourhood the equality occurs *only* at the point (x_0, y_0), then we speak of a *strict relative constrained extreme.*

If $f(x_0, y_0) \geqq f(x, y)$ or $f(x_0, y_0) \leqq f(x, y)$ holds at *every* point (x, y) of the curve $g(x, y) = 0$, then we speak of an (*absolute*) *extreme of the function* $f(x, y)$ *on the curve* $g(x, y) = 0$ (or of an *absolute constrained extreme*) at the point (x_0, y_0). The curve $g(x, y)$ is often called a *constraint.*

REMARK 5. In applications, we find the constrained extremes (at least on a certain part of the given curve) either by computing, for example, y as a function of x, substituting into $z = f(x, y)$ and thus reducing the problem to one of finding extremes of a function of a single variable (Example 2) or by the application of Lagrange's method (Example 3).

Example 2. Let us find the (absolute) maximum and minimum of the function

$$z = x^2 + y^2 - 2x - 4y + 1 \tag{11}$$

in a closed triangle (i.e. its boundary included) with vertices (0, 0), (3, 0), (0, 5).

First, we find the relative extremes by (2):

$$2x - 2 = 0\,, \quad 2y - 4 = 0\,, \quad \text{hence} \quad x = 1\,, \quad y = 2\,. \tag{12}$$

By Theorem 1 we easily verify that at the point (1, 2) there is a (strict) relative minimum, $z = -4$; this point is an interior point of the given triangle.

Constrained extremes: The equations of the sides of the triangle are (§ 5.4, p. 170):

$$y = 0\,, \quad x = 0\,, \quad 5x + 3y - 15 = 0\,. \tag{13}$$

On the segment $y = 0$, $0 \leqq x \leqq 3$, the function (11) is of the form $z = x^2 - 2x + 1$. From $\mathrm{d}z/\mathrm{d}x = 2x - 2 = 0$ it follows that $x = 1$. Because $\mathrm{d}^2z/\mathrm{d}x^2(1, 0) = 2$, there is a (strict) relative minimum, $z = 0$, at the point $x = 1$. There are no other relative extremes of the function $z = x^2 - 2x + 1$. At the first end point $x = 0$, we have $z = 1$, at the second end point $x = 3$, $z = 4$. Thus the function $z = x^2 - 2x + 1$ attains its maximum value on the segment concerned at the point $x = 3$. The absolute extremes of the function (11) on the segment $y = 0$, $0 \leqq x \leqq 3$ are therefore at the points $(1, 0)$, $(3, 0)$.

Similarly, we find that the absolute extremes of the function (11) on the segment $x = 0$, $0 \leqq y \leqq 5$ are at the points $(0, 2)$ (minimum, $z = -3$), $(0, 5)$ (maximum, $z = 6$).

The third side of the triangle is the segment $5x + 3y - 15 = 0$, $0 \leqq x \leqq 3$. Hence, $y = (15 - 5x)/3$. If we put y into (11) we obtain after rearrangement $9z = 34x^2 - 108x + 54$. If we put the first derivative equal to zero, i.e. $68x - 108 = 0$, we get $x = \frac{27}{17}$, $y = \frac{40}{17}$. The second derivative with respect to x is positive; so we have a (strict) relative minimum, $z \doteq -3{\cdot}6$, at this point. At the end points of the segment, $(3, 0)$, $(0, 5)$ we have $z = 4$, $z = 6$, respectively. The absolute extremes on the investigated segment are therefore at the points $(\frac{27}{17}, \frac{40}{17})$ and $(0, 5)$.

The result. The function (11) attains its (absolute) minimum, $z = -4$ at the point $(1, 2)$ and its (absolute) maximum $z = 6$, at the point $(0, 5)$ of the triangle concerned.

Theorem 3 (*Lagrange's Method of Undetermined Coefficients* (*Multipliers*)). *Let $f(x, y)$ and $g(x, y)$ have total differentials in a neighbourhood of the points of the curve $g(x, y) = 0$. Let at least one of the derivatives $\partial g/\partial x$, $\partial g/\partial y$ be non-zero at every point of the curve $g(x, y) = 0$. If the function $z = f(x, y)$ has a relative extreme on the curve $g(x, y) = 0$ at a point (x_0, y_0) of this curve, then there exists a constant λ such that for the function*

$$F(x, y) = f(x, y) + \lambda g(x, y) \tag{14}$$

the equations

$$\frac{\partial F}{\partial x}(x_0, y_0) = 0\,, \quad \frac{\partial F}{\partial y}(x_0, y_0) = 0 \quad (\textit{and also } g(x_0, y_0) = 0) \tag{15}$$

are satisfied at the point (x_0, y_0).

REMARK 6. Constrained extremes can thus be found by constructing the function (14) and then solving equations (15) for the unknowns x_0, y_0, λ. We observe that equations (15) are *necessary* conditions for the existence of a constrained extreme.

REMARK 7. *Sufficient conditions*: Let us construct the second differential of the function (14) at the above-mentioned point (x_0, y_0),

$$\mathrm{d}^2F(x_0, y_0) = \frac{\partial^2 F}{\partial x^2}(x_0, y_0)\,\mathrm{d}x^2 + 2\,\frac{\partial^2 F}{\partial x\,\partial y}(x_0, y_0)\,\mathrm{d}x\,\mathrm{d}y + \frac{\partial^2 F}{\partial y^2}(x_0, y_0)\,\mathrm{d}y^2\,. \tag{16}$$

If for all points $(x_0 + dx, y_0 + dy)$ from a certain neighbourhood of the point (x_0, y_0), and such that $g(x_0 + dx, y_0 + dy) = 0$ and dx and dy are not simultaneously zero, the differential (16) is positive (or negative), then there is a constrained relative minimum (or maximum) at the point (x_0, y_0).

REMARK 8. For this condition to be fulfilled it is sufficient that the form (16) be positive (or negative) definite at the point (x_0, y_0) (§ 1.29, p. 67). For this it is again sufficient that

$$AC - B^2 > 0\,, \tag{17}$$

where

$$A = \frac{\partial^2 F}{\partial x^2}(x_0, y_0)\,, \quad B = \frac{\partial^2 F}{\partial x\,\partial y}(x_0, y_0)\,, \quad C = \frac{\partial^2 F}{\partial y^2}(x_0, y_0)\,,$$

and that

$$\frac{\partial^2 F}{\partial x^2}(x_0, y_0) > 0 \quad \left(\text{or } \frac{\partial^2 F}{\partial x^2}(x_0, y_0) < 0\right). \tag{18}$$

Example 3. Let us find the constrained extremes of the function

$$z = x^2 + y^2 \tag{19}$$

on the curve

$$x^2 + 4y^2 - 1 = 0\,.$$

At least one of the derivatives with respect to x or y is non-zero at every point of the given ellipse. We construct the function (14),

$$F(x, y) = x^2 + y^2 + \lambda(x^2 + 4y^2 - 1) = x^2(1 + \lambda) + y^2(1 + 4\lambda) - \lambda\,. \tag{20}$$

Equations (15) are now

$$2x_0(1 + \lambda) = 0\,, \tag{21}$$
$$2y_0(1 + 4\lambda) = 0\,, \tag{22}$$
$$x_0^2 + 4y_0^2 - 1 = 0\,. \tag{23}$$

If $\lambda \neq -1$ and also $\lambda \neq -\frac{1}{4}$, then by (21) and (22) $x_0 = y_0 = 0$ and (23) is not fulfilled. It follows that either $\lambda = -1$ or $\lambda = -\frac{1}{4}$. For $\lambda = -1$ we obtain from (22) $y_0 = 0$ and from (23) $x_0 = \pm 1$. For $\lambda = -\frac{1}{4}$ we obtain similarly $x_0 = 0$, $y_0 = \pm 0{\cdot}5$. So we have four points:

For $\lambda = -1$: $(1, 0)$, $(-1, 0)$; for $\lambda = -\frac{1}{4}$: $(0, 0{\cdot}5)$, $(0, -0{\cdot}5)$. (24)

For the case $\lambda = -1$ the second differential (16) is (see (20))

$$d^2F(\pm 1; 0) = -6dy^2\,, \tag{25}$$

for $\lambda = -\frac{1}{4}$

$$d^2F(0; \pm 0{\cdot}5) = +\tfrac{3}{2}\,dx^2\,. \tag{26}$$

From the form of the right-hand member of (25) it can be seen that the condition of Remark 7 or 8 is fulfilled at the points $(\pm 1, 0)$ and a constrained relative maximum, $z = 1$, is attained there. Similarly, at the points $(0, \pm 0{\cdot}5)$ a constrained relative minimum, $z = 0{\cdot}25$ is attained. (Evidently, these extremes are also absolute constrained extremes.)

REMARK 9. We could not apply (17) here, because $AC - B^2 = 4(1 + \lambda)(1 + 4\lambda) = 0$ at the points considered.

REMARK 10. The problem of finding the constrained extremes of a function of several variables is solved similarly. For example, a necessary condition for the existence of a relative extreme of the function

$$w = f(x, y, z, u, v)$$

with subsidiary conditions

$$F_1(x, y, z, u, v) = 0\,, \quad F_2(x, y, z, u, v) = 0\,,$$

is that the equations

$$\frac{\partial G}{\partial x} = 0\,, \quad \frac{\partial G}{\partial y} = 0\,, \quad \frac{\partial G}{\partial z} = 0\,, \quad \frac{\partial G}{\partial u} = 0\,, \quad \frac{\partial G}{\partial v} = 0\,, \quad F_1 = 0\,, \quad F_2 = 0\,,$$

where

$$G = f + \lambda_1 F_1 + \lambda_2 F_2\,,$$

be satisfied. (We suppose that at least one of the functional determinants

$$\frac{\partial(F_1, F_2)}{\partial(x, y)}\,, \quad \frac{\partial(F_1, F_2)}{\partial(x, z)}\,, \quad \ldots\,, \quad \frac{\partial(F_1, F_2)}{\partial(u, v)}$$

is non-zero at the points of the hypersurfaces $F_1 = 0$, $F_2 = 0$.) This gives a system of seven equations for seven unknowns $x_0, y_0, z_0, u_0, v_0, \lambda_1, \lambda_2$.

The sufficient conditions are similar to those in Remarks 7 and 8. In particular, we obtain a constrained relative minimum if the second differential of the function G at the point $(x_0, y_0, z_0, u_0, v_0)$ is a positive definite form and a constrained relative maximum if it is a negative definite form.

REMARK 11. *Relative extremes of implicit functions* are found from equations (12.9.2) or (12.9.11), pp. 424 and 427. Since for a function of one variable we require that $y' = 0$, equation (12.9.2) gives the condition

$$\frac{\partial f}{\partial x} = 0\,, \tag{27}$$

while at the same time $f(x, y) = 0$ has to be fulfilled, as the required point (x_0, y_0) has to lie on the curve $f(x, y) = 0$, and

$$\frac{\partial f}{\partial y} \neq 0 .$$

It then follows from equation (12.9.4) (since $y' = 0$) that $y'' \neq 0$ if

$$\frac{\partial^2 f}{\partial x^2} \neq 0 .$$

Generally: The relative extremes of the function $y = \varphi(x)$ given implicitly by the equation $f(x, y) = 0$ (more exactly: of all functions defined implicitly by the equation $f(x, y) = 0$) are to be found at the points at which $\partial f/\partial x = 0$ and simultaneously $f(x, y) = 0$. If at such a point $\partial f/\partial y \neq 0$, then the existence of a relative extreme is assured if the first non-zero derivative in the sequence

$$\frac{\partial f}{\partial x}, \quad \frac{\partial^2 f}{\partial x^2}, \quad \frac{\partial^3 f}{\partial x^3}, \quad \ldots$$

is of an even order. If its sign is the same as that of $\partial f/\partial y$, we obtain a strict relative maximum, if its sign is opposite, we obtain a strict relative minimum.

The case of functions of several variables can be dealt with in a similar manner. For example, the extremes of a function $z = \varphi(x, y)$ given by $f(x, y, z) = 0$ are to be found at the points where, simultaneously,

$$f(x, y, z) = 0 , \quad \frac{\partial f}{\partial x} = 0 , \quad \frac{\partial f}{\partial y} = 0$$

(see equation (12.9.11)) under the assumption that

$$\frac{\partial f}{\partial z} \neq 0 .$$

Example 4. Let us find the extremes of the function (more exactly: of the functions defined implicitly by the equation)

$$x^2 + y^2 - 25 = 0 .$$

According to (27) we solve the system

$$2x = 0 , \quad x^2 + y^2 - 25 = 0 .$$

Solving it, we get two points: $(0, 5)$, $(0, -5)$. At both points $\partial f/\partial y \neq 0$, since

$$\frac{\partial f}{\partial y}(0, 5) = 10 , \quad \frac{\partial f}{\partial y}(0, -5) = -10 .$$

Further

$$\frac{\partial^2 f}{\partial x^2} = 2 .$$

At the point $(0, 5)$ there is a strict relative maximum because

$$\frac{\partial^2 f}{\partial x^2} \quad \text{and} \quad \frac{\partial f}{\partial y}$$

are of the same sign at that point. At the point $(0, -5)$ there is a strict relative minimum because the derivatives are of opposite signs. (This example is only an illustrative one – geometrically the problem is self-evident.)

REMARK 12. From equations (12.9.2) and (12.9.4) and from the corresponding equations for higher derivatives conclusions may also be drawn (on the basis of theorems from § 11.9) on other properties of implicit functions (on convexity, on points of inflexion, etc.), not only on extremes. Implicit functions of several variables may also be studied by means of equations of the type (12.9.11) and of similar equations for higher derivatives.

12.13. Survey of Some Important Formulae from Chapter 12

1. $$\lim_{\substack{x\to x_0\\ y\to y_0}} [f(x, y) \pm g(x, y)] = \lim_{\substack{x\to x_0\\ y\to y_0}} f(x, y) \pm \lim_{\substack{x\to x_0\\ y\to y_0}} g(x, y) ,$$

$$\lim_{\substack{x\to x_0\\ y\to y_0}} kf(x, y) = k \lim_{\substack{x\to x_0\\ y\to y_0}} f(x, y) \quad (k \text{ a constant}) ,$$

$$\lim_{\substack{x\to x_0\\ y\to y_0}} [f(x, y)\, g(x, y)] = \lim_{\substack{x\to x_0\\ y\to y_0}} f(x, y) \lim_{\substack{x\to x_0\\ y\to y_0}} g(x, y) ,$$

$$\lim_{\substack{x\to x_0\\ y\to y_0}} \frac{f(x, y)}{g(x, y)} = \frac{\lim\limits_{\substack{x\to x_0\\ y\to y_0}} f(x, y)}{\lim\limits_{\substack{x\to x_0\\ y\to y_0}} g(x, y)} \quad \text{if} \quad \lim_{\substack{x\to x_0\\ y\to y_0}} g(x, y) \neq 0 \quad \text{(Theorem 12.1.2).}$$

2. $$\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{h\to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h} ,$$

$$\frac{\partial f}{\partial y}(x_0, y_0) = \lim_{k=0} \frac{f(x_0, y_0 + k) - f(x_0, y_0)}{k} \quad \text{(Definition 12.2.1).}$$

3. $$\frac{\partial^2 f}{\partial x\, \partial y} = \frac{\partial^2 f}{\partial y\, \partial x} \quad \text{(Theorem 12.2.1).}$$

4. $dz = \frac{\partial z}{\partial x}\,dx + \frac{\partial z}{\partial y}\,dy$, $d^{(m)} z = \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^m z$

(Definition 12.3.2, Remark 12.3.4),

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial x}, \quad \frac{\partial z}{\partial y} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial y}$$

(differentiation of composite functions, Theorem 12.4.1; see also Remarks 12.4.1 and 12.4.2).

5. $z - z_0 = \left(\frac{\partial f}{\partial x}\right)_0 (x - x_0) + \left(\frac{\partial f}{\partial y}\right)_0 (y - y_0)$

(equation of a tangent plane, Theorem 12.3.4).

6. $\frac{\partial f}{\partial s} = \left(\frac{\partial f}{\partial x}\right)_0 \cos\alpha + \left(\frac{\partial f}{\partial y}\right)_0 \sin\alpha$

(differentiation in a given direction, Theorem 12.5.2).

7. $x\frac{\partial f}{\partial x} + y\frac{\partial f}{\partial y} = nf(x, y)$ (Theorem 12.6.1 on homogeneous functions).

8. $y' = -\frac{\partial f/\partial x}{\partial f/\partial y}$

(differentiation of a function y given implicitly by an equation $f(x, y) = 0$, Theorem 12.9.1).

9. Extremes of a function $z = f(x, y)$: If

$$\frac{\partial f}{\partial x}(x_0, y_0) = 0, \quad \frac{\partial f}{\partial y}(x_0, y_0) = 0,$$

$$\frac{\partial^2 f}{\partial x^2}(x_0, y_0)\frac{\partial^2 f}{\partial y^2}(x_0, y_0) - \left[\frac{\partial^2 f}{\partial x\,\partial y}(x_0, y_0)\right]^2 > 0,$$

then for $\frac{\partial^2 f}{\partial x^2}(x_0, y_0) > 0$ the function has a strict relative minimum at the point (x_0, y_0), for $\frac{\partial^2 f}{\partial x^2}(x_0, y_0) < 0$ a strict relative maximum (Theorem 12.12.1).

13. INTEGRAL CALCULUS OF FUNCTIONS OF ONE VARIABLE

By KAREL REKTORYS

References: [1], [4], [15], [26], [29], [31], [39], [40], [41], [47], [54], [59], [66], [68], [72], [76], [78], [81], [87], [91], [94], [95], [96], [106], [109], [111], [115], [119], [122], [123], [127], [131], [133], [142], [144], [145], [146], [148], [158], [160], [176], [182], [183], [188].

13.1. Primitive Function (Indefinite Integral). Basic Integrals

Definition 1. The function $F(x)$ is said to be a *primitive* (*primitive function, indefinite integral*) of the function $f(x)$ in the interval (a, b), if the relation $F'(x) = f(x)$ holds for all $x \in (a, b)$.

Theorem 1. *For each function $f(x)$ which is continuous in (a, b) there exists a primitive. In fact, there exists an infinite number of them. If $F(x)$ is a primitive, then all others are of the form*

$$F(x) + C\,, \tag{1}$$

where C is an arbitrary constant.

We write

$$\int f(x)\,\mathrm{d}x = F(x) + C\,. \tag{2}$$

The function $f(x)$ itself is called the *integrand* of the integral (2).

REMARK 1. According to the definition we have, keeping C fixed in (2),

$$\frac{\mathrm{d}}{\mathrm{d}x}\int f(x)\,\mathrm{d}x = f(x)\,.$$

(In this sense differentiation and integration may be regarded as inverse operations.)

Example 1. We have

$$\int x^2\,\mathrm{d}x = \frac{x^3}{3} + C \tag{3}$$

(where C is an arbitrary constant, e.g. 7). For, if we differentiate the function on the

right-hand side of equation (3) (keeping C fixed), we obtain x^2. According to Theorem 1, functions of the form $x^3/3 + C$ are the only ones which possess this property.

Theorem 2 (*Standard Integrals*). (*Unless the contrary is stated, the formulae are valid for all real x and for all values of constants involved.*)

1. $$\int x^n \, dx = \frac{x^{n+1}}{n+1} + C \quad (x > 0, \, n \text{ real}, \, n \neq -1).$$

(For some n, the range of validity may be extended. For instance, if k is a positive integer, the relation

$$\int x^k \, dx = \frac{x^{k+1}}{k+1} + C$$

holds for all x.)

2. $$\int \frac{dx}{x} = \ln |x| + C \quad (x \neq 0), \quad \text{or} \quad \int \frac{dx}{x} = \ln kx \quad (kx > 0).$$

3. $$\int e^x \, dx = e^x + C, \qquad \int a^x \, dx = \frac{a^x}{\ln a} + C \quad (a > 0).$$

4. $$\int \sin x \, dx = -\cos x + C, \qquad \int \cos x \, dx = \sin x + C.$$

5. $$\int \frac{dx}{\sin^2 x} = -\cot x + C \quad (x \neq n\pi, \, n \text{ an integer}),$$

$$\int \frac{dx}{\cos^2 x} = \tan x + C \quad (x \neq \tfrac{1}{2}\pi + n\pi, \quad n \text{ an integer}).$$

6. $$\int \frac{dx}{1 + x^2} = \arctan x + C = -\operatorname{arccot} x + k.$$

7. $$\int \frac{dx}{\sqrt{(1 - x^2)}} = \arcsin x + C = -\arccos x + k \quad (-1 < x < 1).$$

8. $$\int \frac{dx}{1 - x^2} = \tfrac{1}{2} \ln \left| \frac{1 + x}{1 - x} \right| + C \quad \text{for} \quad |x| \neq 1,$$

$$= \operatorname{artanh} x + C \quad \text{for} \quad |x| < 1,$$

$$= \operatorname{arcoth} x + C \quad \text{for} \quad |x| > 1.$$

9. $$\int \sinh x \, dx = \cosh x + C, \qquad \int \cosh x \, dx = \sinh x + C.$$

10. $\int \frac{dx}{\sinh^2 x} = -\coth x + C \quad (x \neq 0), \quad \int \frac{dx}{\cosh^2 x} = \tanh x + C.$

11. $\int \frac{dx}{\sqrt{(x^2+1)}} = \ln\left[x + \sqrt{(x^2+1)}\right] + C = \operatorname{arsinh} x + C.$

12. $\int \frac{dx}{\sqrt{(x^2-1)}} = \ln\left|x + \sqrt{(x^2-1)}\right| + C \quad (|x| > 1),$

$= \operatorname{arcosh} x + C \quad (x > 1).$

REMARK 2. arsinh x, arcosh x, artanh x, arcoth x are the functions inverse to the hyperbolic functions ($x = \sinh y$ etc., see § 2.13).

REMARK 3. We have given two expressions for the integral $\int dx/x$. In the domain where x is negative, $\int dx/x = \ln(-x) + C$ (for there we have $|x| = -x$), or $\int dx/x = \ln kx$, where $k < 0$. The fact that $\ln kx$ is the primitive of the function $1/x$, follows from the chain rule (Theorem 11.5.5), since $(\ln kx)' = (1/kx) \cdot k = 1/x$.

Two forms of the indefinite integral do not contradict Theorem 1, for one can be reduced to the other. For instance, $\ln kx = \ln x + \ln k$ holds for $x > 0$ and $k > 0$, and the right-hand side is already of the form $\ln|x| + C$. This situation occurs very frequently. If two forms of an indefinite integral are found, it is always possible to reduce one to the other.

REMARK 4. Even though for each continuous function there exists a primitive, in many cases it is not possible to express it in terms of elementary functions (i.e. algebraic functions and elementary transcendental functions, see § 11.2). An example of such a primitive is the indefinite integral of the function e^{x^2}. Hence a new transcendental function is defined by the integral

$$\int e^{x^2} dx.$$

Similarly, new transcendental functions are defined by the integrals

$$\operatorname{Si} x = \int_0^x \frac{\sin t}{t} dt \quad (\textit{sine integral}), \quad \operatorname{Ci} x = -\int_x^\infty \frac{\cos t}{t} dt \quad (\textit{cosine integral}),$$

$$\operatorname{li} x = \int_0^x \frac{dt}{\ln t} \quad (\textit{logarithm integral}), \quad \operatorname{Ei}(x) = \int_{-\infty}^x \frac{e^t}{t} dt.$$

(At singular points of the integrands, the integrals are understood in the sense of the Cauchy principal value, see Remark 13.8.3.) See also Theorem 13.12.1, p. 550. On elliptic integrals see § 13.12.

13.2. Methods of Integration. Integration by Parts. Method of Substitution. Method of Differentiation with Respect to a Parameter

Theorem 1. *If there exist primitives of the functions* $f_1(x), f_2(x), f(x)$, *then*

$$\int [f_1(x) \pm f_2(x)]\,dx = \int f_1(x)\,dx \pm \int f_2(x)\,dx\,,$$

$$\int k f(x)\,dx = k \int f(x)\,dx \quad (k = \text{const.})\,.$$

(*The relation* $\int f(x)\,g(x)\,dx = \int f(x)\,dx \,.\, \int g(x)\,dx$ *does not hold in general.*)

Example 1.

$$\int \frac{4x-1}{\sqrt{x}}\,dx = \int \left(4\sqrt{(x)} - \frac{1}{\sqrt{(x)}}\right) dx =$$

$$= 4\int x^{1/2}\,dx - \int x^{-1/2}\,dx = \tfrac{8}{3}x^{3/2} - 2x^{1/2} + C = \tfrac{8}{3}\sqrt{(x^3)} - 2\sqrt{(x)} + C \quad (x > 0)\,.$$

Theorem 2 (*Integration by Parts*). *Let us assume that the functions* $u(x)$ *and* $v(x)$ *possess continuous derivatives in the interval* (a, b) (*hence these functions are also continuous in the interval* (a, b) by Theorem 11.5.2). *Then the relation*

$$\int u'v\,dx = uv - \int uv'\,dx \tag{1}$$

holds in the interval (a, b).

REMARK 1. Equation (1) is often written in the form

$$\int v\,du = uv - \int u\,dv\,.$$

REMARK 2. Integration by parts is convenient when integrating functions of the type

$$x^n \sin x\,, \quad x^n \cos x\,, \quad x^n e^x\,, \quad \ln x\,, \quad \arctan x\,,$$

etc. (In the two last cases we put $u' = 1$.) See Examples 2, 3, 4.

Example 2. To evaluate the integral

$$I = \int x \sin x\,dx$$

we put

$$u' = \sin x\,, \quad v = x\,, \quad \text{so that} \quad u = -\cos x\,, \quad v' = 1\,;$$

hence by (1)

$$\int x \sin x \, dx = -x \cos x + \int \cos x \, dx = -x \cos x + \sin x + C .$$

Example 3.

$$\int \ln x \, dx = x \ln x - \int x \frac{1}{x} dx = x \ln x - x + C = x(\ln x - 1) + C \quad (x > 0) ;$$

$$u' = 1 , \quad v = \ln x ,$$

$$u = x , \quad v' = \frac{1}{x} .$$

Example 4.

$$I = \int e^{ax} \sin bx \, dx .$$

First we put

$$u' = e^{ax} , \quad v = \sin bx \quad \text{hence} \quad u = \frac{1}{a} e^{ax} , \quad v' = b \cos bx ,$$

thus reducing the given integral to the integral of the function $e^{ax} \cos bx$. Again integrating by parts, with

$$u' = e^{ax} , \quad v = \cos bx , \quad u = \frac{1}{a} e^{ax} , \quad v' = -b \sin bx ,$$

we obtain the original integral with the opposite sign (and a different constant). From the equation so obtained the required integral may then be easily evaluated:

$$I = \int e^{ax} \sin bx \, dx = \frac{1}{a} e^{ax} \sin bx - \frac{b}{a} \int e^{ax} \cos bx \, dx =$$

$$= \frac{1}{a} e^{ax} \sin bx - \frac{b}{a} \left(\frac{1}{a} e^{ax} \cos bx + \frac{b}{a} \int e^{ax} \sin bx \, dx \right) =$$

$$= \frac{1}{a} e^{ax} \sin bx - \frac{b}{a^2} e^{ax} \cos bx - \frac{b^2}{a^2} I .$$

From this equation (to the right-hand side, we can assign an arbitrary constant) we obtain

$$I = \frac{1}{a^2 + b^2} (a e^{ax} \sin bx - b e^{ax} \cos bx) + C .$$

This method of procedure is often used.

Method of Substitution

A. Substitution $h(x) = z$:

Theorem 3. *Let $f(x)$ be of the form $f(x) = g(h(x))\, h'(x)$ in the interval (a, b), where $h'(x)$ is a continuous function in (a, b) and $g(z)$ is continuous for all $z = h(x)$ when x runs through the interval (a, b). Then*

$$\int f(x)\,\mathrm{d}x = \int g(h(x))\, h'(x)\,\mathrm{d}x = \int g(z)\,\mathrm{d}z = G(z) + C\,, \tag{2}$$

where $h(x)$ is to be substituted for z in the result.

B. Substitution $x = \varphi(z)$:

Theorem 4. *Let $f(x)$ be continuous in the interval (a, b). Let $x = \varphi(z)$ be a function (of the variable z), which is strictly increasing or strictly decreasing in the interval (α, β) and possesses a continuous derivative $\varphi'(z)$. Let us denote by $z = \psi(x)$ the function inverse to the function $x = \varphi(z)$. If, further, $a < \varphi(z) < b$ holds for $z \in (\alpha, \beta)$, then*

$$\int f(x)\,\mathrm{d}x = \int f(\varphi(z))\, \varphi'(z)\,\mathrm{d}z = H(z) + C\,, \tag{3}$$

where $\psi(x)$ is to be substituted for z in the result.

REMARK 3. Application of the method of substitution requires a certain amount of experience, in order to foresee the form of the resulting integral or in order to see in the function $f(x)$ the form $g(h(x))\, h'(x)$. Note that the existence of the inverse function to $h(x) = z$ is not assumed in Theorem 3.

REMARK 4. Note that the right-hand side of equation (3) is formally obtained by substituting $\varphi(z)$ for x and $\varphi'(z)\,\mathrm{d}z$ (which arises as a result of differentiation of the right-hand side of the equation $x = \varphi(z)$) for $\mathrm{d}x$. Similarly in (2).

In the following examples, the range of validity of the result is mentioned only in those cases where it is not evident.

Example 5.

$$\int \sin^2 x \cos x\,\mathrm{d}x = \int z^2\,\mathrm{d}z = \frac{z^3}{3} + C = \frac{\sin^3 x}{3} + C\,.$$

We have made use of the substitution $\sin x = z$ (from which it follows that $\cos x\,\mathrm{d}x = \mathrm{d}z$) and substituted $z = \sin x$ in the result (Theorem 3).

Example 6. Using the substitution

$$\tan x = z\,, \quad \frac{\mathrm{d}x}{\cos^2 x} = \mathrm{d}z \quad \text{or} \quad (1 + \tan^2 x)\,\mathrm{d}x = \mathrm{d}z\,,$$

we obtain (in every interval $(\frac{1}{2}k\pi, \frac{1}{2}(k+1)\pi)$, where k is an integer)

$$\int \tan^4 x\,dx = \int \frac{z^4}{1+z^2}\,dz = \int \frac{(z^4+z^2)-(z^2+1)+1}{z^2+1}\,dz =$$

$$= \int\left(z^2 - 1 + \frac{1}{z^2+1}\right)dz = \frac{z^3}{3} - z + \arctan z + C = \frac{\tan^3 x}{3} - \tan x + x + k$$

(since $\arctan \tan x = x + \text{const.}$).

Example 7 (using Theorem 3).

$$\int \frac{x^2}{\sqrt{(1-x^6)}}\,dx = \frac{1}{3}\int \frac{dz}{\sqrt{(1-z^2)}} = \tfrac{1}{3}\arcsin z + C = \tfrac{1}{3}\arcsin x^3 + C \quad (|x| < 1),$$

where

$$x^3 = z, \quad 3x^2\,dx = dz.$$

Example 8. Using the substitution

$$f(x) = z, \quad f'(x)\,dx = dz$$

we obtain

$$\int \frac{f'(x)}{f(x)}\,dx = \ln|f(x)| + C \quad (f(x) \neq 0).$$

For example

$$\int \frac{5x-1}{x^2+1}\,dx = \frac{5}{2}\int \frac{2x}{x^2+1}\,dx - \int \frac{dx}{x^2+1} = \frac{5}{2}\ln(x^2+1) - \arctan x + C.$$

Example 9. Using the substitution

$$ax + b = z, \quad a\,dx = dz$$

we have (if $F(z)$ denotes a primitive of $f(z)$)

$$\int f(ax+b)\,dx = \frac{1}{a}\int f(z)\,dz = \frac{1}{a}F(z) + C = \frac{1}{a}F(ax+b) + C.$$

For example

$$\int \cos(3x+5)\,dx = \tfrac{1}{3}\int \cos z\,dz = \tfrac{1}{3}\sin z + C = \tfrac{1}{3}\sin(3x+5) + C.$$

Example 10. Using the substitution

$$x = \sin z, \quad dx = \cos z\,dz$$

we obtain

$$\int\sqrt{(1-x^2)}\,dx = \int\sqrt{(1-\sin^2 z)}\cos z\,dz = \int\cos^2 z\,dz = \int\tfrac{1}{2}(1+\cos 2z)\,dz =$$

$$= \tfrac{1}{2}z + \tfrac{1}{4}\sin 2z + C = \tfrac{1}{2}z + \tfrac{1}{2}\sin z\cos z + C = \tfrac{1}{2}\arcsin x +$$

$$+\tfrac{1}{2}x\sqrt{(1-x^2)} + C$$

(Theorem 4). The interval $(-1, 1)$ for x corresponds to the interval $(-\frac{1}{2}\pi, \frac{1}{2}\pi)$ for z, which was denoted by (α, β) in Theorem 4. (The same considerations hold also for closed intervals.) The inverse function to $x = \sin z$ is $z = \arcsin x$; $\sqrt{(1-\sin^2 z)} = +\cos z$, for $\cos z > 0$ if z belongs to the interval $(-\frac{1}{2}\pi, \frac{1}{2}\pi)$. When integrating $\cos 2z$ we employ the substitution $2z = t$; cf. Example 9.

The integral

$$\int\sqrt{(a^2-x^2)}\,dx$$

can be evaluated either by the substitution $x = a\sin z$ or may be reduced by the substitution $x = at$ to the previous case.

For some further typical examples on the method of substitution see § 13.4.

REMARK 5. Theorems on integration by substitution and by parts may be formulated under rather weaker restrictions than those introduced in Theorems 2 and 3.

In practice the two methods are often combined:

Example 11. Integrating by parts,

$$u' = 1\,,\quad v = \arctan x\,,\quad u = x\,,\quad v' = \frac{1}{1+x^2}$$

and using the result of Example 8, we obtain

$$\int\arctan x\,dx = x\arctan x - \int\frac{x}{x^2+1}\,dx = x\arctan x - \tfrac{1}{2}\ln(x^2+1) + C\,.$$

REMARK 6 (*Method of Differentiation of Integrals with Respect to a Parameter*). We have (Example 13.4.6)

$$\int\frac{dx}{\sqrt{(x^2+a)}} = \ln\left[x + \sqrt{(x^2+a)}\right] + C \quad (a > 0)\,.$$

This equation, *formally* differentiated with respect to a (not with respect to x!), gives

$$-\int\frac{dx}{2\sqrt{(x^2+a)^3}} = \frac{1}{x+\sqrt{(x^2+a)}}\,\frac{1}{2\sqrt{(x^2+a)}}\,.$$

Hence, if differentiation with respect to a under the integral sign is "permissible", then this procedure gives the primitive of the function $1/\sqrt{(x^2 + a)^3}$. For the conditions under which this method leads to correct results we refer the reader to § 13.9. Here, we shall treat only the simplest case.

Theorem 5. *Let us denote*

$$\int f(x, a)\,dx = F(x, a) + C\,. \tag{4}$$

If the functions $f(x, a)$, $\frac{\partial f}{\partial a}(x, a)$, *are continuous as functions of two variables (in the region considered), then*

$$\int \frac{\partial f}{\partial a}(x, a)\,dx = \frac{\partial F}{\partial a}(x, a) + k\,. \tag{5}$$

Example 12. Let us determine

$$\int \frac{dx}{(a + bx^2)^2} \quad (a > 0, b > 0)\,.$$

We use the relation (see e.g. § 13.5, formula 33)

$$\int \frac{dx}{a + bx^2} = \frac{1}{\sqrt{(ab)}} \arctan \sqrt{\left(\frac{b}{a}\right)}\, x + C \quad (a > 0, b > 0)\,. \tag{6}$$

Obviously, the function $1/(a + bx^2)$ as well as the function $-1/(a + bx^2)^2$ are continuous functions of x, a, b for all x (since $a > 0$, $b > 0$), hence (4) and (5) are applicable. Differentiating (6) with respect to a, we have

$$-\int \frac{dx}{(a + bx^2)^2} = -\frac{1}{2\sqrt{(a^3 b)}} \arctan \sqrt{\left(\frac{b}{a}\right)}\, x + \frac{1}{\sqrt{(ab)}} \frac{1}{1 + \frac{b}{a} x^2} \left(-\frac{\sqrt{b}}{2\sqrt{a^3}}\, x\right) + k\,;$$

hence on rearranging

$$\int \frac{dx}{(a + bx^2)^2} = \frac{1}{2a\sqrt{(ab)}} \arctan \sqrt{\left(\frac{b}{a}\right)}\, x + \frac{1}{2a} \frac{x}{a + bx^2} + k \quad (a > 0, b > 0)\,.$$

REMARK 7. *Graphical integration* is based on the following idea: If $F(x)$ denotes the primitive of $f(x)$, then $F'(x) = f(x)$. Hence (see § 11.6)

$$F(a + h) - F(a) = h f(a) + h\,\tau(h)\,, \quad \text{where} \quad \tau(h) \to 0 \quad \text{for} \quad h \to 0\,.$$

Neglecting the second term, one gets

$$F(a + h) \approx F(a) + h f(a)\,.$$

Similarly

$$F(a + 2h) - F(a + h) \approx h f(a + h),$$

whence $F(a + 2h)$ is determined, etc.

13.3. Integration of Rational Functions

In this paragraph, we shall deal with integrals of the form

$$\int \frac{P(x)}{Q(x)}\,dx, \tag{1}$$

where $P(x)$ and $Q(x)$ are polynomials (we shall assume throughout that $P(x)$ and $Q(x)$ have *real coefficients*). Basically the method is to split the integrand into the sum of simple functions, which can be integrated directly.

If the degree m of the polynomial $P(x)$ is greater than (or equal to) the degree n of the polynomial $Q(x)$, then the fraction $P(x)/Q(x)$ can be reduced to a sum of a polynomial (of order $m - n$) and of a proper fraction

$$\frac{R(x)}{Q(x)}, \tag{2}$$

where the degree r of the polynomial $R(x)$ is less than the degree n of the polynomial $Q(x)$ (or $R(x) \equiv 0$). The mechanism of division is to be seen from the following example:

Example 1.

$$(x^3 + 4x^2 - x + 2)/(x^2 + x - 3) = x + 3 + \frac{-x + 11}{x^2 + x - 3}$$

$$\begin{array}{r} -x^3 \pm x^2 \mp 3x \\ \hline 3x^2 + 2x + 2 \\ -3x^2 \pm 3x \mp 9 \\ \hline -x + 11 \end{array}$$

Theorem 1. *Every polynomial*

$$Q(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$$

of the n-th order, with real coefficients can be reduced in a unique way, except for the ordering of the factors, into a product of the form (see Example 2)

$$Q(x) = a_n(x - \alpha_1)^{k_1}(x - \alpha_2)^{k_2}\ldots(x - \alpha_i)^{k_i}(x^2 + p_1 x + q_1)^{l_1}(x^2 + p_2 x + q_2)^{l_2}\ldots \ldots(x^2 + p_j x + q_j)^{l_j}, \tag{3}$$

where the numbers $\alpha_1, \alpha_2, \ldots, \alpha_i, p_1, q_1, p_2, q_2, \ldots$ *are real and the quadratic expressions in* (3) *are irreducible to real linear factors* (*i.e. they have negative discriminants,*

$$\frac{p^2}{4} - q < 0). \tag{4}$$

REMARK 1. The reduction (3) is obtained from the well-known method of factorisation in linear factors in the following way: If α is a real zero of $Q(x)$, then the reduction contains the factor $x - \alpha$. If β ($\beta = \beta_1 + i\beta_2$) is a complex zero of $Q(x)$, then the complex conjugate $\bar{\beta} = \beta_1 - i\beta_2$ is also a zero of $Q(x)$. The product of corresponding factors

$$[x - (\beta_1 + i\beta_2)]\,[x - (\beta_1 - i\beta_2)] = (x - \beta_1)^2 + \beta_2^2 =$$
$$= x^2 - 2\beta_1 x + \beta_1^2 + \beta_2^2$$

gives a quadratic expression with *real* coefficients. If the multiplicity of β is l, then $\bar{\beta}$ is of the same multiplicity and we obtain in (3) the factor

$$(x^2 + px + q)^l.$$

REMARK 2. On the practical determination of zeros of a polynomial of the n-th degree see Chaps. 1 and 31. It is often possible to find a zero of the given polynomial by inspection, especially if the polynomial has integral coefficients. Dividing by the corresponding linear factor $x - \alpha$ we reduce the order of the polynomial, as in the following example.

Example 2. The polynomial $Q(x) = x^4 - x^3 - x^2 - x - 2$ has obviously the zero $\alpha_1 = -1$. We have

$$\begin{array}{rl}
(x^4 - x^3 - x^2 - x - 2)/(x + 1) = x^3 - 2x^2 + x - 2\,. \\
\underline{-x^4 \pm x^3} & \\
\quad -2x^3 - x^2 & \\
\quad \underline{\mp 2x^3 \mp 2x^2} & \\
\qquad\quad x^2 - x & \\
\qquad\quad \underline{- x^2 \pm x} & \\
\qquad\qquad -2x - 2 &
\end{array}$$

The resulting polynomial has the zero $\alpha_2 = 2$. Dividing by the linear factor $(x - 2)$ we obtain the polynomial $x^2 + 1$, which is irreducible to real linear factors (see (4); $p = 0$, $q = 1$). Hence

$$Q(x) \equiv x^4 - x^3 - x^2 - x - 2 = (x + 1)(x - 2)(x^2 + 1)$$

and this factorisation is of the form (3).

Theorem 2. *Let*

$$\frac{P(x)}{Q(x)} \tag{5}$$

be a proper rational fraction with real coefficients and let $Q(x)$ *be of the form* (3), i.e.

$$Q(x) = a_n(x - \alpha_1)^{k_1}(x - \alpha_2)^{k_2} \dots (x^2 + p_1x + q_1)^{l_1}(x^2 + p_2x + q_2)^{l_2} \dots .$$

Then there exist real numbers $A_1, A_2, \dots, A_{k_1}, \dots; C_1, C_2, \dots, C_{l_1}; D_1, D_2, \dots, D_{l_1} \dots$ *(uniquely determined by the function* (5)*), such that for all* x *different from the zeros of* $Q(x)$ *the relation*

$$\begin{aligned}\frac{P(x)}{Q(x)} = {} & \frac{A_1}{x - \alpha_1} + \frac{A_2}{(x - \alpha_1)^2} + \dots + \frac{A_{k_1}}{(x - \alpha_1)^{k_1}} + \\ & + \frac{B_1}{x - \alpha_2} + \frac{B_2}{(x - \alpha_2)^2} + \dots + \frac{B_{k_2}}{(x - \alpha_2)^{k_2}} + \dots + \\ & + \frac{C_1x + D_1}{x^2 + p_1x + q_1} + \frac{C_2x + D_2}{(x^2 + p_1x + q_1)^2} + \dots + \frac{C_{l_1}x + D_{l_1}}{(x^2 + p_1x + q_1)^{l_1}} + \\ & + \frac{E_1x + F_1}{x^2 + p_2x + q_2} + \frac{E_2x + F_2}{(x^2 + p_2x + q_2)^2} + \dots + \frac{E_{l_2}x + F_{l_2}}{(x^2 + p_2x + q_2)^{l_2}} + \dots\end{aligned} \tag{6}$$

holds.

REMARK 3. This means that if α_1 is a real k_1-fold root of $Q(x) = 0$, then all fractions with denominators $x - \alpha_1, (x - \alpha_1)^2, \dots, (x - \alpha_1)^{k_1}$ occur in the reduction. If α_1 is a simple root, then only one fraction in the reduction (6) corresponds to it; similarly for all other real roots. The situation for expressions of the form $x^2 + px + q$ is similar; the linear binomials of the form $Cx + D$, however, should be written in the numerators instead of constants.

REMARK 4. The unknown constants $A_1, A_2, \dots$ can be determined by several methods, only two of which will be mentioned here: the *method of undetermined coefficients* and the *substitution method:*

Example 3. Let us reduce the function

$$\frac{x^2 - 2}{x^4 - 2x^3 + 2x^2} \tag{7}$$

according to Theorem 2.

The function (7) is already a proper fraction, hence it is not necessary to perform the preliminary division as in Example 1. The denominator can be written in the form

$$Q(x) \equiv x^4 - 2x^3 + 2x^2 = x^2(x^2 - 2x + 2),$$

where the quadratic expression $x^2 - 2x + 2$ has a negative discriminant (4). Hence, the reduction (6) is of the form ($\alpha = 0$)

$$\frac{x^2 - 2}{x^4 - 2x^3 + 2x^2} = \frac{A_1}{x} + \frac{A_2}{x^2} + \frac{Bx + C}{x^2 - 2x + 2}. \tag{8}$$

1. Determination of the constants A_1, A_2, B, C by the *method of undetermined coefficients.* We multiply equation (8) by $x^2(x^2 - 2x + 2)$, giving

$$x^2 - 2 = A_1 x(x^2 - 2x + 2) + A_2(x^2 - 2x + 2) + (Bx + C)\,x^2 \tag{9}$$

or

$$x^2 - 2 = (A_1 + B)\,x^3 + (-2A_1 + A_2 + C)\,x^2 + (2A_1 - 2A_2)\,x + 2A_2\,. \tag{10}$$

Since equation (8) should be valid for infinitely many values of x, the same is true for equations (9) and (10). Hence, it follows (see Theorem 1.14.1) that coefficients of the same powers of x are equal. Comparing coefficients, we obtain the equations $A_1 + B = 0$, $-2A_1 + A_2 + C = 1$, $2A_1 - 2A_2 = 0$, $2A_2 = -2$, from which

$$A_1 = -1\,, \quad A_2 = -1\,, \quad B = 1\,, \quad C = 0\,.$$

2. *Method of substitution.* Equation (9) is valid for an infinite number of values of x; hence it is valid for all x, in particular for zeros of the polynomial $Q(x)$. If $x = 0$ is substituted into (9), then equation (9) yields

$$-2 = 2A_2 \Rightarrow A_2 = -1\,.$$

To evaluate A_1 we differentiate (9) with respect to x and get

$$2x = A_1(x^2 - 2x + 2) + A_1 x(2x - 2) + A_2(2x - 2) + 2x(Bx + C) + Bx^2\,.$$

If we put $x = 0$ and $A_2 = -1$ (which has been already evaluated), we obtain $0 = 2A_1 + 2 \Rightarrow A_1 = -1$.

If $x = 0$ were a three-fold root of the equation $Q(x) = 0$, then we would determine the corresponding third constant by repeated differentiation of equation (9) (and then substituting $x = 0$). This procedure always leads to the required result.

Next, we substitute $x = 1 + \mathrm{i}$ (which is one of the zeros of the quadratic polynomial $x^2 - 2x + 2$), and obtain (note that $(1 + \mathrm{i})^2 = 2\mathrm{i}$)

$$2\mathrm{i} - 2 = [B(1 + \mathrm{i}) + C]\,2\mathrm{i}$$

or

$$-2 + 2\mathrm{i} = -2B + (2B + 2C)\,\mathrm{i}\,.$$

Comparing the real and the imaginary parts we get $B = 1$, $C = 0$.

(If the expression $x^2 - 2x + 2$ had appeared squared in the reduction, we would have evaluated further coefficients in this case also by differentiation of equation (9) and substitution of $x = 1 + \mathrm{i}$.)

Hence

$$\frac{x^2 - 2}{x^4 - 2x^3 + 2x^2} = \frac{-1}{x} + \frac{-1}{x^2} + \frac{x}{x^2 - 2x + 2} \tag{11}$$

for all x other than 0 and $1 \pm i$.

REMARK 5. The numbers A_1, A_2, B, C may alternatively be determined by substituting any four values for x and solving the resulting four equations for the four unknown constants A_1, A_2, B, C.

REMARK 6. We draw the reader's attention to the fact that a rational function can be reduced to a sum of fractions (6) only if the degree of the polynomial in the numerator is less than that in the denominator. Otherwise division as in Example 1 is first to be performed.

REMARK 7 (concerning practical evaluation of constants). If α is a simple zero of the polynomial $Q(x)$ then the corresponding constant A can be evaluated by the formula

$$A = \frac{P(\alpha)}{Q'(\alpha)}.$$

Example 4. Using (6), we may write

$$\frac{1}{x^3 - 3x^2 + 4} = \frac{1}{(x-2)^2(x+1)} = \frac{A_1}{x-2} + \frac{A_2}{(x-2)^2} + \frac{B}{x+1}.$$

Then

$$B = \frac{P(-1)}{Q'(-1)} = \frac{1}{(3x^2 - 6x)_{x=-1}} = \frac{1}{9},$$

since $\alpha = -1$ is a simple zero of the polynomial $Q(x)$.

REMARK 8 (*Integration after Reduction*). Every proper rational function with real coefficients can be reduced by Theorem 2 in a unique way into the sum (6) (of so-called *partial fractions*). Each term on the right-hand side of equation (6) can be integrated by elementary methods:

$$\int \frac{A_1}{x - \alpha_1}\,dx = A_1 \ln|x - \alpha_1| + C \quad \text{(by substitution } x - \alpha_1 = z\text{)},$$

$$\int \frac{A_k}{(x - \alpha_1)^k}\,dx = \frac{A_k}{-k+1}\frac{1}{(x - \alpha_1)^{k-1}} + C \quad (k \neq 1, \text{ substitution } x - \alpha_1 = z).$$

Terms of the type

$$\frac{Bx + C}{(x^2 + px + q)^k}$$

are integrated as follows (we assume the quadratic polynomial $x^2 + px + q$ has a negative discriminant $p^2/4 - q < 0$):

$$\frac{Bx + C}{(x^2 + px + q)^k} = \frac{B}{2}\frac{2x + p}{(x^2 + px + q)^k} + \left(C - \frac{Bp}{2}\right)\frac{1}{[(x + \frac{1}{2}p)^2 + q - (\frac{1}{2}p)^2]^k}.$$

Putting $x^2 + px + q = z$, $(2x + p)\,dx = dz$ we have:

$$\int \frac{2x + p}{(x^2 + px + q)^k}\,dx = \frac{1}{-k + 1}\frac{1}{(x^2 + px + q)^{k-1}} + C, \quad \text{if} \quad k \neq 1;$$

$$\int \frac{2x + p}{x^2 + px + q}\,dx = \ln(x^2 + px + q) + C, \quad \text{if} \quad k = 1.$$

Next on substituting $x + \frac{1}{2}p = z\sqrt{[q - (\frac{1}{2}p)^2]}$, $dx = \sqrt{[q - (\frac{1}{2}p)^2]}\,dz$,

$$\int \frac{dx}{[(x + \frac{1}{2}p)^2 + q - (\frac{1}{2}p)^2]^k} = \frac{1}{[q - (\frac{1}{2}p)^2]^{k-\frac{1}{2}}}\int \frac{dz}{(z^2 + 1)^k}.$$

For $k = 1$, we have

$$\int \frac{dz}{z^2+1} = \arctan z + C = \arctan \frac{x + \frac{1}{2}p}{\sqrt{[q - (\frac{1}{2}p)^2]}} + C.$$

For $k > 1$ the reduction formula

$$I_{k+1} = \frac{1}{2k}\frac{z}{(1 + z^2)^k} + \frac{2k - 1}{2k} I_k, \tag{12}$$

where

$$I_k = \int \frac{dz}{(1 + z^2)^k}, \quad I_{k+1} = \int \frac{dz}{(1 + z^2)^{k+1}},$$

is valid.

Example 5. Using (12) we have

$$\int \frac{dz}{(1 + z^2)^2} = \frac{1}{2}\frac{z}{1 + z^2} + \frac{2 - 1}{2}\int \frac{dz}{1 + z^2} = \frac{1}{2}\frac{z}{1 + z^2} + \frac{1}{2}\arctan z + C.$$

Example 6. Using (11) and Remark 8 we get

$$\int \frac{x^2 - 2}{x^4 - 2x^3 + 2x^2}\,dx = \int\left(\frac{-1}{x} + \frac{-1}{x^2} + \frac{x}{x^2 - 2x + 2}\right)dx =$$

$$= -\ln|x| + \frac{1}{x} + \tfrac{1}{2}\ln(x^2 - 2x + 2) + \arctan(x - 1) + C.$$

13.4. Integrals which can be rationalized

Some types of integral can be reduced to integrals of rational functions, which are then integrated according to § 13.3. In particular, the following types are considered.

I. $$\int R\left(x, \sqrt[n]{\frac{ax+b}{cx+d}}\right) dx,$$

where $R(x, t)$ is a rational function of the variables x, t.

Integrals like this can be transformed into integrals of rational functions by the substitution

$$\frac{ax+b}{cx+d} = z^n \quad \text{or} \quad x = \frac{dz^n - b}{a - cz^n}.$$

Example 1. The function

$$\int \frac{3x + \sqrt{(2x-1)}}{x - \sqrt{(2x-1)^3}} dx \quad (2x - 1 \geqq 0, \quad x - \sqrt{(2x-1)^3} \neq 0) \tag{1}$$

is of the type mentioned, where

$$R(x, t) = \frac{3x + t}{x - t^3}, \quad t = \sqrt{(2x-1)}.$$

We make the substitution

$$2x - 1 = z^2 \quad (z > 0),$$

from which it follows that $dx = z\,dz$ and

$$\int \frac{3x + \sqrt{(2x-1)}}{x - \sqrt{(2x-1)^3}} dx = \int \frac{3\frac{z^2+1}{2} + z}{\frac{z^2+1}{2} - z^3} z\,dz = -\int \frac{3z^3 + 2z^2 + 3z}{2z^3 - z^2 - 1} dz,$$

and this is now an integral of a rational function. After integration we substitute again $z = \sqrt{(2x-1)}$.

REMARK 1. The fraction $(ax + b)/(cx + d)$ may occur with different (rational) exponents. For instance, the integrand may be a rational function $R(x, t, u)$, where

$$t = \left(\frac{ax+b}{cx+d}\right)^{1/r}, \quad u = \left(\frac{ax+b}{cx+d}\right)^{1/s},$$

is to be substituted. Such an integral can be rationalized by the substitution $(ax + b) : (cx + d) = t^p$, where p stands for the least common multiple of the numbers r and s. The procedure is similar when several roots are involved.

Example 2. The integral

$$\int \frac{dx}{\sqrt{(x+1)} - \sqrt[3]{(x+1)}}$$

is rationalized by the substitution $x + 1 = t^6$ $(t > 0)$; we get

$$\int \frac{dx}{\sqrt{(x+1)} - \sqrt[3]{(x+1)}} = \int \frac{6t^5\,dt}{t^3 - t^2} = 6\int \frac{t^3\,dt}{t-1} =$$

$$= 6\int \frac{(t^3 - t^2) + (t^2 - t) + (t-1) + 1}{t-1}\,dt = 6\left(\frac{t^3}{3} + \frac{t^2}{2} + t + \ln|t-1|\right) + C =$$

$$= 2\sqrt{(x+1)} + 3\sqrt[3]{(x+1)} + 6\sqrt[6]{(x+1)} + 6\ln\left|\sqrt[6]{(x+1)} - 1\right| + C\,.$$

II. *Binomial integrals* are those of the form

$$\int x^m(a + bx^n)^p\,dx \quad (a \neq 0,\ b \neq 0,\ n \neq 0,\ p \neq 0)\,, \tag{2}$$

where m, n, p are rational numbers.

Theorem 1. *Integrals* (2) *can be expressed in terms of elementary functions* (*algebraic functions and elementary transcendental functions,* see § 11.2) *if and only if one of the numbers*

$$p, \quad \frac{m+1}{n}, \quad \frac{m+1}{n} + p$$

is an integer.

REMARK 2. (α) If p is a positive integer we evaluate the expression $(a + bx^n)^p$ by the Binomial Theorem and we obtain an integral of a sum of powers of x.

If p is a negative integer and

$$m = \frac{r}{s}, \quad n = \frac{u}{v}$$

where r, s $(s > 0)$ are integers without a common factor, and similarly for u and v, then the substitution

$$x = z^t\,,$$

where t is the least common multiple of the numbers s, v, will rationalize the integral (2).

(β) If

$$\frac{m+1}{n}$$

is an integer, then the substitution

$$a + bx^n = z$$

will reduce the integral (2) to the previous case.

(γ) If

$$\frac{m+1}{n} + p$$

is an integer then we reduce the integration of (2) to the case (α) by the substitution

$$ax^{-n} + b = z\,. \tag{3}$$

Example 3.

$$\int \frac{(1 + \sqrt[6]{x})^3}{\sqrt[3]{x^2}}\,dx = \int x^{-2/3}(1 + x^{1/6})^3\,dx\,.$$

Case (α), p is a positive integer. Making use of the Binomial Theorem, we have

$$\int x^{-2/3}(1 + x^{1/6})^3\,dx = \int x^{-2/3}(1 + 3x^{1/6} + 3x^{2/6} + x^{3/6})\,dx =$$

$$= \int (x^{-2/3} + 3x^{-1/2} + 3x^{-1/3} + x^{-1/6})\,dx = 3x^{1/3} + 6x^{1/2} + \tfrac{9}{2}x^{2/3} + \tfrac{6}{5}x^{5/6} + C\,.$$

Example 4.

$$\int \frac{dx}{\sqrt{(a + bx^2)^3}} = \int (a + bx^2)^{-3/2}\,dx\,.$$

Here

$$\frac{m+1}{n} + p = \tfrac{1}{2} - \tfrac{3}{2} = -1$$

(case γ). Putting

$$\frac{a}{x^2} + b = z\,, \quad -\frac{2a}{x^3}\,dx = dz$$

we obtain (for $x > 0$)

$$\int \frac{dx}{\sqrt{(a + bx^2)^3}} = \int \frac{1}{x^3} \frac{dx}{\sqrt{\left(\frac{a}{x^2} + b\right)^3}} = -\frac{1}{2a}\int \frac{dz}{z^{3/2}} =$$

$$= \frac{1}{a} z^{-1/2} + C = \frac{1}{a} \frac{1}{\sqrt{\left(\frac{a}{x^2} + b\right)}} + C\,.$$

Example 5. The integral

$$\int x^{-3/4}(1 - x^{1/6})^{-2}\,dx$$

(case α, p is a negative integer) is rationalized by the substitution $x = z^{12}$.

III. Integrals of the form

$$\int R(x, \sqrt{(ax^2 + bx + c)})\,dx \quad (ax^2 + bx + c > 0)\,, \tag{4}$$

where $R(x, t)$ is a rational function of the variables x, t, are rationalized as follows:

(α) If $a > 0$ we make use of the substitution

$$\sqrt{(ax^2 + bx + c)} + \sqrt{(a)}\,x = z\,, \tag{5}$$

from which it follows that

$$ax^2 + bx + c = ax^2 - 2\sqrt{(a)}\,xz + z^2 \quad \text{or} \quad x = \frac{z^2 - c}{b + 2\sqrt{(a)}\,z}\,,$$

$$dx = 2\frac{c\sqrt{(a)} + bz + \sqrt{(a)}\,z^2}{[b + 2\sqrt{(a)}\,z]^2}\,dz\,, \quad \sqrt{(ax^2 + bx + c)} = \frac{c\sqrt{(a)} + bz + \sqrt{(a)}\,z^2}{b + 2\sqrt{(a)}\,z}\,,$$

$$\frac{dx}{\sqrt{(ax^2 + bx + c)}} = \frac{2dz}{b + 2\sqrt{(a)}\,z}\,. \tag{6}$$

(β) If $a < 0$, then there are two distinct zeros x_1 and x_2 of the polynomial $ax^2 + bx + c$ (since we are given that $ax^2 + bx + c > 0$ for some x and the latter polynomial tends to $-\infty$ as $x \to \pm\infty$), hence $ax^2 + bx + c = a(x - x_1)(x - x_2)$. If $x_1 < x_2$, then

$$\sqrt{(ax^2 + bx + c)} = \sqrt{(-a)}\sqrt{[(x - x_1)(x_2 - x)]} = \sqrt{(-a)}\,(x - x_1)\sqrt{\frac{x_2 - x}{x - x_1}}\,.$$

We now proceed as in case I, i.e. we employ the substitution

$$\frac{x_2 - x}{x - x_1} = z^2\,.$$

The given integral can often be transformed to the integral

$$\int \frac{dt}{\sqrt{(1 - t^2)}} \quad \text{or} \quad \int \sqrt{(1 - t^2)}\,dt\,.$$

The first integral is a standard integral and the second one is easily evaluated by the substitution $t = \sin z$ (Example 13.2.10, p. 454).

On some other methods for more complicated cases see e.g. [54].

Example 6. To evaluate

$$\int \frac{dx}{\sqrt{(x^2 + k)}} \quad (k \neq 0,\ x^2 + k > 0).$$

Making use of substitution (5) and equation (6), where $a = 1$, $b = 0$, $c = k$, we get

$$\int \frac{dx}{\sqrt{(x^2 + k)}} = \int \frac{dz}{z} = \ln|z| + C = \ln\left|\sqrt{(x^2 + k)} + x\right| + C.$$

See also § 13.1, formula 11 (p. 450).

IV. Integrals of the form

$$\int R(e^{ax})\,dx,$$

where $R(t)$ is a rational function of the variable t and a is a real constant, are rationalized by the substitution

$$e^{ax} = z.$$

Integrals

$$\int R(\ln x)\frac{dx}{x}, \quad x > 0,$$

where $R(t)$ is again a rational function, are rationalized by putting

$$\ln x = z.$$

V. Integrals of the type

$$\int R(\cos x, \sin x)\,dx,$$

where $R(u, t)$ is a rational function of the variables u and t, can always be rationalized by the substitution

$$\tan \tfrac{1}{2}x = z.$$

From this equation it follows (making use of the relation $\cos^2 \frac{1}{2}x = 1/(1 + \tan^2 \frac{1}{2}x)$) that

$$\cos x = \frac{1 - z^2}{1 + z^2}, \quad \sin x = \frac{2z}{1 + z^2}, \quad dx = \frac{2\,dz}{1 + z^2}.$$

Example 7. The integral

$$\int \frac{1 - \sin x + \cos x}{5 \sin x \cos x}\, dx$$

is transformed by the above substitution into the integral

$$\int \frac{1 - \dfrac{2z}{1+z^2} + \dfrac{1-z^2}{1+z^2}}{5\,\dfrac{2z}{1+z^2}\cdot\dfrac{1-z^2}{1+z^2}} \cdot \frac{2dz}{1+z^2} = \frac{2}{5}\int \frac{1-z}{z(1-z^2)}\, dz = \frac{2}{5}\int \frac{dz}{z(1+z)}.$$

Hence

$$\int \frac{1 - \sin x + \cos x}{5 \sin x \cos x}\, dx = \frac{2}{5}\int \frac{dz}{z(1+z)} = \frac{2}{5}\int\left(\frac{1}{z} - \frac{1}{1+z}\right) dz =$$

$$= \frac{2}{5} \ln\left|\frac{z}{1+z}\right| + C = \frac{2}{5} \ln\left|\frac{\tan \frac{1}{2}x}{1 + \tan \frac{1}{2}x}\right| + C.$$

REMARK 3. One can often choose substitutions which are simpler than the substitution $\tan \frac{1}{2}x = z$. For example, the integral

$$\int \frac{1 + \cos^2 x}{\cos^4 x}\, dx$$

is transformed by the substitution

$$\tan x = z, \quad \text{from which we have} \quad dx/(\cos^2 x) = dz$$

into the integral (making use of the relation $\cos^2 x = 1/(1 + \tan^2 x)$)

$$\int (1 + z^2 + 1)\, dz = \int (2 + z^2)\, dz.$$

Hence

$$\int \frac{1 + \cos^2 x}{\cos^4 x}\, dx = \int (2 + z^2)\, dz = 2z + \frac{z^3}{3} + C = 2 \tan x + \frac{\tan^3 x}{3} + C.$$

In general if the function $R(\cos x, \sin x)$ is odd with respect to the function $\sin x$, i.e. if $R(\cos x, \sin x) = -R(\cos x, -\sin x)$, then it is possible to use the substitution $\cos x = z$ to rationalize the integral $\int R(\cos x, \sin x)\, dx$.

If $R(\cos x, \sin x)$ is odd with respect to $\cos x$, i.e. if $R(\cos x, \sin x) = -R(-\cos x, \sin x)$, the substitution $\sin x = z$ can be used.

If the function $R(\cos x, \sin x)$ is even with respect to both functions, i.e. if $R(\cos x, \sin x) = R(-\cos x, -\sin x)$, the substitution $\tan x = z$ may be employed (see the previous example).

Integrals of the type

$$\int \sin^n x \cos^m x \, dx \tag{7}$$

(where m, n are integers) are evaluated, if n or m is an odd number, by the substitution

$$\cos x = z \quad \text{or} \quad \sin x = z .$$

Example 8 (substitution $\sin x = z$).

$$\int \frac{dx}{\cos x} = \int \frac{\cos x \, dx}{\cos^2 x} = \int \frac{\cos x \, dx}{1 - \sin^2 x} = \int \frac{dz}{1 - z^2} = \frac{1}{2}\int\left(\frac{1}{1 - z} + \frac{1}{1 + z}\right) dz =$$

$$= \tfrac{1}{2} \ln \left|\frac{1 + z}{1 - z}\right| + C = \tfrac{1}{2} \ln \left|\frac{1 + \sin x}{1 - \sin x}\right| + C .$$

(By first transforming the integrand as indicated below this integral can be solved by the substitution $\tan(\frac{1}{4}\pi + \frac{1}{2}x) = z$. Thus:

$$\frac{1}{\cos x} = \frac{1}{\sin(\frac{1}{2}\pi + x)} = \frac{1}{2 \sin(\frac{1}{4}\pi + \frac{1}{2}x) \cos(\frac{1}{4}\pi + \frac{1}{2}x)} =$$

$$= \frac{1}{2 \tan(\frac{1}{4}\pi + \frac{1}{2}x) \cos^2(\frac{1}{4}\pi + \frac{1}{2}x)} ,$$

hence

$$\int \frac{dx}{\cos x} = \ln \left|\tan(\tfrac{1}{4}\pi + \tfrac{1}{2}x)\right| + C .$$

We can show that these two results are equivalent as follows:

$$\tfrac{1}{2} \ln \left|\frac{1 + \sin x}{1 - \sin x}\right| = \tfrac{1}{2} \ln \left|\frac{1 + 2 \sin \frac{1}{2}x \cos \frac{1}{2}x}{1 - 2 \sin \frac{1}{2}x \cos \frac{1}{2}x}\right| =$$

$$= \tfrac{1}{2} \ln \left|\frac{\dfrac{1}{\cos^2 \frac{1}{2}x} + 2 \tan \frac{1}{2}x}{\dfrac{1}{\cos^2 \frac{1}{2}x} - 2 \tan \frac{1}{2}x}\right| = \tfrac{1}{2} \ln \left|\frac{1 + \tan^2 \frac{1}{2}x + 2 \tan \frac{1}{2}x}{1 + \tan^2 \frac{1}{2}x - 2 \tan \frac{1}{2}x}\right| =$$

$$= \tfrac{1}{2} \ln \left(\frac{1 + \tan \frac{1}{2}x}{1 - \tan \frac{1}{2}x}\right)^2 = \ln \left|\frac{1 + \tan \frac{1}{2}x}{1 - \tan \frac{1}{2}x}\right| = \ln \left|\tan(\tfrac{1}{4}\pi + \tfrac{1}{2}x)\right| .)$$

Example 9 (substitution $\sin x = z$):

$$\int \sin^3 x \cos^5 x \, dx = \int z^3(1 - z^2)^2 \, dz = \frac{z^4}{4} - \frac{z^6}{3} + \frac{z^8}{8} + C =$$

$$= \frac{\sin^4 x}{4} - \frac{\sin^6 x}{3} + \frac{\sin^8 x}{8} + C .$$

If both m and n in (7) are even and non-negative, then making use of the formulae $\cos^2 x = \frac{1}{2}(1 + \cos 2x)$, $\sin^2 x = \frac{1}{2}(1 - \cos 2x)$ and if necessary by their further application ($\cos^2 2x = \frac{1}{2}(1 + \cos 4x)$, etc.) we can reduce the degree and hence reduce the integration to the previous case.

Example 10.

$$\int \sin^2 x \, dx = \tfrac{1}{2} \int (1 - \cos 2x) \, dx = \tfrac{1}{2}(x - \tfrac{1}{2}\sin 2x) + C =$$

$$= \tfrac{1}{2}(x - \sin x \cos x) + C .$$

Example 11.

$$\int \cos^4 x \, dx = \tfrac{1}{4} \int (1 + \cos 2x)^2 \, dx = \tfrac{1}{4}x + \tfrac{1}{4} \sin 2x + \tfrac{1}{8} \int (1 + \cos 4x) \, dx =$$

$$= \tfrac{3}{8}x + \tfrac{1}{4} \sin 2x + \tfrac{1}{32} \sin 4x + C .$$

13.5. Table of Indefinite Integrals

See, in particular, [26]. (For standard integrals see § 13.1, p. 449.) Constants of integration are omitted in the Table; m, n are integers, r stands for an arbitrary real number.

The range of validity is indicated only in non-trivial cases. (For example, if the expression $xX = x(ax + b)$ appears in the denominator, we do not draw the reader's attention to the fact that the range of validity is determined by the conditions $x \neq 0$, $ax + b \neq 0$; if the root of the expression $X = ax + b$ is considered, then, of course, the range of validity follows from the condition $ax + b \geqq 0$, i.e. $x \geqq -b/a$ for $a > 0$ and $x \leqq -b/a$ for $a < 0$, etc.)

If $a^2 + x^2$ appears in the integral, then $a > 0$ is assumed, since this case is most often met in applications. The case $a < 0$ is easily reduced to the previous one by putting $a = -b$, $b > 0$.

(a) ***Rational Functions*** (see also Remark 1, p. 510).

Notation: $X = ax + b$ $(a \neq 0, b \neq 0)$.

1. $\int X^n \, dx = \frac{1}{a(n+1)} X^{n+1}$.

2. $\int \frac{dx}{X^n} = \frac{1}{a(1-n)} \cdot \frac{1}{X^{n-1}} \quad (n \neq 1)$.

3. $\int \frac{dx}{X} = \frac{1}{a} \ln |X|$.

4. $\int xX^n \, dx = \frac{1}{a^2(n+2)} X^{n+2} - \frac{b}{a^2(n+1)} X^{n+1}$.

5. $\int x^m X^n \, dx$ (see Remark 13.4.2, case (α), p. 464).

6. $\int \frac{x \, dx}{X} = \frac{x}{a} - \frac{b}{a^2} \ln |X|$.

7. $\int \frac{x \, dx}{X^2} = \frac{b}{a^2 X} + \frac{1}{a^2} \ln |X|$.

8. $\int \frac{x \, dx}{X^3} = \frac{1}{a^2} \left(-\frac{1}{X} + \frac{b}{2X^2} \right)$.

9. $\int \frac{x \, dx}{X^n} = \frac{1}{a^2} \left(\frac{-1}{(n-2) X^{n-2}} + \frac{b}{(n-1) X^{n-1}} \right) \quad (n \neq 1, n \neq 2)$.

10. $\int \frac{x^2 \, dx}{X} = \frac{1}{a^3} (\tfrac{1}{2} X^2 - 2bX + b^2 \ln |X|)$.

11. $\int \frac{x^2 \, dx}{X^2} = \frac{1}{a^3} \left(X - 2b \ln |X| - \frac{b^2}{X} \right)$.

12. $\int \frac{x^2 \, dx}{X^3} = \frac{1}{a^3} \left(\ln |X| + \frac{2b}{X} - \frac{b^2}{2X^2} \right)$.

13. $\int \frac{x^2 \, dx}{X^n} = \frac{1}{a^3} \left[\frac{-1}{(n-3) X^{n-3}} + \frac{2b}{(n-2) X^{n-2}} - \frac{b^2}{(n-1) X^{n-1}} \right]$
$(n \neq 1, n \neq 2, n \neq 3)$.

14. $\int \frac{dx}{xX} = -\frac{1}{b} \ln \left| \frac{X}{x} \right|$.

15. $\int \frac{dx}{xX^2} = -\frac{1}{b^2} \left(\ln \left| \frac{X}{x} \right| + \frac{ax}{X} \right)$.

16. $\int \frac{dx}{xX^3} = -\frac{1}{b^3}\left(\ln\left|\frac{X}{x}\right| + \frac{2ax}{X} - \frac{a^2x^2}{2X^2}\right).$

17. $\int \frac{dx}{xX^n} = -\frac{1}{b^n}\left[\ln\left|\frac{X}{x}\right| - \sum_{k=1}^{n-1}\binom{n-1}{k}\frac{(-a)^k x^k}{kX^k}\right] \quad (n > 1).$

18. $\int \frac{dx}{x^2X} = -\frac{1}{bx} + \frac{a}{b^2}\ln\left|\frac{X}{x}\right|.$

19. $\int \frac{dx}{x^2X^2} = -a\left[\frac{1}{b^2X} + \frac{1}{ab^2x} - \frac{2}{b^3}\ln\left|\frac{X}{x}\right|\right].$

20. $\int \frac{dx}{x^2X^3} = -a\left[\frac{1}{2b^2X^2} + \frac{2}{b^3X} + \frac{1}{ab^3x} - \frac{3}{b^4}\ln\left|\frac{X}{x}\right|\right].$

21. $\int \frac{dx}{x^2X^n} = \frac{1}{b^{n+1}}\left[an\ln\left|\frac{X}{x}\right| - \frac{X}{x} + \sum_{k=2}^{n}\binom{n}{k}\frac{(-a)^k x^{k-1}}{(k-1)X^{k-1}}\right] \quad (n \geqq 2).$

22. $\int \frac{dx}{x^mX^n} = -\frac{1}{b^{m+n-1}}\sum_{k=0}^{m+n-2}\binom{m+n-2}{k}\frac{(-a)^k X^{m-k-1}}{(m-k-1)x^{m-k-1}};$

the term for which $m - k - 1 = 0$ is to be replaced in the sum by the term

$$\binom{m+n-2}{n-1}(-a)^{m-1}\ln\left|\frac{X}{x}\right|.$$

Notation: $\Delta = bf - ag$.

23. $\int \frac{ax+b}{fx+g}dx = \frac{ax}{f} + \frac{\Delta}{f^2}\ln|fx+g| \quad (f \neq 0).$

24. $\int \frac{dx}{(ax+b)(fx+g)} = \frac{1}{\Delta}\ln\left|\frac{fx+g}{ax+b}\right| \quad (\Delta \neq 0).$

25. $\int \frac{x\,dx}{(ax+b)(fx+g)} = \frac{1}{\Delta}\left(\frac{b}{a}\ln|ax+b| - \frac{g}{f}\ln|fx+g|\right)$

$(a \neq 0, f \neq 0, \Delta \neq 0).$

26. $\int \frac{dx}{(ax+b)^2(fx+g)} = \frac{1}{\Delta}\left(\frac{1}{ax+b} + \frac{f}{\Delta}\ln\left|\frac{fx+g}{ax+b}\right|\right) \quad (\Delta \neq 0).$

27. $\int \frac{x\,dx}{(a+x)(b+x)^2} = \frac{b}{(a-b)(b+x)} - \frac{a}{(a-b)^2}\ln\left|\frac{a+x}{b+x}\right| \quad (a \neq b).$

28. $$\int \frac{x^2\,dx}{(a+x)(b+x)^2} =$$

$$= \frac{b^2}{(b-a)(b+x)} + \frac{a^2}{(b-a)^2}\ln|a+x| + \frac{b^2-2ab}{(b-a)^2}\ln|b+x| \quad (a \neq b).$$

29. $$\int \frac{dx}{(a+x)^2(b+x)^2} = \frac{-1}{(a-b)^2}\left(\frac{1}{a+x}+\frac{1}{b+x}\right) + \frac{2}{(a-b)^3}\ln\left|\frac{a+x}{b+x}\right| \quad (a \neq b).$$

30. $$\int \frac{dx}{(a+x)^m(b+x)^n} = \int \frac{dz}{z^m(z+c)^n} \quad \text{(see 22; } z = x+a,\ c = b-a,\ a \neq b).$$

31. $$\int \frac{x\,dx}{(a+x)^2(b+x)^2} =$$

$$= \frac{1}{(a-b)^2}\left(\frac{a}{a+x}+\frac{b}{b+x}\right) + \frac{a+b}{(a-b)^3}\ln\left|\frac{a+x}{b+x}\right| \quad (a \neq b).$$

32. $$\int \frac{x^2\,dx}{(a+x)^2(b+x)^2} =$$

$$= \frac{-1}{(a-b)^2}\left(\frac{a^2}{a+x}+\frac{b^2}{b+x}\right) + \frac{2ab}{(a-b)^3}\ln\left|\frac{a+x}{b+x}\right| \quad (a \neq b).$$

Notation: $X = ax^2 + bx + c$, $\Delta = 4ac - b^2$ $(a \neq 0, \Delta \neq 0)$.

33. $$\int \frac{dx}{X} = \frac{2}{\sqrt{\Delta}}\arctan\frac{2ax+b}{\sqrt{\Delta}} \quad (\Delta > 0),$$

$$= \frac{1}{\sqrt{(-\Delta)}}\ln\left|\frac{2ax+b-\sqrt{(-\Delta)}}{2ax+b+\sqrt{(-\Delta)}}\right| \quad (\Delta < 0).$$

34. $$\int \frac{dx}{X^2} = \frac{2ax+b}{\Delta X} + \frac{2a}{\Delta}\int \frac{dx}{X} \quad \text{(see 33)}.$$

35. $$\int \frac{dx}{X^3} = \frac{2ax+b}{\Delta}\left(\frac{1}{2X^2}+\frac{3a}{\Delta X}\right) + \frac{6a^2}{\Delta^2}\frac{dx}{X} \quad \text{(see 33)}.$$

36. $$\int \frac{dx}{X^n} = \frac{2ax+b}{(n-1)\Delta X^{n-1}} + \frac{(2n-3)2a}{(n-1)\Delta}\int \frac{dx}{X^{n-1}} \quad (n > 1).$$

37. $\int \frac{x\,dx}{X} = \frac{1}{2a}\ln|X| - \frac{b}{2a}\int \frac{dx}{X}$ (see 33).

38. $\int \frac{x\,dx}{X^2} = -\frac{bx + 2c}{\Delta X} - \frac{b}{\Delta}\int \frac{dx}{X}$ (see 33).

39. $\int \frac{x\,dx}{X^n} = -\frac{bx + 2c}{(n-1)\Delta X^{n-1}} - \frac{b(2n-3)}{(n-1)\Delta}\int \frac{dx}{X^{n-1}}$ $(n > 1)$.

40. $\int \frac{x^2\,dx}{X} = \frac{x}{a} - \frac{b}{2a^2}\ln|X| + \frac{b^2 - 2ac}{2a^2}\int \frac{dx}{X}$ (see 33).

41. $\int \frac{x^2\,dx}{X^2} = \frac{(b^2 - 2ac)\,x + bc}{a\,\Delta X} + \frac{2c}{\Delta}\int \frac{dx}{X}$ (see 33).

42. $\int \frac{x^2\,dx}{X^n} = \frac{-x}{(2n-3)\,aX^{n-1}} + \frac{c}{(2n-3)\,a}\int \frac{dx}{X^n} - \frac{(n-2)\,b}{(2n-3)\,a}\int \frac{x\,dx}{X^n}$

(see 36 and 39).

43. $\int \frac{x^m\,dx}{X^n} = -\frac{x^{m-1}}{(2n-m-1)\,aX^{n-1}} + \frac{(m-1)\,c}{(2n-m-1)\,a}\int \frac{x^{m-2}\,dx}{X^n} - \frac{(n-m)\,b}{(2n-m-1)\,a}\int \frac{x^{m-1}\,dx}{X^n}$ $(m \neq 2n-1$; for $m = 2n-1$ see 44).

44. $\int \frac{x^{2n-1}\,dx}{X^n} = \frac{1}{a}\int \frac{x^{2n-3}\,dx}{X^{n-1}} - \frac{c}{a}\int \frac{x^{2n-3}\,dx}{X^n} - \frac{b}{a}\int \frac{x^{2n-2}\,dx}{X^n}$

$(n > 1$; for $n = 1$ see 37).

45. $\int \frac{dx}{xX} = \frac{1}{2c}\ln\frac{x^2}{|X|} - \frac{b}{2c}\int \frac{dx}{X}$ (see 33; $c \neq 0$).

46. $\int \frac{dx}{xX^n} = \frac{1}{2c(n-1)\,X^{n-1}} - \frac{b}{2c}\int \frac{dx}{X^n} + \frac{1}{c}\int \frac{dx}{xX^{n-1}}$ $(c \neq 0, n > 1)$.

47. $\int \frac{dx}{x^2X} = \frac{b}{2c^2}\ln\frac{|X|}{x^2} - \frac{1}{cx} + \left(\frac{b^2}{2c^2} - \frac{a}{c}\right)\int \frac{dx}{X}$ (see 33; $c \neq 0$).

48. $\int \frac{dx}{x^mX^n} = -\frac{1}{(m-1)\,cx^{m-1}X^{n-1}} - \frac{(2n+m-3)\,a}{(m-1)\,c}\int \frac{dx}{x^{m-2}X^n} - \frac{(n+m-2)\,b}{(m-1)\,c}\int \frac{dx}{x^{m-1}X^n}$ $(c \neq 0, m > 1)$.

49. $$\int \frac{dx}{(fx+g)X} = \frac{f}{2(cf^2 - gbf + g^2a)} \ln \frac{(fx+g)^2}{|X|} + \\ + \frac{2ga - bf}{2(cf^2 - gbf + g^2a)} \int \frac{dx}{X} \quad (\text{see } 33;\ cf^2 - gbf + g^2a \neq 0).$$

Notation: $X = a^2 \pm x^2 \quad (a > 0)$,

$Y = \arctan \frac{x}{a}$ for the positive sign,

$Y = \frac{1}{2} \ln \left| \frac{x+a}{x-a} \right| = \begin{cases} \operatorname{artanh}(x/a) \text{ for the negative sign and for } |x| < a, \\ \operatorname{arcoth}(x/a) \text{ for the negative sign and for } |x| > a. \end{cases}$

If both signs occur in a formula, then the upper or lower sign corresponds to the case $X = a^2 + x^2$, or $X = a^2 - x^2$, respectively.

50. $$\int \frac{dx}{X} = \frac{1}{a} Y.$$

51. $$\int \frac{dx}{X^2} = \frac{x}{2a^2X} + \frac{1}{2a^3} Y.$$

52. $$\int \frac{dx}{X^3} = \frac{x}{4a^2X^2} + \frac{3x}{8a^4X} + \frac{3}{8a^5} Y.$$

53. $$\int \frac{dx}{X^{n+1}} = \frac{x}{2na^2X^n} + \frac{2n-1}{2na^2} \int \frac{dx}{X^n} \quad (n \neq 0).$$

54. $$\int \frac{x\,dx}{X} = \pm \tfrac{1}{2} \ln |X|.$$

55. $$\int \frac{x\,dx}{X^2} = \mp \frac{1}{2X}.$$

56. $$\int \frac{x\,dx}{X^3} = \mp \frac{1}{4X^2}.$$

57. $$\int \frac{x\,dx}{X^{n+1}} = \mp \frac{1}{2nX^n} \quad (n \neq 0).$$

58. $$\int \frac{x^2\,dx}{X} = \pm x \mp aY.$$

59. $\int \frac{x^2\,dx}{X^2} = \mp \frac{x}{2X} \pm \frac{1}{2a} Y.$

60. $\int \frac{x^2\,dx}{X^3} = \mp \frac{x}{4X^2} \pm \frac{x}{8a^2X} \pm \frac{1}{8a^3} Y.$

61. $\int \frac{x^2\,dx}{X^{n+1}} = \mp \frac{x}{2nX^n} \pm \frac{1}{2n} \int \frac{dx}{X^n} \quad (n \neq 0).$

62. $\int \frac{dx}{xX} = \frac{1}{2a^2} \ln \frac{x^2}{|X|}.$

63. $\int \frac{dx}{xX^2} = \frac{1}{2a^2X} + \frac{1}{2a^4} \ln \frac{x^2}{|X|}.$

64. $\int \frac{dx}{xX^3} = \frac{1}{4a^2X^2} + \frac{1}{2a^4X} + \frac{1}{2a^6} \ln \frac{x^2}{|X|}.$

65. $\int \frac{dx}{x^2X} = -\frac{1}{a^2x} \mp \frac{1}{a^3} Y.$

66. $\int \frac{dx}{x^2X^2} = -\frac{1}{a^4x} \mp \frac{x}{2a^4X} \mp \frac{3}{2a^5} Y.$

67. $\int \frac{dx}{x^2X^3} = -\frac{1}{a^6x} \mp \frac{x}{4a^4X^2} \mp \frac{7x}{8a^6X} \mp \frac{15}{8a^7} Y.$

68. $\int \frac{dx}{(b + cx)\,X} = \frac{1}{a^2c^2 \pm b^2} \left[c \ln |b + cx| - \frac{c}{2} \ln |X| \pm \frac{b}{a} Y \right] \quad (a^2c^2 \pm b^2 \neq 0).$

Notation: $X = a^3 \pm x^3$ $(a \neq 0)$.
If both signs occur in a formula, then the upper or lower sign corresponds to the case $X = a^3 + x^3$, or $X = a^3 - x^3$, respectively.

69. $\int \frac{dx}{X} = \pm \frac{1}{6a^2} \ln \frac{(a \pm x)^2}{a^2 \mp ax + x^2} + \frac{1}{a^2\sqrt{3}} \arctan \frac{2x \mp a}{a\sqrt{3}}.$

70. $\int \frac{dx}{X^2} = \frac{x}{3a^3X} + \frac{2}{3a^3} \int \frac{dx}{X}$ (see 69).

71. $\int \frac{x\,dx}{X} = \frac{1}{6a} \ln \frac{a^2 \mp ax + x^2}{(a \pm x)^2} \pm \frac{1}{a\sqrt{3}} \arctan \frac{2x \mp a}{a\sqrt{3}}.$

72. $\int \frac{x\,dx}{X^2} = \frac{x^2}{3a^3X} + \frac{1}{3a^3}\int \frac{x\,dx}{X}$ (see 71).

73. $\int \frac{x^2\,dx}{X} = \pm \frac{1}{3}\ln|X|$.

74. $\int \frac{x^2\,dx}{X^2} = \mp \frac{1}{3X}$.

75. $\int \frac{dx}{xX} = \frac{1}{3a^3}\ln\left|\frac{x^3}{X}\right|$.

76. $\int \frac{dx}{xX^2} = \frac{1}{3a^3X} + \frac{1}{3a^6}\ln\left|\frac{x^3}{X}\right|$.

77. $\int \frac{dx}{x^2X} = -\frac{1}{a^3x} \mp \frac{1}{a^3}\int \frac{x\,dx}{X}$ (see 71).

78. $\int \frac{dx}{x^2X^2} = -\frac{1}{a^6x} \mp \frac{x^2}{3a^6X} \mp \frac{4}{3a^6}\int \frac{x\,dx}{X}$ (see 71).

79. $\int \frac{dx}{a^4 + x^4} = \frac{1}{4a^3\sqrt{2}}\ln\frac{x^2 + ax\sqrt{(2)} + a^2}{x^2 - ax\sqrt{(2)} + a^2} + \frac{1}{2a^3\sqrt{2}}\arctan\frac{ax\sqrt{2}}{a^2 - x^2}$

$(a \neq 0)$.

80. $\int \frac{x\,dx}{a^4 + x^4} = \frac{1}{2a^2}\arctan\frac{x^2}{a^2}$ $(a \neq 0)$.

81. $\int \frac{x^2\,dx}{a^4 + x^4} = -\frac{1}{4a\sqrt{2}}\ln\frac{x^2 + ax\sqrt{(2)} + a^2}{x^2 - ax\sqrt{(2)} + a^2} + \frac{1}{2a\sqrt{2}}\arctan\frac{ax\sqrt{2}}{a^2 - x^2}$

$(a \neq 0)$.

82. $\int \frac{x^3\,dx}{a^4 + x^4} = \frac{1}{4}\ln|a^4 + x^4|$.

83. $\int \frac{dx}{a^4 - x^4} = \frac{1}{4a^3}\ln\left|\frac{a + x}{a - x}\right| + \frac{1}{2a^3}\arctan\frac{x}{a}$ $(a \neq 0)$.

84. $\int \frac{x\,dx}{a^4 - x^4} = \frac{1}{4a^2}\ln\left|\frac{a^2 + x^2}{a^2 - x^2}\right|$ $(a \neq 0)$.

85. $\int \frac{x^2\,dx}{a^4 - x^4} = \frac{1}{4a}\ln\left|\frac{a + x}{a - x}\right| - \frac{1}{2a}\arctan\frac{x}{a}$ $(a \neq 0)$.

86. $\int \frac{x^3\,dx}{a^4 - x^4} = -\frac{1}{4}\ln|a^4 - x^4|$.

(b) ***Irrational Functions.***

Notation: $X = a^2 \pm b^2x$ $(a > 0, b > 0)$,

$Y = \arctan\frac{b\sqrt{x}}{a}$ for the positive sign,

$Y = \frac{1}{2}\ln\left|\frac{a + b\sqrt{x}}{a - b\sqrt{x}}\right|$ for the negative sign.

The upper or lower sign in the formulae corresponds to the case $X = a^2 + b^2x$, or $X = a^2 - b^2x$, respectively.

87. $\int \frac{\sqrt{(x)}\,dx}{X} = \pm\frac{2\sqrt{x}}{b^2} \mp \frac{2a}{b^3}Y$.

88. $\int \frac{\sqrt{(x)}\,dx}{X^2} = \mp\frac{\sqrt{x}}{b^2X} \pm \frac{1}{ab^3}Y$.

89. $\int \frac{dx}{X\sqrt{x}} = \frac{2}{ab}Y$.

90. $\int \frac{dx}{X^2\sqrt{x}} = \frac{\sqrt{x}}{a^2X} + \frac{1}{a^3b}Y$.

91. $\int \frac{\sqrt{(x)}\,dx}{a^4 + x^2} = -\frac{1}{2a\sqrt{2}}\ln\frac{x + a\sqrt{(2x)} + a^2}{x - a\sqrt{(2x)} + a^2} + \frac{1}{a\sqrt{2}}\arctan\frac{a\sqrt{(2x)}}{a^2 - x}$

$(a \neq 0)$.

92. $\int \frac{dx}{(a^4 + x^2)\sqrt{x}} = \frac{1}{2a^3\sqrt{2}}\ln\frac{x + a\sqrt{(2x)} + a^2}{x - a\sqrt{(2x)} + a^2} + \frac{1}{a^3\sqrt{2}}\arctan\frac{a\sqrt{(2x)}}{a^2 - x}$

$(a \neq 0)$.

93. $\int \frac{\sqrt{(x)}\,dx}{a^4 - x^2} = \frac{1}{2a}\ln\left|\frac{a + \sqrt{x}}{a - \sqrt{x}}\right| = \frac{1}{a}\arctan\frac{\sqrt{x}}{a}$ $(a \neq 0)$.

94. $\int \frac{dx}{(a^4 - x^2)\sqrt{x}} = \frac{1}{2a^3}\ln\left|\frac{a + \sqrt{x}}{a - \sqrt{x}}\right| + \frac{1}{a^3}\arctan\frac{\sqrt{x}}{a}$ $(a \neq 0)$.

Notation: $X = ax + b \ (a \neq 0, b \neq 0)$.

95. $\int X^r \, dx = \frac{1}{(r+1)a} X^{r+1} \quad (r \neq -1;$ for $r = -1$ see 3).

96. $\int \sqrt{(X)} \, dx = \frac{2}{3a} \sqrt{X^3}$.

97. $\int x \sqrt{(X)} \, dx = \frac{2(3ax - 2b)\sqrt{X^3}}{15a^2}$.

98. $\int x^2 \sqrt{(X)} \, dx = \frac{2(15a^2x^2 - 12abx + 8b^2)\sqrt{X^3}}{105a^3}$.

99. $\int \frac{dx}{\sqrt{X}} = \frac{2\sqrt{X}}{a}$.

100. $\int \frac{x \, dx}{\sqrt{X}} = \frac{2(ax - 2b)}{3a^2} \sqrt{X}$.

101. $\int \frac{x^2 \, dx}{\sqrt{X}} = \frac{2(3a^2x^2 - 4abx + 8b^2)\sqrt{X}}{15a^3}$.

102. $\int \frac{dx}{x\sqrt{X}} = \begin{cases} \frac{1}{\sqrt{b}} \ln \left| \frac{\sqrt{X} - \sqrt{b}}{\sqrt{X} + \sqrt{b}} \right| & \text{for} \quad b > 0, \\ \frac{2}{\sqrt{(-b)}} \arctan \sqrt{\frac{X}{-b}} & \text{for} \quad b < 0. \end{cases}$

103. $\int \frac{\sqrt{X}}{x} \, dx = 2\sqrt{X} + b \int \frac{dx}{x\sqrt{X}}$ (see 102).

104. $\int \frac{dx}{x^2\sqrt{X}} = -\frac{\sqrt{X}}{bx} - \frac{a}{2b} \int \frac{dx}{x\sqrt{X}}$ (see 102).

105. $\int \frac{\sqrt{X}}{x^2} \, dx = -\frac{\sqrt{X}}{x} + \frac{a}{2} \int \frac{dx}{x\sqrt{X}}$ (see 102).

106. $\int \frac{dx}{x^n\sqrt{X}} = -\frac{\sqrt{X}}{(n-1)bx^{n-1}} - \frac{(2n-3)a}{(2n-2)b} \int \frac{dx}{x^{n-1}\sqrt{X}} \quad (n > 1)$.

107. $\int \sqrt{(X^3)}\,dx = \frac{2\sqrt{X^5}}{5a}$.

108. $\int x\sqrt{(X^3)}\,dx = \frac{2}{35a^2}\left[5\sqrt{(X^7)} - 7b\sqrt{(X^5)}\right]$.

109. $\int x^2\sqrt{(X^3)}\,dx = \frac{2}{a^3}\left(\frac{\sqrt{X^9}}{9} - \frac{2b\sqrt{X^7}}{7} + \frac{b^2\sqrt{X^5}}{5}\right)$.

110. $\int \frac{\sqrt{X^3}}{x}\,dx = \frac{2\sqrt{X^3}}{3} + 2b\sqrt{(X)} + b^2\int \frac{dx}{x\sqrt{X}}$ (see 102).

111. $\int \frac{x\,dx}{\sqrt{X^3}} = \frac{2}{a^2}\left(\sqrt{(X)} + \frac{b}{\sqrt{X}}\right)$.

112. $\int \frac{x^2\,dx}{\sqrt{X^3}} = \frac{2}{a^3}\left(\frac{\sqrt{X^3}}{3} - 2b\sqrt{X} - \frac{b^2}{\sqrt{X}}\right)$.

113. $\int \frac{dx}{x\sqrt{X^3}} = \frac{2}{b\sqrt{X}} + \frac{1}{b}\int \frac{dx}{x\sqrt{X}}$ (see 102).

114. $\int \frac{dx}{x^2\sqrt{X^3}} = -\frac{1}{bx\sqrt{X}} - \frac{3a}{b^2\sqrt{X}} - \frac{3a}{2b^2}\int \frac{dx}{x\sqrt{X}}$ (see 102).

115. $\int X^{\pm n/2}\,dx = \frac{2X^{\frac{1}{2}(2\pm n)}}{a(2\pm n)}$ $(n \pm 2 \neq 0)$.

116. $\int xX^{\pm n/2}\,dx = \frac{2}{a^2}\left(\frac{X^{\frac{1}{2}(4\pm n)}}{4\pm n} - \frac{bX^{\frac{1}{2}(2\pm n)}}{2\pm n}\right)$ $(n \pm 2 \neq 0,\ n \pm 4 \neq 0)$.

117. $\int x^2X^{\pm n/2}\,dx = \frac{2}{a^3}\left(\frac{X^{\frac{1}{2}(6\pm n)}}{6\pm n} - \frac{2bX^{\frac{1}{2}(4\pm n)}}{4\pm n} + \frac{b^2X^{\frac{1}{2}(2\pm n)}}{2\pm n}\right)$
$(n \pm 2 \neq 0,\ n \pm 4 \neq 0,\ n \pm 6 \neq 0)$.

118. $\int \frac{X^{n/2}\,dx}{x} = \frac{2X^{n/2}}{n} + b\int \frac{X^{\frac{1}{2}(n-2)}}{x}\,dx$ $(n \neq 0)$.

119. $\int \frac{dx}{xX^{n/2}} = \frac{2}{(n-2)\,bX^{\frac{1}{2}(n-2)}} + \frac{1}{b}\int \frac{dx}{xX^{\frac{1}{2}(n-2)}}$ $(n \neq 2)$.

120. $\int \frac{dx}{x^2X^{n/2}} = -\frac{1}{bxX^{\frac{1}{2}(n-2)}} - \frac{na}{2b}\int \frac{dx}{xX^{n/2}}$.

Notation: $X = ax + b$, $Y = fx + g$, $\Delta = bf - ag$ $(a \neq 0, f \neq 0, \Delta \neq 0)$.

121. $$\int \frac{dx}{\sqrt{(XY)}} \begin{cases} = \dfrac{2}{\sqrt{(-af)}} \arctan \sqrt{-\dfrac{fX}{aY}} & \text{for } af < 0 \quad (aY > 0), \\ = \dfrac{2}{\sqrt{(af)}} \operatorname{artanh} \sqrt{\dfrac{fX}{aY}} = \dfrac{2}{\sqrt{(af)}} \ln \left|\sqrt{(aY)} + \sqrt{(fX)}\right| & \\ \text{for } af > 0 \quad (aY > 0). \end{cases}$$

122. $$\int \frac{x\,dx}{\sqrt{(XY)}} = \frac{\sqrt{(XY)}}{af} - \frac{ag + bf}{2af} \int \frac{dx}{\sqrt{(XY)}} \quad \text{(see 121)}.$$

123. $$\int \frac{dx}{\sqrt{(X)}\sqrt{Y^3}} = -\frac{2\sqrt{X}}{\Delta\sqrt{Y}}.$$

124. $$\frac{dx}{Y\sqrt{X}} = \begin{cases} \dfrac{2}{\sqrt{(-\Delta f)}} \arctan \dfrac{f\sqrt{X}}{\sqrt{(-\Delta f)}} & \text{for } \Delta f < 0, \\ \dfrac{1}{\sqrt{(\Delta f)}} \ln \left|\dfrac{f\sqrt{(X)} - \sqrt{(\Delta f)}}{f\sqrt{(X)} + \sqrt{(\Delta f)}}\right| & \text{for } \Delta f > 0. \end{cases}$$

125. $$\int \sqrt{(XY)}\,dx = \frac{\Delta + 2aY}{4af}\sqrt{(XY)} - \frac{\Delta^2}{8af} \int \frac{dx}{\sqrt{(XY)}} \quad \text{(see 121)}.$$

126. $$\int \sqrt{\left(\frac{Y}{X}\right)}\,dx = \frac{1}{a}\sqrt{(XY)} - \frac{\Delta}{2a} \int \frac{dx}{\sqrt{(XY)}} \quad \text{(see 121; } Y > 0).$$

127. $$\int \frac{\sqrt{(X)}\,dx}{Y} = \frac{2\sqrt{X}}{f} + \frac{\Delta}{f} \int \frac{dx}{Y\sqrt{X}} \quad \text{(see 124)}.$$

128. $$\int \frac{Y^n\,dx}{\sqrt{X}} = \frac{2}{(2n+1)\,a} \left(\sqrt{(X)}\,Y^n - n\Delta \int \frac{Y^{n-1}\,dx}{\sqrt{X}}\right).$$

129. $$\int \frac{dx}{\sqrt{(X)}\,Y^n} = -\frac{1}{(n-1)\,\Delta} \left\{\frac{\sqrt{X}}{Y^{n-1}} + \left(n - \frac{3}{2}\right) a \int \frac{dx}{\sqrt{(X)}\,Y^{n-1}}\right\} \quad (n > 1).$$

130. $$\int \sqrt{(X)}\,Y^n\,dx = \frac{1}{(2n+3)f} \left(2\sqrt{(X)}\,Y^{n+1} + \Delta \int \frac{Y^n\,dx}{\sqrt{X}}\right) \quad \text{(see 128)}.$$

131. $$\int \frac{\sqrt{(X)}\,dx}{Y^n} = \frac{1}{(n-1)f} \left(-\frac{\sqrt{X}}{Y^{n-1}} + \frac{a}{2} \int \frac{dx}{\sqrt{(X)}\,Y^{n-1}}\right) \quad (n > 1).$$

Notation: $X = a^2 - x^2 \quad (a > 0)$.

132. $$\int \sqrt{(X)}\, dx = \frac{1}{2}\left(x\sqrt{(X)} + a^2 \arcsin\frac{x}{a}\right).$$

133. $$\int x\sqrt{(X)}\, dx = -\tfrac{1}{3}\sqrt{X^3}.$$

134. $$\int x^2\sqrt{(X)}\, dx = -\frac{x}{4}\sqrt{(X^3)} + \frac{a^2}{8}\left(x\sqrt{(X)} + a^2 \arcsin\frac{x}{a}\right).$$

135. $$\int \frac{\sqrt{X}}{x}\, dx = \sqrt{(X)} - a \ln\left|\frac{a + \sqrt{X}}{x}\right|.$$

136. $$\int \frac{\sqrt{X}}{x^2}\, dx = -\frac{\sqrt{X}}{x} - \arcsin\frac{x}{a}.$$

137. $$\int \frac{dx}{\sqrt{X}} = \arcsin\frac{x}{a}.$$

138. $$\int \frac{x\, dx}{\sqrt{X}} = -\sqrt{X}.$$

139. $$\int \frac{x^2\, dx}{\sqrt{X}} = -\frac{x}{2}\sqrt{(X)} + \frac{a^2}{2}\arcsin\frac{x}{a}.$$

140. $$\int \frac{dx}{x\sqrt{X}} = -\frac{1}{a}\ln\left|\frac{a + \sqrt{X}}{x}\right|.$$

141. $$\int \frac{dx}{x^2\sqrt{X}} = -\frac{\sqrt{X}}{a^2 x}.$$

142. $$\int \sqrt{(X^3)}\, dx = \frac{1}{4}\left(x\sqrt{(X^3)} + \frac{3a^2x}{2}\sqrt{(X)} + \frac{3a^4}{2}\arcsin\frac{x}{a}\right).$$

143. $$\int x\sqrt{(X^3)}\, dx = -\tfrac{1}{5}\sqrt{X^5}.$$

144. $$\int x^2\sqrt{(X^3)}\, dx = -\frac{x\sqrt{X^5}}{6} + \frac{a^2x\sqrt{X^3}}{24} + \frac{a^4x\sqrt{X}}{16} + \frac{a^6}{16}\arcsin\frac{x}{a}.$$

145. $$\int \frac{\sqrt{X^3}}{x}\, dx = \frac{\sqrt{X^3}}{3} + a^2\sqrt{(X)} - a^3 \ln\left|\frac{a + \sqrt{X}}{x}\right|.$$

146. $\int \frac{\sqrt{X^3}}{x^2}\,dx = -\frac{\sqrt{X^3}}{x} - \frac{3}{2}x\sqrt{X} - \frac{3}{2}a^2 \arcsin\frac{x}{a}$

147. $\int \frac{dx}{\sqrt{X^3}} = \frac{x}{a^2\sqrt{X}}.$

148. $\int \frac{x\,dx}{\sqrt{X^3}} = \frac{1}{\sqrt{X}}.$

149. $\int \frac{x^2\,dx}{\sqrt{X^3}} = \frac{x}{\sqrt{X}} - \arcsin\frac{x}{a}.$

150. $\int \frac{dx}{x\sqrt{X^3}} = \frac{1}{a^2\sqrt{X}} - \frac{1}{a^3}\ln\left|\frac{a+\sqrt{X}}{x}\right|.$

151. $\int \frac{dx}{x^2\sqrt{X^3}} = \frac{1}{a^4}\left(-\frac{\sqrt{X}}{x} + \frac{x}{\sqrt{X}}\right).$

Notation: $X = a^2 + x^2$, $a > 0$.

152. $\int \sqrt{(X)}\,dx = \frac{1}{2}\left(x\sqrt{(X)} + a^2 \operatorname{arsinh}\frac{x}{a}\right) + C =$

$= \frac{1}{2}(x\sqrt{(X)} + a^2 \ln|x + \sqrt{X}|) + C_1.$

153. $\int x\sqrt{(X)}\,dx = \frac{1}{3}\sqrt{X^3}.$

154. $\int x^2\sqrt{(X)}\,dx = \frac{x}{4}\sqrt{(X^3)} - \frac{a^2}{8}\left(x\sqrt{(X)} + a^2 \operatorname{arsinh}\frac{x}{a}\right) + C =$

$= \frac{x}{4}\sqrt{(X^3)} - \frac{a^2}{8}(x\sqrt{(X)} + a^2 \ln|x + \sqrt{X}|) + C_1.$

155. $\int x^3\sqrt{(X)}\,dx = \frac{\sqrt{X^5}}{5} - \frac{a^2\sqrt{X^3}}{3}.$

156. $\int \frac{\sqrt{X}}{x}\,dx = \sqrt{(X)} - a\ln\left|\frac{a+\sqrt{X}}{x}\right|.$

157. $\int \frac{\sqrt{X}}{x^2}\,dx = -\frac{\sqrt{X}}{x} + \operatorname{arsinh}\frac{x}{a} + C = -\frac{\sqrt{X}}{x} + \ln|x + \sqrt{X}| + C_1.$

158. $\int \frac{\sqrt{X}}{x^3}\,dx = -\frac{\sqrt{X}}{2x^2} - \frac{1}{2a}\ln\left|\frac{a+\sqrt{X}}{x}\right|.$

159. $\int \frac{dx}{\sqrt{X}} = \operatorname{arsinh} \frac{x}{a} + C = \ln |x + \sqrt{X}| + C_1 .$

160. $\int \frac{x\,dx}{\sqrt{X}} = \sqrt{X} .$

161. $\int \frac{x^2\,dx}{\sqrt{X}} = \frac{x}{2}\sqrt{(X)} - \frac{a^2}{2} \operatorname{arsinh} \frac{x}{a} + C = \frac{x}{2}\sqrt{(X)} - \frac{a^2}{2} \ln |x + \sqrt{X}| + C_1 .$

162. $\int \frac{x^3\,dx}{\sqrt{X}} = \frac{\sqrt{(X^3)}}{3} - a^2\sqrt{X} .$

163. $\int \frac{dx}{x\sqrt{X}} = -\frac{1}{a} \ln \left| \frac{a + \sqrt{X}}{x} \right| .$

164. $\int \frac{dx}{x^2\sqrt{X}} = -\frac{\sqrt{X}}{a^2 x} .$

165. $\int \frac{dx}{x^3\sqrt{X}} = -\frac{\sqrt{X}}{2a^2x^2} + \frac{1}{2a^3} \ln \left| \frac{a + \sqrt{X}}{x} \right| .$

166. $\int \sqrt{(X^3)}\,dx = \frac{1}{4}\left(x\sqrt{(X^3)} + \frac{3a^2x}{2}\sqrt{(X)} + \frac{3a^4}{2} \operatorname{arsinh} \frac{x}{a} \right) + C =$

$$= \frac{1}{4}\left(x\sqrt{(X^3)} + \frac{3a^2x}{2}\sqrt{(X)} + \frac{3a^4}{2} \ln |x + \sqrt{X}| \right) + C_1 .$$

167. $\int x\sqrt{(X^3)}\,dx = \tfrac{1}{5}\sqrt{X^5} .$

168. $\int x^2\sqrt{(X^3)}\,dx = \frac{x\sqrt{X^5}}{6} - \frac{a^2x\sqrt{X^3}}{24} - \frac{a^4x\sqrt{X}}{16} - \frac{a^6}{16} \operatorname{arsinh} \frac{x}{a} + C =$

$$= \frac{x\sqrt{X^5}}{6} - \frac{a^2x\sqrt{X^3}}{24} - \frac{a^4x\sqrt{X}}{16} - \frac{a^6}{16} \ln |x + \sqrt{X}| + C_1 .$$

169. $\int x^3\sqrt{(X^3)}\,dx = \frac{\sqrt{X^7}}{7} - \frac{a^2\sqrt{X^5}}{5} .$

170. $\int \frac{\sqrt{X^3}}{x}\,dx = \frac{\sqrt{X^3}}{3} + a^2\sqrt{(X)} - a^3 \ln \left| \frac{a + \sqrt{X}}{x} \right| .$

171. $\int \frac{\sqrt{X^3}}{x^2}\,dx = -\frac{\sqrt{X^3}}{x} + \tfrac{3}{2}x\sqrt{(X)} + \tfrac{3}{2}a^2 \operatorname{arsinh} \frac{x}{a} + C =$

$$= -\frac{\sqrt{X^3}}{x} + \tfrac{3}{2}x\sqrt{(X)} + \tfrac{3}{2}a^2 \ln |x + \sqrt{X}| + C_1 .$$

172. $\int \frac{\sqrt{X^3}}{x^3}\,dx = -\frac{\sqrt{X^3}}{2x^2} + \frac{3}{2}\sqrt{(X)} - \frac{3}{2}a \ln\left|\frac{a+\sqrt{X}}{x}\right|.$

173. $\int \frac{dx}{\sqrt{X^3}} = \frac{x}{a^2\sqrt{X}}.$

174. $\int \frac{x\,dx}{\sqrt{X^3}} = -\frac{1}{\sqrt{X}}.$

175. $\int \frac{x^2\,dx}{\sqrt{X^3}} = -\frac{x}{\sqrt{X}} + \operatorname{arsinh}\frac{x}{a} + C = -\frac{x}{\sqrt{X}} + \ln\left|x+\sqrt{X}\right| + C_1.$

176. $\int \frac{x^3\,dx}{\sqrt{X^3}} = \sqrt{(X)} + \frac{a^2}{\sqrt{X}}.$

177. $\int \frac{dx}{x\sqrt{X^3}} = \frac{1}{a^2\sqrt{X}} - \frac{1}{a^3}\ln\left|\frac{a+\sqrt{X}}{x}\right|.$

178. $\int \frac{dx}{x^2\sqrt{X^3}} = -\frac{1}{a^4}\left(\frac{\sqrt{X}}{x} + \frac{x}{\sqrt{X}}\right).$

179. $\int \frac{dx}{x^3\sqrt{X^3}} = -\frac{1}{2a^2x^2\sqrt{X}} - \frac{3}{2a^4\sqrt{X}} + \frac{3}{2a^5}\ln\left|\frac{a+\sqrt{X}}{x}\right|.$

Notation: $X = x^2 - a^2$, $a > 0$.
If arcosh (x/a) occurs in a formula, x belonging to the interval $[a, \infty)$ is assumed. For $x \in (-\infty, -a]$ the function arcosh (x/a) is to be replaced by $-\operatorname{arcosh}(-x/a)$.

180. $\int \sqrt{(X)}\,dx = \frac{1}{2}\left(x\sqrt{(X)} - a^2 \operatorname{arcosh}\frac{x}{a}\right) + C =$

$= \frac{1}{2}(x\sqrt{(X)} - a^2 \ln\left|x+\sqrt{X}\right|) + C_1.$

181. $\int x\sqrt{(X)}\,dx = \frac{1}{3}\sqrt{X^3}.$

182. $\int x^2\sqrt{(X)}\,dx = \frac{x}{4}\sqrt{(X^3)} + \frac{a^2}{8}\left(x\sqrt{(X)} - a^2 \operatorname{arcosh}\frac{x}{a}\right) + C =$

$= \frac{x}{4}\sqrt{(X^3)} + \frac{a^2}{8}\left[x\sqrt{(X)} - a^2 \ln\left|x+\sqrt{X}\right|\right] + C_1.$

183. $\int x^3 \sqrt{(X)}\, dx = \frac{\sqrt{X^5}}{5} + \frac{a^2 \sqrt{X^3}}{3}$.

184. $\int \frac{\sqrt{X}}{x} dx = \sqrt{(X)} - a \arccos \frac{a}{|x|}$.

185. $\int \frac{\sqrt{X}}{x^2} dx = -\frac{\sqrt{X}}{x} + \operatorname{arcosh} \frac{x}{a} + C = -\frac{\sqrt{X}}{x} + \ln |x + \sqrt{X}| + C_1$.

186. $\int \frac{\sqrt{X}}{x^3} dx = -\frac{\sqrt{X}}{2x^2} + \frac{1}{2a} \arccos \frac{a}{|x|}$.

187. $\int \frac{dx}{\sqrt{X}} = \operatorname{arcosh} \frac{x}{a} + C = \ln |x + \sqrt{X}| + C_1$.

188. $\int \frac{x\, dx}{\sqrt{X}} = \sqrt{X}$.

189. $\int \frac{x^2\, dx}{\sqrt{X}} = \frac{x}{2} \sqrt{(X)} + \frac{a^2}{2} \operatorname{arcosh} \frac{x}{a} + C = \frac{x}{2} \sqrt{(X)} + \frac{a^2}{2} \ln |x + \sqrt{X}| + C_1$.

190. $\int \frac{x^3\, dx}{\sqrt{X}} = \frac{\sqrt{X^3}}{3} + a^2 \sqrt{X}$.

191. $\int \frac{dx}{x \sqrt{X}} = \frac{1}{a} \arccos \frac{a}{|x|}$.

192. $\int \frac{dx}{x^2 \sqrt{X}} = \frac{\sqrt{X}}{a^2 x}$.

193. $\int \frac{dx}{x^3 \sqrt{X}} = \frac{\sqrt{X}}{2a^2 x^2} + \frac{1}{2a^3} \arccos \frac{a}{|x|}$.

194. $\int \sqrt{(X^3)}\, dx = \frac{1}{4}\left(x \sqrt{(X^3)} - \frac{3a^2 x}{2} \sqrt{(X)} + \frac{3a^4}{2} \operatorname{arcosh} \frac{x}{a}\right) + C =$

$= \frac{1}{4}\left(x \sqrt{(X^3)} - \frac{3a^2 x}{2} \sqrt{(X)} + \frac{3a^4}{2} \ln |x + \sqrt{X}|\right) + C_1$.

195. $\int x \sqrt{(X^3)}\, dx = \frac{1}{5} \sqrt{X^5}$.

196. $\int x^2 \sqrt{(X^3)}\, dx = \frac{x \sqrt{X^5}}{6} + \frac{a^2 x \sqrt{X^3}}{24} - \frac{a^4 x \sqrt{X}}{16} + \frac{a^6}{16} \operatorname{arcosh} \frac{x}{a} + C =$

$= \frac{x \sqrt{X^5}}{6} + \frac{a^2 x \sqrt{X^3}}{24} - \frac{a^4 x \sqrt{X}}{16} + \frac{a^6}{16} \ln |x + \sqrt{X}| + C_1$.

197. $\int x^3 \sqrt{(X^3)}\, dx = \frac{\sqrt{X^7}}{7} + \frac{a^2 \sqrt{X^5}}{5}$.

198. $\int \frac{\sqrt{X^3}}{x}\, dx = \frac{\sqrt{X^3}}{3} - a^2 \sqrt{(X)} + a^3 \arccos \frac{a}{|x|}$.

199. $\int \frac{\sqrt{X^3}}{x^2}\, dx = -\frac{\sqrt{X^3}}{x} + \frac{3}{2}x \sqrt{(X)} - \frac{3}{2}a^2 \operatorname{arcosh} \frac{x}{a} + C =$

$= -\frac{\sqrt{X^3}}{x} + \frac{3}{2}x \sqrt{(X)} - \frac{3}{2}a^2 \ln |x + \sqrt{X}| + C_1$.

200. $\int \frac{\sqrt{X^3}}{x^3}\, dx = -\frac{\sqrt{X^3}}{2x^2} + \frac{3\sqrt{X}}{2} - \frac{3}{2} a \arccos \frac{a}{|x|}$.

201. $\int \frac{dx}{\sqrt{X^3}} = -\frac{x}{a^2 \sqrt{X}}$.

202. $\int \frac{x\, dx}{\sqrt{X^3}} = -\frac{1}{\sqrt{X}}$.

203. $\int \frac{x^2\, dx}{\sqrt{X^3}} = -\frac{x}{\sqrt{X}} + \operatorname{arcosh} \frac{x}{a} + C = -\frac{x}{\sqrt{X}} + \ln |x + \sqrt{X}| + C_1$.

204. $\int \frac{x^3\, dx}{\sqrt{X^3}} = \sqrt{(X)} - \frac{a^2}{\sqrt{X}}$.

205. $\int \frac{dx}{x \sqrt{X^3}} = -\frac{1}{a^2 \sqrt{X}} - \frac{1}{a^3} \arccos \frac{a}{|x|}$.

206. $\int \frac{dx}{x^2 \sqrt{X^3}} = -\frac{1}{a^4}\left(\frac{\sqrt{X}}{x} + \frac{x}{\sqrt{X}}\right)$.

207. $\int \frac{dx}{x^3 \sqrt{X^3}} = \frac{1}{2a^2x^2 \sqrt{X}} - \frac{3}{2a^4 \sqrt{X}} - \frac{3}{2a^5} \arccos \frac{a}{|x|}$.

Notation: $X = ax^2 + bx + c$, $\Delta = 4ac - b^2$, $k = 4a/\Delta$, $a \neq 0$, $\Delta \neq 0$.

If $\Delta > 0$, then $ax^2 + bx + c$ has the same sign for all x and $\sqrt{X}$ is real either for all x, if $a > 0$, or for no x, if $a < 0$.

If $\Delta < 0$, then the equation $ax^2 + bx + c = 0$ has two distinct real roots $\alpha_1 < \alpha_2$ and $\sqrt{X}$ is real either for $x \in [\alpha_1, \alpha_2]$ if $a < 0$, or for $x \in (-\infty, \alpha_1]$ and $x \in [\alpha_2, \infty)$ if $a > 0$.

208. $$\int \frac{dx}{\sqrt{X}} = \begin{cases} \frac{1}{\sqrt{a}} \ln \left| 2\sqrt{(aX)} + 2ax + b \right| + C & \text{for } a > 0, \\ \frac{1}{\sqrt{a}} \operatorname{arsinh} \frac{2ax + b}{\sqrt{\Delta}} + C_1 & \text{for } a > 0,\ \Delta > 0, \\ -\frac{1}{\sqrt{(-a)}} \arcsin \frac{2ax + b}{\sqrt{(-\Delta)}} & \text{for } a < 0,\ \Delta < 0. \end{cases}$$

209. $$\int \frac{dx}{X\sqrt{X}} = \frac{2(2ax + b)}{\Delta\sqrt{X}}.$$

210. $$\int \frac{dx}{X^2\sqrt{X}} = \frac{2(2ax + b)}{3\Delta\sqrt{X}} \left(\frac{1}{X} + 2k \right).$$

211. $$\int \frac{dx}{X^{\frac{1}{2}(2n+1)}} = \frac{2(2ab + b)}{(2n - 1)\Delta X^{\frac{1}{2}(2n-1)}} + \frac{2k(n - 1)}{2n - 1} \int \frac{dx}{X^{\frac{1}{2}(2n-1)}}.$$

212. $$\int \sqrt{(X)}\, dx = \frac{(2ax + b)\sqrt{X}}{4a} + \frac{1}{2k} \int \frac{dx}{\sqrt{X}} \quad \text{(see 208)}.$$

213. $$\int X\sqrt{(X)}\, dx = \frac{(2ax + b)\sqrt{X}}{8a} \left(X + \frac{3}{2k} \right) + \frac{3}{8k^2} \int \frac{dx}{\sqrt{X}} \quad \text{(see 208)}.$$

214. $$\int X^2\sqrt{(X)}\, dx = \frac{(2ax + b)\sqrt{X}}{12a} \left(X^2 + \frac{5X}{4k} + \frac{15}{8k^2} \right) + \frac{5}{16k^3} \int \frac{dx}{\sqrt{X}} \quad \text{(see 208)}.$$

215. $$\int X^{\frac{1}{2}(2n+1)}\, dx = \frac{(2ax + b)X^{\frac{1}{2}(2n+1)}}{4a(n + 1)} + \frac{2n + 1}{2k(n + 1)} \int X^{\frac{1}{2}(2n-1)}\, dx \quad \text{(see 208 and 212)}.$$

216. $$\int \frac{x\, dx}{\sqrt{X}} = \frac{\sqrt{X}}{a} - \frac{b}{2a} \int \frac{dx}{\sqrt{X}} \quad \text{(see 208)}.$$

217. $$\int \frac{x\, dx}{X\sqrt{X}} = -\frac{2(bx + 2c)}{\Delta\sqrt{X}}.$$

218. $$\int \frac{x\, dx}{X^{\frac{1}{2}(2n+1)}} = -\frac{1}{(2n - 1)aX^{\frac{1}{2}(2n-1)}} - \frac{b}{2a} \int \frac{dx}{X^{\frac{1}{2}(2n+1)}} \quad \text{(see 211)}.$$

219. $$\int \frac{x^2\, dx}{\sqrt{X}} = \left(\frac{x}{2a} - \frac{3b}{4a^2} \right) \sqrt{(X)} + \frac{3b^2 - 4ac}{8a^2} \int \frac{dx}{\sqrt{X}} \quad \text{(see 208)}.$$

220. $$\int \frac{x^2\, dx}{X\sqrt{X}} = \frac{(2b^2 - 4ac)x + 2bc}{a\Delta\sqrt{X}} + \frac{1}{a} \int \frac{dx}{\sqrt{X}} \quad \text{(see 208)}.$$

221. $\int x\sqrt{(X)}\,dx = \frac{X\sqrt{X}}{3a} - \frac{b(2ax+b)}{8a^2}\sqrt{(X)} - \frac{b}{4ak}\int\frac{dx}{\sqrt{X}}$ (see 208).

222. $\int xX\sqrt{(X)}\,dx = \frac{X^2\sqrt{X}}{5a} - \frac{b}{2a}\int X\sqrt{(X)}\,dx$ (see 213).

223. $\int xX^{\frac{1}{2}(2n+1)}\,dx = \frac{X^{\frac{1}{2}(2n+3)}}{(2n+3)a} - \frac{b}{2a}\int X^{\frac{1}{2}(2n+1)}\,dx$ (see 215).

224. $\int x^2\sqrt{(X)}\,dx = \left(x - \frac{5b}{6a}\right)\frac{X\sqrt{X}}{4a} + \frac{5b^2-4ac}{16a^2}\int\sqrt{(X)}\,dx$ (see 212).

225. $$\int\frac{dx}{x\sqrt{X}} = \begin{cases} -\frac{1}{\sqrt{c}}\ln\left|\frac{2\sqrt{(cX)}}{x} + \frac{2c}{x} + b\right| + C & \text{for } c > 0, \\ -\frac{1}{\sqrt{c}}\operatorname{arsinh}\frac{bx+2c}{x\sqrt{\Delta}} + C_1 & \text{for } c > 0,\ \Delta > 0, \\ \frac{1}{\sqrt{(-c)}}\arcsin\frac{bx+2c}{x\sqrt{(-\Delta)}} & \text{for } c < 0,\ \Delta < 0 \\ & (\text{for } c = 0 \text{ see } 231). \end{cases}$$

226. $\int\frac{dx}{x^2\sqrt{X}} = -\frac{\sqrt{X}}{cx} - \frac{b}{2c}\int\frac{dx}{x\sqrt{X}}$ (see 225).

227. $\int\frac{\sqrt{(X)}\,dx}{x} = \sqrt{(X)} + \frac{b}{2}\int\frac{dx}{\sqrt{X}} + c\int\frac{dx}{x\sqrt{X}}$ (see 208 and 225).

228. $\int\frac{\sqrt{(X)}\,dx}{x^2} = -\frac{\sqrt{(X)}}{x} + a\int\frac{dx}{\sqrt{X}} + \frac{b}{2}\int\frac{dx}{x\sqrt{X}}$ (see 208 and 225).

229. $\int\frac{X^{\frac{1}{2}(2n+1)}}{x}\,dx = \frac{X^{\frac{1}{2}(2n+1)}}{2n+1} + \frac{b}{2}\int X^{\frac{1}{2}(2n-1)}\,dx + c\int\frac{X^{\frac{1}{2}(2n-1)}}{x}\,dx$

(see 212 and 227).

230. $$\int\frac{a_nx^n + a_{n-1}x^{n-1} + \ldots + a_1x + a_0}{\sqrt{X}}\,dx =$$

$$= (A_0 + A_1x + \ldots + A_{n-1}x^{n-1})\sqrt{(X)} + A_n\int\frac{dx}{\sqrt{X}};$$

the constants $A_0, A_1, \ldots, A_n$ can be determined by differentiation and then by the method of undetermined coefficients (by equating coefficients).

231. $\int\frac{dx}{x\sqrt{(ax^2+bx)}} = -\frac{2}{bx}\sqrt{(ax^2+bx)}$ $(b \neq 0)$.

232. $$\int \frac{dx}{\sqrt{(2ax - x^2)}} = \arcsin \frac{x - a}{|a|} \quad (a \neq 0).$$

233. $$\int \frac{x\,dx}{\sqrt{(2ax - x^2)}} = -\sqrt{(2ax - x^2)} + a \arcsin \frac{x - a}{|a|} \quad (a \neq 0).$$

234. $$\int \sqrt{(2ax - x^2)}\,dx = \frac{x - a}{2}\sqrt{(2ax - x^2)} + \frac{a^2}{2} \arcsin \frac{x - a}{|a|} \quad (a \neq 0).$$

235. $$\int \frac{dx}{(ax^2 + b)\sqrt{(fx^2 + g)}} = \frac{1}{\sqrt{(b)}\sqrt{(ag - bf)}} \arctan \frac{x\sqrt{(ag - bf)}}{\sqrt{(b)}\sqrt{(fx^2 + g)}}$$
$$(ag - bf > 0),$$
$$= \frac{1}{2\sqrt{(b)}\sqrt{(bf - ag)}} \ln \left| \frac{\sqrt{(b)}\sqrt{(fx^2 + g)} + x\sqrt{(bf - ag)}}{\sqrt{(b)}\sqrt{(fx^2 + g)} - x\sqrt{(bf - ag)}} \right| \quad (ag - bf < 0).$$

236. $$\int \sqrt[n]{(ax + b)}\,dx = \frac{n(ax + b)}{(n + 1)a} \sqrt[n]{(ax + b)} \quad (a \neq 0).$$

237. $$\int \frac{dx}{\sqrt[n]{(ax + b)}}\,dx = \frac{n(ax + b)}{(n - 1)a} \frac{1}{\sqrt[n]{(ax + b)}} \quad (n \neq 1; a \neq 0).$$

238. $$\int \frac{dx}{x\sqrt{(x^n + a^2)}} = -\frac{2}{na} \ln \left| \frac{a + \sqrt{(x^n + a^2)}}{\sqrt{x^n}} \right| \quad (x > 0).$$

239. $$\int \frac{dx}{x\sqrt{(x^n - a^2)}} = \frac{2}{na} \arccos \frac{a}{\sqrt{x^n}} \quad (x > \sqrt[n]{a^2}).$$

240. $$\int \frac{\sqrt{(x)}\,dx}{\sqrt{(a^3 - x^3)}} = \tfrac{2}{3} \arcsin \sqrt{\left(\frac{x}{a}\right)^3} \quad (0 \leqq x < a).$$

Reduction Formulae for Binomial Integrals

241. $$\int x^m(ax^n + b)^p\,dx =$$
$$= \frac{1}{m + np + 1}\left[x^{m+1}(ax^n + b)^p + npb \int x^m(ax^n + b)^{p-1}\,dx\right]$$
$$(m + np + 1 \neq 0),$$
$$= \frac{1}{bn(p + 1)}\left[-x^{m+1}(ax^n + b)^{p+1} + (m + n + np + 1)\int x^m(ax^n + b)^{p+1}\,dx\right]$$
$$(bn(p + 1) \neq 0),$$

$$= \frac{1}{(m+1)\,b}\left[x^{m+1}(ax^n+b)^{p+1} - a(m+n+np+1)\int x^{m+n}(ax^n+b)^p\,dx\right]$$

$$((m+1)\,b \neq 0),$$

$$= \frac{1}{a(m+np+1)}\left[x^{m-n+1}(ax^n+b)^{p+1} - (m-n+1)\,b\int x^{m-n}(ax^n+b)^p\,dx\right]$$

$$(a(m+np+1) \neq 0).$$

(In 241 m, n, p are rational numbers, $x > 0$.)

(c) ***Trigonometric Functions.***

$a \neq 0$ is assumed in all cases.

(See also 402–405, 417–422, 440–444.)

(α) *Integrals Containing the Sine Only.*

242. $$\int \sin ax\,dx = -\frac{1}{a}\cos ax.$$

243. $$\int \sin^2 ax\,dx = \tfrac{1}{2}x - \frac{1}{4a}\sin 2ax.$$

244. $$\int \sin^3 ax\,dx = -\frac{1}{a}\cos ax + \frac{1}{3a}\cos^3 ax.$$

245. $$\int \sin^4 ax\,dx = \tfrac{3}{8}x - \frac{1}{4a}\sin 2ax + \frac{1}{32a}\sin 4ax.$$

246. $$\int \sin^n ax\,dx = -\frac{\sin^{n-1} ax\cos ax}{na} + \frac{n-1}{n}\int \sin^{n-2} ax\,dx.$$

247. $$\int x\sin ax\,dx = \frac{\sin ax}{a^2} - \frac{x\cos ax}{a}.$$

248. $$\int x^2\sin ax\,dx = \frac{2x}{a^2}\sin ax - \left(\frac{x^2}{a} - \frac{2}{a^3}\right)\cos ax.$$

249. $$\int x^3\sin ax\,dx = \left(\frac{3x^2}{a^2} - \frac{6}{a^4}\right)\sin ax - \left(\frac{x^3}{a} - \frac{6x}{a^3}\right)\cos ax.$$

250. $\int x^n \sin ax\, dx = -\frac{x^n}{a}\cos ax + \frac{n}{a}\int x^{n-1}\cos ax\, dx$ (see 289).

251. $\int \frac{\sin ax}{x}\, dx = ax - \frac{(ax)^3}{3 \cdot 3!} + \frac{(ax)^5}{5 \cdot 5!} - \frac{(ax)^7}{7 \cdot 7!} + \ldots$

(the series is convergent for all x; see also §§ 13.12 and 15.7).

252. $\int \frac{\sin ax}{x^2}\, dx = -\frac{\sin ax}{x} + a\int \frac{\cos ax\, dx}{x}$ (see 290).

253. $\int \frac{\sin ax}{x^n}\, dx = -\frac{1}{n-1}\frac{\sin ax}{x^{n-1}} + \frac{a}{n-1}\int \frac{\cos ax}{x^{n-1}}\, dx$ (see 292; $n > 1$).

254. $\int \frac{dx}{\sin ax} = \frac{1}{a}\ln\left|\tan\frac{ax}{2}\right|$.

255. $\int \frac{dx}{\sin^2 ax} = -\frac{1}{a}\cot ax$.

256. $\int \frac{dx}{\sin^3 ax} = -\frac{\cos ax}{2a\sin^2 ax} + \frac{1}{2a}\ln\left|\tan\frac{ax}{2}\right|$.

257. $\int \frac{dx}{\sin^n ax} = -\frac{1}{a(n-1)}\frac{\cos ax}{\sin^{n-1} ax} + \frac{n-2}{n-1}\int \frac{dx}{\sin^{n-2} ax}$ $(n > 1)$.

258. $\int \frac{x\, dx}{\sin ax} = \frac{1}{a^2}\left(ax + \frac{(ax)^3}{3 \cdot 3!} + \frac{7(ax)^5}{3 \cdot 5 \cdot 5!} + \frac{31(ax)^7}{3 \cdot 7 \cdot 7!} + \frac{127(ax)^9}{3 \cdot 5 \cdot 9!} + \ldots + \right.$

$\left. + \frac{2(2^{2n-1}-1)}{(2n+1)!} B_n \cdot (ax)^{2n+1} + \ldots\right)$ $(|x| < \pi/a$; see Remark 3, p. 511).

259. $\int \frac{x\, dx}{\sin^2 ax} = -\frac{x}{a}\cot ax + \frac{1}{a^2}\ln|\sin ax|$.

260. $\int \frac{x\, dx}{\sin^n ax} = -\frac{x\cos ax}{(n-1)\, a \sin^{n-1} ax} - \frac{1}{(n-1)(n-2)\, a^2 \sin^{n-2} ax} +$

$+ \frac{n-2}{n-1}\int \frac{x\, dx}{\sin^{n-2} ax}$ $(n > 2)$.

261. $\int \frac{dx}{1 + \sin ax} = -\frac{1}{a}\tan\left(\frac{\pi}{4} - \frac{ax}{2}\right)$.

262. $\int \frac{dx}{1 - \sin ax} = \frac{1}{a}\tan\left(\frac{\pi}{4} + \frac{ax}{2}\right)$.

263. $\int \frac{x\,dx}{1+\sin ax} = -\frac{x}{a}\tan\left(\frac{\pi}{4}-\frac{ax}{2}\right)+\frac{2}{a^2}\ln\left|\cos\left(\frac{\pi}{4}-\frac{ax}{2}\right)\right|.$

264. $\int \frac{x\,dx}{1-\sin ax} = \frac{x}{a}\cot\left(\frac{\pi}{4}-\frac{ax}{2}\right)+\frac{2}{a^2}\ln\left|\sin\left(\frac{\pi}{4}-\frac{ax}{2}\right)\right|.$

265. $\int \frac{\sin ax\,dx}{1\pm\sin ax} = \pm x+\frac{1}{a}\tan\left(\frac{\pi}{4}\mp\frac{ax}{2}\right).$

266. $\int \frac{dx}{\sin ax(1\pm\sin ax)} = \frac{1}{a}\tan\left(\frac{\pi}{4}\mp\frac{ax}{2}\right)+\frac{1}{a}\ln\left|\tan\frac{ax}{2}\right|.$

267. $\int \frac{dx}{(1+\sin ax)^2} = -\frac{1}{2a}\tan\left(\frac{\pi}{4}-\frac{ax}{2}\right)-\frac{1}{6a}\tan^3\left(\frac{\pi}{4}-\frac{ax}{2}\right).$

268. $\int \frac{dx}{(1-\sin ax)^2} = \frac{1}{2a}\cot\left(\frac{\pi}{4}-\frac{ax}{2}\right)+\frac{1}{6a}\cot^3\left(\frac{\pi}{4}-\frac{ax}{2}\right).$

269. $\int \frac{\sin ax\,dx}{(1+\sin ax)^2} = -\frac{1}{2a}\tan\left(\frac{\pi}{4}-\frac{ax}{2}\right)+\frac{1}{6a}\tan^3\left(\frac{\pi}{4}-\frac{ax}{2}\right).$

270. $\int \frac{\sin ax\,dx}{(1-\sin ax)^2} = -\frac{1}{2a}\cot\left(\frac{\pi}{4}-\frac{ax}{2}\right)+\frac{1}{6a}\cot^3\left(\frac{\pi}{4}-\frac{ax}{2}\right).$

271. $\int \frac{dx}{1+\sin^2 ax} = \frac{1}{2\sqrt{(2)}\,a}\arcsin\left(\frac{3\sin^2(ax)-1}{\sin^2(ax)+1}\right) \quad (\sin 2ax > 0).$

272. $\int \frac{dx}{1-\sin^2 ax} = \int \frac{dx}{\cos^2 ax} = \frac{1}{a}\tan ax.$

273. $\int \sin ax \sin bx\,dx = \frac{\sin(a-b)x}{2(a-b)} - \frac{\sin(a+b)x}{2(a+b)}$

($|a| \neq |b|$; for $|a| = |b|$ see 243).

274. $\int \frac{dx}{b+c\sin ax} = \frac{2}{a\sqrt{(b^2-c^2)}}\arctan\frac{b\tan(\frac{1}{2}ax)+c}{\sqrt{(b^2-c^2)}}$ (for $b^2 > c^2$),

$= \frac{1}{a\sqrt{(c^2-b^2)}}\ln\left|\frac{b\tan(\frac{1}{2}ax)+c-\sqrt{(c^2-b^2)}}{b\tan(\frac{1}{2}ax)+c+\sqrt{(c^2-b^2)}}\right|$ (for $b^2 < c^2$).

275. $\int \frac{\sin ax\,dx}{b+c\sin ax} = \frac{x}{c}-\frac{b}{c}\int\frac{dx}{b+c\sin ax}$ (see 274).

276. $\int \frac{dx}{\sin ax(b+c\sin ax)} = \frac{1}{ab}\ln\left|\tan\frac{ax}{2}\right|-\frac{c}{b}\int\frac{dx}{b+c\sin ax}$ (see 274).

277. $$\int \frac{dx}{(b + c \sin ax)^2} = \frac{c \cos ax}{a(b^2 - c^2)(b + c \sin ax)} + \\ + \frac{b}{b^2 - c^2} \int \frac{dx}{b + c \sin ax} \quad \text{(see 274)}.$$

278. $$\int \frac{\sin ax \, dx}{(b + c \sin ax)^2} = \frac{b \cos ax}{a(c^2 - b^2)(b + c \sin ax)} + \\ + \frac{c}{c^2 - b^2} \int \frac{dx}{b + c \sin ax} \quad \text{(see 274)}.$$

279. $$\int \frac{dx}{b^2 + c^2 \sin^2 ax} = \frac{1}{ab\sqrt{(b^2 + c^2)}} \arctan \frac{\sqrt{(b^2 + c^2)} \tan ax}{b} \quad (b \neq 0).$$

280. $$\int \frac{dx}{b^2 - c^2 \sin^2 ax} = \frac{1}{ab\sqrt{(b^2 - c^2)}} \arctan \frac{\sqrt{(b^2 - c^2)} \tan ax}{b} \\ (b^2 > c^2, \; b \neq 0),$$

$$= \frac{1}{2ab\sqrt{(c^2 - b^2)}} \ln \left| \frac{\sqrt{(c^2 - b^2)} \tan (ax) + b}{\sqrt{(c^2 - b^2)} \tan (ax) - b} \right| \quad (c^2 > b^2, \; b \neq 0).$$

(β) *Integrals Containing the Cosine Only.*

281. $$\int \cos ax \, dx = \frac{1}{a} \sin ax .$$

282. $$\int \cos^2 ax \, dx = \frac{1}{2} x + \frac{1}{4a} \sin 2ax .$$

283. $$\int \cos^3 ax \, dx = \frac{1}{a} \sin ax - \frac{1}{3a} \sin^3 ax .$$

284. $$\int \cos^4 ax \, dx = \frac{3}{8} x + \frac{1}{4a} \sin 2ax + \frac{1}{32a} \sin 4ax .$$

285. $$\int \cos^n ax \, dx = \frac{\cos^{n-1} ax \sin ax}{na} + \frac{n - 1}{n} \int \cos^{n-2} ax \, dx .$$

286. $$\int x \cos ax \, dx = \frac{\cos ax}{a^2} + \frac{x \sin ax}{a} .$$

287. $$\int x^2 \cos ax \, dx = \frac{2x}{a^2} \cos ax + \left(\frac{x^2}{a} - \frac{2}{a^3} \right) \sin ax .$$

288. $\int x^3 \cos ax \, dx = \left(\frac{3x^2}{a^2} - \frac{6}{a^4}\right) \cos ax + \left(\frac{x^3}{a} - \frac{6x}{a^3}\right) \sin ax .$

289. $\int x^n \cos ax \, dx = \frac{x^n \sin ax}{a} - \frac{n}{a} \int x^{n-1} \sin ax \, dx$ (see 250).

290. $\int \frac{\cos ax}{x} dx = \ln |ax| - \frac{(ax)^2}{2 \cdot 2!} + \frac{(ax)^4}{4 \cdot 4!} - \frac{(ax)^6}{6 \cdot 6!} + \ldots$

(the series is convergent for all x; see also §§ 13.12 and 15.7).

291. $\int \frac{\cos ax}{x^2} dx = - \frac{\cos ax}{x} - a \int \frac{\sin ax \, dx}{x}$ (see 251).

292. $\int \frac{\cos ax}{x^n} dx = - \frac{\cos ax}{(n-1) x^{n-1}} - \frac{a}{n-1} \int \frac{\sin ax \, dx}{x^{n-1}}$ $(n > 1$; see 253).

293. $\int \frac{dx}{\cos ax} = \frac{1}{a} \ln \left| \tan \left(\frac{ax}{2} + \frac{\pi}{4}\right) \right| .$

294. $\int \frac{dx}{\cos^2 ax} = \frac{1}{a} \tan ax .$

295. $\int \frac{dx}{\cos^3 ax} = \frac{\sin ax}{2a \cos^2 ax} + \frac{1}{2a} \ln \left| \tan \left(\frac{\pi}{4} + \frac{ax}{2}\right) \right| .$

296. $\int \frac{dx}{\cos^n ax} = \frac{1}{a(n-1)} \frac{\sin ax}{\cos^{n-1} ax} + \frac{n-2}{n-1} \int \frac{dx}{\cos^{n-2} ax}$ $(n > 1)$.

297. $\int \frac{x \, dx}{\cos ax} = \frac{1}{a^2} \left(\frac{(ax)^2}{2} + \frac{(ax)^4}{4 \cdot 2!} + \frac{5(ax)^6}{6 \cdot 4!} + \frac{61(ax)^8}{8 \cdot 6!} + \frac{1{,}385(ax)^{10}}{10 \cdot 8!} + \ldots \right.$

$\left. \ldots + \frac{E_n \cdot (ax)^{2n+2}}{(2n+2)(2n!)} + \ldots \right)$ $\left(|x| < \frac{\pi}{2|a|}$; see Remark 4, p. 511 $\right)$.

298. $\int \frac{x \, dx}{\cos^2 ax} = \frac{x}{a} \tan ax + \frac{1}{a^2} \ln |\cos ax| .$

299. $\int \frac{x \, dx}{\cos^n ax} = \frac{x \sin ax}{(n-1) a \cos^{n-1} ax} - \frac{1}{(n-1)(n-2) a^2 \cos^{n-2} ax} +$

$+ \frac{n-2}{n-1} \int \frac{x \, dx}{\cos^{n-2} ax}$ $(n > 2)$.

300. $\int \frac{dx}{1 + \cos ax} = \frac{1}{a} \tan \frac{ax}{2} .$

301. $\int \frac{dx}{1 - \cos ax} = -\frac{1}{a} \cot \frac{ax}{2}$.

302. $\int \frac{x\,dx}{1 + \cos ax} = \frac{x}{a} \tan \frac{ax}{2} + \frac{2}{a^2} \ln \left| \cos \frac{ax}{2} \right|$.

303. $\int \frac{x\,dx}{1 - \cos ax} = -\frac{x}{a} \cot \frac{ax}{2} + \frac{2}{a^2} \ln \left| \sin \frac{ax}{2} \right|$.

304. $\int \frac{\cos ax\,dx}{1 + \cos ax} = x - \frac{1}{a} \tan \frac{ax}{2}$.

305. $\int \frac{\cos ax\,dx}{1 - \cos ax} = -x - \frac{1}{a} \cot \frac{ax}{2}$.

306. $\int \frac{dx}{\cos ax(1 + \cos ax)} = \frac{1}{a} \ln \left| \tan \left(\frac{\pi}{4} + \frac{ax}{2} \right) \right| - \frac{1}{a} \tan \frac{ax}{2}$.

307. $\int \frac{dx}{\cos ax(1 - \cos ax)} = \frac{1}{a} \ln \left| \tan \left(\frac{\pi}{4} + \frac{ax}{2} \right) \right| - \frac{1}{a} \cot \frac{ax}{2}$.

308. $\int \frac{dx}{(1 + \cos ax)^2} = \frac{1}{2a} \tan \frac{ax}{2} + \frac{1}{6a} \tan^3 \frac{ax}{2}$.

309. $\int \frac{dx}{(1 - \cos ax)^2} = -\frac{1}{2a} \cot \frac{ax}{2} - \frac{1}{6a} \cot^3 \frac{ax}{2}$.

310. $\int \frac{\cos ax\,dx}{(1 + \cos ax)^2} = \frac{1}{2a} \tan \frac{ax}{2} - \frac{1}{6a} \tan^3 \frac{ax}{2}$.

311. $\int \frac{\cos ax\,dx}{(1 - \cos ax)^2} = \frac{1}{2a} \cot \frac{ax}{2} - \frac{1}{6a} \cot^3 \frac{ax}{2}$.

312. $\int \frac{dx}{1 + \cos^2 ax} = \frac{1}{2\sqrt{(2)}\,a} \arcsin \left(\frac{1 - 3\cos^2 ax}{1 + \cos^2 ax} \right)$ $(\sin ax \cos ax > 0)$.

313. $\int \frac{dx}{1 - \cos^2 ax} = \int \frac{dx}{\sin^2 ax} = -\frac{1}{a} \cot ax$.

314. $\int \cos ax \cos bx\,dx = \frac{\sin (a - b)x}{2(a - b)} + \frac{\sin (a + b)x}{2(a + b)}$

$(|a| \neq |b|$; for $|a| = |b|$ see 282).

315. $$\int \frac{dx}{b + c\cos ax} = \frac{2}{a\sqrt{(b^2 - c^2)}} \arctan \frac{(b - c)\tan \frac{1}{2}ax}{\sqrt{(b^2 - c^2)}} \quad (\text{for } b^2 > c^2),$$

$$= \frac{1}{a\sqrt{(c^2 - b^2)}} \ln \left| \frac{(c - b)\tan \frac{1}{2}ax + \sqrt{(c^2 - b^2)}}{(c - b)\tan \frac{1}{2}ax - \sqrt{(c^2 - b^2)}} \right| \quad (\text{for } b^2 < c^2).$$

316. $$\int \frac{\cos ax\, dx}{b + c\cos ax} = \frac{x}{c} - \frac{b}{c}\int \frac{dx}{b + c\cos ax} \quad (\text{see } 315).$$

317. $$\int \frac{dx}{\cos ax(b + c\cos ax)} = \frac{1}{ab} \ln \left| \tan\left(\frac{ax}{2} + \frac{\pi}{4}\right) \right| - \frac{b}{c}\int \frac{dx}{b + c\cos ax} \quad (\text{see } 315).$$

318. $$\int \frac{dx}{(b + c\cos ax)^2} = \frac{c\sin ax}{a(c^2 - b^2)(b + c\cos ax)} - \frac{b}{c^2 - b^2}\int \frac{dx}{b + c\cos ax}$$

(see 315) $(b^2 \neq c^2)$.

319. $$\int \frac{\cos ax\, dx}{(b + c\cos ax)^2} = \frac{b\sin ax}{a(b^2 - c^2)(b + c\cos ax)} - \frac{c}{b^2 - c^2}\int \frac{dx}{b + c\cos ax}$$

(see 315) $(b^2 \neq c^2)$.

320. $$\int \frac{dx}{b^2 + c^2\cos^2 ax} = \frac{1}{ab\sqrt{(b^2 + c^2)}} \arctan \frac{b\tan ax}{\sqrt{(b^2 + c^2)}} \quad (b > 0).$$

321. $$\int \frac{dx}{b^2 - c^2\cos^2 ax} = \frac{1}{ab\sqrt{(b^2 - c^2)}} \arctan \frac{b\tan ax}{\sqrt{(b^2 - c^2)}} \quad (b^2 > c^2 > 0),$$

$$= \frac{1}{2ab\sqrt{(c^2 - b^2)}} \ln \left| \frac{b\tan ax - \sqrt{(c^2 - b^2)}}{b\tan ax + \sqrt{(c^2 - b^2)}} \right| \quad (c^2 > b^2 > 0).$$

(γ) *Integrals Containing both Sine and Cosine.*

322. $$\int \sin ax\cos ax\, dx = \frac{1}{2a}\sin^2 ax.$$

323. $$\int \sin^2 ax\cos^2 ax\, dx = \frac{x}{8} - \frac{\sin 4ax}{32a}.$$

324. $$\int \sin^r ax\cos ax\, dx = \frac{1}{a(r + 1)}\sin^{r+1} ax \quad (r \neq -1).$$

325. $$\int \sin ax\cos^r ax\, dx = -\frac{1}{a(r + 1)}\cos^{r+1} ax \quad (r \neq -1).$$

326. $$\int \sin^n ax \cos^m ax \, dx = -\frac{\sin^{n-1} ax \cos^{m+1} ax}{a(n+m)} + \frac{n-1}{n+m}\int \sin^{n-2} ax \cos^m ax \, dx =$$
$$= \frac{\sin^{n+1} ax \cos^{m-1} ax}{a(n+m)} + \frac{m-1}{n+m}\int \sin^n ax \cos^{m-2} ax \, dx .$$

327. $$\int \frac{dx}{\sin ax \cos ax} = \frac{1}{a} \ln |\tan ax| .$$

328. $$\int \frac{dx}{\sin^2 ax \cos ax} = \frac{1}{a}\left[\ln \left| \tan\left(\frac{\pi}{4} + \frac{ax}{2}\right)\right| - \frac{1}{\sin ax}\right].$$

329. $$\int \frac{dx}{\sin ax \cos^2 ax} = \frac{1}{a}\left(\ln \left| \tan \frac{ax}{2}\right| + \frac{1}{\cos ax}\right).$$

330. $$\int \frac{dx}{\sin^3 ax \cos ax} = \frac{1}{a}\left(\ln |\tan ax| - \frac{1}{2\sin^2 ax}\right).$$

331. $$\int \frac{dx}{\sin ax \cos^3 ax} = \frac{1}{a}\left(\ln |\tan ax| + \frac{1}{2\cos^2 ax}\right).$$

332. $$\int \frac{dx}{\sin^2 ax \cos^2 ax} = -\frac{2}{a} \cot 2ax .$$

333. $$\int \frac{dx}{\sin ax \cos^n ax} = \frac{1}{a(n-1)\cos^{n-1} ax} + \int \frac{dx}{\sin ax \cos^{n-2} ax}$$

(see 327, 329, 331; $n > 1$).

334. $$\int \frac{dx}{\sin^n ax \cos ax} = -\frac{1}{a(n-1)\sin^{n-1} ax} + \int \frac{dx}{\sin^{n-2} ax \cos ax}$$

(see 327, 328, 330; $n > 1$).

335. $$\int \frac{dx}{\sin^n ax \cos^m ax} =$$

$$= -\frac{1}{a(n-1)} \cdot \frac{1}{\sin^{n-1} ax \cos^{m-1} ax} + \frac{n+m-2}{n-1}\int \frac{dx}{\sin^{n-2} ax \cos^m ax}$$

$(n > 1)$,

$$= \frac{1}{a(m-1)} \cdot \frac{1}{\sin^{n-1} ax \cos^{m-1} ax} + \frac{n+m-2}{m-1}\int \frac{dx}{\sin^n ax \cos^{m-2} ax}$$

$(m > 1)$.

336. $\int \frac{\sin ax\,dx}{\cos^2 ax} = \frac{1}{a \cos ax}$.

337. $\int \frac{\sin ax\,dx}{\cos^3 ax} = \frac{1}{2a \cos^2 ax} + C = \frac{1}{2a} \tan^2 ax + C_1$.

338. $\int \frac{\sin ax\,dx}{\cos^n ax} = \frac{1}{a(n-1)\cos^{n-1} ax}$ $(n > 1$; for $n = 1$ see 364$)$.

339. $\int \frac{\sin^2 ax\,dx}{\cos ax} = -\frac{1}{a} \sin ax + \frac{1}{a} \ln \left| \tan \left(\frac{\pi}{4} + \frac{ax}{2} \right) \right|$.

340. $\int \frac{\sin^2 ax\,dx}{\cos^3 ax} = \frac{1}{a} \left[\frac{\sin ax}{2\cos^2 ax} - \frac{1}{2} \ln \left| \tan \left(\frac{\pi}{4} + \frac{ax}{2} \right) \right| \right]$.

341. $\int \frac{\sin^2 ax\,dx}{\cos^n ax} = \frac{\sin ax}{a(n-1)\cos^{n-1} ax} - \frac{1}{n-1} \int \frac{dx}{\cos^{n-2} ax}$

(see 293–296; $n > 1$).

342. $\int \frac{\sin^n ax}{\cos ax} = -\frac{\sin^{n-1} ax}{a(n-1)} + \int \frac{\sin^{n-2} ax\,dx}{\cos ax}$ $(n > 1$; for $n = 1$ see 364$)$.

343. $\int \frac{\sin^n ax}{\cos^m ax} dx = \frac{\sin^{n+1} ax}{a(m-1)\cos^{m-1} ax} - \frac{n-m+2}{m-1} \int \frac{\sin^n ax}{\cos^{m-2} ax} dx$ $(m > 1)$,

$$= -\frac{\sin^{n-1} ax}{a(n-m)\cos^{m-1} ax} + \frac{n-1}{n-m} \int \frac{\sin^{n-2} ax\,dx}{\cos^m ax}$$

$(m \neq n$; for $m = n$ see 367$)$,

$$= \frac{\sin^{n-1} ax}{a(m-1)\cos^{m-1} ax} - \frac{n-1}{m-1} \int \frac{\sin^{n-2} ax\,dx}{\cos^{m-2} ax}$$ $(m > 1)$.

344. $\int \frac{\cos ax\,dx}{\sin^2 ax} = -\frac{1}{a \sin ax}$.

345. $\int \frac{\cos ax\,dx}{\sin^3 ax} = -\frac{1}{2a \sin^2 ax} + C = -\frac{\cot^2 ax}{2a} + C_1$.

346. $\int \frac{\cos ax\,dx}{\sin^n ax} = -\frac{1}{a(n-1)\sin^{n-1} ax}$ $(n > 1$; for $n = 1$ see 373$)$.

347. $\int \frac{\cos^2 ax\,dx}{\sin ax} = \frac{1}{a} \left(\cos ax + \ln \left| \tan \frac{ax}{2} \right| \right)$.

348. $$\int \frac{\cos^2 ax\,dx}{\sin^3 ax} = -\frac{1}{2a}\left(\frac{\cos ax}{\sin^2 ax} - \ln\left|\tan\frac{ax}{2}\right|\right).$$

349. $$\int \frac{\cos^2 ax\,dx}{\sin^n ax} = -\frac{1}{(n-1)}\left(\frac{\cos ax}{a\sin^{n-1} ax} + \int \frac{dx}{\sin^{n-2} ax}\right)$$

(see 254–257; $n > 1$).

350. $$\int \frac{\cos^n ax}{\sin ax}\,dx = \frac{\cos^{n-1} ax}{a(n-1)} + \int \frac{\cos^{n-2} ax\,dx}{\sin ax} \quad (n > 1;\ \text{for } n = 1 \text{ see } 373).$$

351. $$\int \frac{\cos^n ax\,dx}{\sin^m ax} =$$

$$= -\frac{\cos^{n+1} ax}{a(m-1)\sin^{m-1} ax} - \frac{n-m+2}{m-1}\int \frac{\cos^n ax\,dx}{\sin^{m-2} ax} \quad (m > 1),$$

$$= \frac{\cos^{n-1} ax}{a(n-m)\sin^{m-1} ax} + \frac{n-1}{n-m}\int \frac{\cos^{n-2} ax\,dx}{\sin^m ax}$$

$(m \neq n;\ \text{for } m = n \text{ see } 376)$,

$$= -\frac{\cos^{n-1} ax}{a(m-1)\sin^{m-1} ax} - \frac{n-1}{m-1}\int \frac{\cos^{n-2} ax\,dx}{\sin^{m-2} ax} \quad (m > 1).$$

352. $$\int \frac{dx}{\sin ax(1 \pm \cos ax)} = \pm\frac{1}{2a(1 \pm \cos ax)} + \frac{1}{2a}\ln\left|\tan\frac{ax}{2}\right|.$$

353. $$\int \frac{dx}{\cos ax(1 \pm \sin ax)} = \mp\frac{1}{2a(1 \pm \sin ax)} + \frac{1}{2a}\ln\left|\tan\left(\frac{\pi}{4} + \frac{ax}{2}\right)\right|.$$

354. $$\int \frac{\sin ax\,dx}{\cos ax(1 \pm \cos ax)} = \frac{1}{a}\ln\left|\frac{1 \pm \cos ax}{\cos ax}\right|.$$

355. $$\int \frac{\cos ax\,dx}{\sin ax(1 \pm \sin ax)} = -\frac{1}{a}\ln\left|\frac{1 \pm \cos ax}{\sin ax}\right|.$$

356. $$\int \frac{\sin ax\,dx}{\cos ax(1 \pm \sin ax)} = \frac{1}{2a(1 \pm \sin ax)} \pm \frac{1}{2a}\ln\left|\tan\left(\frac{\pi}{4} + \frac{ax}{2}\right)\right|.$$

357. $$\int \frac{dx}{b\sin ax + c\cos ax} = \frac{1}{a\sqrt{(b^2 + c^2)}}\ln\left|\tan\frac{ax + \vartheta}{2}\right|,$$

where $\sin\vartheta = \dfrac{c}{\sqrt{(b^2 + c^2)}}$, $\cos\vartheta = \dfrac{b}{\sqrt{(b^2 + c^2)}}$.

358. $$\int \frac{\sin ax\,dx}{b + c\cos ax} = -\frac{1}{ac}\ln|b + c\cos ax|.$$

359. $$\int \frac{\cos ax\,dx}{b + c\sin ax} = \frac{1}{ac}\ln|b + c\sin ax|.$$

360. $$\int \frac{dx}{b + c\cos ax + f\sin ax} = \int \frac{d(x + \vartheta/a)}{b + \sqrt{(c^2 + f^2)}\sin(ax + \vartheta)},$$

where $\sin\vartheta = \dfrac{c}{\sqrt{(c^2 + f^2)}}$, $\cos\vartheta = \dfrac{f}{\sqrt{(c^2 + f^2)}}$ (see 274).

361. $$\int \frac{dx}{b^2\cos^2 ax + c^2\sin^2 ax} = \frac{1}{abc}\arctan\left(\frac{c}{b}\tan ax\right).$$

362. $$\int \frac{dx}{b^2\cos^2 ax - c^2\sin^2 ax} = \frac{1}{2abc}\ln\left|\frac{c\tan(ax) + b}{c\tan(ax) - b}\right|.$$

363. $$\int \sin ax\cos bx\,dx = -\frac{\cos(a + b)x}{2(a + b)} - \frac{\cos(a - b)x}{2(a - b)}$$

$(a^2 \neq b^2$, for $a = b$ see 322).

(δ) *Integrals Containing the Tangent and Cotangent.*

364. $$\int \tan ax\,dx = -\frac{1}{a}\ln|\cos ax|.$$

365. $$\int \tan^2 ax\,dx = \frac{\tan ax}{a} - x.$$

366. $$\int \tan^3 ax\,dx = \frac{1}{2a}\tan^2 ax + \frac{1}{a}\ln|\cos ax|.$$

367. $$\int \tan^n ax\,dx = \frac{1}{a(n-1)}\tan^{n-1} ax - \int \tan^{n-2} ax\,dx.$$

368. $$\int x\tan ax\,dx =$$

$$= \frac{ax^3}{3} + \frac{a^3x^5}{15} + \frac{2a^5x^7}{105} + \frac{17a^7x^9}{2{,}835} + \ldots + \frac{2^{2n}(2^{2n} - 1)\,B_n a^{2n-1}x^{2n+1}}{(2n+1)!} + \ldots$$

$\left(|x| < \dfrac{\pi}{2|a|}\right.$; see Remark 3, p. 511$\left.\right)$.

369. $$\int \frac{\tan ax\, dx}{x} =$$

$$= ax + \frac{(ax)^3}{9} + \frac{2(ax)^5}{75} + \frac{17(ax)^7}{2{,}205} + \ldots + \frac{2^{2n}(2^{2n}-1)\,B_n\,.\,(ax)^{2n-1}}{(2n-1)\,(2n)!} + \ldots$$

$$\left(|x| < \frac{\pi}{2|a|}\ ;\ \text{see Remark 3, p. 511}\right).$$

370. $$\int \frac{\tan^n ax}{\cos^2 ax}\, dx = \frac{1}{a(n+1)} \tan^{n+1} ax\,.$$

371. $$\int \frac{dx}{\tan ax \pm 1} = \pm \frac{x}{2} + \frac{1}{2a} \ln |\sin ax \pm \cos ax|\,.$$

372. $$\int \frac{\tan ax\, dx}{\tan (ax) \pm 1} = \frac{x}{2} \mp \frac{1}{2a} \ln |\sin ax \pm \cos ax|\,.$$

373. $$\int \cot ax\, dx = \frac{1}{a} \ln |\sin ax|\,.$$

374. $$\int \cot^2 ax\, dx = -\frac{\cot ax}{a} - x\,.$$

375. $$\int \cot^3 ax\, dx = -\frac{1}{2a} \cot^2 ax - \frac{1}{a} \ln |\sin ax|\,.$$

376. $$\int \cot^n ax\, dx = -\frac{1}{a(n-1)} \cot^{n-1} ax - \int \cot^{n-2} ax\, dx \quad (n \neq 1)\,.$$

377. $$\int x \cot ax\, dx = \frac{x}{a} - \frac{ax^3}{9} - \frac{a^3x^5}{225} - \ldots - \frac{2^{2n}B_n a^{2n-1}x^{2n+1}}{(2n+1)!} - \ldots$$

$$(|x| < \pi/|a|;\ \text{see Remark 3, p. 511}).$$

378. $$\int \frac{\cot ax\, dx}{x} = -\frac{1}{ax} - \frac{ax}{3} - \frac{(ax)^3}{135} - \frac{2(ax)^5}{4{,}725} - \ldots - \frac{2^{2n}B_n\,.\,(ax)^{2n-1}}{(2n-1)\,(2n)!} - \ldots$$

$$(|x| < \pi/|a|;\ x \neq 0;\ \text{see Remark 3, p. 511}).$$

379. $$\int \frac{\cot^n ax}{\sin^2 ax}\, dx = -\frac{1}{a(n+1)} \cot^{n+1} ax\,.$$

380. $$\int \frac{dx}{1 \pm \cot ax} = \int \frac{\tan ax\, dx}{\tan (ax) \pm 1} \quad (\text{see } 372)\,.$$

381. $$\int \frac{\tan^r ax}{\cos^2 ax} = \frac{1}{a(r+1)} \tan^{r+1} ax \quad (r \neq -1).$$

382. $$\int \frac{\cot^r ax}{\sin^2 ax} = -\frac{1}{a(r+1)} \cot^{r+1} ax \quad (r \neq -1).$$

(d) ***Other Transcendental Functions.***

$a \neq 0$ is assumed

(α) *Hyperbolic Functions.*

383. $$\int \sinh ax \, dx = \frac{1}{a} \cosh ax .$$

384. $$\int \cosh ax \, dx = \frac{1}{a} \sinh ax .$$

385. $$\int \sinh^2 ax \, dx = \frac{1}{2a} \sinh ax \cosh ax - \tfrac{1}{2}x .$$

386. $$\int \cosh^2 ax \, dx = \frac{1}{2a} \sinh ax \cosh ax + \tfrac{1}{2}x .$$

387. $$\int \sinh^n ax \, dx = \frac{1}{an} \sinh^{n-1} ax \cosh ax - \frac{n-1}{n} \int \sinh^{n-2} ax \, dx .$$

388. $$\int \frac{dx}{\sinh^n ax} = \frac{\cosh ax}{a(1-n) \sinh^{n-1} ax} - \frac{2-n}{1-n} \int \frac{dx}{\sinh^{n-2} ax} \quad (n \neq 1).$$

389. $$\int \cosh^n ax \, dx = \frac{1}{an} \sinh ax \cosh^{n-1} ax + \frac{n-1}{n} \int \cosh^{n-2} ax \, dx .$$

390. $$\int \frac{dx}{\cosh^n ax} = -\frac{\sinh ax}{a(1-n) \cosh^{n-1} ax} + \frac{2-n}{1-n} \int \frac{dx}{\cosh^{n-2} ax} \quad (n \neq 1).$$

391. $$\int \frac{dx}{\sinh ax} = \frac{1}{a} \ln \left| \tanh \frac{ax}{2} \right| .$$

392. $$\int \frac{dx}{\cosh ax} = \frac{2}{a} \arctan e^{ax} .$$

393. $\int x \sinh ax \, dx = \frac{1}{a} x \cosh ax - \frac{1}{a^2} \sinh ax .$

394. $\int x \cosh ax \, dx = \frac{1}{a} x \sinh ax - \frac{1}{a^2} \cosh ax .$

395. $\int \tanh ax \, dx = \frac{1}{a} \ln \cosh ax .$

396. $\int \coth ax \, dx = \frac{1}{a} \ln |\sinh ax| .$

397. $\int \tanh^2 ax \, dx = x - \frac{\tanh ax}{a} .$

398. $\int \coth^2 ax \, dx = x - \frac{\coth ax}{a} .$

399. $\int \sinh ax \sinh bx \, dx = \frac{1}{a^2 - b^2} (a \cosh ax \sinh bx - b \sinh ax \cosh bx)$

400. $\int \cosh ax \cosh bx \, dx = \frac{1}{a^2 - b^2} (a \sinh ax \cosh bx - b \cosh ax \sinh bx)$ $(a^2 \neq b^2).$

401. $\int \cosh ax \sinh bx \, dx = \frac{1}{a^2 - b^2} (a \sinh ax \sinh bx - b \cosh ax \cosh bx)$

$$\int \cosh ax \sinh ax \, dx = \frac{\cosh^2 ax}{2a} + C = \frac{\sinh^2 ax}{2a} + C_1 .$$

402. $\int \sinh ax \sin ax \, dx = \frac{1}{2a} (\cosh ax \sin ax - \sinh ax \cos ax) .$

403. $\int \cosh ax \cos ax \, dx = \frac{1}{2a} (\sinh ax \cos ax + \cosh ax \sin ax) .$

404. $\int \sinh ax \cos ax \, dx = \frac{1}{2a} (\cosh ax \cos ax + \sinh ax \sin ax) .$

405. $\int \cosh ax \sin ax \, dx = \frac{1}{2a} (\sinh ax \sin ax - \cosh ax \cos ax) .$

(β) *Exponential Functions.*

406. $\int e^{ax}\,dx = \frac{1}{a}e^{ax}$.

407. $\int xe^{ax}\,dx = \frac{e^{ax}}{a^2}(ax - 1)$.

408. $\int x^2e^{ax}\,dx = e^{ax}\left(\frac{x^2}{a} - \frac{2x}{a^2} + \frac{2}{a^3}\right)$.

409. $\int x^ne^{ax}\,dx = \frac{1}{a}x^ne^{ax} - \frac{n}{a}\int x^{n-1}e^{ax}\,dx$.

410. $\int \frac{e^{ax}}{x}\,dx = \ln|x| + \frac{ax}{1\,.\,1!} + \frac{(ax)^2}{2\,.\,2!} + \frac{(ax)^3}{3\,.\,3!} + \ldots$ $(x \neq 0$; see also § 15.7).

411. $\int \frac{e^{ax}}{x^n}\,dx = \frac{1}{n-1}\left(-\frac{e^{ax}}{x^{n-1}} + a\int \frac{e^{ax}}{x^{n-1}}\,dx\right)$ $(n > 1)$.

412. $\int \frac{dx}{b + ce^{ax}} = \frac{x}{b} - \frac{1}{ab}\ln|b + ce^{ax}|$ $(b \neq 0)$.

413. $\int \frac{e^{ax}\,dx}{b + ce^{ax}} = \frac{1}{ac}\ln|b + ce^{ax}|$ $(c \neq 0)$.

414. $\int \frac{dx}{be^{ax} + ce^{-ax}} = \frac{1}{a\sqrt{(bc)}}\arctan\left(e^{ax}\sqrt{\frac{b}{c}}\right)$ $(b > 0,\ c > 0)$,

$$= \frac{1}{2a\sqrt{(-bc)}}\ln\left|\frac{c + e^{ax}\sqrt{(-bc)}}{c - e^{ax}\sqrt{(-bc)}}\right| \quad (bc < 0).$$

415. $\int \frac{xe^{ax}\,dx}{(1 + ax)^2} = \frac{e^{ax}}{a^2(1 + ax)}$.

416. $\int e^{ax}\ln x\,dx = \frac{e^{ax}\ln x}{a} - \frac{1}{a}\int \frac{e^{ax}}{x}\,dx$ $(x > 0$; see 410).

417. $\int e^{ax}\sin bx\,dx = \frac{e^{ax}}{a^2 + b^2}(a\sin bx - b\cos bx)$.

418. $\int e^{ax}\cos bx\,dx = \frac{e^{ax}}{a^2 + b^2}(a\cos bx + b\sin bx)$.

419. $$\int e^{ax} \sin^n x \, dx = \frac{e^{ax} \sin^{n-1} x}{a^2 + n^2} (a \sin x - n \cos x) + \frac{n(n-1)}{a^2 + n^2} \int e^{ax} \sin^{n-2} x \, dx \quad \text{(see 406 and 417)}.$$

420. $$\int e^{ax} \cos^n x \, dx = \frac{e^{ax} \cos^{n-1} x}{a^2 + n^2} (a \cos x + n \sin x) + \frac{n(n-1)}{a^2 + n^2} \int e^{ax} \cos^{n-2} x \, dx \quad \text{(see 406 and 418)}.$$

421. $$\int x e^{ax} \sin bx \, dx = \frac{x e^{ax}}{a^2 + b^2} (a \sin bx - b \cos bx) - \frac{e^{ax}}{(a^2 + b^2)^2} [(a^2 - b^2) \sin bx - 2ab \cos bx].$$

422. $$\int x e^{ax} \cos bx \, dx = \frac{x e^{ax}}{a^2 + b^2} (a \cos bx + b \sin bx) - \frac{e^{ax}}{(a^2 + b^2)^2} [(a^2 - b^2) \cos bx + 2ab \sin bx].$$

423. $$\int b^{ax} \, dx = \frac{b^{ax}}{a \ln b} \quad (b > 0, \, b \neq 1).$$

424. $$\int x b^{ax} \, dx = \frac{x b^{ax}}{a \ln b} - \frac{b^{ax}}{a^2 (\ln b)^2} \quad (b > 0, \, b \neq 1).$$

(γ) *Logarithmic Functions.*

$x > 0$ is assumed

425. $$\int \ln x \, dx = x \ln x - x.$$

426. $$\int (\ln x)^2 \, dx = x(\ln x)^2 - 2x \ln x + 2x.$$

427. $$\int (\ln x)^3 \, dx = x(\ln x)^3 - 3x(\ln x)^2 + 6x \ln x - 6x.$$

428. $\int (\ln x)^n \, dx = x(\ln x)^n - n \int (\ln x)^{n-1} \, dx$.

429. $\int \frac{dx}{\ln x} = \ln |\ln x| + \ln x + \frac{(\ln x)^2}{2 \cdot 2!} + \frac{(\ln x)^3}{3 \cdot 3!} + \ldots$

$(x > 0, \; x \neq 1$; see also § 13.12 and 15.7).

430. $\int \frac{dx}{(\ln x)^n} = -\frac{x}{(n-1)(\ln x)^{n-1}} + \frac{1}{n-1} \int \frac{dx}{(\ln x)^{n-1}}$ $(n > 1$; see 429).

431. $\int x^r \ln x \, dx = x^{r+1} \left[\frac{\ln x}{r+1} - \frac{1}{(r+1)^2} \right]$ $(r \neq -1)$.

432. $\int x^r (\ln x)^n \, dx = \frac{x^{r+1} (\ln x)^n}{r+1} - \frac{n}{r+1} \int x^r (\ln x)^{n-1} \, dx$ $(r \neq -1$; see 431).

433. $\int \frac{(\ln x)^n}{x} \, dx = \frac{(\ln x)^{n+1}}{n+1}$.

434. $\int \frac{x^r \, dx}{\ln x} = \int \frac{e^{-y}}{y} \, dy$, where $y = -(r+1) \ln x$ $(r \neq -1$; see 410, 436).

435. $\int \frac{x^r \, dx}{(\ln x)^n} = -\frac{x^{r+1}}{(n-1)(\ln x)^{n-1}} + \frac{r+1}{n-1} \int \frac{x^r \, dx}{(\ln x)^{n-1}}$ $(n > 1)$.

436. $\int \frac{dx}{x \ln x} = \ln |\ln x|$.

437. $\int \frac{dx}{x^n \ln x} = \ln |\ln x| - (n-1) \ln x + \frac{(n-1)^2 (\ln x)^2}{2 \cdot 2!} - \frac{(n-1)^3 (\ln x)^3}{3 \cdot 3!} + \ldots$ $(x > 0, \; x \neq 1)$.

438. $\int \frac{dx}{x(\ln x)^n} = \frac{-1}{(n-1)(\ln x)^{n-1}}$ $(n > 1)$.

439. $\int \frac{dx}{x^r (\ln x)^n} = \frac{-1}{x^{r-1}(n-1)(\ln x)^{n-1}} - \frac{r-1}{n-1} \int \frac{dx}{x^r (\ln x)^{n-1}}$ $(n > 1)$.

440. $\int \ln \sin x \, dx = x \ln x - x - \frac{x^3}{18} - \frac{x^5}{900} - \ldots - \frac{2^{2n-1} B_n x^{2n+1}}{n(2n+1)!} - \ldots$

$(0 < x < \pi$; see Remark 3, p. 511)

441. $$\int \ln \cos x \, dx = -\frac{x^3}{6} - \frac{x^5}{60} - \frac{x^7}{315} - \ldots - \frac{2^{2n-1}(2^{2n}-1)\,B_n x^{2n+1}}{n(2n+1)!} - \ldots$$

$(-\frac{1}{2}\pi < x < \frac{1}{2}\pi$; see Remark 3, p. 511).

442. $$\int \ln \tan x \, dx = x \ln x - x + \frac{x^3}{9} + \frac{7x^5}{450} + \ldots + \frac{2^{2n}(2^{2n-1}-1)\,B_n x^{2n+1}}{n(2n+1)!} + \ldots$$

$(0 < x < \frac{1}{2}\pi$; see Remark 3, p. 511).

443. $$\int \sin \ln x \, dx = \frac{x}{2}(\sin \ln x - \cos \ln x).$$

444. $$\int \cos \ln x \, dx = \frac{x}{2}(\sin \ln x + \cos \ln x).$$

445. $$\int e^{ax} \ln x \, dx = \frac{1}{a} e^{ax} \ln x - \frac{1}{a}\int \frac{e^{ax}}{x} dx \quad \text{(see 410)}.$$

(δ) *Inverse Trigonometric Functions.*

446. $$\int \arcsin \frac{x}{a} dx = x \arcsin \frac{x}{a} + \sqrt{(a^2 - x^2)} \quad (a > 0).$$

447. $$\int x \arcsin \frac{x}{a} dx = \left(\frac{x^2}{2} - \frac{a^2}{4}\right)\arcsin \frac{x}{a} + \frac{x}{4}\sqrt{(a^2 - x^2)} \quad (a > 0).$$

448. $$\int x^2 \arcsin \frac{x}{a} dx = \frac{x^3}{3}\arcsin \frac{x}{a} + \frac{1}{9}(x^2 + 2a^2)\sqrt{(a^2 - x^2)} \quad (a > 0).$$

449. $$\int \frac{\arcsin \frac{x}{a}\, dx}{x} = \frac{x}{a} + \frac{1}{2.3.3}\frac{x^3}{a^3} + \frac{1.3}{2.4.5.5}\frac{x^5}{a^5} + \\ + \frac{1.3.5}{2.4.6.7.7}\frac{x^7}{a^7} + \ldots \quad (|x| \leqq |a|).$$

450. $$\int \frac{\arcsin \frac{x}{a}\, dx}{x^2} = -\frac{1}{x}\arcsin \frac{x}{a} - \frac{1}{a}\ln \left|\frac{a + \sqrt{(a^2 - x^2)}}{x}\right| \quad (a > 0).$$

451. $$\int \arccos \frac{x}{a} dx = x \arccos \frac{x}{a} - \sqrt{(a^2 - x^2)} \quad (a > 0).$$

452. $\int x \arccos \frac{x}{a}\,dx = \left(\frac{x^2}{2} - \frac{a^2}{4}\right) \arccos \frac{x}{a} - \frac{x}{4}\sqrt{(a^2 - x^2)} \quad (a > 0)$.

453. $\int x^2 \arccos \frac{x}{a}\,dx = \frac{x^3}{3} \arccos \frac{x}{a} - \frac{1}{9}(x^2 + 2a^2)\sqrt{(a^2 - x^2)} \quad (a > 0)$.

454. $\int \frac{\arccos \frac{x}{a}\,dx}{x} = \frac{\pi}{2} \ln |x| - \frac{x}{a} - \frac{1}{2.3.3}\frac{x^3}{a^3} - \frac{1.3}{2.4.5.5}\frac{x^5}{a^5} - \frac{1.3.5}{2.4.6.7.7}\frac{x^7}{a^7} - \ldots \quad (|x| \leqq |a|,\ x \neq 0)$.

455. $\int \frac{\arccos \frac{x}{a}\,dx}{x^2} = -\frac{1}{x} \arccos \frac{x}{a} + \frac{1}{a} \ln \left| \frac{a + \sqrt{(a^2 - x^2)}}{x} \right| \quad (a > 0)$.

456. $\int \arctan \frac{x}{a}\,dx = x \arctan \frac{x}{a} - \frac{a}{2} \ln (a^2 + x^2)$.

457. $\int x \arctan \frac{x}{a}\,dx = \frac{1}{2}(x^2 + a^2) \arctan \frac{x}{a} - \frac{ax}{2}$.

458. $\int x^2 \arctan \frac{x}{a}\,dx = \frac{x^3}{3} \arctan \frac{x}{a} - \frac{ax^2}{6} + \frac{a^3}{6} \ln (a^2 + x^2)$.

459. $\int x^n \arctan \frac{x}{a}\,dx = \frac{x^{n+1}}{n+1} \arctan \frac{x}{a} - \frac{a}{n+1} \int \frac{x^{n+1}\,dx}{a^2 + x^2}$.

460. $\int \frac{\arctan \frac{x}{a}\,dx}{x} = \frac{x}{a} - \frac{x^3}{3^2 a^3} + \frac{x^5}{5^2 a^5} - \frac{x^7}{7^2 a^7} + \ldots \quad (|x| \leqq |a|)$.

461. $\int \frac{\arctan \frac{x}{a}\,dx}{x^2} = -\frac{1}{x} \arctan \frac{x}{a} - \frac{1}{2a} \ln \frac{a^2 + x^2}{x^2}$.

462. $\int \frac{\arctan \frac{x}{a}\,dx}{x^n} = -\frac{1}{(n-1)x^{n-1}} \arctan \frac{x}{a} + \frac{a}{n-1} \int \frac{dx}{x^{n-1}(a^2 + x^2)} \quad (n > 1)$.

463. $\int \operatorname{arccot} \frac{x}{a}\,dx = x \operatorname{arccot} \frac{x}{a} + \frac{a}{2} \ln (a^2 + x^2)$.

464. $\int x \operatorname{arccot} \frac{x}{a}\,dx = \frac{1}{2}(x^2 + a^2) \operatorname{arccot} \frac{x}{a} + \frac{ax}{2}$.

465. $\int x^2 \operatorname{arccot} \frac{x}{a}\,dx = \frac{x^3}{3} \operatorname{arccot} \frac{x}{a} + \frac{ax^2}{6} - \frac{a^3}{6} \ln (a^2 + x^2)$.

466. $\int x^n \operatorname{arccot} \frac{x}{a}\,dx = \frac{x^{n+1}}{n+1} \operatorname{arccot} \frac{x}{a} + \frac{a}{n+1} \int \frac{x^{n+1}\,dx}{a^2 + x^2}$.

467. $\int \frac{\operatorname{arccot} \frac{x}{a}\,dx}{x} = \frac{\pi}{2} \ln |x| - \frac{x}{a} + \frac{x^3}{3^2 a^3} - \frac{x^5}{5^2 a^5} + \frac{x^7}{7^2 a^7} - \ldots \quad (|x| \leqq |a|,\ x \neq 0)$.

468. $\int \frac{\operatorname{arccot} \frac{x}{a}\,dx}{x^2} = -\frac{1}{x} \operatorname{arctan} \frac{x}{a} + \frac{1}{2a} \ln \frac{a^2 + x^2}{x^2}$.

469. $\int \frac{\operatorname{arccot} \frac{x}{a}\,dx}{x^n} = -\frac{1}{(n-1)x^{n-1}} \operatorname{arccot} \frac{x}{a} - \frac{a}{n-1} \int \frac{dx}{x^{n-1}(a^2 + x^2)} \quad (n > 1)$.

(ε) *Inverse Hyperbolic Functions.*

470. $\int \operatorname{arsinh} \frac{x}{a}\,dx = x \operatorname{arsinh} \frac{x}{a} - \sqrt{(x^2 + a^2)} \quad (a > 0)$.

471. $\int \operatorname{arcosh} \frac{x}{a}\,dx = x \operatorname{arcosh} \frac{x}{a} - \sqrt{(x^2 - a^2)} \quad (x > a > 0)$.

472. $\int \operatorname{artanh} \frac{x}{a}\,dx = x \operatorname{artanh} \frac{x}{a} + \frac{a}{2} \ln (a^2 - x^2) \quad (|x| < |a|)$.

473. $\int \operatorname{arcoth} \frac{x}{a}\,dx = x \operatorname{arcoth} \frac{x}{a} + \frac{a}{2} \ln (x^2 - a^2) \quad (|x| > |a|)$.

REMARK 1. Some simple reductions of rational functions to partial fractions:

$$\frac{1}{(a + bx)(f + gx)} = \frac{1}{bf - ag} \left(\frac{b}{a + bx} - \frac{g}{f + gx} \right);$$

$$\frac{1}{(x + a)(x + b)(x + c)} = \frac{A}{x + a} + \frac{B}{x + b} + \frac{C}{x + c},$$

where

$$A = \frac{1}{(b-a)(c-a)}, \quad B = \frac{1}{(a-b)(c-b)}, \quad C = \frac{1}{(a-c)(b-c)};$$

$$\frac{1}{(x+a)(x+b)(x+c)(x+d)} = \frac{A}{x+a} + \frac{B}{x+b} + \frac{C}{x+c} + \frac{D}{x+d},$$

where

$$A = \frac{1}{(b-a)(c-a)(d-a)}, \quad B = \frac{1}{(a-b)(c-b)(d-b)},$$

$$C = \frac{1}{(a-c)(b-c)(d-c)}, \quad D = \frac{1}{(a-d)(b-d)(c-d)};$$

$$\frac{1}{(a+bx^2)(f+gx^2)} = \frac{1}{bf-ag}\left(\frac{b}{a+bx^2} - \frac{g}{f+gx^2}\right).$$

REMARK 2. On integrals of the type

$$\int x^m(a+bx^n)^p \, dx$$

(binomial integrals) see §13.4. See also p. 490.

REMARK 3. The Bernoulli coefficients B_n:

B_1	$\frac{1}{6}$	B_4	$\frac{1}{30}$	B_7	$\frac{7}{6}$	B_{10}	$\frac{174,611}{330}$
B_2	$\frac{1}{30}$	B_5	$\frac{5}{66}$	B_8	$\frac{3,617}{510}$	B_{11}	$\frac{854,513}{138}$
B_3	$\frac{1}{42}$	B_6	$\frac{691}{2,730}$	B_9	$\frac{43,867}{798}$		

REMARK 4. The Euler coefficients E_n:

E_1	1	E_3	61	E_5	50,521	E_7	199,360,981
E_2	5	E_4	1,385	E_6	2,702,765		

13.6. Definite Integrals. Cauchy–Riemann Definition. Basic Properties. Mean Value Theorems. Evaluation of a Definite Integral

Suppose we are given a function $y = f(x)$ which is continuous in the interval $[a, b]$. Let us divide the interval $[a, b]$ at the points $x_1, x_2, \ldots, x_{n-1}$ into n (closed) subintervals $\Delta x_1, \Delta x_2, \ldots, \Delta x_n$ (Fig. 13.1) which need not be equally long. Since $f(x)$ is

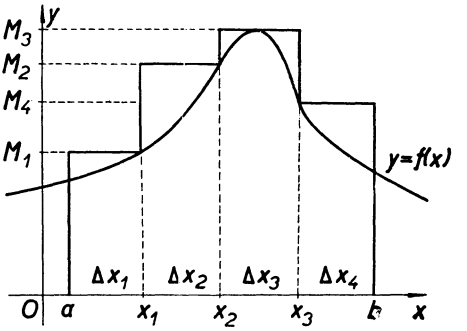


Fig. 13.1.

continuous in $[a, b]$, it assumes its maximum value M and minimum value m in $[a, b]$, and it takes on also its maximum value M_i and minimum value m_i ($M_i \leqq M$, $m_i \geqq m$) in each of the subintervals Δx_i. Let us denote by d the chosen partition of the interval $[a, b]$ and write

$$S(d) = \sum_{i=1}^{n} M_i \, \Delta x_i \,, \quad s(d) = \sum_{i=1}^{n} m_i \, \Delta x_i \,. \tag{1}$$

The numbers $S(d)$ and $s(d)$ are called the *upper Darboux sum* and the *lower Darboux sum* (corresponding to the function $f(x)$ and to the chosen partition), respectively. The geometric meaning of the upper Darboux sum is apparent from Fig. 13.1. The geometric meaning of the lower Darboux sum is similar. If another partition d of $[a, b]$ is chosen, then, generally speaking, other Darboux sums $S(d)$ and $s(d)$ correspond to it. The greatest lower bound (Definition 1.3.3, p. 43) of the values of all upper Darboux sums is called the *upper integral* of the function $f(x)$ in the interval $[a, b]$,

$$\inf_d S(d) = \overline{\int_a^b} f(x) \, \mathrm{d}x \,; \tag{2}$$

similarly the least upper bound of the values of all lower Darboux sums is called the *lower integral* of $f(x)$ in $[a, b]$,

$$\sup_d s(d) = \underline{\int_a^b} f(x) \, \mathrm{d}x \,. \tag{3}$$

REMARK 1. We need not assume the continuity of $f(x)$ in order to be able to define the upper and lower integral; it suffices to assume that $f(x)$ is *bounded* in

$[a, b]$. Instead of maxima and minima of $f(x)$, least upper bounds and greatest lower bounds in the corresponding intervals are then to be considered.

Definition 1. If

$$\overline{\int_a^b} f(x)\,dx = \underline{\int_a^b} f(x)\,dx,$$

then the common value of these integrals is called the (*definite*) *integral of the function* $f(x)$ *over the interval* $[a, b]$ and the function $f(x)$ is said to be *integrable in* $[a, b]$ *according to the Cauchy–Riemann definition.* We write

$$\int_a^b f(x)\,dx.$$

Theorem 1. *Any function piecewise continuous in* $[a, b]$ (Definition 11.3.6) *is integrable in* $[a, b]$. *In particular, any function continuous in* $[a, b]$ *is integrable in* $[a, b]$. (In § 13.14 an example of a function is given which is not integrable according to the Cauchy–Riemann definition.)

REMARK 2. The Lebesgue and Stieltjes definitions of an integral are briefly mentioned in §§13.14 and 13.15.

REMARK 3. The following sums are often considered instead of Darboux sums (cf. Theorem 2): Let us choose, for a fixed partition d in the interval $[a, b]$, an arbitrary point c_k in each interval Δx_k and let us write

$$\sigma(d) = \sum_{k=1}^{n} f(c_k)\,\Delta x_k \tag{4}$$

(this sum depends on how we have chosen d and the points c_k in Δx_k). Obviously

$$s(d) \leqq \sigma(d) \leqq S(d).$$

The greatest of the lengths of the intervals Δx_i is called the *norm* $n(d)$ *of the partition* d.

Theorem 2. *Let us consider a sequence of partitions* $d_1, d_2, d_3, \ldots$ *such that* $\lim_{k\to\infty} n(d_k) = 0$. *If* $f(x)$ *is integrable in* $[a, b]$, *then*

$$\int_a^b f(x)\,dx = \lim_{k\to\infty} S(d_k) = \lim_{k\to\infty} s(d_k) = \lim_{k\to\infty} \sigma(d_k).$$

REMARK 4. Briefly speaking: The integral of the function $f(x)$ in $[a, b]$ is the limit of upper Darboux sums provided that the norms of partitions tend to zero. Similar statements are valid concerning lower Darboux sums and sums (4).

Theorem 3. *If* $f_1(x)$ *and* $f_2(x)$ *are integrable in* $[a, b]$, *then the same is true for the functions* $k_1 f_1(x) + k_2 f_2(x)$, $f_1(x) f_2(x)$, $|f_1(x)|$ (*and of course for* $|f_2(x)|$), *and*

the relations

$$\int_a^b [k_1 f_1(x) + k_2 f_2(x)]\,dx = k_1 \int_a^b f_1(x)\,dx + k_2 \int_a^b f_2(x)\,dx\,,$$

$$\left| \int_a^b f_1(x)\,dx \right| \leqq \int_a^b |f_1(x)|\,dx$$

hold. (The equation

$$\int_a^b f(x)\,g(x)\,dx = \int_a^b f(x)\,dx \int_a^b g(x)\,dx$$

is not, in general, valid! For the function

$$\frac{f_1(x)}{f_2(x)}$$

to be integrable it is sufficient that f_1 and f_2 be integrable and

$$0 < m \leqq f_2(x) \quad \text{or} \quad f_2(x) \leqq M < 0$$

in $[a, b]$, i.e. $f_2(x)$ is bounded below by a positive constant or bounded above by a negative constant.)

Theorem 4. *If $f(x)$ is integrable in $[a, b]$ and if $a < c < b$, then $f(x)$ is integrable in both $[a, c]$ and $[c, b]$ and*

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$

(and conversely).

Definition 2. For $b < a$ the integral

$$\int_a^b f(x)\,dx$$

is defined by the equation

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx\,.$$

REMARK 5. *In the sequel* (Theorems 5, 7, 8, 10, 11, 12) *the functions considered are assumed to be integrable in $[a, b]$.*

Theorem 5. *If*

$$f(x) \geqq 0 \quad \textit{in} \quad [a, b]\,,$$

then

$$\int_a^b f(x)\,dx \geqq 0\,.$$

If, moreover, $f(x)$ *is continuous at least at one point* c *of that interval and if* $f(c) > 0$, *then*

$$\int_a^b f(x)\,dx > 0\,.$$

Theorem 6. *If* $f(x)$ *is continuous in* $[a, b]$ *and* $\int_a^b f^2(x)\,dx = 0$, *then* $f(x) \equiv 0$ *in* $[a, b]$.

Theorem 7. *If* $f(x) \geqq g(x)$ *in* $[a, b]$, *then* $\int_a^b f(x)\,dx \geqq \int_a^b g(x)\,dx$.

Theorem 8. *Let*

$$m \leqq f(x) \leqq M \quad \text{and} \quad g(x) \geqq 0 \quad \text{in} \quad [a, b]\,.$$

Then

$$m\int_a^b g(x)\,dx \leqq \int_a^b f(x)\,g(x)\,dx \leqq M\int_a^b g(x)\,dx\,. \tag{5}$$

REMARK 6. In particular, if $g(x) \equiv 1$, then

$$m(b-a) \leqq \int_a^b f(x)\,dx \leqq M(b-a)\,. \tag{6}$$

If

$$|f(x)| \leqq K \quad \text{in} \quad [a, b]\,,$$

then

$$\left|\int_a^b f(x)\,dx\right| \leqq K(b-a)\,. \tag{7}$$

REMARK 7. The inequality (5) is convenient for estimating the integral

$$\int_a^b f(x)\,g(x)\,dx$$

(for example in the case where the integration of the product $f(x)\,g(x)$ is rather complicated).

The inequality (6) can also be used to estimate an integral.

Example 1. Let us estimate

$$\int_0^1 \frac{dx}{10 + \sqrt{(x^2+3)} - 0{\cdot}1\cos^7 x - x^4}\,.$$

In the interval considered the inequalities

$$\frac{1}{10 + \sqrt{(1 + 3)}} < f(x) < \frac{1}{10 + \sqrt{(3)} - 0{\cdot}1 - 1} < \frac{1}{10}$$

obviously hold. Hence

$$\frac{1}{12} < \int_0^1 \frac{\mathrm{d}x}{10 + \sqrt{(x^2 + 3)} - 0{\cdot}1 \cos^7 x - x^4} < \frac{1}{10}\,.$$

Theorem 9 (*The First Mean Value Theorem*). *If* $f(x)$ *is continuous in* $[a, b]$, *then there is at least one point* $c \in (a, b)$ *such that*

$$\int_a^b f(x)\,\mathrm{d}x = (b - a)\,.\,f(c)\,. \tag{8}$$

(The value $f(c)$ defined by equation (8) is called the *mean value of the function* $f(x)$ *in the interval* $[a, b]$).

More generally: If $f(x)$ is continuous in $[a, b]$, $g(x)$ integrable in $[a, b]$ and $g(x) \geqq \geqq 0$ or $g(x) \leqq 0$, then there exists at least one point $c \in (a, b)$ such that

$$\int_a^b f(x)\,g(x)\,\mathrm{d}x = f(c) \int_a^b g(x)\,\mathrm{d}x\,.$$

Theorem 10 (*The Second Mean Value Theorem*). *Let* $g(x)$ *be monotonic* (*i.e. either increasing or decreasing*) *in* $[a, b]$. *Then there is at least one point* $c \in (a, b)$ *such that*

$$\int_a^b f(x)\,g(x)\,\mathrm{d}x = g(a) \int_a^c f(x)\,\mathrm{d}x + g(b) \int_c^b f(x)\,\mathrm{d}x\,.$$

Theorem 11 (*The Schwarz or Schwarz–Cauchy inequality*).

$$\left[\int_a^b f(x)\,g(x)\,\mathrm{d}x\right]^2 \leqq \int_a^b f^2(x)\,\mathrm{d}x \int_a^b g^2(x)\,\mathrm{d}x\,.$$

Theorem 12. *The function*

$$G(x) = \int_a^x f(t)\,\mathrm{d}t$$

is a continuous function of the variable x *in the interval* $[a, b]$. $G(x)$ *possesses a derivative at every point for which* $f(x)$ *is continuous, and the derivative equals the value of the function* $f(x)$ *at that point*, i.e.

$$\frac{\mathrm{d}G}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x} \int_a^x f(t)\,\mathrm{d}t = f(x)\,.$$

REMARK 8. If $f(x)$ is continuous in (a, b), it follows that $G(x)$ is a primitive of this function in (a, b); under the same assumptions,

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_x^b f(t)\,\mathrm{d}t = -f(x)\,.$$

Theorem 13. *If $f(x)$ is continuous in $[a, b]$ and if $F(x)$ is a primitive of $f(x)$ in (a, b) (continuous in $[a, b]$), then*

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a)\,. \tag{9}$$

REMARK 9. This equation is of fundamental importance for the evaluation of the definite integral. In applications, the right-hand side of equation (9) is usually denoted by

$$[F(x)]_a^b \quad \text{or} \quad F(x)\Big|_a^b, \quad \text{hence} \quad \int_a^b f(x)\,\mathrm{d}x = [F(x)]_a^b = F(x)\Big|_a^b. \tag{10}$$

Example 2.

$$\int_{-2}^{5} x^2\,\mathrm{d}x = \left[\frac{x^3}{3}\right]_{-2}^{5} = \frac{5^3}{3} - \frac{(-2)^3}{3} = \frac{133}{3} = 44\tfrac{1}{3}\,.$$

Theorem 14. *If $f(x)$ is continuous in $[a, b]$ and $f(x) \geqq 0$ in $[a, b]$, then the integral*

$$\int_a^b f(x)\,\mathrm{d}x$$

is equal to the area of the region bounded by the x-axis, by the graph of the function $y = f(x)$ and by the lines parallel to the y-axis through the points $x = a$, $x = b$.

REMARK 10. The area is always positive (or zero). If $f(x) < 0$ holds in the interval $[c, d]$ which is a subinterval of $[a, b]$, then the integral over $[c, d]$ is negative. If we want to determine the area of the region in question, we have to change the sign of the integral in this subinterval. Hence, if the graph of the function $y = f(x)$ crosses the x-axis in the interval $[a, b]$, we determine the coordinates of the intersections and divide the interval $[a, b]$ into intervals in which $f(x)$ is either negative, or positive.

Example 3. To find the area of the region shaded in Fig. 13.2, we write $P = P_1 + P_2$, where

$$P_1 = \int_0^{\pi} \sin x\,\mathrm{d}x = [-\cos x]_0^{\pi} = -(-1 - 1) = 2\,,$$

$$P_2 = -\int_{\pi}^{2\pi} \sin x\,\mathrm{d}x = [\cos x]_{\pi}^{2\pi} = 1 - (-1) = 2\,,$$

hence $P = 4$. The second integral had to be taken with negative sign, since $\sin x \leqq 0$ in $[\pi, 2\pi]$. Direct calculation over the whole interval 0 to 2π gives zero:

$$\int_0^{2\pi} \sin x \, dx = [-\cos x]_0^{2\pi} = -(1 - 1) = 0 .$$

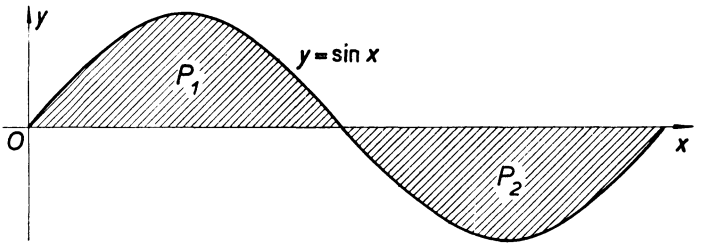


Fig. 13.2.

REMARK 11. If $f(x)$ is an (integrable) *even* function, i.e. if $f(-x) = f(x)$, then $\int_{-a}^{a} f(x)\,dx = 2\int_0^a f(x)\,dx$; if $f(x)$ is *odd*, i.e. if $f(-x) = -f(x)$, then $\int_{-a}^{a} f(x)\,dx = 0$.

Example 4.

$$\int_{-3}^{3} x^2 \, dx = 2\int_0^3 x^2 \, dx = 2\left[\frac{x^3}{3}\right]_0^3 = 18 ;$$

$$\int_{-\pi/4}^{\pi/4} \tan x \, dx = 0 .$$

REMARK 12. The so-called *Newton's definite integral* is defined by equation (9). Hence, Newton's definition assumes (only) the existence of a primitive. The equivalence of Riemann's and Newton's definition for the case where f is continuous follows from Theorem 13.

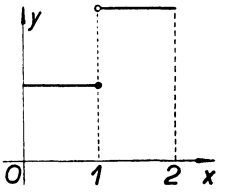


Fig. 13.3.

REMARK 13. The function remains integrable and the value of the integral does not change if the values of the function are changed at a finite number of points of $[a, b]$. For example, the integral of the function

$$f(x) = \begin{cases} 1 & \text{for } 0 \leqq x \leqq 1 , \\ 2 & \text{for } 1 < x \leqq 2 \end{cases}$$

(Fig. 13.3) is calculated as follows:

$$\int_0^2 f(x)\,\mathrm{d}x = \int_0^1 1\,.\,\mathrm{d}x + \int_1^2 2\,.\,\mathrm{d}x = 1 + 2 = 3\,.$$

The second integral is evaluated in the same way as if $f(1)$ were equal to 2.

REMARK 14. On geometrical and physical applications of the definite integral see § 14.9.

13.7. Substitution and Integration by Parts for Definite Integrals

Theorem 1 (*Integration by Parts*). *If* $u'(x)$ *and* $v'(x)$ *are continuous in* $[a, b]$ (*then also* $u(x)$ *and* $v(x)$ *are continuous in* $[a, b]$, Theorem 11.5.2), *then*

$$\int_a^b u'v\,\mathrm{d}x = [uv]_a^b - \int_a^b uv'\,\mathrm{d}x\,.$$

In another form

$$\int_a^b v\,\mathrm{d}u = [uv]_a^b - \int_a^b u\,\mathrm{d}v\,.$$

(The notation $[uv]_a^b$ means $u(b)\,v(b) - u(a)\,v(a)$.)

Example 1.

$$\int_0^\pi x \sin x\,\mathrm{d}x = [-x\cos x]_0^\pi + \int_0^\pi \cos x\,\mathrm{d}x = \pi + [\sin x]_0^\pi = \pi\,.$$

REMARK 1. If $u(x)$, $v(x)$, $u'(x)$, $v'(x)$ are only *piecewise* continuous in $[a, b]$ (Definition 11.3.6) and if c_k $(k = 1, 2, \ldots, n)$ denote the points of discontinuity of the functions $u(x)$, $v(x)$ in (a, b), then

$$\int_a^b u'v\,\mathrm{d}x = [uv]_a^b + \sum_{k=1}^{n} [uv]_{c_k+0}^{c_k-0} - \int_a^b uv'\,\mathrm{d}x = \tag{1}$$

$$= [uv]_a^{c_1-0} + \sum_{k=1}^{n-1} [uv]_{c_k+0}^{c_{k+1}-0} + [uv]_{c_n+0}^b - \int_a^b uv'\,\mathrm{d}x\,, \tag{2}$$

where

$$u(c_k - 0) = \lim_{x \to c_k -} u(x)\,, \quad u(c_k + 0) = \lim_{x \to c_k +} u(x)$$

(Remark 11.4.2) and similarly for the function v. We can alternatively first break up the integral into a sum of integrals

$$\int_a^b = \int_a^{c_1} + \int_{c_1}^{c_2} + \ldots + \int_{c_n}^b\,,$$

and then apply the integration by parts to individual integrals. One has to apply the correct limits to the corresponding functions. For instance

$$\int_{c_1}^{c_2} u'v \, dx = [uv]_{c_1+0}^{c_2-0} - \int_{c_1}^{c_2} uv' \, dx \quad \text{etc.} \tag{3}$$

Example 2. Let

$$f(x) = \begin{cases} x \sin x & \text{for} \quad 0 \leqq x \leqq \frac{1}{2}\pi , \\ 2x \sin x & \text{for} \quad \frac{1}{2}\pi < x \leqq \pi ; \end{cases}$$

Then

$$\int_0^{\pi} f(x) \, dx = \int_0^{\pi/2} x \sin x \, dx + \int_{\pi/2}^{\pi} 2x \sin x \, dx =$$

$$= [-x \cos x]_0^{\pi/2-0} + \int_0^{\pi/2} \cos x \, dx + [-2x \cos x]_{\pi/2+0}^{\pi} + 2 \int_{\pi/2}^{\pi} \cos x \, dx =$$

$$= 0 + 1 + 2\pi - 2 = 2\pi - 1 .$$

Theorem 2 (*Method of Substitution, Case* (a), *Substitution* $h(x) = z$). *Let* $f(x)$ *be of the form* $f(x) = g(h(x)) \, h'(x)$, *where* $h'(x)$ *is continuous in* $[a, b]$ *and* $g(z)$ *is continuous for all* $z = h(x)$ *if* $x \in [a, b]$. *Then*

$$\int_a^b f(x) \, dx = \int_a^b g(h(x)) \, h'(x) \, dx = \int_{h(a)}^{h(b)} g(z) \, dz . \tag{4}$$

Theorem 3 (*Method of Substitution, Case* (b), *Substitution* $x = \varphi(z)$). *Let* $A \leqq a < b \leqq B$, *let* $f(x)$ *be continuous in* $[A, B]$, $\varphi'(z)$ *continuous in* $[\alpha, \beta]$ *and for* $z \in [\alpha, \beta]$ *let* $x = \varphi(z)$ *belong to the interval* $[A, B]$, $\varphi(\alpha) = a$, $\varphi(\beta) = b$. *Then*

$$\int_a^b f(x) \, dx = \int_{\alpha}^{\beta} f(\varphi(z)) \, \varphi'(z) \, dz . \tag{5}$$

REMARK 2. There are often several possible ways of satisfying the equations $\varphi(\alpha) = a$, $\varphi(\beta) = b$ when applying Theorem 3. It is immaterial which one is chosen as long as the conditions of the theorem are satisfied. This point is illustrated in Example 3.

Example 3. Using the substitution $x = z^2$, we obtain

$$\int_1^4 x \, dx = 2 \int_1^2 z^3 \, dz = 2 \int_{-1}^2 z^3 \, dz = 2 \int_1^{-2} z^3 \, dz = 2 \int_{-1}^{-2} z^3 \, dz ,$$

as can be immediately verified by calculation. Either of the roots of the equation $z^2 = 1$ or $z^2 = 4$, respectively, may be chosen as a new limit of integration.

Example 4. Let us evaluate $\int_{-1}^{1} \sqrt{(1 - x^2)} \, dx$.

We make use of the substitution $x = \sin z$ (Theorem 3). Let us choose $\alpha = -\frac{1}{2}\pi$, $\beta = \frac{1}{2}\pi$. Obviously $\varphi(\alpha) = a$, $\varphi(\beta) = b$, since $\sin(-\frac{1}{2}\pi) = -1$, $\sin\frac{1}{2}\pi = 1$. The conditions of Theorem 3 are satisfied and we have

$$\int_{-1}^{1} \sqrt{(1 - x^2)}\,dx = \int_{-\pi/2}^{\pi/2} \sqrt{(1 - \sin^2 z)}\cos z\,dz = \int_{-\pi/2}^{\pi/2} \cos^2 z\,dz =$$

$$= \frac{1}{2}\int_{-\pi/2}^{\pi/2} (1 + \cos 2z)\,dz = \frac{1}{2}\left[z + \frac{\sin 2z}{2}\right]_{-\pi/2}^{\pi/2} = \tfrac{1}{2}\pi\,.$$

($\sqrt{(1 - \sin^2 z)} = +\cos z$, because $\sqrt{(1 - \sin^2 z)}$ is non-negative and $\cos z \geqq 0$ for $z \in [-\frac{1}{2}\pi, \frac{1}{2}\pi]$.)

Example 5. Evaluate $\int_0^{\pi/2} \sin^3 x \cos x\,dx$.

We use Theorem 2, for the function $\sin^3 x \cos x$ is of the form $g(h(x))\,h'(x)$, where $h(x) = \sin x$, $g(z) = z^3$. Hence, by the substitution $\sin x = z$ we get

$$\int_0^{\pi/2} \sin^3 x \cos x\,dx = \int_0^1 z^3\,dz = \left[\frac{z^4}{4}\right]_0^1 = \frac{1}{4}\,.$$

REMARK 3. The following examples indicate common errors in integration by substitution.

Example 6. The integral $\int_1^4 \sqrt{(x^2 - 1)}\,dx$ may not be integrated by the substitution $x = \sin z$ since $x = \sin z$ could not run through the interval $[1, 4]$, when z runs through any interval $[\alpha, \beta]$. However, the given integral may be evaluated by substitution $\sqrt{(x^2 - 1)} = z - x$ (§ 13.4 substitution (5)). In this case we have $1 \leqq x \leqq 4$, $1 \leqq z \leqq 4 + \sqrt{15}$.

Example 7. Evaluate the integral

$$\int_0^{\pi} \frac{1 + \tan^2 x}{1 + k^2 \tan^2 x}\,dx\,, \quad k > 0\,,\ k \neq 1\,.$$

Substitution:

$$\tan x = z\,, \quad \frac{1}{\cos^2 x}\,dx = dz\,, \quad (1 + \tan^2 x)\,dx = dz\,; \quad \tan(0) = 0\,, \quad \tan(\pi) = 0.$$

Hence

$$\int_0^{\pi} \frac{1 + \tan^2 x}{1 + k^2 \tan^2 x}\,dx = \int_0^0 \frac{dz}{1 + k^2 z^2} = 0\,.$$

Obviously, the result is *wrong*, for an integral of a positive function is positive (Theorem 13.6.5) and we cannot obtain zero as a result. The mistake lies in the substitution $\tan x = z$ and in the fact that $\tan x$ is discontinuous at the point $x = \frac{1}{2}\pi$ in the interval $[0, \pi]$. For the correct solution see Example 13.8.14, p. 533.

13.8. Improper Integrals

Improper integrals are a generalization of the Cauchy–Riemann integral (§ 13.6) which has been defined for a *bounded* function and a *finite* interval.

Definition 1. Let $f(x)$ be integrable according to the Cauchy–Riemann definition in any interval $[a, c]$, $a < c < b$. (We do not require that $f(x)$ be bounded in the whole interval $[a, b]$; it may be unbounded in the (left) neighbourhood of the point b.) If there exists the *finite* limit

$$\lim_{c \to b-} \int_a^c f(x)\,dx = A\,, \tag{1}$$

we say that the *integral*

$$\int_a^b f(x)\,dx \tag{2}$$

is convergent (*converges, exists*), and we write

$$\int_a^b f(x)\,dx = A\,.$$

If the limit (1) does not exist or if it is infinite, integral (2) is said to be *divergent* or to *diverge* (or we say that it *does not exist*).

Example 1.

$$\int_0^1 \frac{dx}{\sqrt{(1-x)}} = \lim_{c \to 1-} \int_0^c \frac{dx}{\sqrt{(1-x)}} = \lim_{c \to 1-} \left[-2\sqrt{(1-x)}\right]_0^c =$$

$$= \lim_{c \to 1-} \left[-2\sqrt{(1-c)} + 2\right] = 2\,.$$

Hence this integral is convergent.

Example 2.

$$\int_0^1 \frac{dx}{1-x} = \lim_{c \to 1-} \int_0^c \frac{dx}{1-x} = \lim_{c \to 1-} \left[-\ln(1-x)\right]_0^c = \lim_{c \to 1-} \left[-\ln(1-c)\right] = +\infty\,.$$

The integral is divergent. (We say in this case that its value is $+\infty$.)

REMARK 1. Similarly, if $f(x)$ is integrable in any interval $[c, b]$, $a < c < b$ (in (a, c) it need not be bounded), and if there exists the *finite* limit

$$\lim_{c \to a+} \int_c^b f(x)\,dx = B\,, \tag{3}$$

then the integral

$$\int_a^b f(x)\,\mathrm{d}x$$

is said to be *convergent* (to *converge*, to *exist*); we write

$$\int_a^b f(x)\,\mathrm{d}x = B\,.$$

Otherwise the integral is said to be *divergent*, or to *diverge* (or we say that it *does not exist*).

Definition 2. If d is a fixed point of the interval (a, b) and if $f(x)$ is integrable in arbitrary intervals $[a, a']$, $[b', b]$ with $a < a' < d < b' < b$ (in the neighbourhood of the point d $f(x)$ need not be bounded), then the integral

$$\int_a^b f(x)\,\mathrm{d}x \tag{4}$$

is said to be *convergent* (to *converge*, to *exist*) if the integrals

$$\int_a^d f(x)\,\mathrm{d}x\,, \quad \int_d^b f(x)\,\mathrm{d}x \tag{5}$$

both converge. Their sum is then called the *value of the given integral*. If at least one of the integrals (5) is divergent, the integral (4) is said to be *divergent*.

REMARK 2. If $f(x)$ is not bounded in the neighbourhood of the point a as well as in the neighbourhood of the point b, we consider the integrals

$$\int_a^c f(x)\,\mathrm{d}x\,, \quad \int_c^b f(x)\,\mathrm{d}x\,,$$

where c is an arbitrary point of the interval (a, b). We proceed in a similar way, if $f(x)$ is not bounded in the vicinity of a finite number of points of the interval $[a, b]$.

Example 3.

$$\int_{-1}^2 \frac{\mathrm{d}x}{x} = \int_{-1}^0 \frac{\mathrm{d}x}{x} + \int_0^2 \frac{\mathrm{d}x}{x}$$

provided that the integrals on the right-hand side are both convergent. However, they are both divergent:

$$\int_{-1}^0 \frac{\mathrm{d}x}{x} = \lim_{c\to 0-} \int_{-1}^c \frac{\mathrm{d}x}{x} = \lim_{c\to 0-} [\ln |x|]_{-1}^c = \lim_{c\to 0-} \ln |c| = -\infty$$

and similarly

$$\int_0^2 \frac{dx}{x} = +\infty .$$

Hence the given integral is divergent.

REMARK 3. It is sometimes convenient to deal with the so-called *Cauchy principal value* of the integral. If $f(x)$ is not bounded in the neighbourhood of the point d $(a < d < b)$, then the Cauchy principal value is defined as follows:

$$\int_a^b f(x)\,dx = \lim_{\delta \to 0} \left(\int_a^{d-\delta} f(x)\,dx + \int_{d+\delta}^b f(x)\,dx \right) .$$

Thus, symmetric δ-neighbourhoods of the point d are considered and the limit process for $\delta \to 0$ is carried out. If the improper integral exists, then it also exists considered as the Cauchy principal value, but not conversely, in general.

Example 4.

$$\int_{-1}^2 \frac{dx}{x} = \lim_{\delta \to 0+} \left(\int_{-1}^{-\delta} \frac{dx}{x} + \int_{\delta}^2 \frac{dx}{x} \right) = \lim_{\delta \to 0+} (\ln|-\delta| - \ln|-1| + \ln 2 - \ln \delta) =$$

$$= \lim_{\delta \to 0+} (\ln \delta - 0 + \ln 2 - \ln \delta) = \ln 2 .$$

The given integral is convergent considered as the Cauchy principal value, but it is not convergent if taken in the usual sense. (Example 3).

REMARK 4. If we can determine a primitive of a given function, then we can — as a rule — easily determine the limit (1) or (3) and thus evaluate immediately the integral (Examples 1 and 2). In some cases it may be difficult to find the primitive. We then try to evaluate the integral approximately. To do this, we must first know whether the given integral is convergent or not. The following tests enable us to decide this question. (The corresponding Theorems 1—7 are stated *for the case where $f(x)$ is unbounded only in the neighbourhood of the point b and is integrable in any interval $[a, c]$, $a < c < b$*; other cases are treated similarly.)

Theorem 1 (*The Bolzano–Cauchy Condition*). *The integral*

$$\int_a^b f(x)\,dx \tag{6}$$

is convergent if and only if, for arbitrary $\varepsilon > 0$, a $\delta > 0$ can be found such that for every pair of positive numbers δ_1, δ_2, satisfying $\delta_1 < \delta, \delta_2 < \delta$, the inequality

$$\left| \int_{b-\delta_1}^{b-\delta_2} f(x)\,dx \right| < \varepsilon$$

holds.

Theorem 2. *If*

$$\int_a^b |f(x)|\,dx$$

is convergent, then so is integral (6). In this case integral (6) is said to be *absolutely convergent.*

Theorem 3. *Let us assume* $0 \leqq \psi(x) \leqq \varphi(x)$ *in* $[a, b)$. *If the integral*

$$\int_a^b \varphi(x)\,dx \tag{7}$$

is convergent, then the same is true for the integral

$$\int_a^b \psi(x)\,dx\,. \tag{8}$$

If integral (8) *is divergent, then integral* (7) *is also divergent.*

Theorem 4. *If the inequality* $|f(x)| \leqq \varphi(x)$ *holds in* $[a, b)$ *and if integral* (7) *is convergent, then integral* (6) *is also convergent* (*and its convergence is absolute*). (The function $\varphi(x)$ is called a *majorant of* $f(x)$ *in* $[a, b)$.)

Example 5. The integral

$$\int_0^1 \frac{\sin x}{\sqrt{(1-x)}}\,dx$$

is convergent, for

$$\left|\frac{\sin x}{\sqrt{(1-x)}}\right| \leqq \frac{1}{\sqrt{(1-x)}}$$

and the integral of the right-hand side is convergent by Example 1.

Theorem 5. *Let the finite limit*

$$\lim_{x \to b-} \frac{f(x)}{\varphi(x)} = l$$

exist. If

$$\int_a^b |\varphi(x)|\,dx \tag{9}$$

is convergent, then also

$$\int_a^b |f(x)|\,dx \tag{10}$$

is convergent (and hence the same is true for integral (6)). If $l \neq 0$, then if (9) is divergent so also is (10). (At the same time integral (6) may be convergent.)

Theorem 6. *If integral (6) is convergent and $g(x)$ is a bounded monotonic function in $[a, b]$, then*

$$\int_a^b f(x)\, g(x)\, \mathrm{d}x \tag{11}$$

is also convergent. Further: If for every $c \in (a, b)$ the inequality

$$\left| \int_a^c f(x)\, \mathrm{d}x \right| < K$$

holds and if $g(x)$ is monotonic in $[a, b]$ and

$$\lim_{x \to b-} g(x) = 0\,,$$

then integral (11) is convergent.

Theorem 7. *Let the inequality*

$$|f(x)| \leqq \frac{M}{(b - x)^{\alpha}} \tag{12}$$

hold in a neighbourhood of the point b (for $x < b$), where M is a constant and $\alpha < 1$. Then the integral (6) is (absolutely) convergent. If

$$|f(x)| \geqq \frac{M}{(b - x)^{\alpha}} \quad (M > 0, \alpha \geqq 1)\,, \tag{13}$$

then the integral

$$\int_a^b |f(x)|\, \mathrm{d}x$$

is divergent (while, however, integral (6) may be convergent).

REMARK 5. In the case where $f(x)$ is unbounded in the neighbourhood of the point a (instead of the point b) (cf. the end of Remark 4), then, of course, we write $x - a$ in place of $b - x$ in (12) and (13).

Example 6. Let us consider the convergence of the integral

$$\int_0^1 \ln^2 x\, \mathrm{d}x\,.$$

The function $f(x) = \ln^2 x$ is not bounded in the neighbourhood of the point $x = 0$.

However (by l'Hospital's rule, see Example 11.8.4), if $0 < \alpha < 1$, we have

$$\lim_{x\to 0+} x^\alpha \ln^2 x = \lim_{x\to 0+} \frac{\ln^2 x}{\frac{1}{x^\alpha}} = \lim_{x\to 0+} \frac{\frac{2\ln x}{x}}{-\frac{\alpha}{x^{1+\alpha}}} = -\frac{2}{\alpha}\lim_{x\to 0+} x^\alpha \ln x =$$

$$= -\frac{2}{\alpha}\lim_{x\to 0+} \frac{\frac{1}{x}}{-\frac{\alpha}{x^{1+\alpha}}} = \frac{2}{\alpha^2}\lim_{x\to 0+} x^\alpha = 0 .$$

Hence the function $x^\alpha \ln^2 x$ is bounded in a (right) neighbourhood of the point $x = 0$, i.e.

$$x^\alpha \ln^2 x < M , \quad \ln^2 x < \frac{M}{x^\alpha} \quad (\alpha < 1) ,$$

and therefore the integral considered is convergent by Theorem 7.

REMARK 6. For details and for many examples see e.g. [54], [158]. See also §13.14.

REMARK 7. We also speak about improper integrals in the case where the integration is carried out in an *infinite interval.*

Definition 3. Let $f(x)$ be integrable in every interval $[a, b]$, $b > a$. If there exists the finite limit

$$\lim_{b\to+\infty} \int_a^b f(x)\,dx = A , \tag{14}$$

we say that the *integral*

$$\int_a^\infty f(x)\,dx \tag{15}$$

is convergent (*converges, exists*) and write

$$\int_a^\infty f(x)\,dx = A .$$

If the limit (14) does not exist or is infinite, we say that integral (15) is *divergent* (*diverges, does not exist*).

Similarly, the integral

$$\int_{-\infty}^a f(x)\,dx$$

is defined.

Example 7.

$$\int_2^{\infty} \frac{dx}{x^3} = \lim_{b \to +\infty} \int_2^b \frac{dx}{x^3} = \lim_{b \to +\infty} \left[-\frac{1}{2x^2} \right]_2^b = \lim_{b \to +\infty} \left(-\frac{1}{2b^2} + \frac{1}{2 \cdot 2^2} \right) = \frac{1}{8} .$$

The integral is convergent.

Definition 4. Let us assume $-\infty < a < +\infty$. If both integrals

$$\int_a^{\infty} f(x)\,dx , \quad \int_{-\infty}^a f(x)\,dx \tag{16}$$

are convergent, then the integral

$$\int_{-\infty}^{\infty} f(x)\,dx \tag{17}$$

is said to be *convergent* and the sum of integrals (16) to be its *sum*. If at least one of integrals (16) is divergent, integral (17) is said to be *divergent*.

REMARK 8. If the integral (17) is divergent in the sense of Definition 4, it may be convergent as the Cauchy principal value,

$$\int_{-\infty}^{\infty} f(x)\,dx = \lim_{a \to +\infty} \int_{-a}^a f(x)\,dx \quad (a > 0) \tag{18}$$

(assuming the limit (18) exists and is finite).

Example 8. The integral

$$\int_{-\infty}^{\infty} x\,dx$$

is divergent in the common sense, for e.g.

$$\int_0^{\infty} x\,dx = \lim_{b \to +\infty} \int_0^b x\,dx = \lim_{b \to +\infty} \frac{b^2}{2} = +\infty .$$

However, if we take

$$\int_{-\infty}^{\infty} x\,dx = \lim_{a \to +\infty} \int_{-a}^a x\,dx = \lim_{a \to +\infty} \left(\frac{a^2}{2} - \frac{a^2}{2} \right) = 0 ,$$

it is convergent as the Cauchy principal value.

REMARK 9. Concerning improper integrals in infinite intervals, a remark similar to Remark 4 may be added. Theorems similar to Theorems 1–7 may be used to decide on the convergence or divergence. *In what follows the integrability in every*

finite interval $[a, b]$ *is assumed.* Theorems 2–6 keep exactly the same form, the only difference being in writing ∞ in place of b (the analogy of the second assertion of Theorem 6 will be stated additionally in Theorem 10). Theorems 1 and 7 need a slight modification:

Theorem 8 (*The Bolzano–Cauchy Condition*). *The integral*

$$\int_a^\infty f(x)\,dx$$

is convergent if and only if, for arbitrary $\varepsilon > 0$, *there is a number* B *such that, if* $b_1 > B$, $b_2 > B$, *then the inequality*

$$\left|\int_{b_1}^{b_2} f(x)\,dx\right| < \varepsilon$$

holds.

Theorem 9. *Let*

$$|f(x)| \leqq \frac{M}{x^\alpha}, \quad M = \text{const.}, \quad \alpha > 1,$$

hold for every $x \geqq a$. *Then the integral*

$$\int_a^\infty f(x)\,dx \tag{19}$$

is convergent. If

$$|f(x)| \geqq \frac{M}{x^\alpha}, \quad M = \text{const.} > 0, \quad \alpha \leqq 1,$$

then

$$\int_a^\infty |f(x)|\,dx$$

is divergent (*while integral* (19) *may be convergent*).

REMARK 10. If $f(x)$ is not integrable in every interval $[a, b]$ $(b > a)$ (for example, if it is not bounded in a (right) neighbourhood of the point a), we define

$$\int_a^\infty f(x)\,dx = \int_a^c f(x)\,dx + \int_c^\infty f(x)\,dx$$

(where c is an arbitrary point, $c > a$) provided that the integrals on the right-hand side are both convergent. If at least one of these integrals is divergent, the given integral is said to be divergent.

If the limits of integration are improper and if, moreover, $f(x)$ is not bounded in the vicinity of points $a_1 < a_2 < \ldots < a_n$, we define

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{a_1} f(x)\,dx + \int_{a_1}^{a_2} f(x)\,dx + \ldots + \int_{a_n}^{\infty} f(x)\,dx \tag{20}$$

provided all integrals on the right-hand side of equation (20) are convergent. If at least one is divergent, the integral $\int_{-\infty}^{\infty} f(x)\,dx$ is said to be *divergent.*

Example 9. The integral

$$\int_0^{\infty} \frac{dx}{x^{\alpha}} \tag{21}$$

is divergent for every α. To show this, let us choose $a > 0$. Integral (21) (see Remark 10) is convergent if and only if integrals

$$\int_0^{a} \frac{dx}{x^{\alpha}}, \tag{22}$$

$$\int_a^{\infty} \frac{dx}{x^{\alpha}} \tag{23}$$

are both convergent. By Theorem 9, integral (23) is convergent if $\alpha > 1$ and divergent if $\alpha \leqq 1$. However, if $\alpha > 1$, then by (13) and by Remark 5 integral (22) is divergent.

Theorem 10. *Let us assume that $f(x)$ has a bounded primitive $F(x)$ for $x > a$. (Hence $|F(x)| < K$ holds for all $x > a$). Let $g(x)$ be a monotonic function for $x > a$ such that $\lim_{x\to\infty} g(x) = 0$. Then the integral*

$$\int_a^{\infty} f(x)\,g(x)\,dx$$

is convergent.

Example 10. The integral

$$\int_0^{\infty} \frac{\sin x}{x}\,dx \tag{24}$$

is convergent. First let the function $h(x) = \sin x/x$ be defined at the point $x = 0$ by the equation $h(0) = 1$, then it is continuous at $x = 0$ (see Theorem 11.4.9); hence the point $x = 0$ does not cause difficulty. For all x there is a bounded primitive $F(x) = -\cos x$ of $f(x) = \sin x$; the function $g(x) = 1/x$ is monotonic for $x > 0$ and has zero as its limit as $x \to +\infty$. Hence, by Theorem 10, the integral (24) is convergent. (Its value is $\pi/2$, see Example 13.9.5.)

REMARK 11. Making use of the inequality $|\sin x| \leqq |x|$ in the neighbourhood of the point $x = 0$, and of Theorems 7 and 10, it can be shown that the integral

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

is convergent for all α such that $0 < \alpha < 2$.

REMARK 12. It follows directly from the definitions of improper integrals, that if the integrals of the functions $f_1(x)$ and $f_2(x)$ are convergent, then the same is true for the integral of their sum and of their difference, and also of the functions $k f_1(x)$, $k f_2(x)$, where k is a constant.

REMARK 13. The rules of substitution and integration by parts may often be employed with success for the evaluation of improper integrals.

We state corresponding theorems for the case where the functions considered are unbounded only in the neighbourhood of the point b ($b = +\infty$ is also admitted). The other cases are similar.

Theorem 11. *The equation*

$$\int_a^b f'(x)\, g(x) \, dx = [f(x)\, g(x)]_a^b - \int_a^b f(x)\, g'(x) \, dx \tag{25}$$

is valid provided that the convergence of at least two members of this equation is ensured. The continuity of $f'(x)$ and $g'(x)$ *in* $[a, b)$ *is assumed.*

REMARK 14. The expression $[f(x)\, g(x)]_a^b$ is to be understood as the limit

$$\lim_{c \to b-} f(c)\, g(c) - f(a)\, g(a) . \tag{26}$$

Convergence of the central member of equation (25) is understood to mean the existence of the finite limit (26).

Example 11.

$$\int_0^{\infty} x e^{-x} \, dx = -[x e^{-x}]_0^{\infty} + \int_0^{\infty} e^{-x} \, dx = 0 - [e^{-x}]_0^{\infty} = 0 - 0 + 1 = 1 ,$$

for $\lim\limits_{x \to +\infty} x e^{-x} = 0$ (see Theorem 11.4.9).

Example 12.

$$\int_0^{\infty} \frac{\sin x}{x} \, dx = -\left[\frac{\cos x}{x}\right]_0^{\infty} - \int_0^{\infty} \frac{\cos x}{x^2} \, dx . \tag{27}$$

This equation is nonsense, for

$$\lim_{x\to 0+} \frac{\cos x}{x} = +\infty \quad \text{and the relation} \quad \frac{\cos x}{x^2} > \frac{\frac{1}{2}}{x^2}$$

holds in a sufficiently small neighbourhood of the origin, hence the second integral is divergent by Theorem 7. In spite of this, the integral on the left-hand side of equation (27) is convergent (see Example 10). This example shows how formal use of the method of integration by parts may fail for the evaluation of improper integrals.

Theorem 12. *Let $f(x)$ be continuous in $[a, b)$. Let $x = \varphi(z)$ be an increasing function in the interval (α, β), having a continuous derivative $\varphi'(z)$ in (α, β); further, let $\lim_{z\to\alpha+} \varphi(z) = a$, $\lim_{z\to\beta-} \varphi(z) = b$ (or $\lim_{z\to\beta-} \varphi(z) = +\infty$, if $b = +\infty$). Then the equation*

$$\int_a^b f(x)\,\mathrm{d}x = \int_\alpha^\beta f(\varphi(z))\,\varphi'(z)\,\mathrm{d}z \tag{28}$$

holds, provided at least one of integrals (28) is convergent. If one of them is divergent, so is the second.

Theorem 12 may be stated for other similar cases, for instance for the case when $\varphi(z)$ is decreasing in (α, β), $\lim_{x\to\alpha+} \varphi(z) = b$, $\lim_{z\to\beta-} \varphi(z) = a$ while e.g. $b = +\infty$ may be admitted. See the following example.

Example 13. The integrals

$$\int_0^\infty \frac{\mathrm{d}x}{1+x^4}, \quad \int_0^\infty \frac{x^2\,\mathrm{d}x}{1+x^4} \tag{29}$$

are convergent by Theorem 9, since for x sufficiently large the integrands are both smaller than $1/x^2$. By the substitution

$$x = \frac{1}{z}, \quad \mathrm{d}x = -\frac{\mathrm{d}z}{z^2}$$

we obtain

$$\int_0^\infty \frac{x^2}{1+x^4}\,\mathrm{d}x = -\int_\infty^0 \frac{1/z^2}{1+1/z^4}\cdot\frac{\mathrm{d}z}{z^2} = \int_0^\infty \frac{\mathrm{d}z}{1+z^4}. \tag{30}$$

(Determination of limits: If $x \to 0+$, then $z \to +\infty$; if $x \to +\infty$, then $z \to 0+$.) Hence, both integrals (29) have the same value. This fact may be used for their evaluation. Forming the sum of both integrals, we get

$$\int_0^\infty \frac{\mathrm{d}x}{1+x^4} = \frac{1}{2}\int_0^\infty \frac{1+x^2}{1+x^4}\,\mathrm{d}x = \frac{1}{2}\int_0^\infty \frac{1+1/x^2}{x^2+1/x^2}\,\mathrm{d}x.$$

By the substitution $x - 1/x = t$, $(1 + 1/x^2)\,dx = dt$, $x^2 + 1/x^2 = t^2 + 2$ we obtain

$$\int_0^\infty \frac{1 + 1/x^2}{x^2 + 1/x^2}\,dx = \int_{-\infty}^\infty \frac{dt}{t^2 + 2} = \left[\frac{1}{\sqrt{2}} \arctan \frac{t}{\sqrt{2}}\right]_{-\infty}^\infty = \frac{1}{\sqrt{2}}\left[\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right] = \frac{\pi}{\sqrt{2}}.$$

(Determination of limits: If $x \to 0+$; then $x - 1/x \to -\infty$; if $x \to +\infty$, then $x - 1/x \to +\infty$.) Hence

$$\int_0^\infty \frac{dx}{1 + x^4} = \int_0^\infty \frac{x^2}{1 + x^4}\,dx = \frac{\pi}{2\sqrt{2}}.$$

Example 14. Let us evaluate the integral

$$\int_0^\pi \frac{1 + \tan^2 x}{1 + k^2 \tan^2 x}\,dx, \quad k > 0, \quad k \neq 1.$$

We have

$$I = \int_0^\pi = \int_0^{\pi/2} + \int_{\pi/2}^\pi = I_1 + I_2.$$

By the substitution

$$\tan x = z, \quad \frac{1}{\cos^2 x}\,dx = dz, \quad (1 + \tan^2 x)\,dx = dz$$

and by the further substitution $kz = t$ we obtain

$$I_1 = \int_0^\infty \frac{dz}{1 + k^2 z^2} = \frac{1}{k}\int_0^\infty \frac{dt}{1 + t^2} = \frac{1}{k}\left[\lim_{t \to +\infty} \arctan t - \arctan 0\right] = \frac{1}{k} \cdot \frac{\pi}{2}.$$

In a similar way we obtain (since $\lim_{x \to \pi/2-} \tan x = -\infty$)

$$I_2 = \int_{-\infty}^0 \frac{dz}{1 + k^2 z^2} = \frac{1}{k}\int_{-\infty}^0 \frac{dt}{1 + t^2} = \frac{1}{k}\left[\arctan 0 - \lim_{t \to -\infty} \arctan t\right] =$$

$$= \frac{1}{k}\left[0 - \left(-\frac{\pi}{2}\right)\right] = \frac{1}{k} \cdot \frac{\pi}{2}.$$

Hence

$$I = \frac{\pi}{k}.$$

REMARK 15. The *Schwarz* (or *Cauchy–Schwarz*) *inequality*

$$\left(\int_a^b f(x)\,g(x)\,dx\right)^2 \leqq \int_a^b f^2(x)\,dx \cdot \int_a^b g^2(x)\,dx$$

is often useful when deciding on the convergence of improper integrals. The limits a and b need not be finite. If the integrals on the right-hand side are both convergent, then the integral on the left-hand side is also convergent.

13.9. Integrals Involving a Parameter

It is often convenient to consider integrals depending on a parameter (cf. Remark 13.2.6 where a primitive was found by differentiating an integral with respect to a parameter).

Formal differentiation with respect to a parameter does not always lead to correct results. It can be shown (see Example 5 below) that

$$\int_0^\infty \frac{\sin \alpha x}{x}\,\mathrm{d}x = \frac{\pi}{2} \quad (\alpha \neq 0)\,. \tag{1}$$

If we differentiate equation (1) *with respect to* α, we get

$$\int_0^\infty \cos \alpha x\,\mathrm{d}x = 0 \tag{2}$$

and this is obviously wrong because the integral on the left-hand side of equation (2) is not convergent. (The limit

$$\lim_{b\to+\infty} \int_0^b \cos \alpha x\,\mathrm{d}x = \lim_{b\to+\infty} \left[\frac{1}{\alpha} \sin \alpha x\right]_0^b = \lim_{b\to+\infty} \frac{1}{\alpha} \sin \alpha b$$

does not exist.) However, the following theorems are valid:

Theorem 1. *Let $f(x, \alpha)$ be continuous (as a function of two variables) in the rectangle $\overline{O}(a \leqq x \leqq b,\ \alpha_1 \leqq \alpha \leqq \alpha_2$; a, b, α_1, α_2 are finite numbers). Then the function*

$$g(\alpha) = \int_a^b f(x, \alpha)\,\mathrm{d}x \tag{3}$$

is a continuous function of the variable α in the interval $[\alpha_1, \alpha_2]$ (at α_1 it is continuous from the right, at α_2 from the left), i.e. the relation

$$\lim_{\alpha\to\alpha_0} \int_a^b f(x, \alpha)\,\mathrm{d}x = \int_a^b \lim_{\alpha\to\alpha_0} f(x, \alpha)\,\mathrm{d}x = \int_a^b f(x, \alpha_0)\,\mathrm{d}x \tag{4}$$

holds for every $\alpha_0 \in [\alpha_1, \alpha_2]$.

Theorem 2. *If, in addition, $\partial f/\partial\alpha$ is continuous in $\overline{O}$, then the function $g(\alpha)$ possesses a derivative in $[\alpha_1, \alpha_2]$ (at α_1 the right-hand derivative and at α_2 the left-hand derivative) and*

$$\frac{\mathrm{d}g}{\mathrm{d}\alpha} = \int_a^b \frac{\partial f}{\partial \alpha}(x, \alpha)\,\mathrm{d}x\,, \tag{5}$$

i.e.

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \int_a^b f(x, \alpha)\,\mathrm{d}x = \int_a^b \frac{\partial f}{\partial \alpha}(x, \alpha)\,\mathrm{d}x\,. \tag{6}$$

Theorem 3. *If $f(x, \alpha)$ is continuous in $\overline{O}$, then the relation*

$$\int_{\alpha_1}^{\alpha_0} g(\alpha)\,\mathrm{d}\alpha = \int_a^b \left(\int_{\alpha_1}^{\alpha_0} f(x, \alpha)\,\mathrm{d}\alpha \right) \mathrm{d}x\,, \tag{7}$$

i.e.

$$\int_{\alpha_1}^{\alpha_0} \left(\int_a^b f(x, \alpha)\,\mathrm{d}x \right) \mathrm{d}\alpha = \int_a^b \left(\int_{\alpha_1}^{\alpha_0} f(x, \alpha)\,\mathrm{d}\alpha \right) \mathrm{d}x \tag{8}$$

holds for every $\alpha_0 \in [\alpha_1, \alpha_2]$.

REMARK 1. The assertion of Theorem 3 remains valid under far more general assumptions (see Theorem 14.3.1).

Theorem 4 (*The Limits of Integration Depending on a Parameter*). *Let $x = \varphi_1(\alpha)$, $x = \varphi_2(\alpha)$ be functions having continuous derivatives in $[\alpha_1, \alpha_2]$. Let us denote by $\overline{P}$ the domain $\alpha_1 \leqq \alpha \leqq \alpha_2$, $\varphi_1(\alpha) \leqq x \leqq \varphi_2(\alpha)$, $\varphi_1(\alpha) < \varphi_2(\alpha)$* (see Fig. 13.4). *If $f(x, \alpha)$ and $\partial f / \partial \alpha(x, \alpha)$ are continuous in $\overline{P}$, then the function*

$$g(\alpha) = \int_{\varphi_1(\alpha)}^{\varphi_2(\alpha)} f(x, \alpha)\,\mathrm{d}x \tag{9}$$

has a derivative in $[\alpha_1, \alpha_2]$ (at α_1 the right-hand derivative, at α_2 the left-hand derivative) and the relation

$$\frac{\partial g}{\partial \alpha} = \int_{\varphi_1(\alpha)}^{\varphi_2(\alpha)} \frac{\partial f}{\partial \alpha}(x, \alpha)\,\mathrm{d}x + \varphi_2'(\alpha) \,.\, f(\varphi_2(\alpha), \alpha) - \varphi_1'(\alpha) \,.\, f(\varphi_1(\alpha), \alpha) \tag{10}$$

holds (see Example 2).

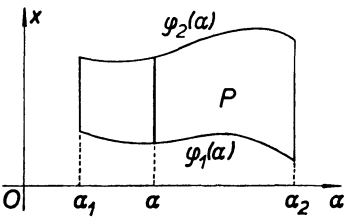


Fig. 13.4.

REMARK 2. The theorems mentioned above, particularly Theorem 2, may advantageously be used for the evaluation of definite integrals. (For the determination of primitives, using the method of a parameter, see Theorem 13.2.5.)

Example 1. We have (§ 13.10, formula 14)

$$\int_0^{\pi/2} \frac{\mathrm{d}x}{a^2 \cos^2 x + b^2 \sin^2 x} = \frac{\pi}{2ab} \quad (a > 0, b > 0)\,.$$

Differentiation with respect to a or b yields formulae for more complicated integrals (the assumptions of Theorem 2 are obviously satisfied for $a > 0$, $b > 0$):

$$-\int_0^{\pi/2} \frac{2a\cos^2 x}{(a^2\cos^2 x + b^2\sin^2 x)^2}\,dx = -\frac{\pi}{2a^2b},$$

$$-\int_0^{\pi/2} \frac{2b\sin^2 x}{(a^2\cos^2 x + b^2\sin^2 x)^2}\,dx = -\frac{\pi}{2ab^2}. \tag{11}$$

Dividing the first equation by $-2a$ and the second by $-2b$ and summing, we get (since $\cos^2 x + \sin^2 x = 1$)

$$\int_0^{\pi/2} \frac{dx}{(a^2\cos^2 x + b^2\sin^2 x)^2} = \frac{\pi}{4ab}\left(\frac{1}{a^2} + \frac{1}{b^2}\right). \tag{12}$$

Example 2.

$$g(\alpha) = \int_\alpha^{\alpha^2+1} (x^3 + \alpha x)\,dx.$$

The assumptions of Theorem 4 are obviously satisfied, hence we obtain by (10)

$$g'(\alpha) = \int_\alpha^{\alpha^2+1} x\,dx + 2\alpha[(\alpha^2+1)^3 + \alpha(\alpha^2+1)] - 1\,.\,[\alpha^3 + \alpha^2] =$$

$$= \frac{(\alpha^2+1)^2}{2} - \frac{\alpha^2}{2} + 2\alpha[(\alpha^2+1)^3 + \alpha(\alpha^2+1)] - [\alpha^3 + \alpha^2].$$

This example is only an illustrative one. The same result may be established by direct evaluation of the integral $g(\alpha)$ and by differentiating the result with respect to α.

REMARK 3. *Improper integrals* involving a parameter are of considerable importance, especially those having infinite limits. In order to formulate the corresponding theorems, we introduce the concept of the uniform convergence of an (improper) integral (in what follows, the integrability of the functions considered in every finite interval is assumed):

Definition 1. The integral

$$\int_a^\infty f(x, \alpha)\,dx \tag{13}$$

is said to be *uniformly convergent in the interval* $\alpha_1 \leqq \alpha \leqq \alpha_2$ (in the case where, for example, $\alpha_2 = +\infty$, we shall consider the semi-open interval $\alpha_1 \leqq \alpha < \alpha_2$),

if to every $\varepsilon > 0$ there exists a number B_0 (depending, in general, on the choice of ε but independent of α) such that the inequality

$$\left| \int_B^\infty f(x, \alpha)\, dx \right| < \varepsilon \tag{14}$$

holds for any $B > B_0$.

Theorem 5. *If for all* $\alpha \in [\alpha_1, \alpha_2]$ *the inequality* $|f(x, \alpha)| \leqq \varphi(x)$ *holds and if* $\int_a^\infty \varphi(x)\, dx$ *converges, then the integral* (13) *is uniformly convergent in the interval* $\alpha_1 \leqq \alpha \leqq \alpha_2$.

Example 3. According to Theorem 5 the integral

$$\int_0^\infty e^{-\alpha x} \sin x\, dx \tag{15}$$

is uniformly convergent in every interval $[\delta, \infty)$, $\delta > 0$. For, if $\alpha \geqq \delta$, then $|e^{-\alpha x} \sin x| \leqq e^{-\delta x}$ and the integral $\int_0^\infty e^{-\delta x}\, dx$ converges. (However, $\delta = 0$ may not be admitted, because $\int_0^\infty \sin x\, dx$ does not converge at all!)

Theorem 6. *Let* $g(x, \alpha)$ *be continuous for* $x \geqq a$, $\alpha_1 \leqq \alpha \leqq \alpha_2$ *and let* $h(x)$ *be continuous and monotonic for* $x \geqq a$, *and* $\lim_{x \to \infty} h(x) = 0$. *Let* $G(x, \alpha)$ *be a primitive (with respect to the variable* x*) of the function* $g(x, \alpha)$. *If* $G(x, \alpha)$ *is bounded (i.e.* $|G(x, \alpha)| \leqq K$ *for all* $\alpha_1 \leqq \alpha \leqq \alpha_2$, $x \geqq a$, K *being a constant), then the integral*

$$\int_a^\infty g(x, \alpha)\, h(x)\, dx$$

is uniformly convergent in the interval $[\alpha_1, \alpha_2]$.

Example 4. We shall prove that the integral

$$\int_a^\infty e^{-\alpha x} \frac{\sin x}{x}\, dx \tag{16}$$

is uniformly convergent in the interval $0 \leqq \alpha < +\infty$. (For $\alpha < 0$ it is obviously divergent. For $\alpha = 0$ it is convergent by Example 13.8.10.)

It suffices to examine the convergence of the integral

$$\int_a^\infty e^{-\alpha x} \frac{\sin x}{x}\, dx\,, \quad a > 0\,, \tag{17}$$

for if we define the function $(\sin x)/x$ at the point $x = 0$ to be equal to 1, then the function $e^{-\alpha x}(\sin x)/x$ is everywhere continuous.

Let us write $g(x, \alpha) = e^{-\alpha x} \sin x$, $h(x) = 1/x$. Then

$$G(x, \alpha) = -\frac{e^{-\alpha x}(\alpha \sin x + \cos x)}{1 + \alpha^2}$$

(Example 13.2.4). The functions $g(x, \alpha)$ and $h(x)$ are continuous if $\alpha \geqq 0$, $x \geqq a > 0$, $h(x)$ is decreasing and $\lim_{x \to \infty} h(x) = 0$. Further, obviously, if $\alpha \geqq 0$ and $x \geqq a$, then $|G(x, \alpha)| < 2$. Hence, by Theorem 6, integral (17), and hence also integral (16) is uniformly convergent for all $\alpha \geqq 0$.

Theorem 7. *Let $f(x, \alpha)$ be continuous in the semi-infinite strip $\alpha_1 \leqq \alpha \leqq \alpha_2$, $x \geqq a$. Assume further that the integral*

$$g(\alpha) = \int_a^\infty f(x, \alpha)\, dx$$

is uniformly convergent for all $\alpha \in [\alpha_1, \alpha_2]$. Then

(a) *$g(\alpha)$ is continuous in $[\alpha_1, \alpha_2]$ (at the point α_1 from the right, at the point α_2 from the left).*

(b) *The relation*

$$\int_{\alpha_1}^{\alpha_0} g(\alpha)\, d\alpha = \int_a^\infty \left(\int_{\alpha_1}^{\alpha_0} f(x, \alpha)\, d\alpha \right) dx , \tag{18}$$

i.e.

$$\int_{\alpha_1}^{\alpha_0} \left(\int_a^\infty f(x, \alpha)\, dx \right) d\alpha = \int_a^\infty \left(\int_{\alpha_1}^{\alpha_0} f(x, \alpha)\, d\alpha \right) dx \tag{19}$$

holds for all $\alpha_0 \in [\alpha_1, \alpha_2]$. (In the case (a) the interval for α need not be finite.)

Theorem 8. *Let $f(x, \alpha)$ be continuous in the region $x \geqq a$, $\alpha \geqq \alpha_1$. Let the integrals*

$$g(\alpha) = \int_a^\infty f(x, \alpha)\, dx , \quad h(x) = \int_{\alpha_1}^\infty f(x, \alpha)\, d\alpha$$

be both uniformly convergent, the first with respect to α, the second with respect to x, from arbitrary finite intervals $[\alpha_1, \alpha_2]$ or $[a, b]$ respectively.

Let at least one of the integrals

$$\int_a^\infty \left(\int_{\alpha_1}^\infty |f(x, \alpha)|\, d\alpha \right) dx , \quad \int_{\alpha_1}^\infty \left(\int_a^\infty |f(x, \alpha)|\, dx \right) d\alpha$$

be convergent. Then both the integrals

$$\int_a^\infty h(x)\, dx , \quad \int_{\alpha_1}^\infty g(\alpha)\, d\alpha$$

are convergent and the equality

$$\int_a^\infty \left(\int_{\alpha_1}^\infty f(x, \alpha)\, \mathrm{d}\alpha \right) \mathrm{d}x = \int_{\alpha_1}^\infty \left(\int_a^\infty f(x, \alpha)\, \mathrm{d}x \right) \mathrm{d}\alpha \tag{20}$$

holds.

Theorem 9. *Let the functions* $f(x, \alpha)$ *and* $\partial f / \partial \alpha (x, \alpha)$ *be continuous in the semi-infinite rectangle* $\alpha_1 \leqq \alpha \leqq \alpha_2$, $x \geqq a$. *If the integral*

$$g(\alpha) = \int_a^\infty f(x, \alpha)\, \mathrm{d}x$$

converges for all $\alpha_1 \leqq \alpha \leqq \alpha_2$ *and if the integral*

$$\int_a^\infty \frac{\partial f}{\partial \alpha}(x, \alpha)\, \mathrm{d}x$$

is uniformly convergent for $\alpha \in [\alpha_1, \alpha_2]$, *then the function* $g(\alpha)$ *has a derivative in* $[\alpha_1, \alpha_2]$ (*at* α_1 *from the right, at* α_2 *from the left*) *and the relation*

$$\frac{\mathrm{d}g}{\mathrm{d}\alpha} = \int_a^\infty \frac{\partial f}{\partial \alpha}(x, \alpha)\, \mathrm{d}x\,, \tag{21}$$

i.e.

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \int_a^\infty f(x, \alpha)\, \mathrm{d}x = \int_a^\infty \frac{\partial f}{\partial \alpha}(x, \alpha)\, \mathrm{d}x\,, \tag{22}$$

holds.

Example 5. We have to evaluate the integral

$$\int_0^\infty \frac{\sin x}{x}\, \mathrm{d}x\,.$$

We shall use the result concerning integral (16). If the function $\varphi(x) = \sin x/x$ is defined at the origin by the equation $\varphi(0) = 1$, then the function $\mathrm{e}^{-\alpha x} \sin x/x$ is continuous for all x and α. The function

$$\frac{\partial}{\partial \alpha}\left(\mathrm{e}^{-\alpha x} \frac{\sin x}{x} \right) = -\, \mathrm{e}^{-\alpha x} \sin x$$

is also continuous for all x and α. According to Example 3, the integral

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

is uniformly convergent for all $\alpha \geqq \delta > 0$ and is equal to $1/(1 + \alpha^2)$, as can be easily verified using integration by parts. Hence, by Theorem 9, we have

$$\frac{d}{d\alpha}\int_0^\infty e^{-\alpha x}\frac{\sin x}{x}\,dx = -\int_0^\infty e^{-\alpha x}\sin x\,dx = -\frac{1}{1+\alpha^2} \tag{23}$$

for all $\alpha \geqq \delta$. Consequently,

$$\int_0^\infty e^{-\alpha x}\frac{\sin x}{x}\,dx = -\arctan\alpha + \frac{\pi}{2} \quad (\alpha \geqq \delta)\,. \tag{24}$$

The value $\frac{1}{2}\pi$ as the constant of integration follows from the relation

$$\int_0^\infty e^{-\alpha x}\frac{\sin x}{x}\,dx \to 0 \quad \text{as} \quad \alpha \to +\infty$$

(which is true since

$$\left|\frac{\sin x}{x}\right| \leqq 1$$

and thus

$$\left|\int_0^\infty e^{-\alpha x}\frac{\sin x}{x}\,dx\right| \leqq \int_0^\infty \left|e^{-\alpha x}\frac{\sin x}{x}\right| dx \leqq \int_0^\infty e^{-\alpha x}\,dx = \frac{1}{\alpha}\Big)\,.$$

By Example 4 the integral

$$g(\alpha) = \int_0^\infty e^{-\alpha x}\frac{\sin x}{x}\,dx \tag{25}$$

is uniformly convergent for all $\alpha \geqq 0$. Hence, by Theorem 7, the function $g(\alpha)$ is continuous from the right at the point $\alpha = 0$, i.e.

$$\lim_{\alpha\to 0+}\int_0^\infty e^{-\alpha x}\frac{\sin x}{x}\,dx = g(0) = \int_0^\infty \frac{\sin x}{x}\,dx\,. \tag{26}$$

It follows from (26) and (24) that

$$\int_0^\infty \frac{\sin x}{x}\,dx = \frac{\pi}{2}\,. \tag{27}$$

REMARK 4. We had to divide the procedure just carried out into two steps: By differentiation with respect to the parameter α we were led to equation (24) (Theorem 9) and then we made use of the continuity (from the right) of the function $g(\alpha)$ at the point $\alpha = 0$ (equation (26)). Theorem 9 was not directly applicable in view of the divergence of the integral $\int_0^\infty e^{-\alpha x}\sin x\,dx$ for $\alpha = 0$.

REMARK 5. Sometimes we meet the case where the limits of integration are finite but the function $f(x, \alpha)$ or $\frac{\partial f}{\partial \alpha}(x, \alpha)$ is *not bounded* in the neighbourhood of the segment $x = b$, $\alpha_1 \leqq \alpha \leqq \alpha_2$. In the same way as in Definition 1, the integral

$$g(\alpha) = \int_a^b f(x, \alpha)\, dx \tag{28}$$

is said to be *uniformly convergent for all* $\alpha \in [\alpha_1, \alpha_2]$ if for every $\varepsilon > 0$ there exists $\delta_0 > 0$ (the same for all $\alpha \in [\alpha_1, \alpha_2]$) such that the inequality

$$\left| \int_{b-\delta}^{b} f(x, \alpha)\, dx \right| < \varepsilon$$

holds for every δ, $0 < \delta < \delta_0$.

In this case, Theorems 5, 7 and 9 are quite similar; one has only to replace ∞ by b and to examine the continuity of the functions $f(x, \alpha)$ and $\frac{\partial f}{\partial \alpha}(x, \alpha)$ in the domain $\alpha_1 \leqq \alpha \leqq \alpha_2$, $a \leqq x < b$.

These theorems may be generalized for the case where $f(x, \alpha)$ or $\frac{\partial f}{\partial \alpha}(x, \alpha)$ is not bounded in the neighbourhood of the curve $x = \varphi(\alpha)$ or in the neighbourhood of several such curves. If the limits of integration are infinite and at the same time the functions $f(x, \alpha)$ or $\frac{\partial f}{\partial \alpha}(x, \alpha)$ are not bounded on some curves, then we divide the given integral into two or more integrals and examine each of them separately, by a method similar to that of Remark 13.8.10.

13.10. Table of Definite Integrals

Throughout this paragraph m, n are positive integers, r a real number, $C = 0.577,215,664,9...$ is the so-called *Euler's constant*, $\Gamma(x)$ is the *gamma function*,

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, dt \quad (x > 0),$$

$B(p, q)$ is the *beta function*,

$$B(p, q) = \frac{\Gamma(p)\,\Gamma(q)}{\Gamma(p+q)} \quad (p > 0,\ q > 0)$$

(see § 13.11).

In particular, for $x = n$

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

holds and for $0 < x < 1$ we have

$$\Gamma(x)\,\Gamma(1 - x) = \frac{\pi}{\sin \pi x}\,.$$

1. $$\int_0^\infty x^r e^{-ax}\,dx = \frac{\Gamma(r + 1)}{a^{r+1}} \quad \text{for} \quad a > 0,\ r > -1\,.$$

In particular for $r = n$ (n is a positive integer) this integral is equal to $n!/a^{n+1}$.

2. $$\int_0^\infty x^r e^{-ax^2}\,dx = \frac{\Gamma(\frac{1}{2}r + \frac{1}{2})}{2a^{\frac{1}{2}(r+1)}} \quad \text{for} \quad a > 0\,,\ r > -1\,.$$

In particular, for r even ($r = 2k$) and positive this integral is equal to

$$\frac{1\,.\,3\,.\,\ldots\,.\,(2k - 1)\sqrt{\pi}}{2^{k+1}a^{(2k+1)/2}}\,,$$

for r odd ($r = 2k + 1$) it is equal to $k!/2a^{k+1}$.

3. $$\int_0^\infty e^{-a^2x^2}\,dx = \frac{\sqrt{\pi}}{2a} \quad (a > 0) \quad (\textit{Laplace–Gauss integral}).$$

4. $$\int_{-\infty}^\infty e^{-a^2x^2+bx}\,dx = \frac{\sqrt{\pi}}{a}\,e^{b^2/4a^2} \quad (a > 0)\,.$$

5. $$\int_0^\infty x^2 e^{-a^2x^2}\,dx = \frac{\sqrt{\pi}}{4a^3} \quad (a > 0)\,.$$

6. $$\int_0^\infty e^{-a^2x^2} \cos bx\,dx = \frac{\sqrt{\pi}}{2a}\,e^{-b^2/(4a^2)} \quad (a > 0)\,.$$

7. $$\int_0^\infty \frac{x\,dx}{e^x - 1} = \frac{\pi^2}{6}\,.$$

8. $$\int_0^\infty \frac{x\,dx}{e^x + 1} = \frac{\pi^2}{12}\,.$$

9. $$\int_0^\infty e^{-ax} \cos bx = \frac{a}{a^2 + b^2} \quad (a > 0)\,.$$

10. $$\int_0^\infty e^{-ax} \sin bx = \frac{b}{a^2 + b^2} \quad (a > 0)\,.$$

11. $$\int_0^\infty \frac{e^{-ax} \sin bx}{x}\, dx = \arctan \frac{b}{a} \quad (a > 0).$$

12. $$\int_0^\infty e^{-x} \ln x\, dx = -C = -0{\cdot}577,215,664,9\ldots .$$

13. $$\int_0^{\pi/2} \sin^{2\alpha+1} x \cos^{2\beta+1} x\, dx = \frac{\Gamma(\alpha+1)\,\Gamma(\beta+1)}{2\Gamma(\alpha+\beta+2)} = \tfrac{1}{2}B(\alpha+1,\ \beta+1).$$

This formula can be used, for example, to evaluate the integrals

$$\int_0^{\pi/2} \sqrt{(\sin x)}\, dx, \quad \int_0^{\pi/2} \sqrt[3]{(\sin x)}\, dx, \quad \int_0^{\pi/2} \frac{dx}{\sqrt[3]{(\cos x)}}, \text{ etc.}$$

If α and β are nonnegative integers, this integral is equal to

$$\frac{\alpha!\,\beta!}{2(\alpha+\beta+1)!} \quad \text{(see 24)}.$$

14. $$\int_0^{\pi/2} \frac{dx}{a^2 \cos^2 x + b^2 \sin^2 x} = \frac{\pi}{2ab}.$$

15. $$\int_0^{\pi} \sin mx \sin nx\, dx = \int_0^{\pi} \cos mx \cos nx\, dx = \begin{cases} 0 & \text{for } m \neq n, \\ \frac{1}{2}\pi & \text{for } m = n. \end{cases}$$

16. $$\int_0^{\pi} \sin mx \cos nx\, dx = \begin{cases} 0 & \text{for } m-n \text{ even}, \\ 2m/(m^2-n^2) & \text{for } m-n \text{ odd}. \end{cases}$$

17. $$\int_{-\pi}^{\pi} \sin mx \sin nx\, dx = \int_0^{2\pi} = \begin{cases} 0 & \text{for } m \neq n, \\ \pi & \text{for } m = n. \end{cases}$$

18. $$\int_{-\pi}^{\pi} \cos mx \cos nx\, dx = \int_0^{2\pi} = \begin{cases} 0 & \text{for } m \neq n, \\ \pi & \text{for } m = n. \end{cases}$$

19. $$\int_{-\pi}^{\pi} \sin mx \cos nx\, dx = \int_0^{2\pi} = 0.$$

20. $$\int_0^{\pi/2} \sin^2 x\, dx = \int_0^{\pi/2} \cos^2 x dx = \frac{\pi}{4}.$$

21. $$\int_0^{\pi/2} \sin^{2n} x\, dx = \int_0^{\pi/2} \cos^{2n} x\, dx = \frac{1 \cdot 3 \cdot 5 \ldots . (2n-1)}{2 \cdot 4 \cdot 6 \ldots . 2n} \frac{\pi}{2}.$$

22. $$\int_0^{\pi/2} \sin^{2n+1} x\, dx = \int_0^{\pi/2} \cos^{2n+1} x\, dx = \frac{2 \cdot 4 \cdot 6 \ldots . 2n}{1 \cdot 3 \cdot 5 \ldots . (2n+1)}.$$

23. $\int_0^{\pi/2} \sin^{2m} x \cos^{2n} x \, dx = \frac{1 \cdot 3 \cdot \ldots \cdot (2m-1) \cdot 1 \cdot 3 \cdot \ldots \cdot (2n-1)}{2 \cdot 4 \cdot \ldots \cdot (2m+2n)} \frac{\pi}{2}$.

24. $\int_0^{\pi/2} \sin^{2m+1} x \cos^{2n+1} x \, dx = \frac{1}{2} \frac{m! \, n!}{(m+n+1)!}$ (m, n nonnegative integers).

25. $\int_0^{\infty} \frac{\sin ax}{x} dx = \begin{cases} \frac{1}{2}\pi & (a > 0), \\ -\frac{1}{2}\pi & (a < 0). \end{cases}$

26. $\int_0^{\alpha} \frac{\cos ax}{x} dx = +\infty$ ($\alpha > 0$ arbitrary; the integral is divergent).

27. $\int_0^{\infty} \frac{\tan ax \, dx}{x} = \begin{cases} \frac{1}{2}\pi & (a > 0) \\ -\frac{1}{2}\pi & (a < 0) \end{cases}$ (taken as the Cauchy principal value).

28. $\int_0^{\infty} \frac{\cos ax - \cos bx}{x} dx = \ln \frac{b}{a} \; (a > 0, \; b > 0)$.

29. $\int_0^{\infty} \frac{\sin x \cos ax}{x} dx = \begin{cases} \frac{1}{2}\pi & \text{for } |a| < 1, \\ \frac{1}{4}\pi & \text{for } |a| = 1, \\ 0 & \text{for } |a| > 1. \end{cases}$

30. $\int_0^{\infty} \frac{\sin x}{\sqrt{x}} dx = \int_0^{\infty} \frac{\cos x}{\sqrt{x}} dx = \sqrt{\frac{\pi}{2}}$ (*Fresnel's integrals*).

31. $\int_0^{\infty} \frac{x \sin bx}{a^2 + x^2} dx = \pm \frac{\pi}{2} e^{-|ab|}$ (the sign is to be taken to agree with that of b).

32. $\int_0^{\infty} \frac{\cos ax}{1 + x^2} dx = \frac{\pi}{2} e^{-|a|}$.

33. $\int_0^{\infty} \frac{\sin^2 ax}{x^2} dx = \frac{\pi}{2} |a|$.

34. $\int_{-\infty}^{+\infty} \sin(x^2) \, dx = \int_{-\infty}^{+\infty} \cos(x^2) \, dx = \sqrt{\frac{\pi}{2}}$ (*Fresnel's integrals*).

35. $\int_0^{\pi/2} \frac{\sin x \, dx}{\sqrt{(1 - k^2 \sin^2 x)}} = \frac{1}{2k} \ln \frac{1+k}{1-k}$ ($|k| < 1$).

36. $\int_0^{\pi/2} \frac{\cos x \, dx}{\sqrt{(1 - k^2 \sin^2 x)}} = \frac{1}{k} \arcsin k$ ($|k| < 1$).

37. $\int_0^{\pi/2} \frac{\sin^2 x \, dx}{\sqrt{(1 - k^2 \sin^2 x)}} = \frac{1}{k^2}(K - E)$ ($|k| < 1$; E, K see § 13.12, p. 552).

38. $\int_0^{\pi/2} \frac{\cos^2 x\,dx}{\sqrt{(1 - k^2 \sin^2 x)}} = \frac{1}{k^2}[\mathrm{E} - (1 - k^2)\mathrm{K}]$ $(|k| < 1$; E, K, see § 13.12, p. 552).

39. $\int_0^{\pi} \frac{\cos ax\,dx}{1 - 2b\cos x + b^2} = \frac{\pi b^a}{1 - b^2}$ (a is a non-negative integer, $|b| < 1$).

40. $\int_0^1 \ln|\ln x|\,dx = -C = -0{\cdot}577,215,664,9\ldots.$

41. $\int_0^1 \frac{\ln x}{x - 1}\,dx = \frac{\pi^2}{6}.$

42. $\int_0^1 \frac{\ln x}{x + 1}\,dx = -\frac{\pi^2}{12}.$

43. $\int_0^1 \frac{\ln x}{x^2 - 1}\,dx = \frac{\pi^2}{8}.$

44. $\int_0^1 \frac{\ln(1 + x)}{x^2 + 1}\,dx = \frac{\pi}{8}\ln 2.$

45. $\int_0^1 \left(\ln\frac{1}{x}\right)^a dx = \Gamma(a + 1)$ $(-1 < a < \infty)$.

46. $\int_0^{\pi/2} \ln\sin x\,dx = \int_0^{\pi/2} \ln\cos x\,dx = -\frac{\pi}{2}\ln 2.$

47. $\int_0^{\pi} x \ln\sin x\,dx = -\frac{\pi^2 \ln 2}{2}.$

48. $\int_0^{\pi/2} \sin x \ln\sin x\,dx = \ln 2 - 1.$

49. $\int_0^{\pi} \ln(a \pm b\cos x)\,dx = \pi \ln\frac{a + \sqrt{(a^2 - b^2)}}{2}$ for $a \geqq b > 0.$

50. $\int_0^{\pi} \ln(a^2 - 2ab\cos x + b^2)\,dx = \begin{cases} 2\pi \ln a & (a \geqq b > 0), \\ 2\pi \ln b & (b \geqq a > 0). \end{cases}$

51. $\int_0^{\pi/2} \ln\tan x\,dx = 0.$

52. $\int_0^{\pi/4} \ln(1 + \tan x)\,dx = \frac{\pi}{8}\ln 2.$

53. $$\int_0^1 x^\alpha (1-x)^\beta \, dx = 2\int_0^1 x^{2\alpha+1}(1-x^2)^\beta \, dx =$$

$$= \frac{\Gamma(\alpha+1)\,\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)} = B(\alpha+1,\ \beta+1) \quad \text{(see 13)}.$$

54. $$\int_0^\infty \frac{dx}{(1+x)\,x^a} = \frac{\pi}{\sin a\pi} \quad (0 < a < 1).$$

55. $$\int_0^\infty \frac{x^{a-1}}{1+x^b}\, dx = \frac{\pi}{b \sin \dfrac{a\pi}{b}} \quad (0 < a < b).$$

56. $$\int_0^1 \frac{dx}{\sqrt{(1-x^a)}} = \frac{\sqrt{(\pi)}\Gamma\left(\dfrac{1}{a}\right)}{a\Gamma\left(\dfrac{2+a}{2a}\right)} \quad (a \neq 0).$$

57. $$\int_0^1 \frac{dx}{1+2x\cos a + x^2} = \frac{a}{2\sin a} \quad (0 < a < \tfrac{1}{2}\pi).$$

58. $$\int_0^\infty \frac{dx}{1+2x\cos a + x^2} = \frac{a}{\sin a} \quad (0 < a < \tfrac{1}{2}\pi).$$

13.11. Euler's Integrals, the Gamma Function, the Beta Function. The Gauss Function. Stirling's Formula

Definition 1. The function

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, dt \tag{1}$$

is called the *gamma function* or *Euler's integral (function) of the second kind.*

Theorem 1. *The gamma function is defined by integral* (1) *for* $x > 0$. *If* $x \leqq 0$ *the integral* (1) *is divergent. The function* (1) *is continuous for* $x > 0$. *It has derivatives of all orders; these derivatives are obtained by formal differentiation with respect to* x *under the integral sign:*

$$\Gamma^{(n)}(x) = \int_0^\infty e^{-t} t^{x-1} \ln^n t\, dt\,. \tag{2}$$

Further, the relations

$$\lim_{x\to+\infty}\Gamma(x) = +\infty\,, \quad \lim_{x\to 0+}\Gamma(x) = +\infty \tag{3}$$

hold. The gamma function has its local minimum between the points $x = 1$ *and* $x = 2$ ($x \doteq 1{\cdot}46$).

Theorem 2. *Basic relations:*

$$\Gamma(x+1) = x\,\Gamma(x)\,; \tag{4}$$

$$\Gamma(x)\,\Gamma(1-x) = \frac{\pi}{\sin \pi x} \quad (0 < x < 1)\,;$$

$$\Gamma(x)\,\Gamma(x+\tfrac{1}{2}) = \frac{\sqrt{\pi}}{2^{2x-1}}\,\Gamma(2x)\,; \tag{5}$$

$\Gamma(1) = 1$, $\Gamma(2) = 1$, *in general*

$$\Gamma(n) = (n-1)! \quad \textit{for every natural } n\,. \tag{6}$$

(With the aid of the gamma function we often extend the definition of the factorial function by the equation $x! = \Gamma(x+1)$ for positive real numbers other than natural.)

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}\,, \quad \Gamma(\tfrac{3}{2}) = \tfrac{1}{2}\sqrt{\pi}\,, \quad \Gamma(\tfrac{5}{2}) = \tfrac{3}{2}\cdot\tfrac{1}{2}\sqrt{\pi}\,, \quad \ldots\,; \tag{7}$$

$$\ln\Gamma(x) = (x-\tfrac{1}{2})\ln x - x + \ln\sqrt{(2\pi)} + \frac{\vartheta}{4x} \quad (x > 0,\ 0 < \vartheta < 1)\,; \tag{8}$$

$$\Gamma(x) = \lim_{n\to\infty} \frac{(n-1)!}{x(x+1)(x+2)\ldots(x+n-1)}\, n^x\,; \tag{9}$$

$$\Gamma(x) = \frac{1}{x\,\mathrm{e}^{Cx}} \cdot \frac{1}{\prod_{n=1}^{\infty} \mathrm{e}^{-x/n}\left(1+\frac{x}{n}\right)}\,, \tag{10}$$

where C *is the so-called Euler's constant;* $C = 0{\cdot}577{,}215{,}664{,}9\ldots$.

REMARK 1. The limit on the right-hand side of equation (9) exists (for the infinite product on the right-hand side of equation (10) is convergent and its value is different from zero) not only for positive x, but for all x other than $0, -1, -2, \ldots$. The function $\Gamma(x)$ is thus defined for all x other than $0, -1, -2, \ldots$. (Sometimes, this extension of the gamma function is denoted by the symbol $\Pi(x-1)$.)

For the gamma function extended in this way the basic relations given above remain valid for all x, with the exception of those values for which the corresponding expressions have no meaning.

The graph of the gamma function is plotted in Fig. 13.5. Some values of the gamma function for $0 < x \leqq 2$ can be found in Table 13.1. By equation (9) or (10), the function $\Gamma(x)$ is defined also for complex values of x other than $0, -1, -2, \ldots$ The gamma function, thus considered as a function of a complex variable, is holomorphic

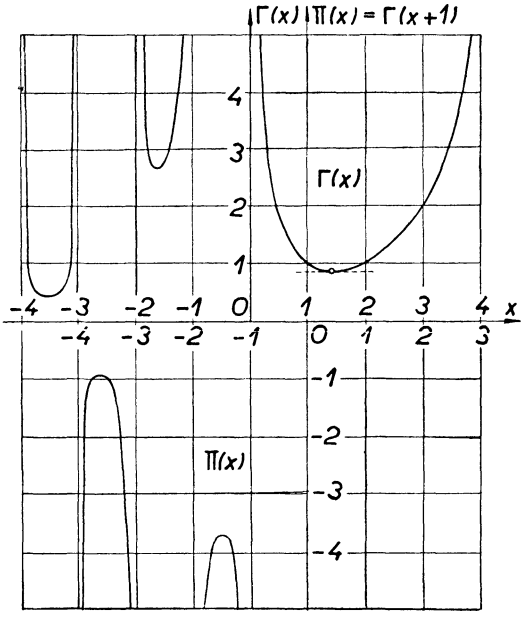


Fig. 13.5.

TABLE 13.1

Table of the gamma function for $0 < x \leqq 2$

x	$\Gamma(x)$	x	$\Gamma(x)$	x	$\Gamma(x)$	x	$\Gamma(x)$
0·00	undefined	0·50	1·772 45	1·00	1·000 00	1·50	0·886 23
05	19·470 09	55	1·616 12	05	0·973 50	55	888 87
10	9·513 51	60	1·489 19	10	951 35	60	893 52
15	6·220 27	65	1·384 80	15	933 04	65	900 12
20	4·590 84	70	1·298 06	20	918 17	70	908 64
0·25	3·625 61	0·75	1·225 42	1·25	0·906 40	1·75	0·919 06
30	2·991 57	80	1·164 23	30	897 47	80	931 38
35	2·546 15	85	1·112 48	35	891 15	85	945 61
40	2·218 16	90	1·068 63	40	887 26	90	961 77
45	1·968 14	95	1·031 45	45	885 66	95	979 88
0·50	1·772 45	1·00	1·000 00	1·46	0·885 60 (minimum)	2·00	1·000 00

(regular) in the complex domain with the exception of the points $0, -1, -2, \ldots$, where it has simple poles.

REMARK 2. The evaluation of many integrals leads to the gamma function (see e.g. [54], [158]). For example

$$\int_0^\infty e^{-x^2}\,dx = \tfrac{1}{2}\Gamma(\tfrac{1}{2}) = \tfrac{1}{2}\sqrt{\pi} \quad \text{(by the substitution } x^2 = z\text{)};$$

$$\int_0^\infty e^{-x^2} x^{2k}\,dx = \tfrac{1}{2}\Gamma\left(\frac{2k+1}{2}\right) = \frac{1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2k-1)}{2^{k+1}}\sqrt{\pi} \quad (k = 1, 2, 3, \ldots);$$

$$\int_0^\infty e^{-nt} t^{x-1}\,dt = \frac{\Gamma(x)}{n^x} \quad (x > 0) \quad \text{(by the substitution } nt = z\text{)};$$

$$\int_0^{\pi/2} \sin^{r-1} x \cos^{s-1} x\,dx = \frac{1}{2}\frac{\Gamma(\frac{1}{2}r)\,\Gamma(\frac{1}{2}s)}{\Gamma(\frac{1}{2}r + \frac{1}{2}s)} \quad (r > 0, s > 0);$$

$$\int_0^{\pi/2} \sin^{r-1} x\,dx = \int_0^{\pi/2} \cos^{r-1} x\,dx = \frac{\sqrt{(\pi)}\,\Gamma(\frac{1}{2}r)}{2\Gamma(\frac{1}{2}r + \frac{1}{2})}, \quad r > 0.$$

Definition 2. The *beta function* is defined by the integral

$$B(p, q) = \int_0^1 x^{p-1}(1-x)^{q-1}\,dx = \int_0^\infty \frac{x^{p-1}\,dx}{(1+x)^{p+q}} \quad (p > 0, q > 0). \tag{11}$$

The beta function is often called *Euler's integral (function) of the first kind.*

Theorem 3. *The relation*

$$B(p, q) = B(q, p), \tag{12}$$

holds.

REMARK 3. This function is also tabulated and the evaluation of some integrals may be reduced to it. The beta function is related to the gamma function by the equation

$$B(p, q) = \frac{\Gamma(p)\,\Gamma(q)}{\Gamma(p+q)}. \tag{13}$$

Example 1. By the substitution $x^m = z$ we have

$$\int_0^1 \frac{1}{\sqrt[n]{(1-x^m)}}\,dx = \frac{1}{m}\int_0^1 (1-z)^{-1/n} z^{(1/m)-1}\,dz = \frac{1}{m} B\left(\frac{1}{m}, 1 - \frac{1}{n}\right).$$

In particular for $n = 2$ and $m = 4$ we obtain (using (13))

$$\int_0^1 \frac{1}{\sqrt{(1-x^4)}}\,dx = \tfrac{1}{4} B(\tfrac{1}{4}, \tfrac{1}{2}) = \frac{1}{4}\frac{\Gamma(\frac{1}{4})\,\Gamma(\frac{1}{2})}{\Gamma(\frac{3}{4})}.$$

Definition 3. The *Gauss function* Ψ is defined by the relation (we write $\Pi(x) = \Gamma(x+1)$)

$$\Psi(x) = \frac{\mathrm{d} \ln \Pi(x)}{\mathrm{d}x} = \frac{\Pi'(x)}{\Pi(x)} \left(= \frac{\Gamma'(x+1)}{\Gamma(x+1)} \right), \quad x > -1 .$$

Theorem 4. *The relation*

$$\Psi(x) = \lim_{n \to \infty} \left(\ln n - \frac{1}{1+x} - \frac{1}{2+x} - \ldots - \frac{1}{n+x} \right)$$

holds. In particular

$$\Psi(0) = \lim_{n \to \infty} \left(\ln n - \frac{1}{1} - \frac{1}{2} - \ldots - \frac{1}{n} \right) = -C ,$$

where $C = 0.577{,}215{,}664{,}9...$ *is Euler's constant.*

$$\Psi(k) = -C + \frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{k} \quad (k \text{ a positive integer}) .$$

Theorem 5. *Stirling's formula:*

$$n! = n^n \sqrt{(2\pi n)}\, \mathrm{e}^{-n+\vartheta/4n} \quad (0 < \vartheta < 1) ;$$

for large n the approximate formula

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

holds.

13.12. Series Expansions of Some Important Integrals. Elliptic Integrals, Elliptic Functions

For general theorems on term-by-term integration of infinite series (including applications to the evaluation of integrals) the reader is referred to §§ 15.2, 15.4, 15.7. On applications of asymptotic expansions to the evaluation of integrals see § 15.7.

Theorem 1. *Sine integral*

$$\operatorname{Si} x = \int_0^x \frac{\sin t}{t}\, \mathrm{d}t = x - \frac{1}{3}\frac{x^3}{3!} + \frac{1}{5}\frac{x^5}{5!} - \ldots \quad (x \text{ arbitrary}).$$

Cosine integral:

$$\operatorname{Ci} x = -\int_x^\infty \frac{\cos t}{t}\, \mathrm{d}t = C + \ln x - \frac{1}{2}\frac{x^2}{2!} + \frac{1}{4}\frac{x^4}{4!} - \ldots \quad (x > 0) .$$

Logarithm integral:

$$\operatorname{li} x = \int_0^x \frac{dt}{\ln t} = C + \ln|\ln x| + \ln x + \frac{1}{2}\frac{(\ln x)^2}{2!} + \frac{1}{3}\frac{(\ln x)^3}{3!} + \dots \quad (0 < x < 1),$$

C denotes Euler's constant

$$C = \lim_{n\to\infty}\left(1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} - \ln n\right) = 0{\cdot}577,\,215,\,664,\,9\dots.$$

Fresnel's integrals:

$$\int_0^x \frac{\cos t}{\sqrt{t}}\,dt = 2\sqrt{(x)}\left(1 - \frac{x^2}{5\,.\,2!} + \frac{x^4}{9\,.\,4!} + \frac{x^6}{13\,.\,6!} + \dots\right) \quad (x > 0);$$

$$\int_0^x \frac{\sin t}{\sqrt{t}}\,dt = 2\sqrt{(x)}\left(\frac{x}{3} - \frac{x^3}{7\,.\,3!} + \frac{x^5}{11\,.\,5!} - \frac{x^7}{15\,.\,7!} + \dots\right) \quad (x > 0);$$

$$\int_0^x e^{t^2}\,dt = x + \frac{x^3}{3\,.\,1!} + \frac{x^5}{5\,.\,2!} + \frac{x^7}{7\,.\,3!}\,.+ \dots \quad (x \textit{ arbitrary});$$

$$\int_0^x e^{-t^2}\,dt = x - \frac{x^3}{3\,.\,1!} + \frac{x^5}{5\,.\,2!} - \frac{x^7}{7\,.\,3!} + \dots \quad (x \textit{ arbitrary})$$

(*the Laplace–Gauss integral*). *Asymptotic expansion for large x* (for details see § 15.7):

$$\int_0^x e^{-t^2}\,dt = \frac{\sqrt{\pi}}{2} - \frac{e^{-x^2}}{2x}\left(1 - \frac{1}{2x^2} + \frac{1\,.\,3}{(2x^2)^2} - \frac{1\,.\,3\,.\,5}{(2x^2)^3} + \dots\right).$$

(The notation

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt, \quad \operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\,dt,$$

is often used in the literature.)

Integrals of the type

$$\int R(x, \sqrt{X(x)})\,dx,$$

where $R(x, y)$ is a rational function of the variables x, y and $X(x)$ is a polynomial of the third or fourth order, are called *elliptic integrals*. (If the order of the polynomial is higher than 4, integrals of this type are called *hyperelliptic*.) The evaluation of elliptic integrals may be reduced by suitable transformations to the evaluation of

the so-called *Legendre integrals* in the normal form:

$$\text{(a)}\quad u = \int_0^x \frac{\mathrm{d}t}{\sqrt{[(1-t^2)\,(1-k^2t^2)]}} = \int_0^\varphi \frac{\mathrm{d}\psi}{\sqrt{(1-k^2\sin^2\psi)}} = F(k,\varphi)$$

$$(0 < k < 1)\,. \tag{1}$$

$$\text{(b)}\quad v = \int_0^x \sqrt{\left(\frac{1-k^2t^2}{1-t^2}\right)}\,\mathrm{d}t = \int_0^\varphi \sqrt{(1-k^2\sin^2\psi)}\,\mathrm{d}\psi = E(k,\varphi)$$

$$(0 < k < 1)\,. \tag{2}$$

Theorem 2.

$$F(k,\varphi) = J_0 + \tfrac{1}{2}k^2J_2 + \frac{1\times 3}{2\times 4}\,k^4J_4 + \frac{1\times 3\times 5}{2\times 4\times 6}\,k^6J_6 + \dots\,,$$

$$E(k,\varphi) = J_0 - \tfrac{1}{2}k^2J_2 - \frac{1\times 1}{2\times 4}\,k^4J_4 - \frac{1\times 1\times 3}{2\times 4\times 6}\,k^6J_6 - \dots$$

hold, where $0 < k < 1$ *and* φ *is arbitrary (it is sufficient, of course, to consider* $0 < \varphi < \frac{1}{2}\pi$*) and*

$$J_{2n} = \int_0^\varphi \sin^{2n}\psi\,\mathrm{d}\psi\,,\quad n = 0,1,2,\dots\,.$$

COMPLETE ELLIPTIC INTEGRALS OF THE FIRST AND SECOND KIND $(0 < k < 1)$:

Theorem 3. *The relations*

$$\mathrm{K} = F(k,\tfrac{1}{2}\pi) = \int_0^{\pi/2} \frac{\mathrm{d}\psi}{\sqrt{(1-k^2\sin^2\psi)}} = \int_0^1 \frac{\mathrm{d}t}{\sqrt{[(1-t^2)\,(1-k^2t^2)]}} =$$

$$= \frac{\pi}{2}\left[1 + (\tfrac{1}{2})^2k^2 + \left(\frac{1\times 3}{2\times 4}\right)^2 k^4 + \dots\right]. \tag{3}$$

$$\mathrm{E} = E(k,\tfrac{1}{2}\pi) = \int_0^{\pi/2} \sqrt{(1-k^2\sin^2\psi)}\,\mathrm{d}\psi = \int_0^1 \sqrt{\left(\frac{1-k^2t^2}{1-t^2}\right)}\,\mathrm{d}t =$$

$$= \frac{\pi}{2}\left[1 - (\tfrac{1}{2})^2\,\frac{k^2}{1} - \left(\frac{1\times 3}{2\times 4}\right)^2\,\frac{k^4}{3} - \dots\right] \tag{4}$$

hold.

Theorem 4. *For the so-called complementary elliptic integrals*

$$\mathrm{K}' = F(k', \tfrac{1}{2}\pi) , \quad \mathrm{E}' = E(k', \tfrac{1}{2}\pi)$$

with the modulus $k' = \sqrt{(1 - k^2)}$, *the relation*

$$\mathrm{KE}' + \mathrm{K}'\mathrm{E} - \mathrm{KK}' = \tfrac{1}{2}\pi$$

holds.

THE LEGENDRE (JACOBI) ELLIPTIC FUNCTIONS

$$\operatorname{sn} u = u - (1 + k^2)\frac{u^3}{3!} + (1 + 14k^2 + k^4)\frac{u^5}{5!} - \dots , \quad |u| < \mathrm{K}' ;$$

$$\operatorname{cn} u = 1 - \frac{u^2}{2!} + (1 + 4k^2)\frac{u^4}{4!} - (1 + 44k^2 + 16k^4)\frac{u^6}{6!} + \dots , \quad |u| < \mathrm{K}' ;$$

$$\operatorname{dn} u = 1 - k^2\frac{u^2}{2!} + k^2(4 + k^2)\frac{u^4}{4!} - k^2(16 + 44k^2 + k^4)\frac{u^6}{6!} + \dots , \quad |u| < \mathrm{K}' .$$

REMARK 1. The function $x = \operatorname{sn} u$, $0 \leqq u \leqq \mathrm{K}$, is the inverse of the function (1), i.e. of the function

$$u = \int_0^x \frac{1}{\sqrt{|(1 - t^2)\,(1 - k^2t^2)|}}\,\mathrm{d}t ,$$

where x runs through the interval $[0, 1]$. The functions $\operatorname{cn} u$, $\operatorname{dn} u$ may then be defined as continuous functions satisfying the relations

$$\operatorname{cn}^2 u = 1 - \operatorname{sn}^2 u , \quad \operatorname{dn}^2 u = 1 - k^2\operatorname{sn}^2 u$$

and $\operatorname{cn} 0 = 1$, $\operatorname{dn} 0 = 1$. For a detailed treatment (in the complex plane) see e.g. [183].

For $k = 0$ the functions $\operatorname{sn} u$ and $\operatorname{cn} u$ become the common trigonometric functions $\sin u$, $\cos u$.

Basic relations:

$$\operatorname{sn}(-u) = -\operatorname{sn} u , \quad \operatorname{cn}(-u) = \operatorname{cn} u , \quad \operatorname{dn}(-u) = \operatorname{dn} u ,$$

$$\operatorname{sn}^2 u + \operatorname{cn}^2 u = 1 , \quad \operatorname{dn}^2 u - k^2\operatorname{cn}^2 u = 1 - k^2 , \quad \operatorname{dn}^2 u + k^2\operatorname{sn}^2 u = 1 ,$$

$$\frac{\mathrm{d}}{\mathrm{d}u}\operatorname{sn} u = \operatorname{cn} u \operatorname{dn} u , \quad \frac{\mathrm{d}}{\mathrm{d}u}\operatorname{cn} u = -\operatorname{sn} u \operatorname{dn} u , \quad \frac{\mathrm{d}}{\mathrm{d}u}\operatorname{dn} u = -k^2\operatorname{sn} u \operatorname{cn} u .$$

The graphs of the elliptic functions are shown in Fig. 13.6 (for $k^2 = \frac{1}{2}$ in Figs. 13.6a, b, c; for $k^2 = \frac{3}{4}$ in Figs. 13.6d, e, f).

The functions $\operatorname{sn} u$, $\operatorname{cn} u$ are periodic with period 4K; the function $\operatorname{dn} u$ has period 2K.

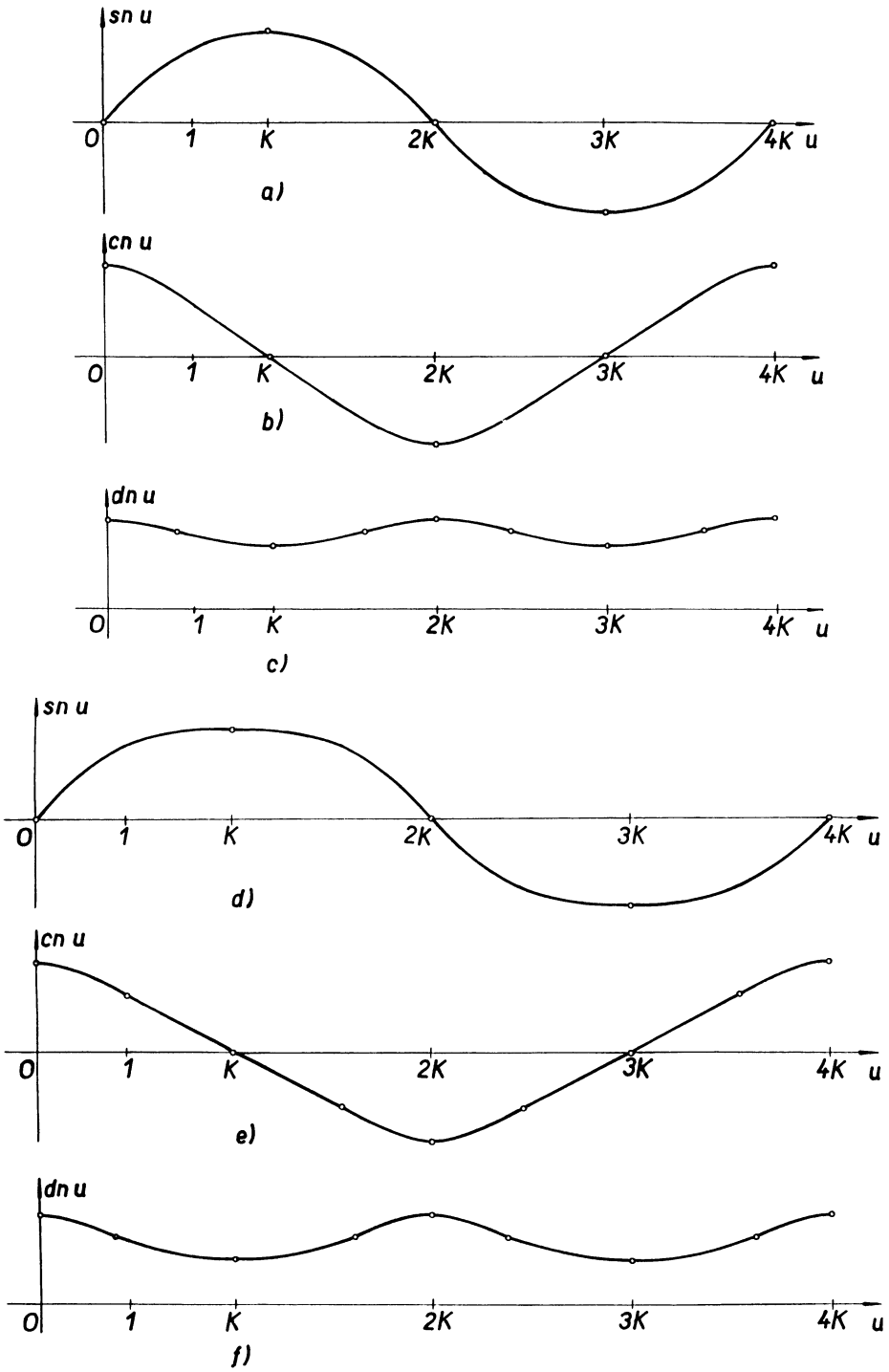

sn u
0
1
K
2K
3K
4K
u
a)
cn u
0
1
K
2K
3K
4K
u
b)
dn u
0
1
K
2K
3K
4K
u
c)
sn u
0
1
K
2K
3K
4K
u
d)
cn u
0
1
K
2K
3K
4K
u
e)
dn u
0
1
K
2K
3K
4K
u
f)

Fig. 13.6

13.13. Approximate Evaluation of Definite Integrals

If a primitive cannot be expressed with the aid of elementary or tabulated functions or if it is difficult or very laborious to find the primitive, definite integrals are usually evaluated by means of quadrature formulae. Such an approach is unavoidable if the integrated function is given by a graph or a table.

By a *quadrature formula* we understand here the approximation of the given integral

$$I(f) = \int_a^b f(x)\,\mathrm{d}x$$

by a linear combination

$$I_{n+1}(f) = \sum_{j=0}^{n} H_j f(a_j) \tag{1}$$

of values of the integrated function. Thus, the values of the function are supposed to be known quantities. The numbers a_j which are supposed to lie in $[a, b]$ are called *abscissae* or *nodes* of a quadrature formula and the numbers H_j are its *weights* or *coefficients*. The weights and abscissae are chosen in such a way that they are independent of the integrated function and satisfy some requirements, especially that the error (or *remainder*) E_{n+1} defined by

$$E_{n+1}(f) = I(f) - I_{n+1}(f)$$

has some useful properties (e.g., that it can be easily estimated, etc.).

In connection with the choice of weights and abscissae of a quadrature formula, the concept of its order is of particular importance. The *order* of a quadrature formula is defined as such integer m that $E_{n+1}(x^k) = 0$ for $k = 0, 1, \ldots, m$ while $E_{n+1}(x^{m+1}) \neq 0$. Thus, a quadrature formula of order m integers *exactly* every polynomial of degree $\leqq m$.

(a) *Gauss' formula* is a quadrature formula (1) where the weights as well as the abscissae are chosen in such a way that its order is $2n + 1$. Such formula exists for any positive integer n and is determined uniquely. Its abscissae a_j are roots of the polynomial ϕ_{n+1} of degree $n + 1$ which is orthogonal, in the interval $[a, b]$, to any polynomial of a lower degree, i.e. for which

$$\int_a^b \phi_{n+1}(x)\, q(x)\,\mathrm{d}x = 0$$

holds for any polynomial of degree $\leqq n$. The weights are then computed by

$$H_j = \int_a^b \frac{\phi_{n+1}(x)}{(x - a_j)\,\phi'_{n+1}(a_j)}\,\mathrm{d}x\ . \tag{2}$$

If the integrated function f has $2n+2$ continuous derivatives in $[a, b]$, then the *error of the Gauss formula* is given by

$$E_{n+1}(f) = \frac{1}{(2n+2)!\, A_{n+1}^2}\, f^{(2n+2)}(\eta) \int_a^b \phi_{n+2}^2(x)\, \mathrm{d}x\ , \tag{3}$$

where η is some point from $[a, b]$ and A_{n+1} is the coefficient at x^{n+1} in ϕ_{n+1}.

In particular, if the interval of integration is $[-1, 1]$, then ϕ_{n+1} is equal (up to a multiplicative constant, eventually) to the Legendre polynomial P_{n+1} (see §16.5). Thus the abscissae of this special Gauss quadrature formula (called often also the *Gauss-Legendre formula*) are roots of the Legendre polynomial of corresponding degree. For example, for $n = 2$ we have $\mathrm{P}_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x$ with roots 0 and $\pm 0.774\ 597$. If we use, moreover, (2) (naturally with $a = -1$, $b = 1$ and $\phi_3 = \mathrm{P}_3$), we obtain

$$\int_{-1}^{1} f(x)\, \mathrm{d}x \approx \tfrac{5}{9} f(-0.774\ 597) + \tfrac{8}{9} f(0) + \tfrac{5}{9} f(0.774\ 597) \tag{4}$$

and the corresponding error is not greater than $6.35 \times 10^{-5} M_6$, where

$$M_6 = \max_{[-1,1]} |f^{(6)}(x)|$$

as follows from (3).

(b) The *Newton-Cotes formulae* have equidistant abscissae, $a_j = a + jh$, $j = 0, \ldots, n$, where $h = (b-a)/n$. The corresponding weights are given again by (2), where we set $(x-a_0)\ldots(x-a_n)$ for the function ϕ_{n+1}. For the *error of the Newton-Cotes formula* we have

$$E_{n+1}(f) = \frac{1}{(n+2)!}\, f^{(n+2)}(\eta) \int_a^b x\ \phi_{n+1}(x)\, \mathrm{d}x$$

for n even and

$$E_{n+1} = \frac{1}{(n+1)!}\, f^{(n+1)}(\eta) \int_a^b \phi_{n+1}(x)\, \mathrm{d}x$$

for n odd. Consequently, for even n the corresponding order is $n+1$ whilst for n odd it is only n. In the just presented form, the Newton-Cotes formulae are used only seldom, their importance lies in their application when constructing the so-called composite quadrature formulae.

(c) A *composite quadrature formula* is constructed in such a way that we first divide the given interval into m subintervals, then use a certain quadrature formula

(the Newton-Cotes, Gauss, or, eventually another one) of a low order n in any of these subintervals and finally add the results.

(α) The *trapezoidal rule* is a quadrature formula given by

$$\int_a^b f(x)\,\mathrm{d}x = h[\tfrac{1}{2}f(a_0) + f(a_1) + \ldots + f(a_{m-1}) + \tfrac{1}{2}f(a_m)] + E(f)\ , \qquad (5)$$

where

$$h = (b-a)/m\ ,\quad a_j = a + jh\ ,\quad j = 0, 1, \ldots, m\ ,$$
$$E(f) = -(b-a)\,h^2 f''(\eta)/12\ .$$

It is a composite quadrature formula in which the Newton-Cotes formula with $n = 2$ is used in each subinterval.

(β) *Simpson's rule.* Here, the given interval is divided into an even number of subintervals by equidistant points, similarly as in the case of the trapezoidal rule, and in any interval $[a_{2i}, a_{2i+2}]$, $i = 0, 1, \ldots, \frac{m}{2} - 1$, of the length $2h$ the Newton-Cotes formula is used. The resulting formula is

$$\int_a^b f(x)\,\mathrm{d}x = \tfrac{1}{3}h[f(a_0) + 4f(a_1) + 2f(a_2) + 4f(a_3) + \ldots$$
$$\ldots + 4f(a_{m-3}) + 2f(a_{m-2}) + 4f(a_{m-1}) + f(a_m)] + E(f)\ ,$$

where

$$h = \frac{b-a}{m}\ (m \text{ even}) \quad \text{and} \quad E(f) = -(b-a)\,h^4 f^{(4)}(\eta)/180\ .$$

The number h occuring in both last quadrature formulae is usually called the *integration step.*

(d) *Romberg's formula.* The evaluation of the given integral proceeds here according to the following so-called *T-scheme*:

$$\begin{array}{llll} T_{00} & & & \\ T_{01} & T_{10} & & \\ T_{02} & T_{11} & T_{20} & \\ \vdots & \vdots & \vdots & \\ T_{0m} & T_{1,m-1} & T_{2,m-2} & \ldots, T_{m0}\ . \end{array}$$

The quantities T_{0k} in the first column of this scheme are approximate values of the integral computed by the trapezoidal rule with the integration step $h = (b-a)/2^k$. Further columns are filled up subsequently for $m = 1, 2, \ldots$ by the formula

$$T_{mk} = \frac{1}{4^m - 1}\,(4^m T_{m-1,k+1} - T_{m-1,k})\ ,\quad k = 0, 1, \ldots\ .$$

Thus, from the sequence $\{T_{0m}\}$ of approximations of the given integral by the trapezoidal rule, one constructs the "diagonal" sequence $\{T_{m0}\}$. This new sequence converges substantially more rapidly than the original sequence, as usual. Rate of convergence is the higher the smoother is the integrated function.

REMARK 1. Provided the integrated function is periodic, with period $b-a$, the trapezoidal rule itself exhibits the same properties as the Romberg method. Consequently, the use of the T scheme does not bring any further effect here.

Example 1. Let us evaluate approximately

$$I = \int_0^{0.8} \frac{\sin x}{x}\, \mathrm{d}x \,. \tag{6}$$

(a) Romberg's formula: The corresponding T-scheme is

0.758 678		
0.768 757	0.772 116	
0.771 262	0.772 097	0.772 096

Since six-digits accurate value of I is 0.772 095, the entry T_{20} is really better than any of approximations by trapezoidal rule used for its construction.

(b) Gauss' formula with $n = 2$: First, we transform the interval (0, 0.8) into the interval $(-1, 1)$ by the transformation $y = (2x - 0.8)/0.8$. We obtain

$$I = 0.4 \int_{-1}^{1} \frac{\sin \frac{0.8y+0.8}{2}}{\frac{0.8y+0.8}{2}}\, \mathrm{d}y \,.$$

Using (4), we have

$$I_3 = 0.772\ 096 \,,$$

thus again a very good result.

Example 2. Let us approximate

$$I = \int_0^{\pi} \sin^2 x \, \mathrm{d}x$$

by Romberg's method. (Note that the exact value of this integral is $\pi/2 = 1.570\ 796\ldots$). From the T-scheme,

0		
1.570 796	2.094 395	
1.570 796	1.570 796	1.535 890

we see that, in case of periodic functions, the trapezoidal rule really converges at least as well as Romberg's method.

13.14. The Lebesgue Integral

In pure as well as applied mathematics, the so-called *Lebesgue integral* became a powerful tool. In this paragraph, the definition and fundamental properties of this integral are given, in a surveyable form, for the case of functions of one variable. At the end, the extension of these concepts and results to the more-dimensional case is briefly sketched.

In what follows, only bounded sets of points (in E_1, or in E_N, respectively) are considered, what is quite sufficient for the aim of this book.

Let M be a set of points in E_1. Let us remind (see §22.1) that M is called *bounded* in E_1 if it lies entirely in an interval $(-R, R)$ with R sufficiently large. It is called *open* if every point $x \in M$ is an interior (= inner) point of this set and *closed* if it contains all its points of accumulation. The union of an arbitrary number of open sets is an open set, the intersection of an arbitrary number of closed sets is a closed set. Two sets are called *disjoint* if their intersection is an empty set (i.e. if they have no common points).

Theorem 1. *Every nonempty bounded open set M (in E_1) can be obtained as a union of a finite or countable set of disjoint open intervals I_k the endpoints of which do not belong to M.*

Theorem 2. *Every nonempty bounded closed set N is either a closed interval, or it can be obtained from a closed interval by removing a finite or countable set of disjoint open intervals I_k the endpoints of which belong to N.*

REMARK 1. The definition of the Lebesgue integral is based on the concept of the so-called *Lebesgue measure* which is a suitable generalization of the concept of length. To every so-called *Lebesgue measurable* set M the so-called *Lebesgue measure* mM (or μM, or mes M, or meas M) is assigned with properties similar to those of the length (additivity, etc.). The measure of an empty set is equal to zero, by definition. The measure of a bounded interval (open, closed, or half-open) is equal to its length. Thus $m(a,b) = m[a,b] = m(a,b] = m[a,b) = b - a$. If M is the open set discussed in Theorem 1 and I_k the open intervals from the same theorem, we define

$$mM = \sum_k mI_k \ . \tag{1}$$

Thus, by (1), the Lebesgue measure of an arbitrary nonempty bounded open set is defined (uniquely, as can be shown), as the sum of the lengths of open intervals I_k by which that set is constituted. (In accordance with Theorem 1, this sum can be finite or infinite; thus, the summation in (1) runs between the limits $k = 1$ and $k = n$, where n is a positive integer, or between 1 and ∞.)

Now, let N be a closed set mentioned in Theorem 2, I the closed interval and I_k the open intervals from the same theorem. Similarly, as above, we define

$$mN = mI - \sum_k mI_k \ . \tag{2}$$

In this way, the measure of an arbitrary nonempty bounded closed set is uniquely defined.

Now, let M be an arbitrary bounded set which need be neither open nor closed.

Definition 1. By an *outer Lebesgue measure* m^*M of a nonempty bounded set M we understand the greatest lower bound of measures of all bounded open sets P which contain the set M, thus

$$m^*M = \inf_{M \subset P} mP, \quad P \text{ bounded open sets} . \tag{3}$$

By the *inner Lebesgue measure* m_*M of this set we call the least upper bound of measures of all bounded closed sets Q contained in M, thus

$$m_*M = \sup_{Q \subset M} mQ, \quad Q \text{ bounded closed sets} . \tag{4}$$

For every bounded set M we have

$$0 \leqq m_*M \leqq m^*M . \tag{5}$$

Definition 2. If $m_*M = m^*M$ we say that M is *Lebesgue measurable* (briefly *measurable*); the common value of the outer and inner measures is called the *(Lebesgue) measure* mM of this set,

$$mM = m^*M = m_*M . \tag{6}$$

Example 1. It can be shown that

(i) if M is a set consisting of a finite number of points, then its measure is equal to zero;

(ii) the set of all points $x \in [a, b]$ with rational coordinates is also of zero measure;

(iii) the same holds for every countable set of points (which even need not be bounded);

(iv) there exist sets which are not countable and are of measure zero.

REMARK 2. It can be shown that there exist sets which are not measurable. However, the way how to construct such sets is rather difficult. The class of (Lebesgue) measurable sets (as well as of measurable functions, see below) is very broad. It can be said – very roughly speaking – that all sets (and functions) we meet in applications are measurable.

REMARK 3. In mathematics as well as in applications it is customary to say briefly "*almost everywhere*" instead of the more lengthy "with the possible exception of points constituting a set of measure zero". For example, if we say that a function f is continuous in $[a, b]$ almost everywhere, it is to be understood that ei-

ther it is continuous in the whole interval $[a, b]$, or that it is discontinuous at some points of that interval, but that these points constitute a set of measure zero.

Definition 3. Let f be a (real) function defined on a (bounded) measurable set M. If the set of all points $x \in M$, for which $f(x) < C$, is measurable for every choice of C, we say that the *function* f is *(Lebesgue) measurable on* M.

In particular, every function continuous or piecewise continuous in an interval $[a, b]$ is measurable on $[a, b]$.

REMARK 4. If f and g are measurable functions on M, then also the functions $af + bg$ (a, b arbitrary real numbers) and fg are measurable on M. Measurability of f implies measurability of $|f|$; the converse is not true, in general.

REMARK 5 *(The Lebesgue Integral of Bounded Measurable Functions).* While in the case of the Cauchy-Riemann definition of an integral, the basic interval $[a, b]$ is divided into "small parts" when upper and lower integral sums are constructed, in the case of the Lebesgue definition the partition into "small parts" concerns the range of the given function:

Let, on a (bounded) measurable set M a measurable function f be given, bounded on M

$$A \leqq f(x) \leqq B \quad \text{for all} \quad x \in M \ . \tag{7}$$

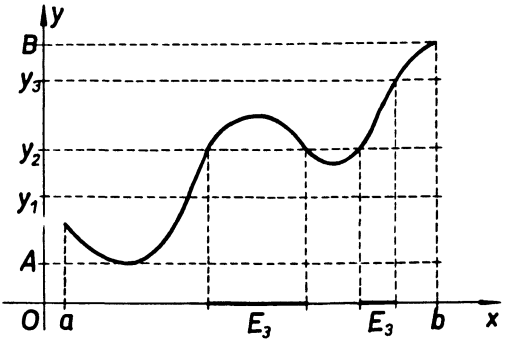


Fig. 13.7

(See Fig. 13.7, where $M = [a, b]$ has been chosen for the sake of simplicity). Let a partition d of the interval $[A, B]$ by the points

$$A = y_0 < y_1 < y_2 < \ldots < y_{n-1} < y_n = B \tag{8}$$

be given. Let us denote by E_k the set of such points $x \in [a, b]$ for which $y_{k-1} \leqq f(x) < y_k$ and let mE_k be its (Lebesgue) measure. (See Fig. 13.7, where the set E_3 consisting of two intervals is indicated.) Let us note that measurability of E_k follows from the

assumption on measurability of the function f. Let us construct the *upper*, or *lower integral sum* $S(d)$, or $s(d)$, respectively, corresponding to the given partition d,

$$S(d) = \sum_{k=1}^{n} y_k m E_k \ , \quad s(d) = \sum_{k=1}^{n} y_{k-1} m E_k \ . \tag{9}$$

It can be shown – and this is one of the basic results of the Lebesgue theory – that for every bounded measurable function on M, the greatest lower bound of the set of all upper integral sums (obtained when considering all possible partitions of the interval $[A, B]$) is equal to the least upper bound of the set of all lower integral sums, i.e. that

$$\inf_d S(d) = \sup_d s(d) \ . \tag{10}$$

Definition 4. The common value (10) of the greatest lower bound of upper integral sums and of the least upper bound of lower integral sums is called the *Lebesgue integral of a bounded measurable function on a (bounded) measurable set* M. We write

$$\int_M f(x)\,\mathrm{d}x \ . \tag{11}$$

Or, to distinguish this integral from an integral according to another definition, e.g. from that by Cauchy-Riemann, we denote it by

$$(L) \int_M f(x)\,\mathrm{d}x \ . \tag{12}$$

In special cases we write

$$\int_a^b f(x)\,\mathrm{d}x \ , \tag{13}$$

etc. We also speak about a *Lebesgue integrable* (*L-integrable*) function in contrast to a *Riemann integrable* (*R-integrable*) one, etc.

Thus: *Every bounded Lebesgue measurable function is (Lebesgue) integrable.*

REMARK 6. As concerns bounded functions, the Lebesgue integral represents a substantial generalization of that of Riemann. The first generalization consists in the fact that the domain of integration can be an arbitrary bounded measurable set. Further, any bounded R- integrable function is L-integrable (while both integrals – the Riemann and the Lebesgue one – are equal). The converse is not true: Let us consider the so-called *Dirichlet function* defined in the interval $[0, 1]$ as follows:

$$\begin{aligned} f(x) &= 1 \ \text{ if } \ x \ \text{ is rational} \ , \\ f(x) &= 0 \ \text{ if } \ x \ \text{ is irrational} \ . \end{aligned} \tag{14}$$

This function is not Riemann integrable, because the upper integral is equal to one, while the lower one equals to zero. On the other hand, the function (14) is Lebesgue integrable (the set of rational $x \in [0,1]$ is of measure zero), the integral is equal to zero.

In applications, a frequent case is that the functions in question are both Riemann and Lebesgue integrable. Their values as well as methods of their evaluation are then the same.

REMARK 7 *(The Lebesgue Integral of Unbounded Functions).* Let f be a measurable function on a (bounded measurable) set M, unbounded, in general. First, let us consider the case $f(x) \geqq 0$ on M. Choose an arbitrary $K > 0$ and define, on M, a function f_K by

$$f_K(x) = f(x) \quad \text{if} \quad f(x) \leqq K \ ,$$

$$f_K(x) = K \qquad \text{if} \quad f(x) > K$$

(Fig. 13.8). Because f is measurable on M (by assumption), so is the function f_K. Further, since f_K is bounded, the Lebesgue integral (11) exists,

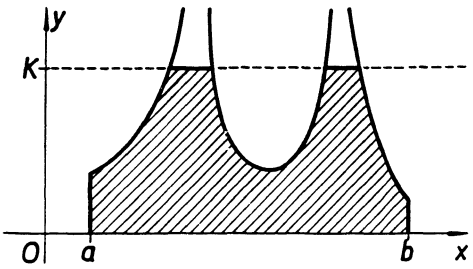


Fig. 13.8

$$\int_M f_K(x)\,\mathrm{d}x \ . \tag{15}$$

We define

$$\int_M f(x)\,\mathrm{d}x = \lim_{K\to\infty} \int_M f_K(x)\,\mathrm{d}x \ . \tag{16}$$

The limit in (16) exists (since the integral (15) is an increasing function of K). If this limit is finite, we say that the integral (16) is *convergent.* If it is infinite, we say that it is *divergent* and write

$$\int_M f(x)\,\mathrm{d}x = +\infty \ . \tag{17}$$

For example,

$$\int_0^1 \frac{1}{x}\,\mathrm{d}x = +\infty \ ,$$

because here (assuming $K > 1$ already)

$$\int_0^1 f_K(x)\,\mathrm{d}x = \int_0^{1/K} K\,\mathrm{d}x + \int_{1/K}^1 \frac{1}{x}\,\mathrm{d}x =$$

$$= 1 + \int_{1/K}^1 \frac{1}{x}\,\mathrm{d}x =$$

$$= 1 - \ln\frac{1}{K} = 1 + \ln K$$

and

$$\lim_{K\to\infty} (1 + \ln K) = +\infty\ .$$

Now, let f be an arbitrary unbounded measurable function on M, not necessarily nonnegative. Let us divide it into its "positive" and "negative" parts,

$$f = f_+ - f_-\ , \tag{18}$$

where

$$f_+(x) = \begin{cases} f(x) & \text{if } f(x) \geqq 0\ , \\ 0 & \text{if } f(x) < 0\ , \end{cases} \tag{19}$$

$$f_-(x) = \begin{cases} 0 & \text{if } f(x) \geqq 0\ , \\ -f(x) & \text{if } f(x) < 0\ . \end{cases} \tag{20}$$

Since the functions (19), (20) are nonnegative, there exist integrals in the sense (16),

$$\int_M f_+(x)\,\mathrm{d}x\ , \tag{21}$$

$$\int_M f_-(x)\,\mathrm{d}x\ , \tag{22}$$

each of them being either convergent, or divergent with the value $+\infty$. In accordance with (18) we define

$$\int_M f(x)\,\mathrm{d}x = \int_M f_+(x)\,\mathrm{d}x - \int_M f_-(x)\,\mathrm{d}x \tag{23}$$

provided the sum on the right-hand side has a sense, thus except the case that both integrals (21) and (22) have the value $+\infty$. In that case we say that the integral

$$\int_M f(x)\,\mathrm{d}x \tag{24}$$

does not exist. If both integrals (21) and (22) are convergent (thus have a finite value), we say that the integral (24) (defined by the sum (23)) is *convergent.* If (21) is divergent and (22) convergent, or conversely, we say that the integral is *divergent* and has the value $+\infty$, or $-\infty$, respectively.

REMARK 8. The Lebesgue integral has many properties similar to those of the Riemann one. For example,

$$f(x) \leqq g(x) \text{ in } M \Longrightarrow \int_M f(x)\,\mathrm{d}x \leqq \int_M g(x)\,\mathrm{d}x\ , \tag{25}$$

$$\left|\int_M f(x)\,\mathrm{d}x\right| \leqq \int_M |f(x)|\,\mathrm{d}x\ , \tag{26}$$

$$\int_M [af(x)+bg(x)]\,\mathrm{d}x = a\int_M f(x)\,\mathrm{d}x + b\int_M g(x)\,\mathrm{d}x \tag{27}$$

(a, b arbitrary real numbers), provided the sum on the right-hand side of (27) has a sense, i.e. except the case that this sum is of the form $+\infty+(-\infty)$, or $-\infty+(+\infty)$.

However, some properties of the Lebesgue integral are substantially different from those of the Riemann one, and this is why in modern mathematical disciplines the Lebesgue integral is almost exclusively used. This concerns, in particular, functional spaces, typical example of which are the L_2-spaces (Chaps. 16 and 22) and the Sobolev spaces (Chap. 22), where application of the Lebesgue integral plays a fundamental role in the question of their completeness (Chap. 22), upon which further theoretical considerations as well as applications are based. The Lebesgue integral plays also an important role in "classical" problems of analysis (see, e.g., theorems on interchange of limit and integration, §15.1).

Definition 5. We say that a function f, measurable on a set M, is *(Lebesgue) square integrable on* M, if the integral

$$\int_M f^2(x)\,\mathrm{d}x \tag{28}$$

is convergent (= finite).

Example 2. Every bounded measurable function on an interval $[a, b]$ is square integrable on this interval. In particular, every function which is continuous, or piecewise continuous in $[a, b]$ is square integrable on that interval. An example of an unbounded function which is square integrable on $[0, 1]$, is the function

$$f(x) = x^{-1/3}\ ;$$

we have

$$\int_0^1 f^2(x)\,\mathrm{d}x = \int_0^1 x^{-2/3}\,\mathrm{d}x = 3 < +\infty\ . \tag{29}$$

On the other hand, the function $g(x) = x^{-1/2}$ is not square integrable on that interval, since

$$\int_0^1 g^2(x)\,\mathrm{d}x = \int_0^1 x^{-1}\,\mathrm{d}x = +\infty\ . \tag{30}$$

REMARK 9. It can be shown that convergence of the integral (28) implies convergence of the integral

$$\int_M |f(x)|\,\mathrm{d}x$$

and, by (26), also the convergence of the integral

$$\int_M f(x)\,\mathrm{d}x\ . \tag{31}$$

Further, if f and g are square integrable on M, then

(i) an arbitrary linear combination

$$af + bg \tag{32}$$

of them is square integrable on M,

(ii) the integral

$$\int_M f(x)\ g(x)\,\mathrm{d}x \tag{33}$$

(by which the scalar product in the space L_2 is defined, see §16.1), is convergent (= has a finite value).

As concerns the above mentioned "classical" problems, we have, using the Lebesgue integral:

Theorem 3. *If a sequence $\{f_n\}$ of measurable functions converges almost everywhere on M* (Remark 3) *to a function f (then also f is measurable on M, as can be shown) and $|f_n(x)| \leqq K$ for all n and all $x \in M$, then we have (even if the convergence is not uniform)*

$$\lim_{n\to\infty} \int_M f_n(x)\,\mathrm{d}x = \int_M f(x)\,\mathrm{d}x\ . \tag{34}$$

By Remark 6, this theorem implies:

Theorem 4. *If the sequence $\{f_n\}$ of Riemann integrable functions converges in an interval $[a,b]$ to a Riemann integrable function f and if $|f_n(x)| \leqq K$ holds for all n and all $x \in [a,b]$, then*

$$\lim_{n\to\infty} \int_a^b f_n(x)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x\ . \tag{35}$$

REMARK 10 *(The Lebesgue Integral of Functions of More Variables).* The defini-

tion of the Lebesgue integral of function of N variables is similar to that of functions of one variable.

Let us show the basic ideas of the theory for the case $N = 2$. For $N > 2$ the procedure is similar.

Instead of an open interval (a, b), the open square $Q = (a, b) \times (a, b)$ is considered in the two-dimensional case (the open N-dimensional cube $(a, b) \times (a, b) \times \ldots \times (a, b)$ in the N-dimensional case). A certain modification is necessary when defining the measure of a (nonempty) bounded open set, because – in contrast to the one-dimensional case – it is not possible here to express every such set as a union of a finite or countable set of disjoint open squares. However, it is possible to express it as a union of a countable set of closed squares without common interior points. If we define the measure of a square to which there may belong a part of its boundary, or the whole boundary, as the square of its side, then the measure of a nonempty bounded open set can be defined as the sum of measures of the just mentioned closed squares which constitute this set (independently of the choice of its decomposition, as can be shown). (A similar procedure could have been used also for the case $N = 1$, of course.) The measure of a nonempty bounded closed set M can then be defined (uniquely, again) as the difference of the measure of an open square M_1 in which the set M is contained and of the measure of the (open) set $M_1 - M$. On the basis of measures of a nonempty bounded open, or closed sets it is possible to define, successively, the outer and inner measures of an arbitrary nonempty bounded set, its Lebesgue measure, measurable functions, the Lebesgue integral of a measurable function (first for bounded, then for unbounded functions) and square integrable functions in a quite similar way as it has been done in Definitions 1–4, Remark 7 and Definition 5 for the case of functions of one variable.

As concerns methods of evaluation of the Lebesgue integral for the case $N \geq 2$, it is possible to apply, e.g., the Fubini Theorem which may be formulated in a much more general way than in the case of the Riemann integral. See, e.g., [133].

13.15. The Stieltjes integral

When solving some technical (as well as mathematical) problems it is useful to apply the so-called *Stieltjes integral*. Here we mention briefly its definition and fundamental properties and show an example of its application.

Let $f(x)$ and $g(x)$ be two (finite real) functions defined in the interval $[a, b]$. Let us divide this interval by the points

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b$$

into n subintervals, in each of them choose an arbitrary point ξ_k and construct the sum (depending on the chosen partition d and on the choice of the points ξ_k)

$$\sigma(d,\xi_k)=\sum_{k=1}^{n} f(\xi_k)\,[g(x_k)-g(x_{k-1})]\ . \tag{1}$$

By the *norm* $\nu(d)$ of the chosen partition we understand the maximal length of the intervals considered, i.e.

$$\nu(d)=\max_k\,(x_k-x_{k-1})\ .$$

Let, for $\nu(d)\to 0$, the sums (1) converge to a (finite) number I, independently of how the points ξ_k have been chosen in individual subintervals. In more detail: Let such a number I exist that to every $\varepsilon>0$ such a $\delta>0$ can be found that for every partition d of the interval $[a,b]$ with $\nu(d)<\delta$ and for every choice of the points $\xi_k\in[x_{k-1},x_k]$ we have $|I-\sigma(d,\xi_k)|<\varepsilon$. Then we say that I is the *Stieltjes integral of the function f with respect to the function g* and write

$$I=\int_a^b f(x)\,\mathrm{d}g(x)\ ,\quad \text{briefly}\quad I=\int_a^b f\,\mathrm{d}g\ . \tag{2}$$

REMARK 1. For $g(x)=x$ we get the Cauchy-Riemann integral as a special case. Of course, from the point of view of applications, the Stieltjes integral is most interesting in the case when the function g has jumps (discontinuities of the first kind, Definition 11.3.5, p. 368) in the interval considered. See Theorem 3 below.

Theorem 1. *If the function f is continuous in the interval $[a,b]$ and if the function g is of bounded variation in this interval* (Definition 11.3.7, p. 370), *then the Stieltjes integral* (2) *exists.*

Theorem 2. *If f is continuous in $[a,b]$ and if g has a continuous derivative in that interval, then the integral* (2) *exists and the relation*

$$\int_a^b f\,\mathrm{d}g=\int_a^b f(x)\,g'(x)\,\mathrm{d}x \tag{3}$$

holds.

Theorem 3. *If f is continuous in $[a,b]$ and if g and g' are piecewise continuous in that interval, then the interval* (2) *exists. Let us denote by*

$$a=x_0<x_1<x_2<\ldots<x_p=b$$

points of discontinuity of the function g in $[a,b]$, by s_k the jump of this function at the point x_k, i.e.

$$\begin{aligned} &s_k=g(x_{k+})-g(x_{k-})=\lim_{x\to x_{k+}} g(x)-\lim_{x\to x_{k-}} g(x)\ ,\\ &s_0=g(a+)-g(a)\ ,\quad s_p=g(b)-g(b-)\ . \end{aligned} \tag{4}$$

Then

$$\int_a^b f \, \mathrm{d}g = \int_a^b f(x)\, g'(x)\, \mathrm{d}x + \sum_{k=0}^{p} f(x_k)\, s_k \, . \tag{5}$$

Thus the Stieltjes integral is the sum of a Riemann integral and of the jumps of the function g at the points x_k multiplied by the values of the function f at these points.

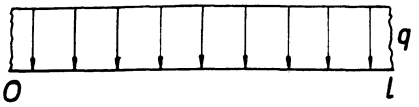


Fig. 13.9

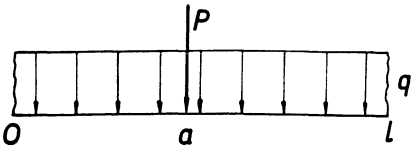


Fig. 13.10

Example 1. Let us consider a bar of length l loaded by a continuous load $q(x)$, $0 \leqq x \leqq l$ (Fig. 13.9 where a bar with a constant load is illustrated). Let us denote by $Q(x)$ the total load in the interval $[0, x]$, thus

$$Q(x) = \int_0^x q(t)\, \mathrm{d}t \, . \tag{6}$$

Then the total moment (all over the bar) with respect to the origin O (the reactions are not considered) can be written in the form

$$M = \int_0^l x \, \mathrm{d}Q \, ; \tag{7}$$

in fact, by Theorem 2 we have

$$\int_0^l x \, \mathrm{d}Q = \int_0^l xq(x)\, \mathrm{d}x \, .$$

Let, moreover, a concentrated load P be acting at the point a (Fig. 13.10). The function (6) by which the total load in the interval $[0, x]$ is given, changes now into

$$\begin{aligned} Q(x) &= \int_0^x q(t)\, \mathrm{d}t \quad \text{if} \quad x < a \, , \\ Q(x) &= \int_0^x q(t)\, \mathrm{d}t + P \quad \text{if} \quad x \geqq a \, . \end{aligned} \tag{8}$$

Also in this case M can be expressed in the closed form (7),

$$M = \int_0^l x \, \mathrm{d}Q \, ,$$

what is useful in many considerations. In fact, by (8) and Theorem 3 we have

$$\int_0^l x \, \mathrm{d}Q = \int_0^l x \, q(x) \, \mathrm{d}x + aP \ ,$$

what is the well-known expression for M.

REMARK 2. More general than the Stieltjes integral is the so-called *Lebesgue-Stieltjes* integral

$$\int_a^b f \, \mathrm{d}\mu$$

(integral of the function f with respect to the measure μ). See e.g. [133]: This concept can well be extended to the case when f is to be integrated over a more general set than an interval, as well as to the more-dimensional case.

13.16. Survey of Some Important Formulae from Chapter 13

(See also Theorem 13.1.2, §§13.5 and 13.10 and some formulae in §§13.11 and 13.12. See also applications to geometry and physics in §14.9.)

1. $$\int [c_1 f_1(x) + c_2 f_2(x)] \, \mathrm{d}x = c_1 \int f_1(x) \, \mathrm{d}x + c_2 \int f_2(x) \, \mathrm{d}x$$

   (Theorem 13.2.1).

2. $$\int u' v \, \mathrm{d}x = uv - \int uv' \, \mathrm{d}x \quad \text{or} \quad \int v \, \mathrm{d}u = uv - \int u \, \mathrm{d}v$$

   (integration by parts, Theorem 13.2.2).

3. $$\int g(h(x)) \, h'(x) \, \mathrm{d}x = \int g(z) \, \mathrm{d}z \ , \quad \int f(x) \, \mathrm{d}x = \int f(\varphi(z)) \, \varphi'(z) \, \mathrm{d}z$$

   (integration by substitution, Theorems 13.2.3, 13.2.4).

4. $$\frac{\mathrm{d}}{\mathrm{d}x} \int_a^x f(t) \, \mathrm{d}t = f(x) \ , \quad \frac{\mathrm{d}}{\mathrm{d}x} \int_x^b f(t) \, \mathrm{d}t = -f(x)$$

   (Theorem 13.6.12, Remark 13.6.8).

5. $$\int_a^b f(x) \, \mathrm{d}x = F(b) - F(a)$$ (Theorem 13.6.13).

6. $$\int_a^b [c_1 f_1(x) + c_2 f_2(x)] \, \mathrm{d}x = c_1 \int_a^b f_1(x) \, \mathrm{d}x + c_2 \int_a^b f_2(x) \, \mathrm{d}x$$

(Theorem 13.6.3).

7. $\int_a^b u'v\,\mathrm{d}x = [uv]_a^b - \int_a^b uv'\,\mathrm{d}x$ or $\int_a^b v\,\mathrm{d}u = [uv]_a^b - \int_a^b u\,\mathrm{d}v$

(integration by parts, Theorem 13.7.1).

8. $\int_a^b g(h(x))\,h'(x)\,\mathrm{d}x = \int_{h(a)}^{h(b)} g(z)\,\mathrm{d}z$,
$\int_a^b f(x)\,\mathrm{d}x = \int_\alpha^\beta f(\varphi(z))\,\varphi'(z)\,\mathrm{d}z$, where $\varphi(\alpha) = a$, $\varphi(\beta) = b$

(integration by substitution, Theorems 13.7.2, 13.7.3).

9. $\int_b^a f(x)\,\mathrm{d}x = -\int_a^b f(x)\,\mathrm{d}x$.

10. $\int_a^a f(x)\,\mathrm{d}x = 0$.

11. $\int_{-a}^a f(x)\,\mathrm{d}x = 2\int_0^a f(x)\,\mathrm{d}x$ if f is even ,
$= 0$ if f is odd (Remark 13.6.11).

12. $\frac{\mathrm{d}}{\mathrm{d}\alpha}\int_a^b f(x,\alpha)\,\mathrm{d}x = \int_a^b \frac{\partial f}{\partial \alpha}(x,\alpha)\,\mathrm{d}x$ (Theorem 13.9.2).

13. $\frac{\mathrm{d}}{\mathrm{d}\alpha}\int_{\varphi_1(\alpha)}^{\varphi_2(\alpha)} f(x,\alpha)\,\mathrm{d}x = \int_{\varphi_1(\alpha)}^{\varphi_2(\alpha)} \frac{\partial f}{\partial \alpha}(x,\alpha)\,\mathrm{d}x + \varphi_2'(\alpha)\,f(\varphi_2(\alpha),\alpha) -$
$-\varphi_1'(\alpha)\,f(\varphi_1(\alpha),\alpha)$ (Theorem 13.9.4).

14. $\int_a^b f(x)\,\mathrm{d}g(x) = \int_a^b f(x)\,g'(x)\,\mathrm{d}x + \sum_{k=0}^{p} f(x_k)\,s_k$

(Theorem 13.15.3).

14. INTEGRAL CALCULUS OF FUNCTIONS OF TWO AND MORE VARIABLES

By KAREL REKTORYS

References: [4], [26], [31], [54], [59], [68], [91], [96], [112], [119], [122], [123], [127], [142], [146], [148], [158].

When using the concepts of a curve, a surface, a region and a function in this chapter, we suppose that they are of the type defined in §14.1. Closed domains (i.e. which contain the boundary S) are denoted by a bar: $\overline{\Omega} = \Omega + S$.

14.1. Basic Definitions and Notation

Definition 1. By a *simple finite piecewise smooth curve (arc)* in the xy-plane we mean a set of points (x, y) given in parametric form by the equations

$$x = \varphi(t) , \quad y = \psi(t) \qquad (\alpha \leqq t \leqq \beta) , \tag{1}$$

where

1. $\varphi(t)$ and $\psi(t)$ are continuous in $[\alpha, \beta]$ and have piecewise continuous derivatives in $[\alpha, \beta]$;

2. $x'(t)$ and $y'(t)$ do not vanish simultaneously for any $t \in [\alpha, \beta]$ (at points of discontinuity and at the end points of the interval $[\alpha, \beta]$ we understand here by $x'(t)$ and $y'(t)$ the values of their continuous extensions from the right or the left);

3. for any pair $t_1 \neq t_2$ from $[\alpha, \beta]$ (with the possible exception of the pair $t_1 = \alpha$, $t_2 = \beta$), the equations

$$\varphi(t_1) = \varphi(t_2) , \quad \psi(t_1) = \psi(t_2)$$

do not simultaneously hold.

REMARK 1. Condition 3 expresses the *simplicity* of the curve, which means that the curve does not intersect itself. If $\varphi(\alpha) = \varphi(\beta)$ and if at the same time $\psi(\alpha) = \psi(\beta)$, we say that the curve is *closed.*

The terms *curve* and *arc* are not uniformly used in the literature. Often the term *curve* is used for a closed curve and the term *arc* for an open curve.

The continuity of functions φ, ψ which is required in condition 1 expresses the fact that the curve is connected; the piecewise continuous derivatives express that it consists of a finite number of arcs with continuously changing tangent. If the derivatives are continuous *everywhere* (and in the case of a closed curve moreover the relations $\varphi'(\alpha) = \varphi'(\beta)$, $\psi'(\alpha) = \psi'(\beta)$ hold), then the curve has everywhere a continuously changing tangent and is said to be *smooth*. Condition 2 excludes various singularities, e.g. the "curve" $x = a$, $y = b$ (degenerating into a point), etc.

The above mentioned curves are *rectifiable*, i.e. they have a finite length

$$l = \int_{\alpha}^{\beta} \sqrt{[\varphi'^2(t) + \psi'^2(t)]}\, \mathrm{d}t \ .$$

Thus, geometrically, a simple finite piecewise smooth curve is a curve of finite length, not intersecting itself and consisting of a finite number of arcs with continuously changing tangent. In applications, we deal almost exclusively with such types of curves (as far as curves of finite length are considered).

Example 1. The circumference of a square is a simple finite piecewise smooth closed curve. The circumference of a circle or of an ellipse is a simple finite smooth closed curve.

In a similar way a simple finite piecewise smooth curve in three-dimensional space can be defined, given parametrically by the equations

$$x = \varphi(t) \ , \quad y = \psi(t) \ , \quad z = \chi(t) \qquad (\alpha \leqq t \leqq \beta) \ .$$

REMARK 2. For $x = t$ we get from (1) $y = \psi(x)$ and the curve is the graph of the function $y = \psi(x)$.

REMARK 3. A rather more general concept is the so-called *Jordan curve* in the plane which is a simple rectifiable closed curve (not necessarily smooth or piecewise smooth). Every Jordan curve divides the plane into two parts, the bounded part of which is said to be the *interior of the curve* (the so-called *Jordan region*), and the unbounded part the *exterior of the curve*.

Definition 2. Let Ω be a bounded region in the plane (see Definitions 22.1.6 and 22.1.9; Ω need not be simply connected). If its boundary consists of a finite number of simple finite piecewise smooth closed curves, then the region Ω is said to be *of type* A.

A *closed region of type* A is defined similarly.

REMARK 4. In applications we almost exclusively meet regions of type A, as far as bounded regions are considered.

Examples: the square, the polygon, the interior of an ellipse, the annulus, etc.

Definition 3. Let a function $f(x,y)$ be defined on a region Ω of type A. The function $f(x,y)$ is said to be *of type* B in Ω, if it is bounded in Ω and continuous in Ω with the possible exception of a finite number of points or of points constituting a finite number of simple finite piecewise smooth curves. Similarly we define the function $f(x,y)$ of type B in a *closed* region $\bar{\Omega}$ of type A.

REMARK 5. If $\bar{\Omega}$ is a closed region of type A, then obviously every function $f(x,y)$ which is continuous in $\bar{\Omega}$ is of type B in $\bar{\Omega}$.

REMARK 6. Clearly, if $f(x,y)$ is of type B in $\bar{\Omega}$, it is of type B also in Ω. Conversely: If $f(x,y)$ is of type B in Ω and if we define it on the boundary of the region Ω in such a way that it remains bounded (otherwise arbitrarily), then $f(x,y)$ is of type B in $\bar{\Omega}$.

Example 2. The rectangle $\bar{\Omega}$ $(-a \leqq x \leqq a,\ -b \leqq y \leqq b)$ is a closed region of type A. We define the function $f(x,y)$ as follows (Fig. 14.1):

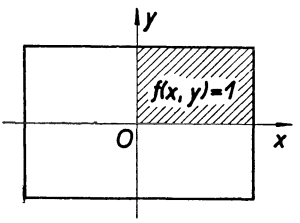


Fig. 14.1.

$$f(x,y) = 1 \quad \text{for} \quad 0 \leqq x \leqq a\ ,\quad 0 \leqq y \leqq b\ ,$$

$$f(x,y) = 0 \quad \text{for the other points of the rectangle } \bar{\Omega}\ .$$

The function $f(x,y)$ is of type B in $\bar{\Omega}$, for it is bounded and continuous everywhere in $\bar{\Omega}$ with exception of the line segments

$$x = 0\ ,\quad 0 \leqq y \leqq b \quad \text{and} \quad y = 0\ ,\quad 0 \leqq x \leqq a\ .$$

REMARK 7. In a manner similar to that of Definition 1, we define a *simple finite piecewise smooth surface*

$$x = \varphi(u, v) , \quad y = \psi(u, v) , \quad z = \chi(u, v) . \tag{2}$$

The interval $[\alpha, \beta]$ in Definition 1 is replaced here by a closed region $\overline{\Omega}$ of type A (Definition 2) in which the functions (2) are continuous and have piecewise continuous (Remark 12.1.8) partial derivatives of the first order. The condition 2 in Definition 1 is replaced by the condition that at no point (u, v) in the region $\overline{\Omega}$ the determinants

$$\left| \begin{matrix} \dfrac{\partial x}{\partial u} , & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} , & \dfrac{\partial y}{\partial v} \end{matrix} \right| , \quad \left| \begin{matrix} \dfrac{\partial x}{\partial u} , & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial z}{\partial u} , & \dfrac{\partial z}{\partial v} \end{matrix} \right| , \quad \left| \begin{matrix} \dfrac{\partial y}{\partial u} , & \dfrac{\partial y}{\partial v} \\ \dfrac{\partial z}{\partial u} , & \dfrac{\partial z}{\partial v} \end{matrix} \right| \tag{3}$$

vanish simultaneously.

REMARK 8. Intuitively: A simple finite piecewise smooth surface is a surface of finite area, not intersecting itself and consisting of a finite number of parts with a continuously changing tangent plane.

REMARK 9. If in (2) $x = u$, $y = v$, we obtain the equation of the surface in the explicit form $z = \chi(x, y)$. This surface is naturally simple, finite and piecewise smooth if the function $\chi(x, y)$ is continuous in $\overline{\Omega}$ and has there piecewise continuous partial derivatives of the first order.

Definition 4. A bounded three-dimensional region Ω (not necessarily simply connected) is said to be *of type* A if its boundary is formed by a finite number of simple finite piecewise smooth closed surfaces. Similarly we define a *closed three-dimensional region of type* A. Instead of "three-dimensional region of type A" we shall often say "*solid of type* A".

Example 3. Example of solids of type A are the cube, the sphere, the ellipsoid, etc.

Definition 5. Let a function $u = f(x, y, z)$ be defined in a solid Ω of type A. If $f(x, y, z)$ is bounded in Ω and, at the same time continuous with the possible exception of a finite number of points or of points constituting a finite number of simple finite piecewise smooth curves or surfaces, we say that $f(x, y, z)$ is *of type* B *in* Ω. Similarly we define the function $f(x, y, z)$ of type B in a *closed* solid $\overline{\Omega}$.

REMARK 10. Remarks similar to Remarks 5 and 6 hold in this case as well.

REMARK 11. *In the following text we use the concepts of a curve, a surface, a region and a function in the above mentioned sense* (we consider regions of type A, functions of type B, etc.).

REMARK 12. If we say that $f(x,y)$ is *continuous on a curve* c, we mean that the function $f(x,y)$ is continuous at each point (x_0,y_0) of the curve c in the usual sense (according to the ε- and δ-definition), with the difference that in the δ-neighbourhood of the point (x_0,y_0) in question we consider only the points of the curve c. Similarly we consider the *continuity of the function* $f(x,y,z)$ *on a surface.*

14.2. The Double Integral

Let a continuous function $z = f(x,y)$ be defined in a closed rectangle $\overline{R}$ $(a \leqq x \leqq b,\ c \leqq y \leqq d)$, let $k \leqq f(x,y) \leqq K$ on $\overline{R}$. Let us divide the interval $[a,b]$ into subintervals $\Delta x_1, \ldots, \Delta x_m$, the interval $[c,d]$ into subintervals $\Delta y_1, \ldots, \Delta y_n$ (as in Fig. 14.2, where we have chosen $m = 4$, $n = 3$). We shall denote the closed rectangle with the base Δx_i and the height Δy_j by $\overline{R}_{ij}$. Let the just constructed partition of the rectangle $\overline{R}$ into rectangles $\overline{R}_{ij}$ be denoted by p. Let us construct the so-called *upper* and *lower integral* (or *Darboux*) *sums* $S(p)$, $s(p)$, corresponding to that partition,

$$S(p) = \sum_{i,j} K_{ij}\,\Delta x_i\,\Delta y_j\ ,\quad s(p) = \sum_{i,j} k_{ij}\,\Delta x_i\,\Delta y_j\ , \tag{1}$$

where K_{ij}, or k_{ij} is the maximal, or the minimal value of the function $f(x,y)$ in $\overline{R}_{ij}$, respectively, and $\Delta x_i\,\Delta y_k$ is the area of the rectangle $\overline{R}_{ij}$. The set of all upper sums for all possible partitions p is bounded below (for the inequality $S(p) \geqq m(b-a)\times$ $\times(d-c)$ always holds). Its greatest lower bound (Definition 1.3.3) is called the *upper double integral* of the function $f(x,y)$ in $\overline{R}$,

$$\inf_p\ S(p) = \overline{\iint_R} f(x,y)\,\mathrm{d}x\,\mathrm{d}y\ ;$$

the least upper bound of the set of all lower sums,

$$\sup_p\ s(p) = \underline{\iint_R} f(x,y)\,\mathrm{d}x\,\mathrm{d}y\ ,$$

is called the *lower double integral.*

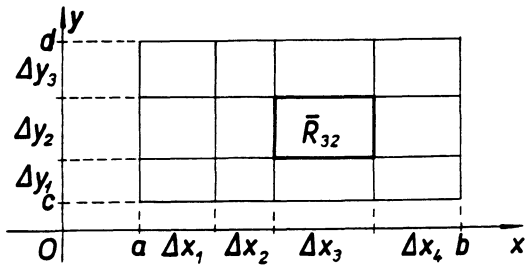


Fig. 14.2

Similarly the upper and the lower integrals are defined in the case where we assume mere boundedness instead of continuity of $f(x, y)$ in $\overline{R}$. Then in the sums (1) K_{ij}, or k_{ij} need not be the maximum, or minimum, but, in general, the least upper bound, or the greatest lower bound of the values of the function $f(x, y)$ in $\overline{R}_{ik}$, respectively.

Definition 1. If the upper and lower integrals are equal, then their common value is called the *double integral of the function* $f(x, y)$ *in* (or *on*, or *over*) *the rectangle* R,

$$\iint_R f(x, y)\,\mathrm{d}x\,\mathrm{d}y\ , \tag{2}$$

and the function $f(x, y)$ is said to be *integrable in* (or *on*, or *over*) R *in the Cauchy-Riemann sense.*

REMARK 1. The geometric meaning of the integral (2) for the case that $f(x, y)$ is continuous and positive in $\overline{R}$: (2) is the volume of the solid whose lower base is formed by the rectangle $\overline{R}$, upper base by the surface $z = f(x, y)$ above the region $\overline{R}$ and the lateral surface by lines parallel to the z-axis (Fig. 14.5, §14.3).

Theorem 1. *Every function* $z = f(x, y)$ *which is continuous in* $\overline{R}$ $(a \leqq x \leqq b,\ c \leqq y \leqq d)$ *is integrable in* R.

Theorem 2. *Every function* $f(x, y)$ *which is of type* B *in* $\overline{R}$ (Definition 14.1.3) *is integrable in* R.

REMARK 2. Theorem 2 offers the possibility of defining the integral for other regions than a rectangular one. Let $\overline{\Omega}$ be a closed region of type A (Definition 14.1.2) and let a function $f(x, y)$ of type B be defined in $\overline{\Omega}$. As Ω is a bounded region, we can construct a closed rectangle $\overline{R}$ $(a \leqq x \leqq b,\ c \leqq y \leqq d)$ in which this region is contained (Fig. 14.3). Let us define the function $g(x, y)$ in $\overline{R}$ as follows:

$$g(x, y) = f(x, y) \quad \text{for} \quad (x, y) \in \overline{\Omega}\ ,$$

$g(x, y) = 0$ for other points of $\overline{R}$ (shaded area in Fig. 14.3). (Thus, the function

$g(x,y)$ coincides with $f(x,y)$ in $\overline{\Omega}$ and at other points of $\overline{R}$ is equal to zero.) We define

$$\iint_{\Omega} f(x,y)\,dx\,dy = \iint_{R} g(x,y)\,dx\,dy\ . \tag{3}$$

Thus, in this sense, every function of type B is integrable on Ω. In particular, every function continuous in $\overline{\Omega}$ is integrable on Ω.

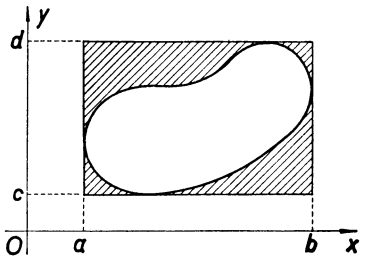


Fig. 14.3.

REMARK 3. The geometric meaning of the integral $\iint_{R} f(x,y)\,dx\,dy$ if $f(x,y)$ is continuous and positive in $\overline{\Omega}$ is that it is the volume of the body the "lower base" of which is formed by the region $\overline{\Omega}$, the "upper base" by the surface $z = f(x,y)$ and the lateral surface by lines parallel to the z-axis.

REMARK 4. The integral

$$\iint_{\Omega} f(x,y)\,dx\,dy$$

may be defined in many other ways different from that used above. Another way of defining the integral is clear from Fig. 14.4. In the sums (1) only such rectangles R_{ij} occur which lie entirely in the region $\overline{\Omega}$ (shaded area in Fig. 14.4).

Other definitions do not use a rectangular network, but divide the region $\overline{\Omega}$ into (small) regions which are in a certain sense arbitrary.

The definition of the *Lebesgue integral* is based, as in the case of the one-dimensional integral, on the definition of the measure of a two-dimensional region. See Remark 13.14.10.

We give now the basic theorems for the Cauchy-Riemann integral. First we define:

Definition 2. By the *norm* $\nu(p)$ of the partition p we mean the greatest of the lengths of the intervals Δx_i, Δy_k,

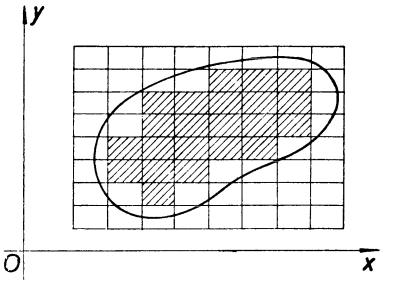


Fig. 14.4.

$$\nu(p) = \max(\Delta x_1, \Delta x_2, \ldots, \Delta x_m, \Delta y_1, \Delta y_2, \ldots, \Delta y_n) .$$

REMARK 5. Let us choose, in every rectangle $\overline{R}_{ij}$, an arbitrary point $P_{ij}(\xi_i, \eta_j)$. Form the sum

$$\sigma(p) = \sum_{i,j} f(\xi_i, \eta_j)\, \Delta x_i\, \Delta y_j$$

depending on the chosen partition p and on the choice of the points (ξ_i, η_j) in $\overline{R}_{ij}$.

Theorem 3. *Let $p_1, p_2, \ldots$ be a sequence of partitions such that*

$$\lim_{n\to\infty} \nu(p_n) = 0 .$$

Then, if $f(x,y)$ is integrable in R, we have

$$\iint_R f(x,y)\, dx\, dy = \lim_{n\to\infty} S(p_n) = \lim_{n\to\infty} s(p_n) = \lim_{n\to\infty} \sigma(p_n) ,$$

i.e. the integral is the limit of the upper sums, of the lower sums, or of the sums σ, respectively, if the norm of the partition converges to zero. Regarding the sums σ, the choice of the point (ξ_i, η_j) in $\overline{R}_{ij}$ is immaterial.

It follows from Remark 2 that a similar theorem holds true for an arbitrary region of type A.

Theorem 4. *If $f(x,y)$ is integrable in a region Ω of type* A (Definition 14.1.2), *then it is integrable also in every part of the region Ω which is of type* A. *If, in particular, $\overline{\Omega} = \overline{\Omega}_1 + \overline{\Omega}_2$, where $\Omega, \Omega_1, \Omega_2$ are regions of type* A *and Ω_1, Ω_2 have no common points, then*

$$\iint_\Omega f(x,y)\, dx\, dy = \iint_{\Omega_1} f(x,y)\, dx\, dy + \iint_{\Omega_2} f(x,y)\, dx\, dy .$$

Theorem 5. *If the functions $f_1(x,y)$, $f_2(x,y)$ are integrable in Ω, then the functions*

$$c_1 f_1 + c_2 f_2 , \quad f_1 f_2 , \quad |f_1|, \ |f_2|$$

are also integrable in Ω *and the relations*

$$\iint_\Omega [c_1 f_1(x,y) + c_2 f_2(x,y)]\,dx\,dy =$$

$$= c_1 \iint_\Omega f_1(x,y)\,dx\,dy + c_2 \iint_\Omega f_2(x,y)\,dx\,dy\ ,$$

$$\left|\iint_\Omega f_1(x,y)\,dx\,dy\right| \leqq \iint_\Omega |f_1(x,y)|\,dx\,dy$$

hold.

The equation

$$\iint_\Omega f_1(x,y)\ f_2(x,y)\,dx\,dy = \iint_\Omega f_1(x,y)\,dx\,dy\ .\ \iint_\Omega f_2(x,y)\,dx\,dy$$

does *not* hold, in general! A sufficient condition for the integrability of the function

$$\frac{f_1(x,y)}{f_2(x,y)}$$

is that the functions f_1 and f_2 are integrable and either

$$0 < k \leqq f_2(x,y)\ ,\quad \text{or}\quad f_2(x,y) \leqq K < 0\ ,$$

i.e. if f_2 is positive and bounded below in Ω by a positive constant or negative and bounded above by a negative constant, respectively.

Theorem 6. *Let* $f(x,y)$, $g(x,y)$ *be integrable in* Ω,

$$k \leqq f(x,y) \leqq K\ ,\quad g(x,y) \geqq 0\ \ \textit{in}\ \ \Omega\ .$$

Then

$$k \iint_\Omega g(x,y)\,dx\,dy \leqq \iint_\Omega f(x,y)\ g(x,y)\,dx\,dy \leqq K \iint_\Omega g(x,y)\,dx\,dy\ .$$

Theorem 7 *(Mean-Value Theorem). If the function* $f(x,y)$ *is continuous in a closed region* $\overline{\Omega}$ *of type* A, *then there exists at least one point* $(x_0, y_0) \in \overline{\Omega}$ *such that*

$$\iint_\Omega f(x,y)\,dx\,dy = P f(x_0, y_0)\ ,$$

where P is the area of the region $\overline{\Omega}$.

Theorem 8. *Neither the integrability of the function nor the value of the integral is changed if the value of the function $f(x,y)$ is changed at a finite number of points or on a finite number of piecewise smooth finite curves.*

14.3. Evaluation of a Double Integral by Repeated Integration

Theorem 1 (often called *Fubini's Theorem*). *Let $\bar{R}$ be a closed rectangle $a \leqq x \leqq b, c \leqq y \leqq d$. If $f(x,y)$ is of type* B *in $\bar{R}$* (see Definition 14.1.3), *then*

$$\iint_R f(x, y)\,dx\,dy = \int_a^b \left[\int_c^d f(x, y)\,dy\right] dx = \int_c^d \left[\int_a^b f(x, y)\,dx\right] dy\,. \qquad (1)$$

REMARK 1. Thus, in particular, Theorem 1 holds for functions continuous in $\bar{R}$. It should be noted that in the general case the function

$$F(x) = \int_c^d f(x, y)\,dy$$

need not be defined for *all* $x \in [a, b]$ (the function $f(x,y)$ need not be integrable as a function of the variable y for all these x). The same holds for the corresponding integral in the last term in (1).

REMARK 2. In accordance with (1) the integral

$\iint_R f(x, y)\,dx\,dy$ is often denoted by $\int_a^b\int_c^d f(x, y)\,dx\,dy$ or $\int_c^d\int_a^b f(x, y)\,dy\,dx$.

Instead of

$$\int_a^b \left[\int_c^d f(x, y)\,dy\right] dx \quad \text{or} \quad \int_c^d \left[\int_a^b f(x, y)\,dx\right] dy$$

we employ the notation

$$\int_a^b dx \int_c^d f(x, y)\,dy \quad \text{or} \quad \int_c^d dy \int_a^b f(x, y)\,dx\,,$$

respectively.

REMARK 3. *Geometric interpretation* of Theorem 1 (Fig. 14.5; for simplicity let us suppose that $f(x, y)$ is continuous and positive in $\bar{R}$; see Remark 14.2.1). For a constant x, the integral

$$\int_c^d f(x, y)\,dy = F(x)$$

(where we have integrated only with respect to y) gives the area of the cross-section of the solid shown in Fig. 14.5. For small Δx,

$$F(x)\,\Delta x \quad \text{or} \quad \Delta x \int_c^d f(x, y)\,dy$$

denotes the volume of a small layer of the solid under consideration. Summing up the volumes of these layers ($\Sigma F(x_i)\,\Delta x_i$) we get the approximate volume of the solid. By going to the limit, this sum changes into the integral.

We arrive at the same result (see the second of equations (1)) if we "cut" the solid into layers by the planes $y = \text{const}$.

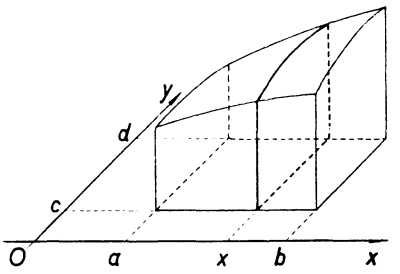


Fig. 14.5.

REMARK 4. The integrals on the right-hand side of equation (1) are called *repeated integrals*. Theorem 1 is a generalization of Theorem 13.9.3. .

REMARK 5. According to Remark 14.2.2, Theorem 1 holds for arbitrary closed regions $\bar{\Omega}$ of type A and for functions $f(x, y)$ of type B in $\bar{\Omega}$. Usually we deal with the following simple case: Let $y = h_1(x)$, $y = h_2(x)$ be continuous functions with continuous or piecewise continuous derivatives in $[a, b]$ such that $h_2(x) > h_1(x)$ in (a, b). Let $k = \min h_1(x)$ in $[a, b]$, $K = \max h_2(x)$ in $[a, b]$. Let $\bar{\Omega}$ be the region (obviously of type A) the boundary of which is formed by the functions $y = h_1(x)$, $y = h_2(x)$ and the lines $x = a$, $x = b$ parallel to the y-axis. Denote the rectangle

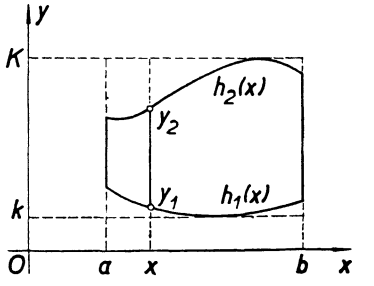


Fig. 14.6.

$a \leqq x \leqq b$, $k \leqq y \leqq K$ by $\bar{R}$ (Fig. 14.6). By equation (3) of Remark 14.2.2 and Theorem 1, we have

$$\iint_{\Omega} f(x, y)\,\mathrm{d}x\,\mathrm{d}y = \iint_{R} g(x, y)\,\mathrm{d}x\,\mathrm{d}y = \int_a^b \left[\int_k^K g(x, y)\,\mathrm{d}y\right]\mathrm{d}x\,.$$

The definition of the function $g(x, y)$ (see Remark 14.2.2), however, implies

$$\int_k^K g(x, y)\,\mathrm{d}y = \int_{y_1}^{y_2} f(x, y)\,\mathrm{d}y\,, \quad \text{where} \quad y_1 = h_1(x)\,, \quad y_2 = h_2(x)$$

(Fig. 14.6). Thus

$$\iint_{\Omega} f(x, y)\,\mathrm{d}x\,\mathrm{d}y = \int_a^b \left[\int_{h_1(x)}^{h_2(x)} f(x, y)\,\mathrm{d}y\right]\mathrm{d}x\,. \tag{2}$$

Similarly: If the boundary of the region Ω consists of the curves $x = \varphi_1(y)$, $x = \varphi_2(y)$ (with similar properties as before) and of the lines $y = c$, $y = d$ parallel to the x-axis then

$$\iint_{\Omega} f(x, y)\,\mathrm{d}x\,\mathrm{d}y = \int_c^d \left[\int_{\varphi_1(y)}^{\varphi_2(y)} f(x, y)\,\mathrm{d}x\right]\mathrm{d}y\,. \tag{3}$$

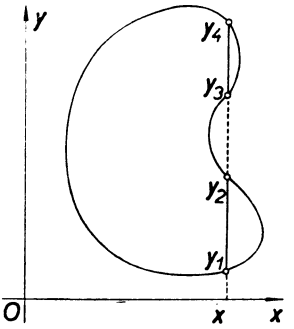


Fig. 14.7.

We can go on similarly if the boundary of the region Ω is more complicated. For example, for x chosen as in Fig. 14.7, we have

$$\int_c^d g(x, y)\,\mathrm{d}y = \int_{y_1}^{y_2} f(x, y)\,\mathrm{d}y + \int_{y_3}^{y_4} f(x, y)\,\mathrm{d}y\,.$$

In cases like this we evaluate the limits $y_1, y_2, \ldots$ from the equation of the boundary of the region Ω (see Example 2).

Example 1. Determine the volume V of the ellipsoid with semi-axes a, b, c.

The volume V is equal to twice the volume of the upper half-ellipsoid which is bounded by the plane $z = 0$ and by the surface

$$z = c\sqrt{\left(1 - \frac{x^2}{a^2} - \frac{y^2}{b^2}\right)}, \tag{4}$$

the equation of which is obtained from the equation of the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

for $z \geqq 0$.

The base Ω of the half-ellipsoid is bounded by the ellipse

$$1 - \frac{x^2}{a^2} - \frac{y^2}{b^2} = 0 \tag{5}$$

the equation of which is obtained from (4) for $z = 0$ (because the base lies in this plane). Thus

$$V = 2\iint_{\Omega} c\sqrt{\left(1 - \frac{x^2}{a^2} - \frac{y^2}{b^2}\right)}\,\mathrm{d}x\,\mathrm{d}y =$$

$$= 2c\int_{-a}^{a}\left[\int_{-b\sqrt{(1-x^2/a^2)}}^{b\sqrt{(1-x^2/a^2)}}\sqrt{\left(1 - \frac{x^2}{a^2} - \frac{y^2}{b^2}\right)}\,\mathrm{d}y\right]\mathrm{d}x\,. \tag{6}$$

Using (1) we transformed the double integral into a repeated integral. The limits y_1, y_2 have been determined (for a given x) from (5) (see Fig. 14.8).

In (6), we integrate with respect to y keeping x *fixed*. Writing $\sqrt{(1 - x^2/a^2)} = m \geqq 0$, we obtain (using the substitution $y = bm\sin t$, $\mathrm{d}y = bm\cos t\,\mathrm{d}t$; see Example 13.2.10)

$$\int_{-b\sqrt{(1-x^2/a^2)}}^{b\sqrt{(1-x^2/a^2)}}\sqrt{\left(1 - \frac{x^2}{a^2} - \frac{y^2}{b^2}\right)}\,\mathrm{d}y = \int_{-bm}^{bm}\sqrt{\left(m^2 - \frac{y^2}{b^2}\right)}\,\mathrm{d}y = bm^2\int_{-\pi/2}^{\pi/2}\cos^2 t\,\mathrm{d}t =$$

$$= \tfrac{1}{2}\pi bm^2 = \tfrac{1}{2}\pi b\left(1 - \frac{x^2}{a^2}\right).$$

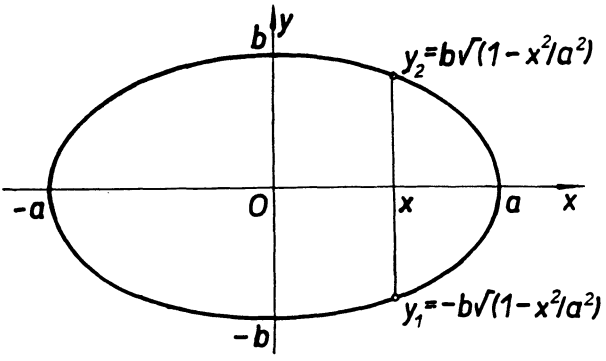


Fig. 14.8.

Substituting into (6), we have

$$V = 2c \int_{-a}^{a} \tfrac{1}{2}\pi b \left(1 - \frac{x^2}{a^2}\right) \mathrm{d}x = \pi bc \left[x - \frac{x^3}{3a^2}\right]_{-a}^{a} = \tfrac{4}{3}\pi abc\,.$$

REMARK 6. It would have been *wrong* to proceed in the following way:

$$\iint_{\Omega} \sqrt{\left(1 - \frac{x^2}{a^2} - \frac{y^2}{b^2}\right)} \,\mathrm{d}x\,\mathrm{d}y = \int_{-a}^{a} \left[\int_{-b}^{b} \sqrt{\left(1 - \frac{x^2}{a^2} - \frac{y^2}{b^2}\right)} \mathrm{d}y\right] \mathrm{d}x\,,$$

for according to Remark 14.2.2 $g(x, y) = 0$ if the point (x, y) does not lie in Ω. Here, of course, this circumstance is obvious from the fact that at these points the function $f(x, y)$ ceases to have meaning as a real function. For example, at the point with coordinates $x = a$, $y = b$ we would have got

$$\sqrt{\left(1 - \frac{x^2}{a^2} - \frac{y^2}{b^2}\right)} = \sqrt{(1 - 1 - 1)} = \sqrt{(-1)}\,.$$

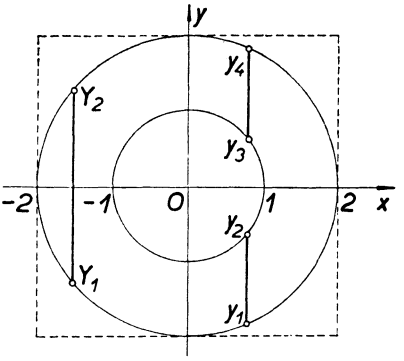


Fig. 14.9.

Example 2. Somewhat more difficult is to determine the limits of integration in the case of the integration of a function $f(x, y)$ of type B in the annulus Ω shown in Fig. 14.9. For the inner radius r and outer radius R we have $r = 1$, $R = 2$. If $-2 < x < -1$ or $1 < x < 2$, we obtain, for a chosen x, two values Y_1, Y_2:

$$Y_1 = -\sqrt{(4 - x^2)}\,, \quad Y_2 = \sqrt{(4 - x^2)}\,.$$

If, however, $-1 < x < 1$, then for a given x we have four values y_1, y_2, y_3, y_4:

$$y_1 = -\sqrt{(4 - x^2)}\,, \quad y_2 = -\sqrt{(1 - x^2)}\,, \quad y_3 = \sqrt{(1 - x^2)}\,, \quad y_4 = \sqrt{(4 - x^2)}\,.$$

Thus

$$\iint_{\Omega} f(x, y)\,\mathrm{d}x\,\mathrm{d}y = \int_{-2}^{-1} \left[\int_{-\sqrt{(4-x^2)}}^{\sqrt{(4-x^2)}} f(x, y)\,\mathrm{d}y\right] \mathrm{d}x +$$

$$+\int_{-1}^{1}\left[\int_{-\sqrt{(4-x^2)}}^{-\sqrt{(1-x^2)}} f(x, y)\,dy + \int_{\sqrt{(1-x^2)}}^{\sqrt{(4-x^2)}} f(x, y)\,dy\right]dx + \int_{1}^{2}\left[\int_{-\sqrt{(4-x^2)}}^{\sqrt{(4-x^2)}} f(x, y)\,dy\right]dx .$$

14.4. Method of Substitution for Double Integrals

Theorem 1. *Let the closed region $\bar{N}$ (of variables u, v) be mapped in a one-to-one correspondence by the equations*

$$x = x(u, v), \quad y = y(u, v) \tag{1}$$

on the closed region $\bar{M}$ (of variables x, y). Let $\bar{M}$ and $\bar{N}$ be of type A (see Definition 14.1.2) *and $f(x, y)$ of type* B *in $\bar{M}$* (see Definition 14.1.3). *If the functions $x(u, v)$, $y(u, v)$ have continuous first partial derivatives in $\bar{N}$ and the Jacobian*

$$D(u, v) = \begin{vmatrix} \dfrac{\partial x}{\partial u}, & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u}, & \dfrac{\partial y}{\partial v} \end{vmatrix} \tag{2}$$

in $\bar{N}$ is different from zero, then

$$\iint_{M} f(x, y)\,dx\,dy = \iint_{N} f(x(u, v), y(u, v))\,|D(u, v)|\,du\,dv . \tag{3}$$

REMARK 1. Note that the *absolute value* of $D(u, v)$ appears in (3). For the most frequently used substitution

$$x = \rho\cos\varphi, \quad y = \rho\sin\varphi \tag{4}$$

(polar coordinates), the Jacobian is

$$D(\rho, \varphi) = \begin{vmatrix} \dfrac{\partial x}{\partial \rho}, & \dfrac{\partial x}{\partial \varphi} \\ \dfrac{\partial y}{\partial \rho}, & \dfrac{\partial y}{\partial \varphi} \end{vmatrix} = \begin{vmatrix} \cos\varphi, & -\rho\sin\varphi \\ \sin\varphi, & \rho\cos\varphi \end{vmatrix} = \rho \geqq 0 \tag{5}$$

so that, in this case, it is not necessary to pay special attention to the absolute value.

Example 1. Let $f(x, y)$ be of type B in the closed circle $\bar{M}$, $x^2 + y^2 \leqq 25$. We make use of the substitution (4); ρ will run between the limits 0, 5, φ between the limits 0, 2π (Fig. 14.10). Let us investigate if equation (3), i. e. the equation

$$\iint_M f(x, y)\,dx\,dy = \int_0^5 \int_0^{2\pi} f(\rho\cos\varphi, \rho\sin\varphi)\,\rho\,d\rho\,d\varphi \tag{6}$$

holds. The assumptions of Theorem 1 are not satisfied. For, firstly, if the mapping has to be one-to-one for $\rho > 0$, then φ has to run through the interval $[0, 2\pi)$ and not through the interval $[0, 2\pi]$ so that $\bar{N}$ is not a closed region. Secondly, to the point $(0, 0) \in \bar{M}$ there corresponds in $\bar{N}$ the entire line segment $r = 0$, $0 \leqq \varphi < 2\pi$. Moreover, $D(\rho, \varphi) = 0$ for $\rho = 0$. In spite of this we can show that (6) holds. To the shaded sector S of the annulus in Fig. 14.11 there corresponds the closed rectangle $\sigma \leqq \rho \leqq 5, 0 \leqq \varphi \leqq 2\pi - \delta$ $(\sigma > 0,\ \delta > 0)$. Now, all the assumptions of Theorem 1 are satisfied, so that the relation

$$\iint_S f(x, y)\,dx\,dy = \int_\sigma^5 \int_0^{2\pi-\delta} f(\rho\cos\varphi, \rho\sin\varphi)\,\rho\,d\rho d\varphi \tag{7}$$

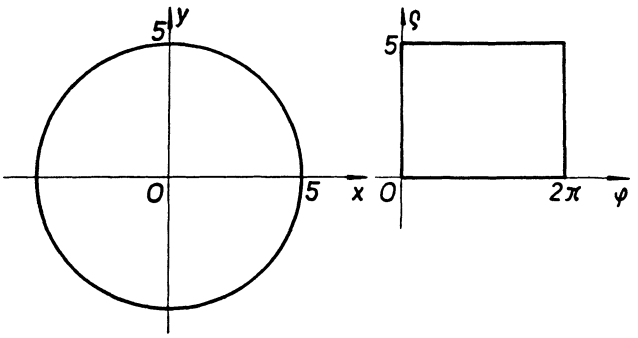


Fig. 14.10.

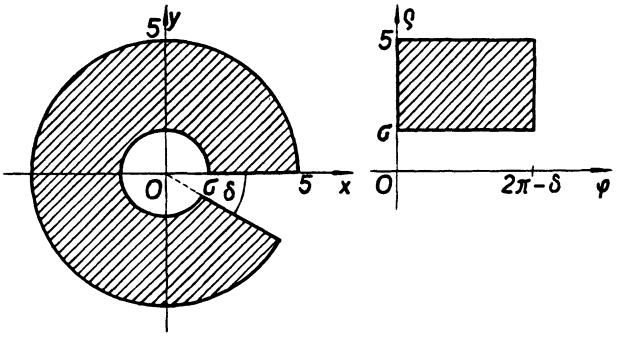


Fig. 14.11.

holds. Since $f(x, y)$ is of type B it is bounded in $\bar{M}$, hence the integral $\iint_S$ differs from $\iint_M$ by as little as we please if δ and σ are sufficiently small. Letting $\delta \to 0$ and then $\sigma \to 0$, we obtain (6). Thus, if $f(x, y)$ is of type B, then equation (6) holds.

Example 2. Let us determine the moment of inertia I of the homogeneous sphere of constant density ϱ_0 and of radius R with respect to an axis passing through its centre.

We choose the centre of the sphere as the origin of the coordinate system and the axis mentioned as the z-axis. The equation of the spherical surface is

$$x^2 + y^2 + z^2 = R^2$$

whence, for the upper hemispherical surface, we have

$$z = \sqrt{(R^2 - x^2 - y^2)}\,.$$

The moment of inertia I with respect to the z-axis will be equal to twice the moment of the upper hemisphere. Thus (by formula (14.9.70), p. 628), we have

$$I = 2\varrho_0 \iint_M (x^2 + y^2)\sqrt{(R^2 - x^2 - y^2)}\,\mathrm{d}x\,\mathrm{d}y\,.$$

We integrate over the circle $\overline{M}$ with centre at the origin and with radius R.

Using polar coordinates

$$x = \rho\cos\varphi\,, \quad y = \rho\sin\varphi \quad (0 \leqq \rho \leqq R,\ 0 \leqq \varphi < 2\pi)\,, \tag{8}$$

we have by (6)

$$I = 2\varrho_0 \int_0^R \int_0^{2\pi} \rho^2 \sqrt{(R^2 - \rho^2)}\,\rho\,\mathrm{d}\rho\mathrm{d}\varphi = 2\varrho_0 \int_0^R \left(\int_0^{2\pi} \rho^2 \sqrt{(R^2 - \rho^2)}\,\rho\,\mathrm{d}\varphi \right) \mathrm{d}\rho =$$

$$= 4\pi\varrho_0 \int_0^R \rho^2 \sqrt{(R^2 - \rho^2)}\,\rho\,\mathrm{d}\rho = 4\pi\varrho_0 \int_0^R (R^2 - t^2)\,t^2\,\mathrm{d}t =$$

$$= 4\pi\varrho_0 \left[R^2 \frac{t^3}{3} - \frac{t^5}{5} \right]_0^R = \frac{8\pi\varrho_0}{15} R^5\,,$$

where we have used the substitution $\sqrt{(R^2 - \rho^2)} = t$.

REMARK 2. In Example 2, we integrated over the *circle* and used substitution (8). Often we integrate over the interior of the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1\,, \quad a > 0\,, \quad b > 0\,. \tag{9}$$

Then, substitution (8) is not convenient as the upper limit for r turns out to be dependent on φ. In this case we use the substitution

$$x = ar\cos\varphi\,, \quad y = br\sin\varphi \quad (0 \leqq r \leqq 1,\ 0 \leqq \varphi < 2\pi) \tag{10}$$

(r having here a constant upper limit, equal to one, since for $r = 1$ equations (10) give just the parametric equations of the ellipse (9); if $r < 1$, then $x^2/a^2 + y^2/b^2 < 1$ and the point (x, y) lies inside the ellipse, for $r > 1$ outside the ellipse.) For the Jacobian (2) we obtain

$$D = abr\,. \tag{11}$$

The remarks made about the justification of equation (6) apply here also (see Example 1).

Example 3. For the volume V_0 of the upper semiellipsoid with the semi-axes a, b, c we obtain by (10) and (11)

$$V_0 = \iint_\Omega c\sqrt{\left(1 - \frac{x^2}{a^2} - \frac{y^2}{b^2}\right)}\,\mathrm{d}x\,\mathrm{d}y = c\int_0^1\int_0^{2\pi}\sqrt{(1-r^2)}\,abr\,\mathrm{d}r\,\mathrm{d}\varphi =$$

$$= abc \times 2\pi\int_0^1 t^2\,\mathrm{d}t = \tfrac{2}{3}\pi abc\,.$$

(We made use of the substitution $1 - r^2 = t^2$, $t > 0$.) Thus, the total volume V is

$$V = \tfrac{4}{3}\pi abc\,.$$

The evaluation is obviously easier than in Example 14.3.1.

14.5. Triple Integrals

Triple integrals are defined in a way similar to double integrals: Let a bounded function $u = f(x, y, z)$ be given in a rectangular parallelepiped $\overline{Q}$ ($a \leqq x \leqq b$, $c \leqq y \leqq d$, $e \leqq z \leqq f$). We divide this parallelepiped by planes parallel to the coordinate planes into small parallelepipeds and denote the least upper bound and greatest lower bound of the function $f(x, y, z)$ on the parallelepipeds $\Delta x_i\,\Delta y_k\,\Delta z_l$ by M_{ikl} and m_{ikl}, respectively. (If $f(x, y, z)$ is continuous in $\overline{Q}$, then the least upper bound is also the maximum and the greatest lower bound is also the minimum of the values of the function.) We construct the *upper* and *lower integral sums*

$$S(p) = \sum M_{ikl}\,\Delta x_i\,\Delta y_k\,\Delta z_l\,,\quad s(p) = \sum m_{ikl}\,\Delta x_i\,\Delta y_k\,\Delta z_l\,,$$

respectively, summing up over all the parallelepipeds into which the parallelepiped Q is divided by the chosen partition p. The greatest lower bound of the set of all upper sums (for all possible partitions p) is called the *upper integral*, the least upper bound of the set of all lower sums – the *lower integral of the function u in* (or *over*) Q.

Definition 1. If the upper and lower integrals are equal, then the function $u = f(x,y,z)$ is said to be *integrable in* (or *over*) Q *in the Cauchy-Riemann sense* and their common value is called the *triple integral of the function* f *in the parallelepiped* Q. We write

$$\iiint_Q f(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\ , \quad \text{often also} \quad \iiint_Q f(x,y,z)\,\mathrm{d}V\ .$$

The integral of a function of more than three variables may be defined like-wise.

REMARK 1. The interpretation of the triple integral if $f(x,y,z) > 0$: the mass of the parallelepiped Q the density of which is given by the function $f(x,y,z)$.

REMARK 2. The basic properties of triple integrals (or of higher-dimensional integrals) are similar to those of double integrals (see Theorems 14.2.3–14.2.8).

Theorem 1. *Every function of type B in* $\overline{Q}$ (Definition 14.1.5) *is integrable in* Q.

REMARK 3. In particular, any function continuous in $\overline{Q}$ is integrable in Q.

REMARK 4. If $f(x,y,z)$ is defined in a region $\overline{\Omega}$ which is a closed region of type A (Definition 14.1.4) but not a (rectangular) parellelepiped, then similarly as in Remark 14.2.2 we define

$$\iiint_\Omega f(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = \iiint_Q g(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\ , \tag{1}$$

where Q is a parallelepiped containing the region $\overline{\Omega}$, $g(x,y,z) = f(x,y,z)$ for all points $(x,y,z) \in \overline{\Omega}$ and $g(x,y,z) = 0$ for the other points of Q.

REMARK 5. The triple integral can be defined also in many other ways. In particular it is possible, using the concept of a measurable set, to define the triple (or multi-dimensional) integral *in the Lebesgue sense*. See §13.14.

Theorem 2. *If* $f(x,y,z)$ *is of type* B (Definition 14.1.5) *in the parallelepiped* $\overline{Q}$ $(a \leqq x \leqq b,\ c \leqq y \leqq d,\ e \leqq z \leqq f)$, *then*

$$\iiint_Q f(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = \int_e^f \left[\int_a^b \int_c^d f(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\right] \mathrm{d}z =$$

$$= \int_a^b \int_c^d \left[\int_e^f f(x,y,z)\,\mathrm{d}z\right] \mathrm{d}x\,\mathrm{d}y = \int_a^b \left[\int_c^d \int_e^f f(x,y,z)\,\mathrm{d}y\,\mathrm{d}z\right] \mathrm{d}x =$$

$$= \int_c^d \int_e^f \left[\int_a^b f(x,y,z)\,\mathrm{d}x\right] \mathrm{d}y\,\mathrm{d}z = \int_c^d \left[\int_a^b \int_e^f f(x,y,z)\,\mathrm{d}x\,\mathrm{d}z\right] \mathrm{d}y =$$

$$= \int_a^b \int_e^f \left[\int_c^d f(x,y,z)\,dy \right] dx\,dz = \int_a^b \left\{ \int_c^d \left[\int_e^f f(x,y,z)\,dz \right] dy \right\} dx\,, \quad (2)$$

where in the last integral it is again possible to change the order of integration (see, however, also Remark 14.3.1).

REMARK 6. If $f(x,y,z)$ is defined in the domain $\overline{\Omega}$ which need not be a prism, then the evaluation of this integral according to (1) and (2) requires careful determination of the limits of integration in the same way as in Remark 14.3.5. Here also we often meet the following case: In a closed region $\overline{G}$ of type A (Definition 14.1.2) there are given smooth or continuous and piecewise smooth (Remark 12.1.8, p. 405) functions $z = z_1(x,y)$, $z = z_2(x,y)$ such that throughout G we have $z_2(x,y) > z_1(x,y)$. Let $\overline{\Omega}$ be the closed solid with "upper base" $z = z_2(x,y)$ and with "lower base" $z = z_1(x,y)$. The lateral surface is formed by lines parallel to the z-axis passing through the boundary of the domain $\overline{G}$ (Fig. 14.12). Thus, the domain $\overline{\Omega}$ is a closed solid of type A (Definition 14.1.4). Let a function of type B (Definition 14.1.5) be given in $\overline{\Omega}$. Then

$$\iiint_\Omega f(x,y,z)\,dx\,dy\,dz = \iint_G \left[\int_{z_1(x,y)}^{z_2(x,y)} f(x,y,z)\,dz \right] dx\,dy \quad (3)$$

as can easily be derived from (1) and from the second of the relations (2) (see also Fig. 14.12).

Another procedure, based on the first of the relations (2), is as follows: $\overline{\Omega}$ lies between the planes $z = e$, $z = f$. Let us denote by $\overline{K}_{z_0}$ the cross-section of the solid $\overline{\Omega}$ by the plane $z = z_0$ ($e \leqq z_0 \leqq f$). Then

$$\iiint_\Omega f(x,y,z)\,dx\,dy\,dz = \int_e^f \left[\iint_{K_z} f(x,y,z)\,dx\,dy \right] dz \quad (4)$$

(Fig. 14.13, where Ω is a sphere).

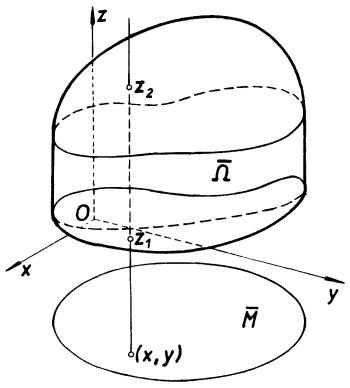


Fig. 14.12.

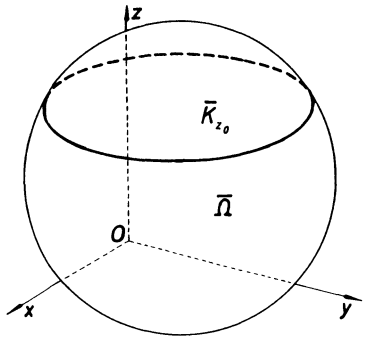


Fig. 14.13.

Example 1. Let us determine the mass m of the sphere Ω whose surface has the equation

$$x^2 + y^2 + z^2 = R^2 \ , \tag{5}$$

and whose density increases with the square of the distance from the z-axis, i.e. $\rho = (x^2 + y^2)\,\rho_0$, where ρ_0 is a constant. By Remark 1

$$m = \iiint_\Omega \rho_0 (x^2 + y^2)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \ . \tag{6}$$

By Remark 6 we have

$$\iiint_\Omega \rho_0 (x^2 + y^2)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = \iint_G \Big(\int_{z_1}^{z_2} \rho_0 (x^2 + y^2)\,\mathrm{d}z\Big)\,\mathrm{d}x\,\mathrm{d}y \ ,$$

where G is a circle with centre at the origin and radius R, $z_2 = \sqrt{(R^2 - x^2 - y^2)}$, $z_1 = -\sqrt{(R^2 - x^2 - y^2)}$. After integration with respect to z and substitution of the limits we obtain

$$m = \rho_0 \iint_G (x^2 + y^2) \times 2\,\sqrt{(R^2 - x^2 - y^2)}\,\mathrm{d}x\,\mathrm{d}y \ .$$

This integral can be evaluated either by repeated integration in the same way as in Example 14.3.1, or by transformation to polar coordinates as in Example 14.4.2. Result: $m = \frac{8}{15}\,\pi\rho_0 R^5$ (cf. Example 2).

REMARK 7. The given problem is evidently equivalent to that of the determination of the moment of inertia of the homogeneous sphere of density ρ_0 with respect to the z-axis. Note that in Example 1 ρ_0 has the dimension kg m^{-5} as follows from the equation $\rho = (x^2 + y^2)\,\rho_0$.

Theorem 3 *(The Method of Substitution for Triple Integrals). Let $\overline{M}$ and $\overline{N}$ be closed regions of type* A (Definition 14.1.4) *in variables x, y, z and u, v, w, respectively. Let there exist a one-to-one mapping between $\overline{M}$ and $\overline{N}$ expressed by the equations*

$$x = x(u, v, w) \ , \quad y = y(u, v, w) \ , \quad z = z(u, v, w) \tag{7}$$

and assume that the functions (7) *have continuous partial derivatives of first order in $\overline{M}$ and that the Jacobian*

$$D(u, v, w) = \begin{vmatrix} \dfrac{\partial x}{\partial u} , & \dfrac{\partial x}{\partial v} , & \dfrac{\partial x}{\partial w} \\ \dfrac{\partial y}{\partial u} , & \dfrac{\partial y}{\partial v} , & \dfrac{\partial y}{\partial w} \\ \dfrac{\partial z}{\partial u} , & \dfrac{\partial z}{\partial v} , & \dfrac{\partial z}{\partial w} \end{vmatrix}$$

is different from zero in $\overline{N}$. *Further, Let* $f(x,y,z)$ *be of type* B (Definition 14.1.5) *in* $\overline{M}$. *Then*

$$\iiint_M f(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z =$$

$$= \iiint_N f(x(u,v,w),\ y(u,v,w),\ z(u,v,w))\ |D(u,v,w)|\,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}w\ . \tag{8}$$

REMARK 8. The transformation theorem can be extended under similar assumptions to n-dimensional integrals. The Jacobian will then be n-dimensional.

REMARK 9. The transformation to spherical coordinates given by the equations

$$x = r\sin\vartheta\cos\varphi\ ,\quad y = r\sin\vartheta\sin\varphi\ ,\quad z = r\cos\vartheta \tag{9}$$

(§6.1) is used very frequently. These coordinates are employed when integrating over a sphere with centre at the origin and radius R (then $0 \leqq r \leqq R, 0 \leqq \varphi < 2\pi, 0 \leqq \vartheta \leqq \pi$), over a hemisphere (if we deal with the "upper" hemisphere, then $0 \leqq \vartheta \leqq \frac{1}{2}\pi$) etc. The Jacobian is

$$D(r,\varphi,\vartheta) = r^2\sin\vartheta\ .$$

It is possible to establish by a limiting process similar to that used in Example 14.4.1 the validity of the equation

$$\iiint_K f(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z =$$

$$= \int_0^R\int_0^{2\pi}\int_0^{\pi} f(r\sin\vartheta\cos\varphi,\ r\sin\vartheta\sin\varphi,\ r\cos\vartheta)\,r^2\sin\vartheta\,\mathrm{d}r\,\mathrm{d}\varphi\,\mathrm{d}\vartheta\ , \tag{10}$$

where K is a sphere with centre at the origin and radius R. We assume that $f(x,y,z)$ is of type B in K.

REMARK 10. If we integrate over the ellipsoid with semi-axes a, b, c, we use instead of (9) the substitution

$$x = ar\sin\vartheta\cos\varphi\ ,\quad y = br\sin\vartheta\sin\varphi\ ,\quad z = cr\cos\vartheta \tag{11}$$

with the Jacobian $D = abc\,r^2\sin\vartheta$. In this case $0 \leqq r \leqq 1, 0 \leqq \varphi < 2\pi, 0 \leqq \vartheta \leqq \pi$.

REMARK 11. For substitutions (9) and (11) $D \geqq 0$. In the general case, the absolute value of D appears in (8).

Example 2. By transforming to spherical coordinates we can easily evaluate integral (6) from Example 1. We have

$$x^2 + y^2 = r^2\sin^2\vartheta\cos^2\varphi + r^2\sin^2\vartheta\sin^2\varphi = r^2\sin^2\vartheta\ ,$$

thus by (10),

$$\iiint_O \varrho_0(x^2 + y^2)\,dx\,dy\,dz = \int_0^R \int_0^{2\pi} \int_0^{\pi} \varrho_0 r^2 \sin^2 \vartheta \,.\, r^2 \sin\vartheta\,dr\,d\varphi\,d\vartheta =$$

$$= 2\pi\varrho_0 \int_0^R r^4 \left(\int_0^{\pi} \sin^3 \vartheta\,d\vartheta \right) dr = 2\pi\varrho_0 \,.\, \tfrac{4}{3} \,.\, \tfrac{1}{5} R^5 = \tfrac{8}{15}\pi\varrho_0 R^5 \,.$$

14.6. Improper Double and Triple Integrals

By the term *improper* integrals we mean integrals where either the integrand, or the domain of integration is unbounded. The simplest cases are those of improper double integrals, where the integrand is unbounded in the neighbourhood of only one point or where the domain of integration is the entire xy-plane.

Definition 1. Let $f(x, y)$ be defined in the closed region $\bar{\Omega}$ of type A (Definition 14.1.2) and let it become unbounded in the neighbourhood of the point $(x_0, y_0) \in \Omega$. At the point (x_0, y_0) itself f need not be defined. Assume, further, that $f(x, y)$ has the following property: If we remove from $\bar{\Omega}$ an arbitrary (open) region ω of type A ($\bar{\omega} \in \Omega$) containing the point (x_0, y_0), then $f(x, y)$ is of type B (Definition 14.1.3) in $\bar{\Omega} - \omega$ (and thus, the integral

$$I_\omega = \iint_{\Omega - \omega} f(x, y)\,dx\,dy \tag{1}$$

exists). If there exists a number I such that for every $\varepsilon > 0$ it is possible to find a rectangle R so small that for all regions ω with the above mentioned property, lying inside R (and containing the point (x_0, y_0)), the inequality

$$|I_\omega - I| < \varepsilon$$

holds, then the integral

$$\iint_\Omega f(x, y)\,dx\,dy \tag{2}$$

is said to be *convergent* and to have the value I.

REMARK 1. Similarly we define the three-dimensional improper integral (or the m-dimensional improper integral) for the case of one singular point (x_0, y_0, z_0).

Theorem 1. *The integral* (2) *is convergent if and only if* $\iint |f(x, y)|\,dx\,dy$ *is convergent.* (*The convergence of the integral is absolute unlike the one-dimensional case of improper integrals.*)

Theorem 2. *If in a given neighbourhood of the point* (x_0, y_0) *we have*

$$|f(x, y)| \leqq \frac{M}{r^\alpha} \tag{3}$$

where M *is a positive constant,* $r = \sqrt{[(x - x_0)^2 + (y - y_0)^2]}$ (i.e. *the distance*

of the point (x, y) *from the point* (x_0, y_0)) *and* $\alpha < 2$, *then the integral*

$$I = \iint_{\Omega} f(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

is convergent and

$$I = \lim_{n\to\infty} \iint_{\Omega-\omega_n} f(x, y)\,\mathrm{d}x\,\mathrm{d}y\,, \tag{4}$$

where ω_n *is an arbitrary sequence of circles contained in* Ω *having centres at the point* (x_0, y_0) *and radius* r_n, *with* $\lim_{n\to\infty} r_n = 0$.

Moreover, under the assumption (3) *the theorem for the replacement of the double integral by the repeated integral* (Theorem 14.3.1) *and the theorem on substitution* (Theorem 14.4.1) *are valid.*

REMARK 2. A similar assertion holds for triple (or m-dimensional) integrals; it is sufficient to require $\alpha < 3$ (or $\alpha < m$, respectively); in the three-dimensional case r is given by the expression

$$r = \sqrt{[(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2]}$$

(and similarly in the m-dimensional case).

REMARK 3. The advantage of using equation (4) lies in the fact that the integrals (4) can be easily evaluated as the domains ω_n are circles.

REMARK 4. In Theorems 1 and 2 is understood, of course, that the basic assumptions on the existence of the integral, mentioned in Definition 1 (concerning the type B of the function $f(x, y)$ in each region $\Omega - \omega$) are satisfied.

REMARK 5. By the limit (4) the integral in the sense of the *principal value* (see Remark 13.8.3) is defined. A finite limit (4) may exist even if integral (2) does not converge.

REMARK 6. The conclusions of Theorem 2 remain valid if condition (3) is replaced by the condition that at least one of the repeated integrals (over the domain Ω)

$$\int\left[\int|f(x, y)|\,\mathrm{d}y\right]\mathrm{d}x\,, \quad \int\left[\int|f(x, y)|\,\mathrm{d}x\right]\mathrm{d}y$$

converges.

Example 1. Let us examine the integral

$$I = \iint_{K} \frac{\mathrm{d}x\,\mathrm{d}y}{\sqrt{(x^2 + y^2)}}\,,$$

where K is the circle with centre at the origin and radius $R = 1$.

The singularity is at the origin. Our function is of the form $f(x, y) = 1/\rho$, thus $M = 1$, $\alpha = 1$ in (3). By Theorem 2, the integral is convergent. Substituting polar

coordinates, $x = \rho \cos \varphi$, $y = \rho \sin \varphi$, we obtain (since $D(\rho, \varphi) = \rho$, see equation (14.4.5))

$$\iint_K \frac{\mathrm{d}x\,\mathrm{d}y}{\sqrt{(x^2 + y^2)}} = \int_0^1 \int_0^{2\pi} \frac{1}{\rho} \cdot \rho\,\mathrm{d}\rho\mathrm{d}\varphi = 2\pi .$$

REMARK 7. The second important case of improper integrals concerns double integrals where the domain of integration is the entire xy-plane. We write

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x\,\mathrm{d}y . \tag{5}$$

The exact meaning is as follows:

Definition 2. Let $f(x, y)$ be of type B in every bounded closed region $\bar{\Omega}$ of type A (Definition 14.1.2). If there exists a number I such that for every $\varepsilon > 0$ it is possible to find a circle K with centre at the origin and radius R so large that for every closed region $\bar{\Omega}$ of type A, containing this circle, we have

$$\left| I - \iint_{\Omega} f(x, y) \right| < \varepsilon ,$$

then the integral (5) is said to be *convergent* and to have the value I.

REMARK 8. Improper triple or m-dimensional integrals can be defined similarly.

Theorem 3. *Integral* (5) *is convergent if and only if* $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |f(x, y)|\,\mathrm{d}x\,\mathrm{d}y$ *is convergent.* (*The convergence of the integral is absolute.*)

Theorem 4. *If everywhere outside a certain circle with the centre at the origin we have*

$$|f(x, y)| < \frac{M}{r^{\alpha}}, \quad M = \text{const}, \quad r = \sqrt{(x^2 + y^2)}, \quad \alpha > 2 , \tag{6}$$

then integral (5) *is convergent and it is possible to use the theorem on repeated integration* (Theorem 14.3.1), *and the theorem on substitution* (Theorem 14.4.1).

In addition,

$$I = \lim_{R \to +\infty} \iint_{K_R} f(x, y)\,\mathrm{d}x\,\mathrm{d}y , \tag{7}$$

where K_R *is the circle with centre at the origin and of radius* R.

REMARK 9. Remarks similar to Remarks 2–6 are also valid here. In particular, for the m-dimensional integral it is sufficient to require $\alpha > m$ in (6). If using a substitution, we transform the integral (7) and take the limit for $R \to +\infty$.

Example 2. Evaluate the integral

$$A = \int_{-\infty}^{\infty} e^{-x^2}\,dx\,. \tag{8}$$

To this purpose we examine the integral

$$I = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2}e^{-y^2}dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\,dx\,dy\,. \tag{9}$$

The integral (9) is convergent by Theorem 4, since, for sufficiently large r, we have $e^{-r^2} < M/r^k$ ($M > 0$, $k > 0$ being arbitrary). Thus

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2}e^{-y^2}\,dx\,dy =$$

$$= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} e^{-x^2}e^{-y^2}\,dy\right]dx = \int_{-\infty}^{\infty} e^{-y^2}\,dy\int_{-\infty}^{\infty} e^{-x^2}\,dx = A\,.\,A = A^2\,. \tag{10}$$

Further, by Theorem 4, transformation to polar coordinates $x = \rho\cos\varphi$, $y = \rho\sin\varphi$ may be carried out. We obtain

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\,dx\,dy = \lim_{R\to+\infty}\iint_{K_R} e^{-(x^2+y^2)}\,dx\,dy =$$

$$= \lim_{R\to+\infty}\int_0^R\int_0^{2\pi} e^{-\rho^2}\rho\,d\rho d\varphi = \lim_{R\to+\infty} 2\pi\,.\,\tfrac{1}{2}\int_0^{R^2} e^{-z}\,dz = \pi\,.$$

Thus, by (10)

$$A = \int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}\,.$$

Symmetry implies that

$$\int_0^{\infty} e^{-x^2}\,dx = \frac{\sqrt{\pi}}{2}\,.$$

REMARK 10. The definition of the improper integral with infinite domain of integration can be extended to the case where the domain of integration is not the entire plane but an arbitrary (in a certain sense) unbounded region M. In definition 2 one has to take instead of the whole region $\bar{\Omega}$ containing the circle K only that part which the closed regions $\bar{M}$ and $\bar{\Omega}$ have in common (i.e. the intersection $\bar{M}\cap\bar{\Omega}$, Fig. 14.14).

REMARK 11. Definition 1 may be extended as well. The following case is of importance: The function $f(x, y)$ is unbounded in $\bar{\Omega}$ in the neighbourhood of a simple

finite piecewise smooth curve k. Then the domains ω in Definition 1 are required to contain the curve k and the function $f(x, y)$ is assumed to be of type B in $\bar{\Omega} - \omega$. For this case the following theorems are valid:

Theorem 5. *If at least one of the repeated integrals over the region* Ω

$$\int\left(\int|f(x, y)|\,dy\right)dx\,, \quad \int\left(\int|f(x, y)|\,dx\right)dy \tag{11}$$

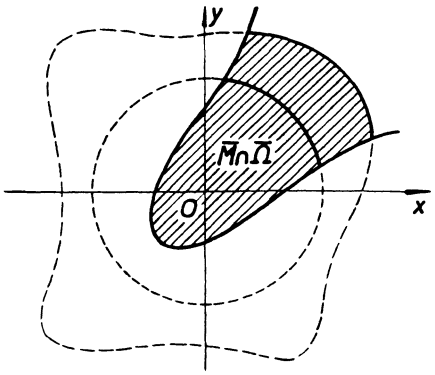


Fig. 14.14.

is convergent, then the integral

$$\iint_{\Omega} f(x, y)\,dx\,dy \tag{12}$$

is also convergent.

Theorem 6. *If integral* (12) *is convergent and if the integral*

$$\int\left[\int f(x, y)\,dy\right]dx \quad \text{or} \quad \int\left[\int f(x, y)\,dx\right]dy \tag{13}$$

converges, then this integral is equal to integral (12).

Example 3. It is required to decide whether the integral

$$\iint_{\Omega} \frac{x}{\sqrt{(x^2 - y^2)}}\,dx\,dy\,, \tag{14}$$

where Ω is the domain shaded in Fig. 14.15, is convergent.

This integral is improper since the integrand is unbounded in the (right) neighbourhood of the line segment $y = x$, $0 \leqq x \leqq 2p$. According to Theorem 5 it is sufficient to prove the convergence of one of integrals (11), where the sign of absolute value may

be omitted for in the domain considered we have $x \geqq 0$ and $x^2 - y^2 \geqq 0$. We get

$$\int_0^{2p}\left(\int_y^{\sqrt{(2py)}} \frac{x}{\sqrt{(x^2-y^2)}}\,dx\right)dy = \int_0^{2p}\left([\sqrt{(x^2-y^2)}]_y^{\sqrt{(2py)}}\right)dy = \int_0^{2p}\sqrt{(2py-y^2)}\,dy\,. \tag{15}$$

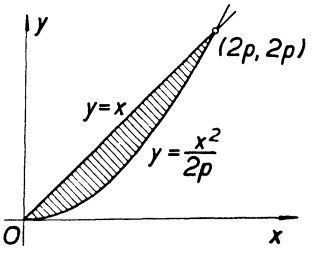


Fig. 14.15.

The last integral is no longer an improper one so that it has a finite value (and thus (14) is convergent by Theorem 5). Let us evaluate it. By the substitution $y - p = u$ and then $u = pz$ we obtain

$$\int_0^{2p}\sqrt{(2py-y^2)}\,dy = \int_0^{2p}\sqrt{[p^2-(y-p)^2]}\,dy = \int_{-p}^{p}\sqrt{(p^2-u^2)}\,du =$$

$$= p^2\int_{-1}^{1}\sqrt{(1-z^2)}\,dz = \tfrac{1}{2}\pi p^2\,.$$

(Example 13.7.4, p. 520). By Theorem 6 this is also the value of integral (14). If we had evaluated the integral

$$\int_0^{2p}\left(\int_{x^2/2p}^{x} \frac{x}{\sqrt{(x^2-y^2)}}\,dy\right)dx$$

we would have arrived at the same value.

14.7. Curvilinear Integrals. Green's Theorem

REMARK 1. Unless the contrary is stated in this paragraph, the term *curve* means a *simple finite piecewise smooth curve* according to Definition 14.1.1. If the curve k is oriented and is the boundary of a region Ω, then this curve is said to be *positively oriented with respect to* Ω if Ω remains on the left-hand side of the point which runs through the curve k in the positive direction. We shall always choose the parametric representation in such a way that if $t_1 < t_2$ then the point A with the parameter t_1 lies on k before the point B with parameter t_2. (Briefly, we say that the curve k is *oriented in the sense of increasing parameter*).

REMARK 2. We say that in the equations of the curve k

$$x = \varphi(s), \quad y = \psi(s)$$

s denotes the *length of arc* (or that s is the *parameter of the length of arc*) if for each s (with the possible exception of a finite number of points)

$$\varphi'^2(s) + \psi'^2(s) = 1 .$$

Geometrically: If points A, B are given by $s = 0$ and $s = s_0$ respectively, then the arc AB has length s_0. For an *arbitrary* parametric representation of the curve k,

$$x = f(t), \quad y = g(t) \quad (\alpha \leqq t \leqq \beta),$$

the length of arc is given by

$$s_0 = \int_{\alpha}^{t_0} \sqrt{[f'^2(t) + g'^2(t)]}\, dt .$$

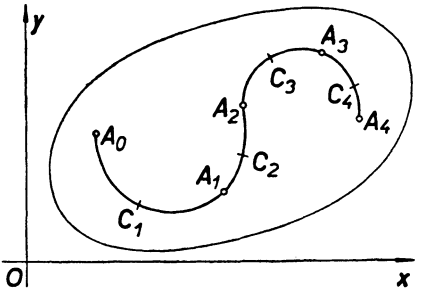


Fig. 14.16.

Definition 1. Let the oriented curve k be given by the equations

$$x = \varphi(t), \quad y = \psi(t), \quad \alpha \leqq t \leqq \beta .$$

Let a function $z = f(x, y)$ be given on the curve k. Let us subdivide k into n arcs $o_1, o_2, \ldots, o_n$ (Fig. 14.16) by points $A_1, A_2, \ldots, A_{n-1}$ with parameters $t_1 < t_2 < \ldots < t_{n-1}$, let us choose arbitrarily on each arc o_k a point C_k (ξ_k, η_k) and write the sums

$$S_x = \sum_{k=1}^{n} f(\xi_k, \eta_k)\,(x_k - x_{k-1}) = \sum_{k=1}^{n} f(\xi_k, \eta_k)\,[\varphi(t_k) - \varphi(t_{k-1})], \tag{1}$$

$$S_y = \sum_{k=1}^{n} f(\xi_k, \eta_k)\,(y_k - y_{k-1}) = \sum_{k=1}^{n} f(\xi_k, \eta_k)\,[\psi(t_k) - \psi(t_{k-1})], \tag{2}$$

$$S_s = \sum_{k=1}^{n} f(\xi_k, \eta_k)\,(s_k - s_{k-1}), \tag{3}$$

where

$$s_k = \int_{\alpha}^{t_k} \sqrt{[\varphi'^2(t) + \psi'^2(t)]}\, \mathrm{d}t\,.$$

Let us denote by l_k the length of the arc o_k and by $\nu(p)$ the greatest of these lengths for the chosen partition p, i.e. $\nu(p) = \max l_k$; $\nu(p)$ is called the *norm of the partition p.*

If there exists a number I_x, or I_y, or I_s such that for an arbitrary $\varepsilon > 0$ it is possible to find a $\delta > 0$ such that

$$|I_x - S_x| < \varepsilon\,, \quad \text{or} \quad |I_y - S_y| < \varepsilon\,, \quad \text{or} \quad |I_s - S_s| < \varepsilon \tag{4}$$

for each partition p for which $\nu(p) < \delta$ – independently of the choice of the points C_k on o_k – we say that there *exists a curvilinear (line) integral* of the function $f(x, y)$ along the oriented curve k with respect to x, or y, or s, respectively and write

$$I_x = \int_k f(x, y)\, \mathrm{d}x\,, \quad \text{or} \quad I_y = \int_k f(x, y)\, \mathrm{d}y\,, \quad \text{or} \quad I_s = \int_k f(x, y)\, \mathrm{d}s\,. \tag{5}$$

(The first two of the integrals (5) are often called *curvilinear integrals of the second kind*, the third of them is called the *integral of the first kind.*)

REMARK 3. Intuitively: The integral I_x is the limit of the sums (1) as $\nu(p) \to 0$. A similar assertion is true for I_y and I_s.

For geometric and physical meaning see Remark 7.

Theorem 1. *If $f(x, y)$ is continuous in the region Ω containing the curve k, then the integrals (5) exist. It is even sufficient if $f(x, y)$ is continuous on k* (Remark 14.1.12).

Theorem 2. *If $f(x, y)$ is continuous in Ω (or on k), then*

$$\begin{aligned}
\int_k f(x, y)\, \mathrm{d}x &= \int_{\alpha}^{\beta} f(\varphi(t), \psi(t))\, \varphi'(t)\, \mathrm{d}t\,, \\
\int_k f(x, y)\, \mathrm{d}y &= \int_{\alpha}^{\beta} f(\varphi(t), \psi(t))\, \psi'(t)\, \mathrm{d}t\,, \\
\int_k f(x, y)\, \mathrm{d}s &= \int_{\alpha}^{\beta} f(\varphi(t), \psi(t)) \sqrt{[\varphi'^2(t) + \psi'^2(t)]}\, \mathrm{d}t\,.
\end{aligned} \tag{6}$$

Theorem 3. *If k is given by the equation $y = g(x)$ in $[a, b]$, then*

$$\int_k f(x, y)\, \mathrm{d}x = \int_a^b f(x, g(x))\, \mathrm{d}x \tag{7}$$

if k is oriented in such a way that $(a, g(a))$ is its initial point, and

$$\int_k f(x, y)\,dx = -\int_a^b f(x)g(x)dx \tag{8}$$

if $(b, g(b))$ is its initial point. Furthermore

$$\int_k f(x, y)\,ds = \int_a^b f(x, g(x))\sqrt{[1 + g'^2(x)]}\,dx \quad (a < b) \tag{9}$$

(no matter how the curve is oriented).

Similarly: If k is given by the equation $x = h(y)$ in $[c, d]$, then

$$\int_k f(x, y)\,dy = \pm\int_c^d f(h(y), y)\,dy\,,$$

where the plus or minus sign is taken depending on whether the initial point of the curve k is the point $(h(c), c)$, or the point $(h(d), d)$, respectively. Furthermore

$$\int_k f(x, y)\,ds = \int_c^d f(h(y), y)\sqrt{[1 + h'^2(y)]}dy \quad (c < d).$$

REMARK 4. If the curve k is a segment parallel to the x-axis, then $\int_k f(x, y)\,dy = 0$; if it is parallel to the y-axis, then $\int_k f(x, y)\,dx = 0$.

Theorem 4. *If k is composed of two arcs o_1, o_2, then*

$$\int_k f(x, y)\,dx = \int_{o_1} f(x, y)\,dx + \int_{o_2} f(x, y)\,dx\,;$$

similarly for

$$\int_k f(x, y)\,dy\,, \quad \int_k f(x, y)\,ds\,.$$

Theorem 5.

$$\int_k [c_1 f_1(x, y) + c_2 f_2(x, y)]\,dx = c_1\int_k f_1(x, y)\,dx + c_2\int_k f_2(x, y)\,dx\,,$$

if the integrals on the right-hand side exist; similarly for the line integrals with respect to y and with respect to s.

Theorem 6. *Let us denote by k' the curve which is inversely oriented to k. Then*

$$\int_{k'} f(x, y)\,dx = -\int_k f(x, y)\,dx\,, \quad \int_{k'} f(x, y)\,dy = -\int_k f(x, y)\,dy\,.$$

REMARK 5. For line integrals with respect to s the sign remains the same.

REMARK 6. By the sum

$$\int_k [f(x, y)\,dx + g(x, y)\,dy]$$

we mean the sum of the integrals

$$\int_k f(x, y)\,dx + \int_k g(x, y)\,dy\,.$$

Example 1. Let k be the circle $x = \cos t$, $y = \sin t$ $(0 \leqq t < 2\pi)$, $f(x, y) = 2 + y$. Then (by (6))

$$\int_k f(x, y)\,dx = \int_0^{2\pi} (2 + \sin t)(-\sin t)\,dt = -\pi\,.$$

This example can also be solved as follows: Let $k = k_1 + k_2$, where k_1 is the upper and k_2 the lower semicircle (Fig. 14.17). On k_1 we have $y = \sqrt{(1 - x^2)}$, on k_2, $y = -\sqrt{(1 - x^2)}$. By Theorem 4 and by (7), (8) we have

$$\int_k (2 + y)\,dx = \int_{k_1} (2 + y)\,dx + \int_{k_2} (2 + y)\,dx =$$

$$= -\int_{-1}^{1} [2 + \sqrt{(1 - x^2)}]\,dx + \int_{-1}^{1} [2 - \sqrt{(1 - x^2)}]\,dx =$$

$$= -2\int_{-1}^{1} \sqrt{(1 - x^2)}\,dx = -\pi\,.$$

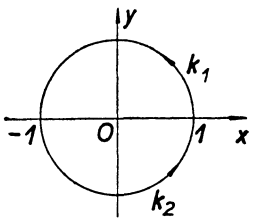


Fig. 14.17.

REMARK 7. The *geometric meaning of a curvilinear integral:* I_s is the area of the surface shown in Fig. 14.18. I_x or I_y is equal to the area of the (oriented) projection of this surface onto the xz-plane, or onto the yz-plane (Fig. 14.18), respectively.

The *physical meaning:* If a field of force $\mathbf{F} = \mathbf{i}\,P(x, y) + \mathbf{j}\,Q(x, y)$ ($\mathbf{i}$, $\mathbf{j}$ are the unit coordinate vectors) is given and if we have to calculate the amount of work

done in this field when moving a particle along the oriented curve k, then this work – as we know from physics – is given by the integral (i.e. by the limit of sums, in a certain sense) of the scalar product

$$L = \int_k \mathbf{F}\,\mathrm{d}l = \int_k [\mathbf{i}\,P(x, y) + \mathbf{j}\,Q(x, y)]\,(\mathbf{i}\,\mathrm{d}x + \mathbf{j}\,\mathrm{d}y) =$$

$$= \int_k P(x, y)\,\mathrm{d}x + \int_k Q(x, y)\,\mathrm{d}y\,. \tag{10}$$

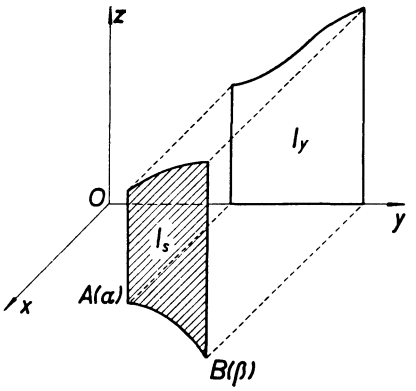


Fig. 14.18.

Thus, the evalution of the amount of work done reduces to the evaluation of two curvilinear integrals.

REMARK 8. The definition of a *curvilinear integral along a curve in space* is similar to that of line integrals in the *plane*. In particular curvilinear integrals of a function continuous in the region Ω in which the given curve lies exist, and

$$\int_k f(x, y, z)\,\mathrm{d}x = \int_\alpha^\beta f(\varphi(t), \psi(t), \chi(t))\,\varphi'(t)\,\mathrm{d}t\,,$$

$$\int_k f(x, y, z)\,\mathrm{d}y = \int_\alpha^\beta f(\varphi(t), \psi(t), \chi(t))\,\psi'(t)\,\mathrm{d}t\,,$$

$$\int_k f(x, y, z)\,\mathrm{d}z = \int_\alpha^\beta f(\varphi(t), \psi(t), \chi(t))\,\chi'(t)\,\mathrm{d}t\,,$$

$$\int_k f(x, y, z)\,\mathrm{d}s = \int_\alpha^\beta f(\varphi(t), \psi(t), \chi(t))\,\sqrt{[\varphi'^2(t) + \psi'^2(t) + \chi'^2(t)]}\,\mathrm{d}t\,. \tag{11}$$

The physical meaning of the first three integrals is similar to that of (10).

REMARK 9. It can be shown that the value of integrals (11) (as well as that of integrals (6), previously introduced) does not depend on the choice of the parametric representation of the curve k (naturally respecting the convention of Remark 1).

Theorem 7 (*Green's Theorem*). *Let* $\bar{\Omega}$ *be a closed region of type* A (Definition 14.1.2) *with the boundary* k *positively oriented with respect to* Ω (see Remark 1). *Let the functions*

$$P(x, y), \quad Q(x, y), \quad \frac{\partial P}{\partial y}(x, y), \quad \frac{\partial Q}{\partial x}(x, y)$$

be continuous in $\bar{\Omega}$ (*i.e. in* $\Omega + k$). *Then* (see Remark 6)

$$\int_k (P\,dx + Q\,dy) = \iint_\Omega \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dx\,dy. \tag{12}$$

If Ω *is a multiply connected region, then* $\int_k$ *means the sum of the curvilinear integrals along individual parts of the boundary.*

REMARK 10. If

$$Q(x, y) \equiv 0 \quad \text{or} \quad P(x, y) \equiv 0,$$

we obtain particular cases of Green's theorem

$$\int_k P\,dx = -\iint_\Omega \frac{\partial P}{\partial y}\,dx\,dy, \quad \int_k Q\,dy = \iint_\Omega \frac{\partial Q}{\partial x}\,dx\,dy. \tag{13}$$

Green's theorem is often written in the following form:

$$\int_k [-P(x, y)\cos\beta + Q(x, y)\cos\alpha]\,ds = \iint_\Omega \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dx\,dy,$$

or, writing $-P(x, y) = T(x, y)$, $Q(x, y) = S(x, y)$, in the form

$$\int_k [S(x, y)\cos\alpha + T(x, y)\cos\beta]\,ds = \iint_\Omega \left(\frac{\partial S}{\partial x} + \frac{\partial T}{\partial y}\right) dx\,dy,$$

where $\cos\alpha$, $\cos\beta$ are the direction cosines of the outward normal.

Definition 2. Let k be an oriented curve in Ω with initial point A and end point B. If the value of the integral

$$\int_k [f(x, y)\,dx + g(x, y)\,dy] \tag{14}$$

depends only on the choice of the points A, B, and not on the choice of the curve connecting the points A, B (and naturally, lying in the considered region Ω), then the integral (14) is said to be *independent of the path of integration in* Ω.

Theorem 8. *Let*

$$P(x, y), \quad Q(x, y), \quad \frac{\partial P(x, y)}{\partial y}, \quad \frac{\partial Q(x, y)}{\partial x}$$

be continuous functions in a simply connected region Ω. Then a necessary and sufficient condition that the integral

$$\int_k [P(x, y)\,dx + Q(x, y)\,dy] \tag{15}$$

is independent of the path of integration in Ω is that the equation

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x} \tag{16}$$

is satisfied in Ω.

REMARK 11. In particular: If condition (16) is satisfied, then the integral (15) along each *closed* curve k is equal to zero.

Theorem 9. *If P, Q, $\partial P/\partial y$, $\partial Q/\partial x$ are continuous functions in a simply connected region Ω, then a necessary and sufficient condition that the expression*

$$P\,dx + Q\,dy \tag{17}$$

is the total differential of some function $F(x, y)$ in Ω is that the equation

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x} \tag{18}$$

holds in Ω. If this condition is satisfied, then the value of the integral (15) is given by the difference $F(x_2, y_2) - F(x_1, y_1)$, where (x_1, y_1) is the initial and (x_2, y_2) the end point of the curve k.

Theorem 10. *Let $P(x,y)$, $Q(x,y)$ be continuous in a region Ω. Then a necessary and sufficient condition that the integral*

$$\int_k (P\,dx + Q\,dy)$$

is independent of the path of integration in Ω is that there exists a function $F(x, y)$ such that the expression

$$P\,dx + Q\,dy$$

is its total differential.

Example 2. According to Theorem 9 the expression

$$(x^2 - y^2)\,dx + (5 - 2xy)\,dy \tag{19}$$

is a total differential (even in the entire xy-plane). The corresponding function $F(x, y)$ may be found from the condition

$$dF = \frac{\partial F}{\partial x}\,dx + \frac{\partial F}{\partial y}\,dy = P\,dx + Q\,dy$$

(which has to be satisfied for each dx, dy):

$$\frac{\partial F}{\partial x} = P = x^2 - y^2\,, \quad \frac{\partial F}{\partial y} = Q = 5 - 2xy\,. \tag{20}$$

From the first equation it follows that

$$F(x, y) = \frac{x^3}{3} - xy^2 + f(y)\,. \tag{21}$$

By differentiation of (21) with respect to y and by comparison with the second of equations (20), we have

$$-2xy + f'(y) = -2xy + 5\,, \quad \text{hence} \quad f(y) = 5y + C\,.$$

Thus

$$F(x, y) = \frac{x^3}{3} - xy^2 + 5y + C\,, \tag{22}$$

where C is an arbitrary constant.

The function $F(x, y)$ may also be obtained by the formula

$$F(x, y) = \int_{x_0}^{x} P(x, y_0)\,dx + \int_{y_0}^{y} Q(x, y)\,dy\,. \tag{23}$$

In our case

$$F(x, y) = \int_{x_0}^{x} (x^2 - y_0^2)\,dx + \int_{y_0}^{y} (5 - 2xy)\,dy =$$

$$= \left(\frac{x^3}{3} - y_0^2 x\right) - \left(\frac{x_0^3}{3} - y_0^2 x_0\right) + (5y - xy^2) - (5y_0 - xy_0^2) =$$

$$= \frac{x^3}{3} + 5y - xy^2 - \frac{x_0^3}{3} + y_0^2 x_0 - 5y_0 \tag{24}$$

and this result is of the form (22).

REMARK 12. Theorems 8, 9 and 10 are true also for line integrals along curves in space, with the only difference that instead of integral (15) we have to write the integral

$$\int_k [P(x, y, z)\,dx + Q(x, y, z)\,dy + R(x, y, z)\,dz] \tag{25}$$

and to replace equation (16), or (18) by the equations (which have to be simultaneously fulfilled)

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}\,, \quad \frac{\partial P}{\partial z} = \frac{\partial R}{\partial x}\,, \quad \frac{\partial Q}{\partial z} = \frac{\partial R}{\partial y} \tag{26}$$

(cf. Theorem 14.8.6, p. 614). The function $F(x, y, z)$ can be calculated by both the methods of Example 2, in particular

$$F(x, y, z) = \int_{x_0}^{x} P(x, y_0, z_0)\,dx + \int_{y_0}^{y} Q(x, y, z_0)\,dy + \int_{z_0}^{z} R(x, y, z)\,dz .$$

REMARK 13 (concerning practical evaluation). From (13) it follows that if P is a function of x only in a simply connected region Ω, then the integral $\int_k P\,dx$, along each closed curve k in Ω, is equal to zero. Similarly, if Q is a function of the variable y only.

REMARK 14. Using (13) it is possible to show that the area p of a bounded simply connected region Ω with the boundary k (positively oriented with respect to Ω) may be expressed as follows:

$$p = \int_k x\,dy, \quad p = -\int_k y\,dx$$

or,

$$p = \tfrac{1}{2}\int_k [x\,dy - y\,dx]. \tag{27}$$

The last of the integrals (27) is often used for the evaluation of the area of a sector (Fig. 14.19). On OA and BO the integral is equal to zero (for on these line segments $x\,dy = y\,dx$) so that

$$p = \tfrac{1}{2}\int_{k_0} [x\,dy - y\,dx],$$

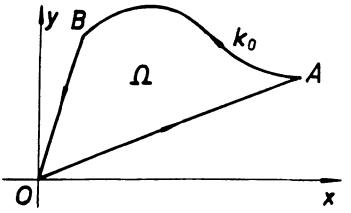


Fig. 14.19.

where $k_0 = \overset{\frown}{AB}$. If the equations of the curve k_0 are given in the form

$$x = \varphi(t), \quad y = \psi(t) \quad (\alpha \leqq t \leqq \beta) \tag{28}$$

(with positive orientation in the sense of increasing parameter), then

$$p = \tfrac{1}{2}\int_{\alpha}^{\beta} [\varphi(t)\,\psi'(t) - \psi(t)\,\varphi'(t)]\,dt. \tag{29}$$

14.8. Surface Integrals. The Gauss–Ostrogradski Theorem, Stokes's Theorem, Green's Identities

REMARK 1. In this paragraph we shall deal only with surfaces having some simple properties. If we use the term *surface*, we shall always mean a *simple finite piecewise smooth surface* (Remark 14.1.7).

REMARK 2. A surface is said to be *oriented* if its two sides can be distinguished as the *exterior* and *interior side*, respectively. This concept is intuitively very clear for the case of closed surfaces that constitute boundaries of solids of type A (Definition 14.1.4). In the same sense we speak about the *oriented outward* and *inward normal of a surface*. Of course, for surfaces that constitute boundaries of the above-mentioned solids, the inward normal is oriented inside the solid, while the outward normal has the reverse orientation. (Not all surfaces can be oriented (e.g. the so-called Mobius leaf).

REMARK 3. Let us consider an oriented surface $z = f(x, y)$. The area of the projection of the surface onto the xy-plane, assigned with the plus or minus sign according as the exterior side or the interior side is seen when observing the surface in the negative direction of the z-axis,* is called the *oriented projection of the surface onto the xy-plane*. The oriented projection of a surface onto the xy- plane is (by definition) equal to zero if the surface is formed by parallels to the z-axis.

Similarly we define the oriented projection of the oriented surfaces $x = g(y, z)$ and $y = h(x, z)$ on to the yz-plane and xz-plane, respectively. The plus or minus sign is again chosen according as the exterior or interior side is seen when observing the surface in the negative direction of the x- or y-axis, respectively.

REMARK 4. The concepts introduced in Remarks 2 and 3 are based to a large extent on intuition, and were therefore not given as definitions.

Definition 1. Let two points A, B be given on a surface. Let us connect them with a rectifiable curve k lying on the surface. The greatest lower bound of the lengths of all such curves is called the *distance between the points A, B measured on the given surface*. The least upper bound of the distances between all possible pairs of points on the surface is called the *interior diameter of the surface.*

REMARK 5. Roughly speaking: the distance between the points A, B measured on the surface is given by the length of the shortest curve lying on the surface and connecting these points. The interior diameter of a finite surface is the greatest possible distance between two points measured on this surface.

For example, the spherical surface with radius r has diameter $2r$; however, its *interior* diameter is equal to πr (the length of the meridian between "the north and south poles").

* i.e. according as to whether the outward normal makes an angle with the positive z-axis less or greater than $\pi/2$.

Definition 2. Let a function $u = f(x, y, z)$ be given on the surface S (see Remark 1). Let us divide S into n parts S_i and denote the greatest of the interior diameters of the parts S_i (Definition 1) by $\nu(p)$. Choose a point (x_i, y_i, z_i) in each part S_i and write down the sum

$$\sigma = \sum_{i=1}^{n} f(x_i, y_i, z_i)\, s_i\,, \tag{1}$$

where s_i is the area of the part S_i. If there exists a number I_S such that for every $\varepsilon > 0$ there is another number $\delta > 0$ such that for each partition p for which $\nu(p) < \delta$ the inequality

$$|I_S - \sigma| < \varepsilon$$

holds, independently of the choice of the points (x_i, y_i, z_i) in S_i, then we say that the *surface integral*

$$\iint_S f(x, y, z)\, \mathrm{d}S \tag{2}$$

over the surface S exists and that its value is I_S.

REMARK 6. Roughly speaking: (2) is the limit of the sums (1) as $\nu(p) \to 0$.

REMARK 7. The integral (2) is often called the *surface integral of the first kind.*

Definition 3. Let a function $u = f(x, y, z)$ be given on an *oriented* surface S. Let S be divisible into a finite number n of parts S_i which either may be represented in the form $z = f(x, y)$, or are formed by parallels to the z-axis. Let p be such a partition. By the norm $\nu(p)$ of that partition we mean the greatest of the interior diameters of the parts S_i (Definition 1). Let us choose, in each part S_i, a point (x_i, y_i, z_i) and write the sum

$$\sigma_{xy} = \sum_{i=1}^{n} f(x_i, y_i, z_i)\, p_i\,, \tag{3}$$

where p_i is the oriented projection of the part S_i on the xy-plane (Remark 3). If there exists a number I_{xy} such that for an arbitrary $\varepsilon > 0$ there is a $\delta > 0$ such that for each partition p (with the above-mentioned properties) for which $\nu(p) < \delta$ the inequality

$$|I_{xy} - \sigma_{xy}| < \varepsilon$$

holds, independently of the choice of the points (x_i, y_i, z_i) in S_i, then we say that the *surface integral over the oriented surface S with respect to the coordinates x, y,*

$$\iint_S f(x, y, z)\, \mathrm{d}x\, \mathrm{d}y\,, \tag{4}$$

exists and that its value is I_{xy}.

REMARK 8. Roughly speaking: (4) is the limit of the sums (3) as $\nu(p) \to 0$.

REMARK 9. The integrals

$$\iint_S g(x, y, z)\,dy\,dz\,, \quad \iint_S h(x, y, z)\,dz\,dx \tag{5}$$

are defined in a similar way. Integrals (4), (5) are often called *surface integrals of the second kind*. Surface integrals are a generalization of curvilinear integrals in a certain sense. In the same way as in the case of curvilinear integrals of the second kind, the orientation of the surface must be known when considering surface integrals of the second kind.

By the integral

$$\iint_S [f(x, y, z)\,dx\,dy + g(x, y, z)\,dy\,dz + h(x, y, z)\,dz\,dx]$$

we mean the sum of the integrals (4) and (5).

REMARK 10. To ensure the existence of the integrals (2), (4), (5) it is sufficient, under the above-mentioned assumptions regarding the surface S (Remark 1), to assume the continuity of the functions under consideration in the region Ω in which the surface S lies (in fact continuity on the surface S itself is sufficient).

REMARK 11. The basic properties of surface integrals are similar to those of curvilinear integrals (Theorems 14.7.4, 14.7.5). If the orientation of the surface is changed (i.e. the exterior and the interior sides of the surface are interchanged), then the integrals (4) and (5) change their signs.

Theorem 1. *If the surface S is given in the explicit form $z = \varphi(x, y)$, where $\varphi(x, y)$ is a continuous piecewise smooth function in M, then the integral* (2) *is equal to the integral*

$$\iint_M f(x, y, \varphi(x, y))\sqrt{\left[1 + \left(\frac{\partial \varphi}{\partial x}\right)^2 + \left(\frac{\partial \varphi}{\partial y}\right)^2\right]}\,dx\,dy\,. \tag{6}$$

Theorem 2. *If the surface S is given parametrically by the equations*

$$x = x(u, v)\,, \quad y = y(u, v)\,, \quad z = z(u, v)\,, \quad (u, v) \in Q$$

(where x, y, z are assumed to be continuous piecewise smooth functions in Q), then the integral (2) *is equal to the integral*

$$\iint_Q f(x(u, v), y(u, v), z(u, v))\sqrt{(EG - F^2)}\,du\,dv\,. \tag{7}$$

(The functions E, F, G are defined in equations (14.9.75), (14.9.76.)

Theorem 3. *If S is given in the form $z = \varphi(x, y)$ for $(x, y) \in M$, then*

$$\iint_S f(x, y, z)\,\mathrm{d}x\,\mathrm{d}y = \pm \iint_M f(x, y, \varphi(x, y))\,\mathrm{d}x\,\mathrm{d}y \tag{8}$$

where the plus or minus sign is to be chosen according to Remark 3 (see Example 1).

REMARK 12. Theorem 3 is useful for the practical evaluation of the integral (4). We divide the surface S into surfaces which may be represented in the form $z = \varphi(x, y)$ and surfaces formed by lines parallel to the z-axis. In the second case the corresponding integrals are equal to zero (Remark 3), in the first case they are evaluated using (8).

Theorem 4. *Let the surface S be given parametrically by $x = x(u, v)$, $y = y(u, v)$, $z = z(u, v)$ for $u, v \in Q$. Then*

$$\iint_S f(x, y, z)\,\mathrm{d}x\,\mathrm{d}y = \pm \iint_Q f(x(u, v), y(u, v), z(u, v)) \frac{D(x, y)}{D(u, v)}\,\mathrm{d}u\,\mathrm{d}v\,, \tag{9}$$

where

$$\frac{D(x, y)}{D(u, v)} = \begin{vmatrix} \dfrac{\partial x}{\partial u}, & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u}, & \dfrac{\partial y}{\partial v} \end{vmatrix}. \tag{10}$$

The sign in (9) *can be determined as follows: We choose a point $(u_0, v_0) \in Q$ such that in its neighbourhood Q', $D(x, y)/D(u, v) \neq 0$. Then the corresponding part S' of the surface S may be represented in the form $z = \varphi(x, y)$. Since S is an oriented surface, we are able, according to Theorem* 3 *or Remark* 3, *to determine the correct sign in the equation*

$$\iint_{S'} f(x, y, z)\,\mathrm{d}x\,\mathrm{d}y = \pm \iint_{M'} f(x, y, \varphi(x, y))\,\mathrm{d}x\,\mathrm{d}y\,. \tag{11}$$

If $D(x, y)/D(u, v) > 0$ in Q', we choose the same sign in (9) *as in* (11); *if $D(x, y)/|D(u, v)| < 0$ in Q', we choose the opposite sign.* (See Example 1.)

REMARK 13. Theorems 3, 4 and Remark 12 are valid in a similar form for the integrals (5). Expression (10) has to be successively replaced by the expressions

$$\frac{D(y, z)}{D(u, v)} = \begin{vmatrix} \dfrac{\partial y}{\partial u}, & \dfrac{\partial y}{\partial v} \\ \dfrac{\partial z}{\partial u}, & \dfrac{\partial z}{\partial v} \end{vmatrix} \quad \text{or} \quad \frac{D(z, x)}{D(u, v)} = \begin{vmatrix} \dfrac{\partial z}{\partial u}, & \dfrac{\partial z}{\partial v} \\ \dfrac{\partial x}{\partial u}, & \dfrac{\partial x}{\partial v} \end{vmatrix}. \tag{12}$$

Example 1. Let us evaluate

$$\iint_S x^2 z \, dx \, dy , \tag{13}$$

where S is the sphere with centre at the origin and radius 1. S is a closed surface by which its orientation is thus given. S obviously consists of two surfaces S_1 and S_2 with equations $z = \sqrt{(1 - x^2 - y^2)}$ and $z = -\sqrt{(1 - x^2 - y^2)}$, respectively. If we make use of (8), we choose, according to Theorem 3 and Remark 3, the plus sign for S_1 and the minus sign for S_2. Hence

$$\iint_S x^2 z \, dx \, dy = + \iint_K x^2 \sqrt{(1 - x^2 - y^2)} \, dx \, dy - \iint_K x^2 [-\sqrt{(1 - x^2 - y^2)}] \, dx \, dy ,$$

where K is the circle $x^2 + y^2 \leqq 1$. The right-hand side of the equation is easily evaluated by transforming to polar coordinates

$$x = r \cos \varphi , \quad y = r \sin \varphi \quad \text{(see Examples 14.4.1, 14.4.2)} .$$

The result:

$$\iint_S x^2 z \, dx \, dy = 2 \iint_K x^2 \sqrt{(1 - x^2 - y^2)} \, dx \, dy = \tfrac{4}{15} \pi .$$

If S is given in parametric form,

$$x = \sin u \cos v , \quad y = \sin u \sin v , \quad z = \cos u \quad (0 \leqq u \leqq \pi, \ 0 \leqq v < 2\pi) ,$$

we use (9). According to Theorem 4, let us choose, for example, $(u_0, v_0) = (\frac{1}{6}\pi, \frac{1}{3}\pi)$. There is obviously a point on S_1 which corresponds to this point so that in (11) we have the plus sign. Furthermore, for $u = \frac{1}{6}\pi$, $v = \frac{1}{3}\pi$ (and — as follows from continuity — also in a certain neighbourhood of this point)

$$\frac{D(x, y)}{D(u, v)} = \begin{vmatrix} \cos u \cos v, & -\sin u \sin v \\ \cos u \sin v, & \sin u \cos v \end{vmatrix} = \sin u \cos u > 0$$

so that according to Theorem 4 we choose in (9) the *same* sign as in (11), i.e. the plus sign. Thus we have

$$\iint_S x^2 z \, dx \, dy = \int_0^{\pi} \int_0^{2\pi} \sin^2 u \cos^2 v \cos u \sin u \cos u \, du \, dv = \tfrac{4}{15} \pi$$

in accordance with the previous result.

Theorem 5 (*The Gauss–Ostrogradski Theorem*). *Let the functions*

$$P(x, y, z), \quad Q(x, y, z), \quad R(x, y, z) \quad \text{and} \quad \frac{\partial P}{\partial x} , \quad \frac{\partial Q}{\partial y} , \quad \frac{\partial R}{\partial z}$$

be continuous in a closed solid $\overline{V}$ *of type* A (Definition 14.1.4) (*which need not necessarily be simply connected*) *with the oriented boundary* S *so that the outward normal* n *points out of the solid. Then*

$$\iiint_V \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z}\right) dx\, dy\, dz = \iint_S (P\, dy\, dz + Q\, dz\, dx + R\, dx\, dy), \quad (14)$$

or, if we denote the direction cosines of the outward normal by $\cos\alpha$, $\cos\beta$, $\cos\gamma$,

$$\iiint_V \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z}\right) dx\, dy\, dz = \iint_S (P\cos\alpha + Q\cos\beta + R\cos\gamma)\, dS \quad (15)$$

or, if we take into consideration that $\partial x/\partial n = \cos\alpha$ *etc.*

$$\iiint_V \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z}\right) dx\, dy\, dz = \iint_S \left(P\frac{\partial x}{\partial n} + Q\frac{\partial y}{\partial n} + R\frac{\partial z}{\partial n}\right) dS. \quad (16)$$

If V *is multiply connected, then* $\iint_S$ *means the sum of the surface integrals over the individual parts of the boundary.*

Definition 4. Let S be an oriented surface and let k be a *closed* simple finite piecewise smooth (generally spatial) curve on S. The curve k is said to be *positively oriented with respect to* S, if – observed from the side of the outward normal (thus erected on the exterior part of the surface) – when traversing the curve k in the positive direction, the part of the surface S enclosed by the curve k, remains on the left-hand side (Fig. 14.20).

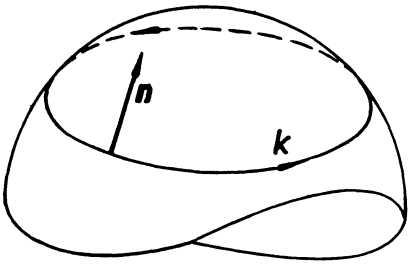


Fig. 14.20.

Theorem 6 (*The Stokes Theorem*). *Let* S *be an oriented open surface the boundary of which is formed by a closed simple finite piecewise smooth curve* k *which is positively oriented with respect to* S. *Then*

$$\int_k (P\, dx + Q\, dy + R\, dz) =$$

$$= \int\int_S \left[\left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dx dy + \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right) dy dz + \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right) dz dx\right]. \quad (17)$$

REMARK 14. Stokes's theorem is a generalization of Green's theorem (Theorem 14.7.7) for simply connected regions in the xy-plane.

From the Gauss–Ostrogradski theorem there easily follows:

Theorem 7. *If the functions P, Q, R, $\partial P/\partial x$, $\partial Q/\partial y$, $\partial R/\partial z$ are continuous in Ω, then a necessary and sufficient condition that the surface integral* (14) (*or* (15), *or* (16)) *is equal to zero for each surface S, which is the boundary of a closed simply connected solid $\bar{V}$ of type* A *lying in the region Ω, is given by the equation*

$$\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z} \equiv 0 \quad in \quad \Omega . \tag{18}$$

Theorem 8. *If (18) is satisfied for Ω simply connected, then, in Ω, the surface integral* (14) (*or* (15), *or* (16) *over a surface S', the boundary of which is a curve k which is positively oriented with respect to S', depends only on k and is independent of the form of the surface S'.*

Theorem 9 (*Green's Identities*). *Let us write, as usual, $\Delta u = \partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 + \partial^2 u/\partial z^2$ and let $\partial u/\partial n$ be the derivative of u in the direction of the outward normal,* i.e.

$$\frac{\partial u}{\partial n} = \frac{\partial u}{\partial x}\frac{\partial x}{\partial n} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial n} + \frac{\partial u}{\partial z}\frac{\partial z}{\partial n},$$

where

$$\frac{\partial x}{\partial n} = \cos\alpha , \quad \frac{\partial y}{\partial n} = \cos\beta , \quad \frac{\partial z}{\partial n} = \cos\gamma$$

and $\cos\alpha$, $\cos\beta$, $\cos\gamma$ *are the direction cosines of the (oriented) outward normal. Let $u(x, y, z)$, $v(x, y, z)$, together with their derivatives up to the second order, be continuous in a closed solid $\bar{V}$ of type* A *with an oriented boundary S. Then*

$$\iiint_V \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} + \frac{\partial u}{\partial z}\frac{\partial v}{\partial z}\right) dx\, dy\, dz =$$

$$= -\iiint_V u\, \Delta v\, dx\, dy\, dz + \iint_S u \frac{\partial v}{\partial n}\, dS = \tag{19}$$

$$= -\iiint_V v\, \Delta u\, dx\, dy\, dz + \iint_S v \frac{\partial u}{\partial n}\, dS . \tag{20}$$

From (19) *and* (20) *it follows that*

$$\iiint_V (u\, \Delta v - v\, \Delta u)\, dx\, dy\, dz = \iint_S \left(u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n}\right) dS . \tag{21}$$

For $v \equiv 1$ we obtain

$$\iiint_V \Delta u \, dx \, dy \, dz = \iint_S \frac{\partial u}{\partial n} \, dS . \tag{22}$$

REMARK 15. Using the symbolism of vector analysis (Chap. 7) and the notation $dV = dx \, dy \, dz$, the previous theorems may be rewritten in a rather more concise form:

$$\boldsymbol{A} = \boldsymbol{i}P + \boldsymbol{j}Q + \boldsymbol{k}R , \quad d\boldsymbol{S} = \boldsymbol{i} \, dS \cos\alpha + \boldsymbol{j} \, dS \cos\beta + \boldsymbol{k} \, dS \cos\gamma ,$$

$$d\boldsymbol{s} = \boldsymbol{i} \, dx + \boldsymbol{j} \, dy + \boldsymbol{k} \, dz .$$

Equation (15) (or (14), or (16)):

$$\iiint_V \operatorname{div} \boldsymbol{A} \, dV = \iint_S \boldsymbol{A} \, d\boldsymbol{S} .$$

Equation (17):

$$\int_C \boldsymbol{A} \, d\boldsymbol{s} = \iint_S \operatorname{curl} \boldsymbol{A} \, d\boldsymbol{S} .$$

Equations (19), (20):

$$\iiint_V \operatorname{grad} u \operatorname{grad} v \, dV = - \iiint_V u \, \Delta v \, dV + \iint_S u \operatorname{grad} v \, d\boldsymbol{S} =$$

$$= - \iiint_V v \, \Delta u \, dV + \iint_S v \operatorname{grad} u \, d\boldsymbol{S} .$$

Equation (21):

$$\iiint_V (u \, \Delta v - v \, \Delta u) \, dV = \iint_S (u \operatorname{grad} v - v \operatorname{grad} u) \, d\boldsymbol{S} .$$

Equation (22):

$$\iiint_V \Delta u \, dV = \iint_S \operatorname{grad} u \, d\boldsymbol{S}$$

(see Chap. 7).

14.9. Applications of the Integral Calculus in Geometry and Physics

(Curves, plane figures, solids, surfaces – lengths, areas, volumes, masses, statical moments, centres of gravity, moments of inertia; the work of a force along a given curve; some special formulae; Guldin's rules; Steiner's theorem; examples.)

REMARK 1. Exact definitions of the length of a curve, of the area of a surface, etc., may be found in many textbooks of integral calculus.

REMARK 2. All formulae given below may easily be obtained from a geometrical or physical conception of the problem. For example, the formula for calculation of the volume of a solid of revolution, the lateral surface of which is obtained by rotating the curve $y = f(x)$ ($f(x) \geqq 0$, $a \leqq x \leqq b$) around the x-axis (Fig. 14.21) can be derived in an intuitive way as follows: We divide the interval $[a, b]$ into n subintervals $\Delta x_1, \Delta x_2, \ldots, \Delta x_n$. Choosing a point ξ_k in each interval Δx_k, we consider the function $y = f(x)$ to be constant in Δx_k, i.e. $y = f(\xi_k)$. By rotating the segment $y = f(\xi_k)$ over Δx_k around the x-axis, we get a circular cylinder with the volume

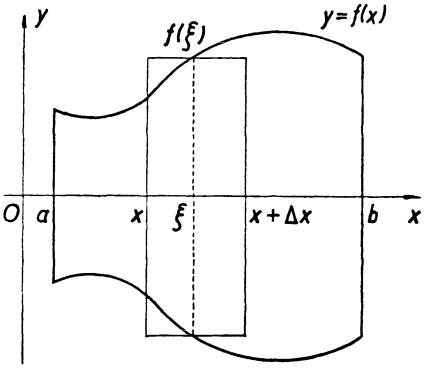


Fig. 14.21.

$\pi f^2(\xi_k)\,\Delta x_k$. For the sum of the volumes of all these cylinders in the interval $[a, b]$ we have

$$\sum_{k=1}^{n} \pi f^2(\xi_k)\,\Delta x_k\,. \tag{1}$$

Now, it seems to be evident that as $n \to \infty$ and $\Delta x_k \to 0$, we get the exact volume of the solid of revolution, while the sum (1) turns into the integral

$$\pi \int_a^b f^2(x)\,\mathrm{d}x\,. \tag{2}$$

If the function $f(x)$ is continuous in $[a, b]$, the heuristic consideration just indicated yields — by Theorem 13.6.2 — the correct result. The same is true if upper or lower sums are used instead of the sums (1). The idea just explained may be employed for the derivation of all formulae given below.

In all these formulae we assume the continuity of the function considered. If some of the functions are, for example, piecewise continuous, we proceed according to Remark 13.6.13, p. 518.

NOTE. The majority of formulae are contained in the first four sections, (**a**) curves, (**b**) plane figures, (**c**) solids, (**d**) surfaces. At the beginning of each of these sections, the types of objects are listed for which formulae are presented. For example, in section (**c**) formulae for five types of solids (a) to (e) are given; these types

are described at the beginning of section (c). Thus, for the volume of solids, five formulae corresponding to types (a) to (e) are given; the formulae for the mass, statical moment, etc., are treated similarly.

(a) Curves.

(α) *Plane curves.* The curve k is given:

(a) by the graph of a function $y = f(x)$, $a \leqq x \leqq b$,

(b) parametrically, $x = \varphi(t)$, $y = \psi(t)$, $t_1 \leqq t \leqq t_2$
(notation: $\mathrm{d}\varphi/\mathrm{d}t = \dot{\varphi}$, $\mathrm{d}\psi/\mathrm{d}t = \dot{\psi}$),

(c) in polar coordinates, $r = r(\varphi)$, $\varphi_1 \leqq \varphi \leqq \varphi_2$, $r \geqq 0$.

The specific density (mass per unit of length, $\mathrm{kg\,m^{-1}}$) ϱ is given as a function of x or t or φ. (If ϱ is constant, then, of course, ϱ may be put in front of the integral sign.)

LENGTH:

(a) $$l = \int_a^b \sqrt{[1 + f'^2(x)]}\,\mathrm{d}x = \int_a^b \sqrt{(1 + y'^2)}\,\mathrm{d}x\,, \tag{3}$$

(b) $$l = \int_{t_1}^{t_2} \sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t)]}\,\mathrm{d}t = \int_{t_1}^{t_2} \sqrt{(\dot{x}^2 + \dot{y}^2)}\,\mathrm{d}t\,, \tag{4}$$

(c) $$l = \int_{\varphi_1}^{\varphi_2} \sqrt{[r^2(\varphi) + r'^2(\varphi)]}\,\mathrm{d}\varphi\,. \tag{5}$$

MASS:

(a) $$M = \int_a^b \varrho(x) \sqrt{[1 + f'^2(x)]}\,\mathrm{d}x = \int_a^b \varrho \sqrt{(1 + y'^2)}\,\mathrm{d}x\,, \tag{6}$$

(b) $$M = \int_{t_1}^{t_2} \varrho(t) \sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t)]}\,\mathrm{d}t = \int_{t_1}^{t_2} \varrho \sqrt{(\dot{x}^2 + \dot{y}^2)}\,\mathrm{d}t\,, \tag{7}$$

(c) $$M = \int_{\varphi_1}^{\varphi_2} \varrho(\varphi) \sqrt{[r^2(\varphi) + r'^2(\varphi)]}\,\mathrm{d}\varphi\,. \tag{8}$$

STATICAL MOMENT WITH RESPECT TO THE x- OR y-AXIS:

(a) $$S_x = \int_a^b f(x)\,\varrho(x) \sqrt{[1 + f'^2(x)]}\,\mathrm{d}x = \int_a^b y\varrho \sqrt{(1 + y'^2)}\,\mathrm{d}x\,, \tag{9}$$

$$S_y = \int_a^b x\varrho(x) \sqrt{[1 + f'^2(x)]}\,\mathrm{d}x = \int_a^b x\varrho \sqrt{(1 + y'^2)}\,\mathrm{d}x\,, \tag{9'}$$

(b) $$S_x = \int_{t_1}^{t_2} \psi(t)\,\varrho(t)\,\sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t)]}\,\mathrm{d}t = \int_{t_1}^{t_2} y\varrho\,\sqrt{(\dot{x}^2 + \dot{y}^2)}\,\mathrm{d}t\,, \tag{10}$$

$$S_y = \int_{t_1}^{t_2} \varphi(t)\,\varrho(t)\,\sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t)]}\,\mathrm{d}t = \int_{t_1}^{t_2} x\varrho\,\sqrt{(\dot{x}^2 + \dot{y}^2)}\,\mathrm{d}t\,, \tag{10'}$$

(c) $$S_x = \int_{\varphi_1}^{\varphi_2} \varrho(\varphi)\,r(\varphi)\sin\varphi\,\sqrt{[r^2(\varphi) + r'^2(\varphi)]}\,\mathrm{d}\varphi\,, \tag{11}$$

$$S_y = \int_{\varphi_1}^{\varphi_2} \varrho(\varphi)\,r(\varphi)\cos\varphi\,\sqrt{[r^2(\varphi) + r'^2(\varphi)]}\,\mathrm{d}\varphi\,. \tag{11'}$$

COORDINATES OF THE CENTRE OF GRAVITY:

$$x_T = \frac{S_y}{M}\,, \quad y_T = \frac{S_x}{M}\,; \tag{12}$$

for example,

$$y_T = \frac{\displaystyle\int_{\varphi_1}^{\varphi_2} \varrho(\varphi)\,r(\varphi)\cos\varphi\,\sqrt{[r^2(\varphi) + r'^2(\varphi)]}\,\mathrm{d}\varphi}{\displaystyle\int_{\varphi_1}^{\varphi_2} \varrho(\varphi)\,\sqrt{[r^2(\varphi) + r'^2(\varphi)]}\,\mathrm{d}\varphi}\,. \tag{13}$$

MOMENT OF INERTIA WITH RESPECT TO THE x- OR y-AXIS:

(a) $$I_x = \int_a^b f^2(x)\,\varrho(x)\,\sqrt{[1 + f'^2(x)]}\,\mathrm{d}x = \int_a^b y^2\varrho\,\sqrt{(1 + y'^2)}\,\mathrm{d}x\,, \tag{14}$$

$$I_y = \int_a^b x^2\varrho(x)\,\sqrt{[1 + f'^2(x)]}\,\mathrm{d}x = \int_a^b x^2\varrho\,\sqrt{(1 + y'^2)}\,\mathrm{d}x\,, \tag{15}$$

(b) $$I_x = \int_{t_1}^{t_2} \psi^2(t)\,\varrho(t)\,\sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t)]}\,\mathrm{d}t = \int_{t_1}^{t_2} y^2\varrho\,\sqrt{(\dot{x}^2 + \dot{y}^2)}\,\mathrm{d}t\,, \tag{16}$$

$$I_y = \int_{t_1}^{t_2} \varphi^2(t)\,\varrho(t)\,\sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t)]}\,\mathrm{d}t = \int_{t_1}^{t_2} x^2\varrho\,\sqrt{(\dot{x}^2 + \dot{y}^2)}\,\mathrm{d}t\,, \tag{17}$$

(c) $$I_x = \int_{\varphi_1}^{\varphi_2} \varrho(\varphi)\,r^2(\varphi)\sin^2\varphi\,\sqrt{[r^2(\varphi) + r'^2(\varphi)]}\,\mathrm{d}\varphi\,, \tag{18}$$

$$I_y = \int_{\varphi_1}^{\varphi_2} \varrho(\varphi)\,r^2(\varphi)\cos^2\varphi\,\sqrt{[r^2(\varphi) + r'^2(\varphi)]}\,\mathrm{d}\varphi\,. \tag{18'}$$

When computing moments of inertia with respect to the origin (i.e. about the z-axis) we must replace x^2 or y^2 by the sum $x^2 + y^2$; particularly, in (18) $r^2\sin^2\varphi$ or $r^2\cos^2\varphi$ must be replaced by r^2.

(β) *Curves in space.* A curve c is given parametrically by

$$x = \varphi(t), \quad y = \psi(t), \quad z = \chi(t), \quad t_1 \leqq t \leqq t_2 : \quad \frac{\mathrm{d}\varphi}{\mathrm{d}t} = \dot{\varphi}, \quad \text{etc.}$$

LENGTH:

$$l = \int_{t_1}^{t_2} \sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t) + \dot{\chi}^2(t)]}\, \mathrm{d}t = \int_{t_1}^{t_2} \sqrt{(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)}\, \mathrm{d}t\,. \tag{19}$$

MASS:

$$M = \int_{t_1}^{t_2} \varrho(t) \sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t) + \dot{\chi}^2(t)]}\, \mathrm{d}t = \int_{t_1}^{t_2} \varrho \sqrt{(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)}\, \mathrm{d}t\,. \tag{20}$$

STATICAL MOMENT WITH RESPECT TO THE xy- OR xz- OR yz-PLANE:

$$S_{xy} = \int_{t_1}^{t_2} \chi(t)\, \varrho(t) \sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t) + \dot{\chi}^2(t)]}\, \mathrm{d}t = \int_{t_1}^{t_2} z\varrho \sqrt{(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)}\, \mathrm{d}t\,, \tag{21}$$

$$S_{xz} = \int_{t_1}^{t_2} \psi(t)\, \varrho(t) \sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t) + \dot{\chi}^2(t)]}\, \mathrm{d}t = \int_{t_1}^{t_2} y\varrho \sqrt{(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)}\, \mathrm{d}t\,, \tag{22}$$

$$S_{yz} = \int_{t_1}^{t_2} \varphi(t)\, \varrho(t) \sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t) + \dot{\chi}^2(t)]}\, \mathrm{d}t = \int_{t_1}^{t_2} x\varrho \sqrt{(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)}\, \mathrm{d}t\,. \tag{23}$$

COORDINATES OF THE CENTRE OF GRAVITY:

$$x_T = \frac{S_{yz}}{M}, \quad y_T = \frac{S_{xz}}{M}, \quad z_T = \frac{S_{xy}}{M}\,. \tag{24}$$

MOMENT OF INERTIA WITH RESPECT TO THE x- OR y- OR z-AXIS:

$$I_x = \int_{t_1}^{t_2} [\psi^2(t) + \chi^2(t)]\, \varrho(t) \sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t) + \dot{\chi}^2(t)]}\, \mathrm{d}t =$$

$$= \int_{t_1}^{t_2} (y^2 + z^2)\, \varrho \sqrt{(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)}\, \mathrm{d}t\,, \tag{25}$$

$$I_y = \int_{t_1}^{t_2} [\varphi^2(t) + \chi^2(t)]\, \varrho(t) \sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t) + \dot{\chi}^2(t)]}\, \mathrm{d}t =$$

$$= \int_{t_1}^{t_2} (x^2 + z^2)\, \varrho \sqrt{(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)}\, \mathrm{d}t\,, \tag{26}$$

$$I_z = \int_{t_1}^{t_2} [\varphi^2(t) + \psi^2(t)]\, \varrho(t) \sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t) + \dot{\chi}^2(t)]}\, \mathrm{d}t =$$

$$= \int_{t_1}^{t_2} (x^2 + y^2)\, \varrho \sqrt{(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)}\, \mathrm{d}t\,. \tag{27}$$

If a curve is given as the intersection of two surfaces, then for the above computations it is usually better to establish its parametric representation.

(b) Plane Figures.

(a) Ω is a region bounded by a closed curve k positively oriented with respect to Ω (Fig. 14.22; see Remark 14.7.1).

If k is given parametrically, i.e. if

$$x = \varphi(t), \quad y = \psi(t), \quad t_1 \leqq t \leqq t_2,$$

then as t increases from t_1 to t_2, the point (x, y) moves along the curve k so that Ω remains on the left-hand side (Fig. 14.22).

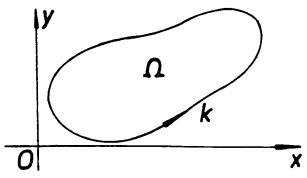


Fig. 14.22.

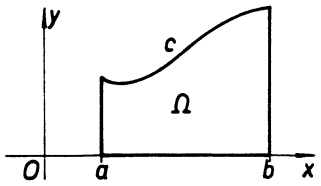


Fig. 14.23.

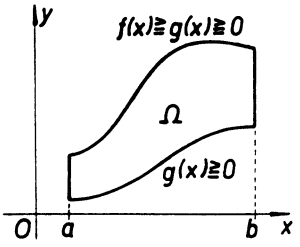


Fig. 14.24.

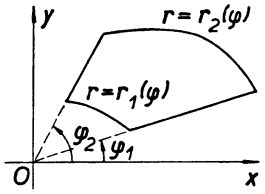


Fig. 14.25.

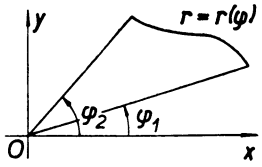


Fig. 14.26.

The specific density (the mass per unit area, $\mathrm{kg\,m^{-2}}$) is denoted by $\sigma(x, y)$.

(b) The region Ω is given by a curve c according to Fig. 14.23; the curve c is the graph of a continuous function $y = f(x)$, $f(x) \geqq 0$ in $[a, b]$, or is given parametrically,

$$x = \varphi(t), \quad y = \psi(t), \quad t_1 \leqq t \leqq t_2,$$
$$\psi(t) \geqq 0, \quad \varphi(t_1) = a, \quad \varphi(t_2) = b.$$

The specific density is denoted by $\sigma(x)$ ($\mathrm{kg\,m^{-2}}$).

(The area of the region Ω in Fig. 14.24 is obtained, of course, as the difference between areas of regions of the type shown in Fig. 14.23. Similarly for masses, statical moments, coordinates of the centre of gravity, moments of inertia, etc. The case of parametric representation is treated in a similar way.)

(c) The region Ω is given according to Fig. 14.25, $0 < r_1(\varphi) < r_2(\varphi)$ in (φ_1, φ_2). The specific density is denoted by $\sigma(r, \varphi)$.

(d) If $r_1 \equiv 0$ in (c) we get a sector (Fig. 14.26). (We omit the index in $r_2(\varphi)$ by writing $r(\varphi)$.)

If $\sigma = \sigma(x)$ or $\sigma = \text{const.}$, some formulae may be simplified. (If $\sigma(x) = \text{const.}$, or $\sigma(x, y)) = \text{const.}$, or $\sigma(r, \varphi) = \text{const.}$, then, of course, σ can be put in front of the integral sign.)

AREA:

(a)
$$P = \iint_{\Omega} \mathrm{d}x\,\mathrm{d}y = \tag{28}$$

$$= \int_k x\,\mathrm{d}y = \int_{t_1}^{t_2} \varphi(t)\,\dot{\psi}(t)\,\mathrm{d}t = \tag{29}$$

$$= -\int_k y\,\mathrm{d}x = -\int_{t_1}^{t_2} \psi(t)\,\dot{\varphi}(t)\,\mathrm{d}t = \tag{30}$$

$$= \tfrac{1}{2}\int_k (x\,\mathrm{d}y - y\,\mathrm{d}x) = \tfrac{1}{2}\int_{t_1}^{t_2} [\varphi(t)\,\dot{\psi}(t) - \psi(t)\,\dot{\varphi}(t)]\,\mathrm{d}t. \tag{31}$$

In particular, for a sector, see Remark 14.7.14, p. □ .

(b)
$$P = \int_a^b f(x)\,\mathrm{d}x = \int_a^b y\,\mathrm{d}x, \tag{32}$$

$$P = \int_{t_1}^{t_2} \psi(t)\,\dot{\varphi}(t)\,\mathrm{d}t = \int_{t_1}^{t_2} y\dot{x}\,\mathrm{d}t. \tag{33}$$

(c) $$P = \int_{\varphi_1}^{\varphi_2}\left[\int_{r_1(\varphi)}^{r_2(\varphi)} r\,\mathrm{d}r\right]\mathrm{d}\varphi = \tfrac{1}{2}\int_{\varphi_1}^{\varphi_2}\left[r_2^2(\varphi) - r_1^2(\varphi)\right]\mathrm{d}\varphi\,. \tag{34}$$

(d) $$P = \tfrac{1}{2}\int_{\varphi_1}^{\varphi_2} r^2(\varphi)\,\mathrm{d}\varphi\,. \tag{34'}$$

MASS:

(a) $$M = \iint_{\Omega}\sigma(x, y)\,\mathrm{d}x\,\mathrm{d}y\,. \tag{35}$$

(b) $$M = \int_a^b \sigma(x) f(x)\,\mathrm{d}x = \int_a^b \sigma y\,\mathrm{d}x\,. \tag{36}$$

(c) $$M = \int_{\varphi_1}^{\varphi_2}\left[\int_{r_1(\varphi)}^{r_2(\varphi)} r\sigma(r, \varphi)\,\mathrm{d}r\right]\mathrm{d}\varphi\,. \tag{37}$$

(d) For $\sigma = \sigma(\varphi)$ we have

$$M = \tfrac{1}{2}\int_{\varphi_1}^{\varphi_2}\sigma(\varphi)\, r^2(\varphi)\,\mathrm{d}\varphi\,. \tag{37'}$$

STATICAL MOMENT WITH RESPECT TO THE x- OR y-AXIS:

(a) $$S_x = \iint_{\Omega} y\sigma(x, y)\,\mathrm{d}x\,\mathrm{d}y\,, \quad S_y = \iint_{\Omega} x\sigma(x, y)\,\mathrm{d}x\,\mathrm{d}y\,. \tag{38}$$

(b) $$S_x = \int_a^b \sigma(x)\frac{f^2(x)}{2}\,\mathrm{d}x = \tfrac{1}{2}\int_a^b \sigma y^2\,\mathrm{d}x\,, \quad S_y = \int_a^b x\sigma(x) f(x)\,\mathrm{d}x = \int_a^b \sigma x y\,\mathrm{d}x\,. \tag{39}$$

(c) $$S_x = \int_{\varphi_1}^{\varphi_2}\sin\varphi\left[\int_{r_1(\varphi)}^{r_2(\varphi)} r^2\sigma(r, \varphi)\,\mathrm{d}r\right]\mathrm{d}\varphi\,, \tag{40}$$

$$S_y = \int_{\varphi_1}^{\varphi_2}\cos\varphi\left[\int_{r_1(\varphi)}^{r_2(\varphi)} r^2\sigma(r, \varphi)\,\mathrm{d}r\right]\mathrm{d}\varphi\,. \tag{40'}$$

(d) For $\sigma = \sigma(\varphi)$ we have

$$S_x = \tfrac{1}{3}\int_{\varphi_1}^{\varphi_2}\sigma(\varphi)\, r^3(\varphi)\sin\varphi\,\mathrm{d}\varphi\,, \quad S_y = \tfrac{1}{3}\int_{\varphi_1}^{\varphi_2}\sigma(\varphi)\, r^3(\varphi)\cos\varphi\,\mathrm{d}\varphi\,. \tag{40''}$$

COORDINATES OF THE CENTRE OF GRAVITY:

$$x_T = \frac{S_y}{M}\,, \quad y_T = \frac{S_x}{M}\,.$$

MOMENT OF INERTIA WITH RESPECT TO THE x- OR y-AXIS, OR TO THE ORIGIN (I. E. ABOUT THE z-AXIS):

(a)
$$I_x = \iint_\Omega y^2\sigma(x, y)\,\mathrm{d}x\,\mathrm{d}y\,, \quad I_y = \iint_\Omega x^2\sigma(x, y)\,\mathrm{d}x\,\mathrm{d}y\,, \tag{41}$$

$$I_z = \iint_\Omega (x^2 + y^2)\,\sigma(x, y)\,\mathrm{d}x\,\mathrm{d}y = I_x + I_y\,. \tag{42}$$

(b)
$$I_x = \int_a^b \sigma(x)\frac{f^3(x)}{3}\,\mathrm{d}x = \tfrac{1}{3}\int_a^b \sigma y^3\,\mathrm{d}x\,, \tag{43}$$

$$I_y = \int_a^b x^2\sigma(x)f(x)\,\mathrm{d}x = \int_a^b \sigma x^2 y\,\mathrm{d}x\,, \tag{44}$$

$$I_z = \int_a^b \sigma(x)\left[\frac{f^3(x)}{3} + x^2 f(x)\right]\mathrm{d}x =$$

$$= \int_a^b \sigma\left(\frac{y^3}{3} + x^2 y\right)\mathrm{d}x = I_x + I_y\,. \tag{45}$$

(c)
$$I_x = \int_{\varphi_1}^{\varphi_2} \sin^2\varphi\left[\int_{r_1(\varphi)}^{r_2(\varphi)} r^3\sigma(r, \varphi)\,\mathrm{d}r\right]\mathrm{d}\varphi\,,$$

$$I_y = \int_{\varphi_1}^{\varphi_2} \cos^2\varphi\left[\int_{r_1(\varphi)}^{r_2(\varphi)} r^3\sigma(r, \varphi)\,\mathrm{d}r\right]\mathrm{d}\varphi\,, \tag{46}$$

$$I_z = \int_{\varphi_1}^{\varphi_2}\left[\int_{r_1(\varphi)}^{r_2(\varphi)} r^3\sigma(r, \varphi)\,\mathrm{d}r\right]\mathrm{d}\varphi = I_x + I_y\,. \tag{47}$$

(d) For $\sigma = \sigma(\varphi)$ we have

$$I_x = \tfrac{1}{4}\int_{\varphi_1}^{\varphi_2} \sigma(\varphi)\,r^4(\varphi)\sin^2\varphi\,\mathrm{d}\varphi\,,$$

$$I_y = \tfrac{1}{4}\int_{\varphi_1}^{\varphi_2} \sigma(\varphi)\,r^4(\varphi)\cos^2\varphi\,\mathrm{d}\varphi\,, \tag{46'}$$

$$I_z = \tfrac{1}{4}\int_{\varphi_1}^{\varphi_2} \sigma(\varphi)\,r^4(\varphi)\,\mathrm{d}\varphi = I_x + I_y\,. \tag{47'}$$

(c) Solids.

(a) The solid is a three-dimensional region T of type A (Definition 14.1.4) (Fig. 14.27). The specific density (mass per unit volume) is denoted by $\varrho(x, y, z)$ ($\mathrm{kg\,m^{-3}}$).

(b) The solid T is a special region of type A: The lower base Ω lies in the xy-plane, the upper one is the surface $z = f(x, y)$ and the lateral surface is formed by parallels to the z-axis (Fig. 14.28). The specific mass is denoted by $\varrho = \varrho(x, y)$.

(c) T is a solid with the lower base $\Omega(r_1(\varphi) < r < r_2(\varphi),\ \varphi_1 < \varphi < \varphi_2)$ and the upper base $z = z(r, \varphi)$; the lateral surface is formed by parallels to the z-axis (Fig. 14.29). The specific mass is denoted by $\varrho(r, \varphi)$.

(d) T is a solid with the bases P_1, P_2 perpendicular to the x-axis. P_1 lies in the plane $x = a$, P_2 in the plane $x = b$. The areas p of all cross-sections perpendicular to the x-axis are known, $p = p(x)$ (Fig. 14.30). The specific mass is denoted by $\varrho(x)$.

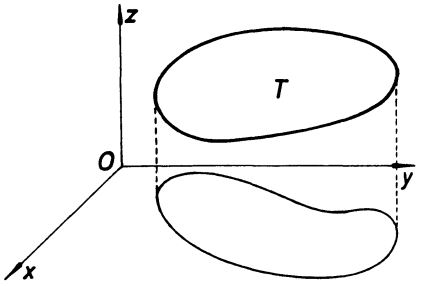


Fig. 14.27.

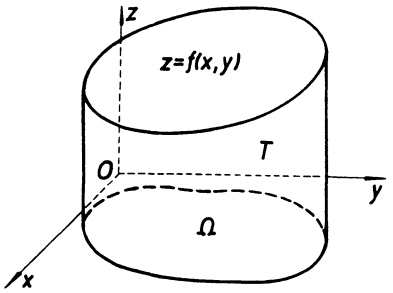


Fig. 14.28.

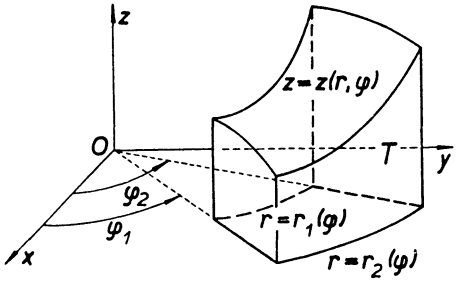


Fig. 14.29.

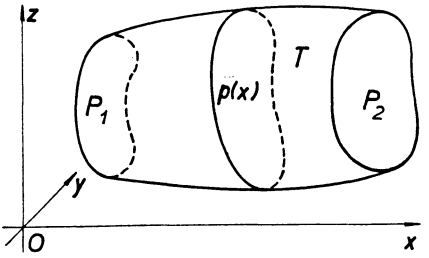


Fig. 14.30.

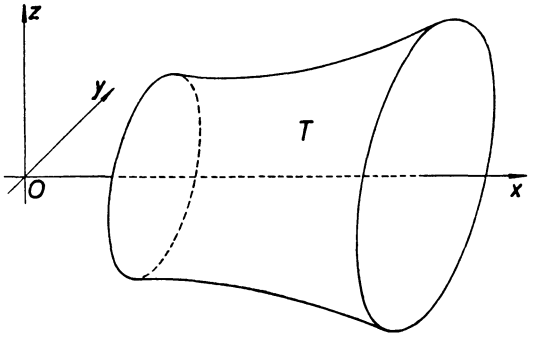


Fig. 14.31.

(e) T is a solid of revolution, whose lateral surface is formed by rotating the curve $y = f(x)$, $a \leqq x \leqq b$, $f(x) \geqq 0$, around the x-axis (Fig. 14.31). The specific mass is denoted by $\varrho(x)$.

In the cases where $\varrho = \text{const.}$, ϱ can be put in front of the integral sign, of course.

(a) $$V = \iiint_T \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,. \tag{48}$$

(b) $$V = \iint_\Omega f(x, y)\,\mathrm{d}x\,\mathrm{d}y\,. \tag{49}$$

(c) $$V = \int_{\varphi_1}^{\varphi_2} \left[\int_{r_1(\varphi)}^{r_2(\varphi)} rz(r, \varphi)\,\mathrm{d}r \right] \mathrm{d}\varphi\,. \tag{50}$$

(d) $$V = \int_a^b p(x)\,\mathrm{d}x\,. \tag{51}$$

If, especially, $p(x)$ is a polynomial of at most the third degree, then

$$V = \frac{b-a}{6}(p_1 + 4p_m + p_2)\,, \tag{52}$$

where p_1, p_2 are areas of the bases P_1, P_2 and p_m is the area of the mean cross-section (i.e. for $x = \frac{1}{2}(a + b)$).

(e) $$V = \pi \int_a^b f^2(x)\,\mathrm{d}x = \pi \int_a^b y^2\,\mathrm{d}x\,. \tag{53}$$

If the lateral surface of a solid of revolution is formed by rotating a curve c around the x-axis and c is given parametrically by $x = \varphi(t)$, $y = \psi(t)$, $t_1 \leqq t \leqq t_2$, $\psi(t) \geqq 0$, $\mathrm{d}\varphi/\mathrm{d}t = \dot{\varphi}(t) > 0$, then

$$V = \pi \int_{t_1}^{t_2} \psi^2(t)\,\dot{\varphi}(t)\,\mathrm{d}t\,. \tag{53'}$$

If $\dot{\varphi} < 0$, then

$$V = \pi \int_{t_2}^{t_1} \psi^2(t)\dot{\varphi}(t)\,\mathrm{d}t\,. \tag{53''}$$

MASS:

(a) $$M = \iiint_T \varrho(x, y, z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,. \tag{54}$$

(b) $$M = \iint_\Omega \varrho(x, y) f(x, y)\,\mathrm{d}x\,\mathrm{d}y\,. \tag{55}$$

(c) $$M = \int_{\varphi_1}^{\varphi_2} \left[\int_{r_1(\varphi)}^{r_2(\varphi)} r\varrho(r, \varphi)\, z(r, \varphi)\,\mathrm{d}r \right] \mathrm{d}\varphi\,. \tag{56}$$

(d) $$M = \int_a^b \varrho(x)\, p(x)\, \mathrm{d}x\,. \tag{57}$$

(e) $$M = \pi \int_a^b \varrho(x) f^2(x)\, \mathrm{d}x = \pi \int_a^b \varrho y^2\, \mathrm{d}x\,. \tag{58}$$

STATICAL MOMENT WITH RESPECT TO THE xy-, yz- OR zx-PLANE:

(a) $$S_{xy} = \iiint_T z\varrho(x, y, z)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z\,, \tag{59}$$

$$S_{yz} = \iiint_T x\varrho(x, y, z)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z\,, \tag{59'}$$

$$S_{zx} = \iiint_T y\varrho(x, y, z)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z\,. \tag{59''}$$

(b) $$S_{xy} = \tfrac{1}{2} \iint_\Omega \varrho(x, y) f^2(x, y)\, \mathrm{d}x\, \mathrm{d}y\,, \tag{60}$$

$$S_{yz} = \iint_\Omega x\varrho(x, y) f(x, y)\, \mathrm{d}x\, \mathrm{d}y\,, \quad S_{zx} = \iint_\Omega y\varrho(x, y) f(x, y)\, \mathrm{d}x\, \mathrm{d}y\,. \tag{61}$$

(c) $$S_{xy} = \tfrac{1}{2} \int_{\varphi_1}^{\varphi_2} \left[\int_{r_1(\varphi)}^{r_2(\varphi)} r\varrho(r, \varphi)\, z^2(r, \varphi)\, \mathrm{d}r \right] \mathrm{d}\varphi\,, \tag{62}$$

$$S_{yz} = \int_{\varphi_1}^{\varphi_2} \left[\int_{r_1(\varphi)}^{r_2(\varphi)} r^2 \varrho(r, \varphi)\, z(r, \varphi)\, \mathrm{d}r \right] \cos\varphi\, \mathrm{d}\varphi\,, \tag{63}$$

$$S_{zx} = \int_{\varphi_1}^{\varphi_2} \left[\int_{r_1(\varphi)}^{r_2(\varphi)} r^2 \varrho(r, \varphi)\, z(r, \varphi)\, \mathrm{d}r \right] \sin\varphi\, \mathrm{d}\varphi\,. \tag{64}$$

(d) $$S_{yz} = \int_a^b x\varrho(x)\, p(x)\, \mathrm{d}x\,. \tag{65}$$

(e) $$S_{yz} = \pi \int_a^b x\varrho(x) f^2(x)\, \mathrm{d}x = \pi \int_a^b x\varrho y^2\, \mathrm{d}x\,. \tag{66}$$

COORDINATES OF THE CENTRE OF GRAVITY:

$$x_T = \frac{S_{yz}}{M}\,, \quad y_T = \frac{S_{zx}}{M}\,, \quad z_T = \frac{S_{xy}}{M}\,.$$

MOMENT OF INERTIA WITH RESPECT TO THE x- OR y- OR z-AXIS (see also Steiner's theorem, formula (118)):

(a)
$$I_x = \iiint_T (y^2 + z^2)\,\varrho(x, y, z)\,dx\,dy\,dz\,, \tag{67}$$

$$I_y = \iiint_T (x^2 + z^2)\,\varrho(x, y, z)\,dx\,dy\,dz\,, \tag{68}$$

$$I_z = \iiint_T (x^2 + y^2)\,\varrho(x, y, z)\,dx\,dy\,dz\,. \tag{69}$$

(b)
$$I_z = \iint_O (x^2 + y^2)\,\varrho(x, y,)\,f(x, y)\,dx\,dy\,. \tag{70}$$

(c)
$$I_z = \int_{\varphi_1}^{\varphi_2} \left[\int_{r_1(\varphi)}^{r_2(\varphi)} r^3 \varrho(r, \varphi)\, z(r, \varphi)\,dr \right] d\varphi\,. \tag{71}$$

(e)
$$I_x = \tfrac{1}{2}\pi \int_a^b \varrho(x) f^4(x)\,dx = \tfrac{1}{2}\pi \int_a^b \varrho y^4\,dx\,. \tag{72}$$

(d) Surfaces.

(a) The finite simple piecewise smooth surface (Remark 14.1.7) is given parametrically:
$$x = x(u, v)\,, \quad y = y(u, v)\,, \quad z = z(u, v) \tag{73}$$
$(u, v) \in \Omega$. The density (mass per unit area, kg m^{-2}) is denoted by $\sigma = \sigma(u, v)$.

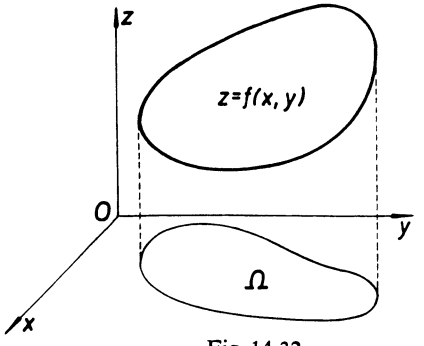


Fig. 14.32.

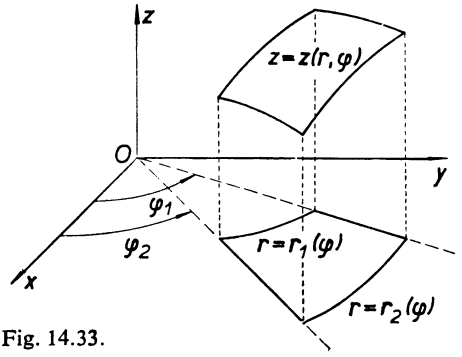


Fig. 14.33.

(b) The surface $z = f(x, y)$ over a region Ω (Fig. 14.32). The density is denoted by $\sigma = \sigma(x, y)$.

(c) The surface $z = z(r, \varphi)$ over a region $\Omega(\varphi_1 < \varphi < \varphi_2, r_1(\varphi) < r < r_2(\varphi))$, Fig. 14.33. The density is denoted by $\sigma(r, \varphi)$.

(d) The surface of revolution obtained by rotating a curve $y = f(x)$, $a \leqq x \leqq b$, $f(x) \geqq 0$ around the x-axis. The density is denoted by $\sigma(x)$.

(e) The surface of revolution obtained by rotating a simple curve $x = \varphi(t)$, $y = \psi(t)$, $t_1 \leqq t \leqq t_2$, $\psi(t) \geqq 0$, $\mathrm{d}\varphi/\mathrm{d}t = \dot{\varphi}(t) > 0$ around the x-axis. The density is denoted by $\sigma(t)$.

(f) The surface of revolution obtained by rotating a curve $r = r(\varphi), 0 \leqq \varphi_1 \leqq \varphi \leqq \leqq \varphi_2 \leqq \pi$ around the x-axis. The density is denoted by $\sigma(\varphi)$.

(If $\sigma = \text{const.}$, σ can be put in front of the integral sign.)

AREA:

(a)
$$P = \iint_{\sigma} \sqrt{(EG - F^2)}\, \mathrm{d}u\, \mathrm{d}v\,, \tag{74}$$

where

$$E = \left(\frac{\partial x}{\partial u}\right)^2 + \left(\frac{\partial y}{\partial u}\right)^2 + \left(\frac{\partial z}{\partial u}\right)^2, \quad G = \left(\frac{\partial x}{\partial v}\right)^2 + \left(\frac{\partial y}{\partial v}\right)^2 + \left(\frac{\partial z}{\partial v}\right)^2, \tag{75}$$

$$F = \frac{\partial x}{\partial u}\frac{\partial x}{\partial v} + \frac{\partial y}{\partial u}\frac{\partial y}{\partial v} + \frac{\partial z}{\partial u}\frac{\partial z}{\partial v}. \tag{76}$$

(b)
$$P = \iint_{\Omega} \sqrt{\left[1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2\right]}\, \mathrm{d}x\, \mathrm{d}y\,. \tag{77}$$

(c)
$$P = \int_{\varphi_1}^{\varphi_2} \left[\int_{r_1(\varphi)}^{r_2(\varphi)} \sqrt{\left[r^2 + r^2\left(\frac{\partial z}{\partial r}\right)^2 + \left(\frac{\partial z}{\partial \varphi}\right)^2\right]}\, \mathrm{d}r\right] \mathrm{d}\varphi\,. \tag{78}$$

(d)
$$P = 2\pi \int_a^b f(x) \sqrt{[1 + f'^2(x)]}\, \mathrm{d}x = 2\pi \int_a^b y \sqrt{(1 + y'^2)}\, \mathrm{d}x\,. \tag{79}$$

(e)
$$P = 2\pi \int_{t_1}^{t_2} \psi(t) \sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t)]}\, \mathrm{d}t = 2\pi \int_{t_1}^{t_2} y \sqrt{(\dot{x}^2 + \dot{y}^2)}\, \mathrm{d}t\,. \tag{80}$$

(f)
$$P = 2\pi \int_{\varphi_1}^{\varphi_2} r(\varphi) \sin\varphi \sqrt{[r^2(\varphi) + r'^2(\varphi)]}\, \mathrm{d}\varphi\,. \tag{81}$$

MASS:

(a)
$$M = \iint_{\Omega} \sigma(u, v) \sqrt{(EG - F^2)}\, \mathrm{d}u\, \mathrm{d}v \quad \text{(see (75), (76))}\,. \tag{82}$$

(b)
$$M = \iint_{\Omega} \sigma(x, y) \sqrt{\left[1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2\right]}\, \mathrm{d}x\, \mathrm{d}y\,. \tag{83}$$

(c) $$M = \int_{\varphi_1}^{\varphi_2} \left[\int_{r_1(\varphi)}^{r_2(\varphi)} \sigma(r, \varphi) \sqrt{\left[r^2 + r^2 \left(\frac{\partial z}{\partial r} \right)^2 + \left(\frac{\partial z}{\partial \varphi} \right)^2 \right]} \, dr \right] d\varphi . \tag{84}$$

(d) $$M = 2\pi \int_a^b \sigma(x) f(x) \sqrt{[1 + f'^2(x)]} \, dx = 2\pi \int_a^b \sigma y \sqrt{(1 + y'^2)} \, dx . \tag{85}$$

(e) $$M = 2\pi \int_{t_1}^{t_2} \sigma(t) \psi(t) \sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t)]} \, dt = 2\pi \int_{t_1}^{t_2} \sigma y \sqrt{(\dot{x}^2 + \dot{y}^2)} \, dt . \tag{86}$$

(f) $$M = 2\pi \int_{\varphi_1}^{\varphi_2} \sigma(\varphi) r(\varphi) \sin \varphi \sqrt{[r^2(\varphi) + r'^2(\varphi)]} \, d\varphi . \tag{87}$$

STATICAL MOMENT WITH RESPECT TO THE xy- OR yz- OR zx-PLANE:

(a) $$S_{xy} = \iint_\Omega \sigma(u, v) z(u, v) \sqrt{(EG - F^2)} \, du \, dv \quad \text{(see (75), (76))} , \tag{88}$$

$$S_{yz} = \iint_\Omega \sigma(u, v) x(u, v) \sqrt{(EG - F^2)} \, du \, dv ,$$

$$S_{zx} = \iint_\Omega \sigma(u, v) y(u, v) \sqrt{(EG - F^2)} \, du \, dv . \tag{89}$$

(b) $$S_{xy} = \iint_\Omega \sigma(x, y) f(x, y) \sqrt{\left[1 + \left(\frac{\partial f}{\partial x} \right)^2 + \left(\frac{\partial f}{\partial y} \right)^2 \right]} \, dx \, dy , \tag{90}$$

$$S_{yz} = \iint_\Omega x\sigma(x, y) \sqrt{\left[1 + \left(\frac{\partial f}{\partial x} \right)^2 + \left(\frac{\partial f}{\partial y} \right)^2 \right]} \, dx \, dy ,$$

$$S_{zx} = \iint_\Omega y\sigma(x, y) \sqrt{\left[1 + \left(\frac{\partial f}{\partial x} \right)^2 + \left(\frac{\partial f}{\partial y} \right)^2 \right]} \, dx \, dy . \tag{91}$$

(c) $$S_{xy} = \int_{\varphi_1}^{\varphi_2} \left[\int_{r_1(\varphi)}^{r_2(\varphi)} \sigma(r, \varphi) z(r, \varphi) \sqrt{\left[r^2 + r^2 \left(\frac{\partial z}{\partial r} \right)^2 + \left(\frac{\partial z}{\partial \varphi} \right)^2 \right]} \, dr \right] d\varphi , \tag{92}$$

$$S_{yz} = \int_{\varphi_1}^{\varphi_2} \left[\int_{r_1(\varphi)}^{r_2(\varphi)} \sigma(r, \varphi) r \sqrt{\left[r^2 + r^2 \left(\frac{\partial z}{\partial r} \right)^2 + \left(\frac{\partial z}{\partial \varphi} \right)^2 \right]} \, dr \right] \cos \varphi \, d\varphi , \tag{93}$$

$$S_{zx} = \int_{\varphi_1}^{\varphi_2} \left[\int_{r_1(\varphi)}^{r_2(\varphi)} \sigma(r, \varphi) r \sqrt{\left[r^2 + r^2 \left(\frac{\partial z}{\partial r} \right)^2 + \left(\frac{\partial z}{\partial \varphi} \right)^2 \right]} \, dr \right] \sin \varphi \, d\varphi . \tag{94}$$

(d) $$S_{xy} = 0 , \quad S_{zx} = 0 , \quad S_{yz} = 2\pi \int_a^b x\sigma(x) f(x) \sqrt{[1 + f'^2(x)]} \, dx =$$

$$= 2\pi \int_a^b x\sigma y \sqrt{(1 + y'^2)} \, dx . \tag{95}$$

(e) $S_{xy} = 0\,, \quad S_{zx} = 0\,, \quad S_{yz} = 2\pi \int_{t_1}^{t_2} \sigma(t)\, \varphi(t)\, \psi(t) \sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t)]}\, \mathrm{d}t =$

$$= 2\pi \int_{t_1}^{t_2} \sigma x y \sqrt{(\dot{x}^2 + \dot{y}^2)}\, \mathrm{d}t\,. \tag{96}$$

(f) $S_{xy} = 0\,, \quad S_{zx} = 0\,, \quad S_{yz} = 2\pi \int_{\varphi_1}^{\varphi_2} \sigma(\varphi) r^2(\varphi) \sin\varphi \cos\varphi \sqrt{[r^2(\varphi) + r'^2(\varphi)]}\, \mathrm{d}\varphi\,.$ (97)

COORDINATES OF THE CENTRE OF GRAVITY:

$$x_T = \frac{S_{yz}}{M}\,, \quad y_T = \frac{S_{zx}}{M}\,, \quad z_T = \frac{S_{xy}}{M}\,. \tag{98}$$

MOMENT OF INERTIA WITH RESPECT TO THE x- OR y- OR z-AXIS:

(a) $$I_x = \iint_\Omega \sigma(u, v) \left[y^2(u, v) + z^2(u, v)\right] \sqrt{(EG - F^2)}\, \mathrm{d}u\, \mathrm{d}v \quad \text{(see (75), (76))}, \tag{99}$$

$$I_y = \iint_\Omega \sigma(u, v) \left[x^2(u, v) + z^2(u, v)\right] \sqrt{(EG - F^2)}\, \mathrm{d}u\, \mathrm{d}v\,,$$

$$I_z = \iint_\Omega \sigma(u, v) \left[x^2(u, v) + y^2(u, v)\right] \sqrt{(EG - F^2)}\, \mathrm{d}u\, \mathrm{d}v\,. \tag{100}$$

(b) $$I_x = \iint_\Omega \sigma(x, y) \left[y^2 + f^2(x, y)\right] \sqrt{\left[1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2\right]}\, \mathrm{d}x\, \mathrm{d}y\,, \tag{101}$$

$$I_y = \iint_\Omega \sigma(x, y) \left[x^2 + f^2(x, y)\right] \sqrt{\left[1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2\right]}\, \mathrm{d}x\, \mathrm{d}y\,, \tag{102}$$

$$I_z = \iint_\Omega \sigma(x, y) \left(x^2 + y^2\right) \sqrt{\left[1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2\right]}\, \mathrm{d}x\, \mathrm{d}y\,. \tag{103}$$

(c)

$$I_x = \int_{\varphi_1}^{\varphi_2} \left[\int_{r_1(\varphi)}^{r_2(\varphi)} \sigma(r, \varphi) \left[r^2 \sin^2\varphi + z^2(r, \varphi)\right] \sqrt{\left[r^2 + r^2 \left(\frac{\partial z}{\partial r}\right)^2 + \left(\frac{\partial z}{\partial \varphi}\right)^2\right]}\, \mathrm{d}r\right] \mathrm{d}\varphi\,, \tag{104}$$

$$I_y = \int_{\varphi_1}^{\varphi_2} \left[\int_{r_1(\varphi)}^{r_2(\varphi)} \sigma(r, \varphi) \left[r^2 \cos^2\varphi + z^2(r, \varphi)\right] \sqrt{\left[r^2 + r^2 \left(\frac{\partial z}{\partial r}\right)^2 + \left(\frac{\partial z}{\partial \varphi}\right)^2\right]}\, \mathrm{d}r\right] \mathrm{d}\varphi\,, \tag{105}$$

$$I_z = \int_{\varphi_1}^{\varphi_2} \left[\int_{r_1(\varphi)}^{r_2(\varphi)} \sigma(r, \varphi)\, r^2 \sqrt{\left[r^2 + r^2 \left(\frac{\partial z}{\partial r} \right)^2 + \left(\frac{\partial z}{\partial \varphi} \right)^2 \right]} \, dr \right] d\varphi . \tag{106}$$

(d) $$I_x = 2\pi \int_a^b \sigma(x) f^3(x) \sqrt{[1 + f'^2(x)]}\, dx = 2\pi \int_a^b \sigma y^3 \sqrt{(1 + y'^2)}\, dx . \tag{107}$$

(e) $$I_x = 2\pi \int_{t_1}^{t_2} \sigma(t)\, \psi^3(t) \sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t)]}\, dt = 2\pi \int_{t_1}^{t_2} \sigma y^3 \sqrt{(\dot{x}^2 + \dot{y}^2)}\, dt . \tag{108}$$

(f) $$I_x = 2\pi \int_{\varphi_1}^{\varphi_2} \sigma(\varphi)\, r^3(\varphi) \sin^3 \varphi \sqrt{[r^2(\varphi) + r'^2(\varphi)]}\, d\varphi . \tag{109}$$

(e) The Work Done by a Force Moving Along a Given Curve. The work L done by a force $\boldsymbol{P}$ in moving along an oriented curve c in a field of force

$$\boldsymbol{P} = \boldsymbol{i} P_1(x, y, z) + \boldsymbol{j} P_2(x, y, z) + \boldsymbol{k} P_3(x, y, z)$$

is given by the sum of curvilinear integrals

$$L = \int_c P_1(x, y, z)\, dx + \int_c P_2(x, y, z)\, dy + \int_c P_3(x, y, z)\, dz . \tag{110}$$

If c is given parametrically and is positively oriented with increasing parameter (Remark 14.7.1, p. 599), we have

$$L = \int_{t_1}^{t_2} P_1(\varphi(t), \psi(t), \chi(t))\, \varphi'(t)\, dt + \int_{t_1}^{t_2} P_2(\varphi(t), \psi(t), \chi(t))\, \psi'(t)\, dt + \\ + \int_{t_1}^{t_2} P_3(\varphi(t), \psi(t), \chi(t))\, \chi'(t)\, dt . \tag{111}$$

In the plane (where $\boldsymbol{P} = \boldsymbol{i} P_1(x, y) + \boldsymbol{j} P_2(x, y)$):

$$L = \int_c P_1(x, y)\, dx + \int_c P_2(x, y)\, dy = \\ = \int_{t_1}^{t_2} P_1(\varphi(t), \psi(t))\, \varphi'(t)\, dt + \int_{t_1}^{t_2} P_2(\varphi(t), \psi(t))\, \psi'(t)\, dt . \tag{112}$$

(f) Some Special Formulae. The area of a cylindrical surface $y = f(x)$ cut off by the cylindrical surface $z = g(x)$ and by the planes $z = 0$, $x = a$, $x = b$ ($g(x) \geqq 0$ for $a \leqq x \leqq b$) is given by

$$P = \int_a^b g(x) \sqrt{[1 + f'^2(x)]}\, dx = \int_a^b z \sqrt{(1 + y'^2)}\, dx . \tag{113}$$

The area P of a conical surface with the vertex at the origin and whose base is the curve

$$x = x(t), \quad y = y(t), \quad z = z(t) \quad (t_1 \leqq t \leqq t_2)$$

(i.e. the surface composed of lines joining the origin to points on this curve):

$$P = \tfrac{1}{2} \int_{t_1}^{t_2} \sqrt{[(x\dot{y} - y\dot{x})^2 + (x\dot{z} - z\dot{x})^2 + (y\dot{z} - z\dot{y})^2]}\, \mathrm{d}t . \tag{114}$$

(g) Guldin's Rules.* Let V be a solid of revolution obtained by rotating a region Ω (of type A, Definition 14.1.2, p. 573) with boundary H (which does not intersect the x-axis), around the x-axis. Then we have

$$S = 2\pi l y_T, \quad V = 2\pi P Y_T, \tag{115}$$

where S denotes the surface area of the solid, V the volume of the solid, P the area of the region Ω, l the length of the boundary H, y_T the y-coordinate of the centre of gravity of the boundary H and Y_T the y-coordinate of the centre of gravity of the region Ω (see Example 2).

Guldin's rules remain true (under obvious assumptions) even in the following more general case: A given profile with the area P and with the length l of boundary is moved so that its plane remains perpendicular to the (spatial) curve described by its centre of gravity. Then

(a) the area of the lateral surface is equal to ld, (116)

(b) the volume of the solid is equal to PD, (117)

where D is the length of the trajectory traversed by the centre of gravity of the profile and d the length of the trajectory traversed by the centre of gravity of the boundary of the profile.

(h) Steiner's Theorem (Parallel Axes Theorem). The moment of inertia I_p of a (not necessarily homogeneous) solid with respect to a given straight line p is equal to the moment of inertia I_r of this solid with respect to the straight line r parallel to p and passing through the centre of gravity of the solid, plus a^2M, where M is the mass of the solid and a is the distance between the straight lines p and r; i.e.

$$I_p = I_r + a^2 M . \tag{118}$$

(i) Examples.

Example 1. Let us calculate the moment of inertia of a homogeneous cone ($\varrho = \varrho_0 = \text{const.}$) with respect to its axis, where v is the height of the cone and r is the radius of the base.

* Otherwise known as Pappus's Rules.

Let the axis of the cone be the x-axis, and let the vertex be at the origin. The lateral surface of the cone is obtained by rotating the segment

$$y = \frac{r}{v} x \quad (0 \leqq x \leqq v)$$

around the x-axis. By (72) we have

$$I_x = \frac{\pi \varrho_0}{2} \int_0^v \frac{r^4}{v^4} x^4 \, dx = \frac{\pi \varrho_0}{10} r^4 v \, .$$

Example 2. Let us find the volume and surface area of a *torus* (Fig. 14.34; the torus is obtained by rotating the circle shaded on the figure, around the x-axis).

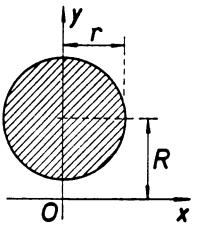


Fig. 14.34.

According to (115) we have

$$S = 2\pi \, . \, 2\pi r \, . \, R = 4\pi^2 r R \, ,$$

$$V = 2\pi \, . \, \pi r^2 \, . \, R = 2\pi^2 r^2 R \, .$$

14.10. Survey of Some Important Formulae in Chapter 14

(See also physical and geometrical applications in § 14.9, also Theorem 14.8.9 and Remark 14.8.15.)

1. $$\iint_\Omega [c_1 f_1(x, y) + c_2 f_2(x, y)] \, dx \, dy =$$

$$= c_1 \iint_\Omega f_1(x, y) \, dx \, dy + c_2 \iint_\Omega f_2(x, y) \, dx \, dy$$

and similarly for triple integrals (Theorem 14.2.5).

2. $$\iint_\Omega f(x, y) \, dx \, dy = \int_a^b dx \int_{h_1(x)}^{h_2(x)} f(x, y) \, dy = \int_c^d dy \int_{\varphi_1(y)}^{\varphi_2(y)} f(x, y) \, dx \, .$$

(Evaluation of a double integral by successive integration, Remark 14.3.5, p. 582. Similar formulae valid for triple integrals are given in Remark 14.5.6, p. 591)

3. $$\iint_M f(x, y)\,\mathrm{d}x\,\mathrm{d}y = \iint_N f(x(u, v), y(u, v))\,|D(u, v)|\,\mathrm{d}u\,\mathrm{d}v\,.$$

(Substitution in a double integral, Theorem 14.4.1, p. 586; for polar coordinates $x = \rho\cos\varphi$, $y = \rho\sin\varphi$ we have $|D(\rho, \varphi)| = \rho$.)

4. $$\iiint_M f(x, y, z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z =$$

$$= \iiint_N f(x(u, v, w), y(u, v, w), z(u, v, w))\,|D(u, v, w)|\,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}w$$

(Theorem 14.5.3); for spherical coordinates $x = r\sin\vartheta\cos\varphi$, $y = r\sin\vartheta\sin\varphi$, $z = r\cos\vartheta$ we have $|D(r, \vartheta, \varphi)| = r^2\sin\vartheta$.

5. $$\int_k f(x, y)\,\mathrm{d}x = \int_\alpha^\beta f(\varphi(t), \psi(t))\,\dot{\varphi}(t)\,\mathrm{d}t\,,$$

$$\int_k f(x, y)\,\mathrm{d}y = \int_\alpha^\beta f(\varphi(t), \psi(t))\,\dot{\psi}(t)\,\mathrm{d}t\,,$$

$$\int_k f(x, y)\,\mathrm{d}s = \int_\alpha^\beta f(\varphi(t), \psi(t))\,\sqrt{[\dot{\varphi}^2(t) + \dot{\psi}^2(t)]}\,\mathrm{d}t$$

(Theorem 14.7.2); the curve k is given parametrically by equations

$$x = \varphi(t)\,,\quad y = \psi(t)\quad (\alpha \leqq t \leqq \beta)$$

and is oriented positively for t increasing.

For similar formulae valid for curves in space see Remark 14.7.8, p. 604.

6. $$\int_k f(x, y)\,\mathrm{d}x = \pm\int_a^b f(x, g(x))\,\mathrm{d}x,$$

provided c is given by the equation $y = f(x)$. For more details see Theorem 14.7.3.

7. $$\int_k (P\,\mathrm{d}x + Q\,\mathrm{d}y) = \iint_\Omega \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\mathrm{d}x\,\mathrm{d}y$$

(Green's theorem, Theorem 14.7.7, p. 605).

8. $$\iiint_V \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z}\right) dx\, dy\, dz = \iint_S (P\, dy\, dz + Q\, dz\, dx + R\, dx\, dy)$$

(Gauss's theorem, Theorem 14.8.5, p. 613).

9. $$\int_k (P dx + Q dy + R dz) =$$

$$\int\int_S \left[\left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dx dy + \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right) dy dz + \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right) dz dx\right]$$

(Stokes's theorem, Theorem 14.8.6, p. 614).

15. SEQUENCES AND SERIES WITH VARIABLE TERMS (SEQUENCES AND SERIES OF FUNCTIONS)

By KAREL REKTORYS

References: [1], [17], [26], [28], [31], [37], [48], [54], [59], [74], [75], [84], [89], [91], [92], [99], [111], [112], [119], [122], [148], [183].

15.1. Sequences with Variable Terms, Uniform Convergence, the Arzelà-Ascoli Theorem. Interchange of Limiting Processes. Integration and Differentiation of Sequences with Variable Terms. Limiting Process under the Integration and Differentiation Signs

Definition 1. Let a sequence of functions

$$f_1(x), f_2(x), f_3(x), \dots \tag{1}$$

defined in an interval I be given. The sequence is said to *converge* (or to *tend) pointwise (briefly to converge) to the (limiting) function* $f(x)$ *in* I, if for every $x_0 \in I$ a finite limit

$$\lim_{n\to\infty} f_n(x_0) = f(x_0) \tag{2}$$

exists.

REMARK 1. The interval I may be either open or closed, finite or infinite, etc. The definition remains unchanged even if the sequence (1) is given on another set than on an interval.

Definition 2. The sequence (1) will be called *uniformly convergent in* I, if for every $\varepsilon > 0$ a number n_0, *independent of the choice of* $x \in I$, can be found such that for every $n > n_0$ and every $x \in I$ we have

$$|f_n(x) - f(x)| < \varepsilon . \tag{3}$$

REMARK 2. Roughly speaking, sequence (1) is uniformly convergent in I, if the functions $f_n(x)$ converge to $f(x)$ "at approximately the same rate" in the whole interval I.

Theorem 1 (*The Bolzano–Cauchy Condition of Convergence*). *The sequence of functions* (1) *is uniformly convergent in* I *if and only if for every* $\varepsilon > 0$ *a positive*

integer n_0 exists such that

$$|f_n(x) - f_m(x)| < \varepsilon$$

holds for every $x \in I$ and for any pair of numbers m, n with $n > n_0$, $m > n_0$.

Theorem 2. *Let the functions* (1) *be continuous in I and let the sequence* (1) *be uniformly convergent in I. Then the limiting function $f(x)$ is also continuous in I.*

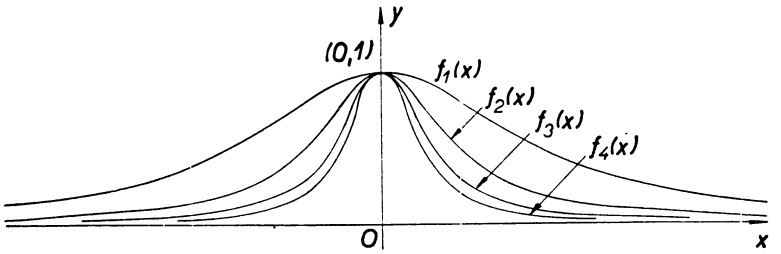


Fig 15.1.

Example 1. Consider the sequence

$$f_n(x) = \frac{1}{1 + n^2x^2} \quad (n = 1, 2, 3, \ldots) \tag{4}$$

in the interval $I = (-\infty, \infty)$ (Fig. 15.1). For every $x \neq 0$, we have

$$\lim_{n\to\infty} f_n(x) = 0$$

since the denominator tends to infinity as $n \to \infty$. For $x = 0$, we have $f_n(0) = 1$ for any n, and consequently,

$$\lim_{n\to\infty} f_n(0) = 1 .$$

Hence, the limiting function $f(x)$ is equal to unity for $x = 0$ and to zero for $x \neq 0$ so that it is not continuous at $x = 0$. Each function (4) is, however, continuous in I. Thus (according to Theorem 2), the sequence (4) cannot be uniformly convergent in I; this fact is also apparent from Fig. 15.1. However, it can be shown that (4) is uniformly convergent in any closed interval which does not contain the point $x = 0$.

Definition 3. The sequence (1) will be called *uniformly bounded in I* if a constant $M > 0$ exists such that

$$|f_n(x)| \leqq M \quad \text{for every } n \text{ and every } x \in I . \tag{5}$$

(For example, the terms of the sequence (4) are uniformly bounded by the constant $M = 1$.)

Definition 4. The terms $f_n(x)$ of the sequence (1) are called *equicontinuous in* I if for every $\varepsilon > 0$ a positive δ can be found such that

$$|f_n(x_1) - f_n(x_2)| < \varepsilon \tag{6}$$

holds for every n and every pair of points $x_1, x_2 \in I$ such that $|x_2 - x_1| < \delta$.

Theorem 3 *(The Arzelà-Ascoli Theorem). Let* (1) *be a uniformly bounded sequence of equicontinuous functions in* I. *Then we can choose from* (1) *a subsequence of functions*

$$f_{k_1}(x), f_{k_2}(x), f_{k_3}(x), \ldots \quad (k_1 < k_2 < k_3 < \ldots) \tag{7}$$

which converges uniformly in I.

REMARK 3. The functions (4) are uniformly bounded in the interval $(-\infty, \infty)$, but they are not equicontinuous in this interval; the latter fact follows easily from Theorems 2, 3 and Example 1 and is also geometrically obvious, since as n increases these functions become "steeper and steeper" in the neighbourhood of the origin.

Theorem 4 (*Interchange of Limiting Processes*). *Let the sequence* (1) *converge uniformly in the interval* $(a, a + \delta)$ $(\delta > 0)$ *to the function* $f(x)$. *Let each of the functions* $f_n(x)$ *tend to a finite limit,*

$$\lim_{x \to a+} f_n(x) = c_n . \tag{8}$$

Then there exist finite limits

$$\lim_{n \to \infty} c_n , \quad \lim_{x \to a+} f(x)$$

and these are equal, i.e.

$$\lim_{n \to \infty} \lim_{x \to a+} f_n(x) = \lim_{x \to a+} \lim_{n \to \infty} f_n(x) . \tag{9}$$

Under similar assumptions, Theorem 4 *is also true for* $x \to a-$ *and for* $x \to a$.

Theorem 5 (*Limiting Process under the Integral Sign*). *Let the sequence* (1) *converge uniformly in* $[a, b]$ *and let the functions* $f(t), f_n(t)$ *be integrable in* $[a, b]$. *Then*

$$\int_a^x f(t)\,\mathrm{d}t = \lim_{n \to \infty} \int_a^x f_n(t)\,\mathrm{d}t \quad \textit{for every} \quad x \in [a, b] , \tag{10}$$

i.e.

$$\int_a^x \lim_{n \to \infty} f_n(t)\,\mathrm{d}t = \lim_{n \to \infty} \int_a^x f_n(t)\,\mathrm{d}t . \tag{11}$$

Moreover, the sequence of functions

$$F_n(x) = \int_a^x f_n(t)\,\mathrm{d}t$$

is also uniformly convergent in $[a, b]$. (Cf. also Theorem 7.)

Theorem 6 (*Limiting Process under the Differentiation Sign*). *Let* (1) *be convergent at least at one point of the interval* $[a, b]$. *Further, let the functions* (1) *have* (*finite*) *derivatives in* $[a, b]$. (*At the points a and b, the derivative from the right or from the left are understood respectively*). *Let the sequence*

$$f_1'(x), f_2'(x), f_3'(x), \ldots \tag{12}$$

be uniformly convergent in $[a, b]$. *Then the sequence* (1) *is also uniformly convergent in* $[a, b]$ *and its limiting function* $f(x)$ (Definition 1) *has a derivative in* $[a, b]$, *while*

$$f'(x) = \lim_{n\to\infty} f_n'(x)\,, \tag{13}$$

i.e.

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[\lim_{n\to\infty} f_n(x)\right] = \lim_{n\to\infty} \frac{\mathrm{d}f_n(x)}{\mathrm{d}x}\,. \tag{14}$$

REMARK 4. Note that the uniform convergence assumed here is that of sequence (12) and not of sequence (1), i.e. it is the *derived* sequence that is assumed to be uniformly convergent.

REMARK 5. If the assumption of uniform convergence is not satisfied in Theorems 5 and 6, then equations (10), (11) and (13), (14) need not hold.

Example 2. For the sequence

$$f_n(x) = n^2 x \mathrm{e}^{-nx} \tag{15}$$

we have, in the interval $[0, 1]$,

$$f(x) = \lim_{n\to\infty} f_n(x) = 0\,. \tag{16}$$

Furthermore

$$\int_0^1 f_n(x)\,\mathrm{d}x = \int_0^1 n^2 x \mathrm{e}^{-nx} = \left[-nx\mathrm{e}^{-nx}\right]_0^1 + n\int_0^1 \mathrm{e}^{-nx}\,\mathrm{d}x = -n\mathrm{e}^{-n} - \mathrm{e}^{-n} + 1\,.$$

Hence

$$\lim_{n\to\infty} \int_0^1 f_n(x)\,\mathrm{d}x = 1\,. \tag{17}$$

However, $f(x) \equiv 0$ in $[0, 1]$ so that

$$\int_0^1 f(x)\,dx = 0 \tag{18}$$

and (10) is not true. (By Theorem 5, the sequence (15) does not converge uniformly in $[0, 1]$; this fact may, of course, be demonstrated in an other way.)

REMARK 6. The condition of uniform convergence may be replaced by other conditions; for example, we have:

Theorem 7. *Let the functions* (1) *together with the limiting function* $f(x)$ *be integrable in* $[a, b]$ *and let* $|f_n(x)| < M$ *in* $[a, b]$ (i.e. the *functions* $f_n(x)$ *are uniformly bounded in* $[a, b]$, see Definition 3). *Then* (10) *and* (11) *hold.* (Cf. also Theorem 13.14.4.)

REMARK 7. Theorem 7 cannot be applied to the sequence (15), since (15) is not uniformly bounded; for example, for $x = 1/n$ we have

$$f_n\left(\frac{1}{n}\right) = n^2 \frac{1}{n} e^{-n/n} = ne^{-1}.$$

15.2. Series with Variable Terms. Uniform Convergence. Integration and Differentiation of Series with Variable Terms

Definition 1. Let a sequence of functions

$$f_1(x), f_2(x), f_3(x), \ldots \tag{1}$$

in an interval I be given, and for each $x_0 \in I$ let the series of numbers

$$f_1(x_0) + f_2(x_0) + f_3(x_0) + \ldots = S(x_0) \tag{2}$$

be convergent. Then the series

$$f_1(x) + f_2(x) + f_3(x) + \ldots \tag{3}$$

is said to be (pointwise) *convergent in* I and to *have the sum* $S(x)$.

REMARK 1. For a fixed $x_0 \in I$, (2) is a series of numbers (since $f_1(x_0)$, $f_2(x_0), \ldots$ are numbers). As usual, its *partial sum* $s_n(x_0)$ is defined by

$$s_n(x_0) = f_1(x_0) + f_2(x_0) + \ldots + f_n(x_0).$$

Thus, for any $x_0 \in I$ the relation

$$S(x_0) = \lim_{n\to\infty} s_n(x_0) \tag{4}$$

holds.

The convergence of the series (3) may also be defined as follows: We construct the sequence of partial sums

$$s_1(x) = f_1(x),\ s_2(x) = f_1(x) + f_2(x),\ \ldots,\ s_n(x) = f_1(x) + f_2(x) + \ldots + f_n(x),\ \ldots.$$

If the sequence of functions $s_n(x)$ is convergent in I (see § 15.1) and has the limit $S(x)$, then the series (3) is said to be *convergent in I* and to *have the sum* $S(x)$. Obviously, the two definitions are equivalent.

Example 1. Let I be the open interval $(-1, 1)$. For every $x \in I$ we have

$$S(x) = 1 + x + x^2 + x^3 + \ldots = \frac{1}{1-x}. \tag{5}$$

In fact, it is

$$s_n(x) = 1 + x + x^2 + \ldots + x^{n-1} = \frac{1-x^n}{1-x}$$

in I and

$$S(x) = \lim_{n\to\infty} s_n(x) = \frac{1}{1-x}.$$

Theorem 1. *If the series*

$$|f_1(x)| + |f_2(x)| + |f_3(x)| + \ldots \tag{6}$$

is convergent in I, then the series (3) *is also convergent in I.* (If (6) converges, then the series (3) is said to be *absolutely convergent in I.*)

Definition 2. The series (3) is called *uniformly convergent in I*, if the sequence of partial sums $s_n(x)$,

$$s_n(x) = f_1(x) + f_2(x) + \ldots + f_n(x), \tag{7}$$

is uniformly convergent in I (cf. Definition 15.1.2).

Theorem 2 (*The Bolzano–Cauchy Condition of Convergence*). *The series* (3) *converges uniformly in I, if and only if corresponding to every* $\varepsilon > 0$ *a positive integer* n_0 (*independent of* x) *can be found such that everywhere in I*

$$|f_n(x) + f_{n+1}(x) + \ldots + f_{n+p}(x)| < \varepsilon$$

holds for every $n > n_0$ *and every positive integer* p.

Theorem 3 (*Weierstrass's M-Test*). *If*

$$|f_1(x)| \leqq A_1,\ |f_2(x)| \leqq A_2,\ |f_3(x)| \leqq A_3,\ \ldots$$

holds for every $x \in I$ and if the series of numbers

$$A_1 + A_2 = A_3 + \dots \tag{8}$$

is convergent, then the series

$$f_1(x) + f_2(x) + f_3(x) + \dots$$

is uniformly convergent in I.

The series (8) is called a *majorant of the series* (3).

Example 2. The geometrical series (5) converges uniformly, e.g. in the interval $[-0{\cdot}9;\ 0{\cdot}9]$. As the majorant series (8), we may obviously take the series

$$1 + 0{\cdot}9 + 0{\cdot}9^2 + \dots .$$

On the other hand, it can be shown that the series (5) does *not* converge uniformly in the whole interval $(-1, 1)$.

Example 3. The series

$$\zeta(x) = \frac{1}{1^x} + \frac{1}{2^x} + \dots + \frac{1}{n^x} + \dots$$

(the so-called Riemann *zeta-function*, frequently used in the theory of numbers) converges uniformly in the interval $[a, \infty)$, where a is any number greater than 1. In fact, for every $x \geqq a$ we then have $1/n^x \leq 1/n^a$, so that the series

$$\frac{1}{1^a} + \frac{1}{2^a} + \dots + \frac{1}{n^a} + \dots$$

(convergent according to Example 10.2.7, p. 348) is a majorant of the given series. (The given series, however, does not converge uniformly in the whole interval $(1, \infty)$.)

Theorem 4. *If the series* (3) *converges uniformly in I and the functions $f_1(x)$, $f_2(x)$, ... are continuous in I, then $S(x)$ is also a continuous function in I.*

Theorem 5. *Let the functions (1) be continuous in I and let $f_n(x) \geqq 0$ for every n and $x \in I$. Then $S(x)$ is continuous in I if and only if the series* (3) *converges uniformly in I. (Thus, if $f_n(x) \geqq 0$, then not only the continuity of $S(x)$ follows from the uniform convergence of the series* (3)*, but also the convers is true.)*

Theorem 6 (*Theorem on Integration of Series with Variable Terms*). *Let*

$$f_1(x) + f_2(x) + f_3(x) + \dots = S(x) \tag{9}$$

be a series of integrable functions in $[a, b]$, which converges uniformly in $[a, b]$; then the series

$$F_1(x) + F_2(x) + F_3(x) + \dots \tag{10}$$

with

$$F_n(x) = \int_a^x f_n(t)\,\mathrm{d}t \quad (x \in [a, b])$$

also converges uniformly in $[a, b]$ *and has the sum*

$$\int_a^x S(t)\,\mathrm{d}t\,,$$

i.e.

$$\sum_{n=1}^{\infty} \int_a^x f_n(t)\,\mathrm{d}t = \int_a^x \left[\sum_{n=1}^{\infty} f_n(t)\right] \mathrm{d}t\,. \tag{11}$$

Theorem 7 (*Theorem on Differentiation of Series with Variable Terms*). *Let the series*

$$f_1'(x) + f_2'(x) + f_3'(x) + \ldots = f(x) \tag{12}$$

converge uniformly in $[a, b]$ *and let the series*

$$f_1(x) + f_2(x) + f_3(x) + \ldots \tag{13}$$

converge for at least one $x_0 \in [a, b]$; *then* (13) *converges uniformly in* $[a, b]$. *Denoting the sum of the series* (13) *by* $S(x)$, *we have*

$$S'(x) = f(x)\,,$$

i.e.

$$\frac{\mathrm{d}}{\mathrm{d}x} \sum_{n=1}^{\infty} f_n(x) = \sum_{n=1}^{\infty} \frac{\mathrm{d}}{\mathrm{d}x} f_n(x) \tag{14}$$

(cf. also Theorem 8).

REMARK 2. Observe that the uniform convergence of the *derived* series (12), not that of the series (13), is required.

REMARK 3. In Theorems 6 and 7, the condition of uniform convergence, which is rather strong, may often be replaced by a weaker condition, as for example:

Theorem 8. *Let the functions* (1) *together with the sum* $S(x)$ *of the series* (3) *be integrable functions in* $[a, b]$. *Moreover, let the sequence of partial sums* $s_n(x)$,

$$s_n(x) = f_1(x) + f_2(x) + \ldots + f_n(x)\,,$$

be uniformly bounded in $[a, b]$, *i.e. let*

$$|s_n(x)| < M$$

for every n *and every* $x \in [a, b]$ (M *being a constant*). *Then* (11) *holds, i.e. the given series can be integrated term by term.*

REMARK 4. Sometimes it is convenient to define the sum of a series in another way than as a limit of partial sums. Such series are then called *summable*, in a certain sense (see Remark 10.2.16, p. 354, where summability in the sense of Cesàro was considered). Some series may be summable while they are divergent in the usual sense. This concept of summability is of considerable importance, particularly in the theory of Fourier series.

Example 4. The series

$$a_1 + a_2 + a_3 + \dots \tag{15}$$

is called *summable in the sense* of Euler (or of Abel), if a finite limit

$$\lim_{x \to 1-} \sum_{n=1}^{\infty} a_n x^n \tag{16}$$

exists. This limit is called the *sum of the series* (15) *in the sense of Euler.*

The series

$$1 - 1 + 1 - 1 + 1 - \dots \tag{17}$$

has the sum $1/2$ in the sense of Euler, because

$$x - x^2 + x^3 - x^4 + x^5 - \dots = \frac{x}{1 + x} \tag{18}$$

and

$$\lim_{x \to 1-} \frac{x}{1 + x} = \frac{1}{2}.$$

In the usual sense, however, the series (17) is divergent. According to Abel's theorem (Theorem 15.3.4), any series which is convergent in the usual sense is summable in the sense of Euler (and has the same sum).

15.3. Power Series

REMARK 1. A *power series* is a particular case of a series with variable terms; it is a series of the form

$$a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \dots, \tag{1}$$

where the a_n are constants (generally complex).

For $x_0 = 0$ the power series assumes the form

$$a_0 + a_1 x + a_2 x^2 + \dots. \tag{2}$$

Theorem 1. *For any series* (1) *a number* $r \geqq 0$ (*the so-called radius of convergence; the possibility* $r = +\infty$ *is not excluded*) *exists such that* (1) *converges absolutely for each* $x \in (x_0 - r, x_0 + r)$, *and diverges for all* x *lying outside the interval* $[x_0 - r, x_0 + r]$. (*By the notation* $r = +\infty$ *we mean that the series converges for all finite* x.) *Moreover, if* q *is an arbitrary number such that* $0 < q < r$, *then* (1) *converges uniformly in the interval* $[x_0 - q, x_0 + q]$.

REMARK 2. In geometrical terms, (1) converges in an interval which is symmetric about the point x_0. The series (1) is therefore sometimes called a *series with the centre at the point* x_0. For the same reason, (2) is called a *series with the centre at the origin.*

REMARK 3. At the end points of the interval of convergence, i.e. at the points $x_0 - r$, $x_0 + r$, the series may be either convergent or divergent. For example, the series

$$1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \ldots \quad (r = 1)$$

is divergent at $x = 1$, but convergent at $x = -1$ (Example 10.2.4, p. 345).

REMARK 4. The radius of convergence may also be zero. For example, it can be shown that the series

$$1 + 1!\,x + 2!\,x^2 + 3!\,x^3 + \ldots$$

converges only for $x = 0$ (see also Theorem 2).

Theorem 2. *Let either the limit*

$$\lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} = l \quad \textit{or the limit} \quad \lim_{n\to\infty} \sqrt[n]{|a|} = l \tag{3}$$

exist. Then

$$r = \frac{1}{l}. \tag{4}$$

(If $l = 0$, then $r = +\infty$; if $l = +\infty$, then $r = 0$.)

Example 1. For the series

$$1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$$

we have

$$\lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} = \lim_{n\to\infty} \frac{\frac{1}{(n+1)!}}{\frac{1}{n!}} = \lim_{n\to\infty} \frac{1}{n+1} = 0.$$

Hence, $r = +\infty$. The series in question converges for all finite x.

REMARK 5. If the limits (3) do not exist, Theorem 2 remains true provided the limits (3) are replaced by the limits (see Theorem 10.1.8, p. 340)

$$\overline{\lim_{n\to\infty}}\left|\frac{a_{n+1}}{a_n}\right|, \quad \overline{\lim_{n\to\infty}}\sqrt[n]{|a_n|}.$$

Theorem 3. *The series*

$$a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \ldots, \tag{5}$$

$$|a_0| + |a_1|\,|x - x_0| + |a_2|\,|x - x_0|^2 + \ldots \tag{6}$$

have the same radius of convergence. (In more detail: A single r exists such that both the series (5), (6) are convergent for $|x - x_0| < r$ and divergent for $|x - x_0| > r$.)

REMARK 6. Theorem 3 is very significant from the practical point of view, since the series (6) has positive (or non-negative) terms and its convergence may be tested using the criteria of § 10.2, which are valid for series with positive (or non-negative) terms. One of consequences of Theorem 3 is also the preceding Theorem 2.

Theorem 4 (*Abel's Theorem*). *If the series* (1) *converges for* $x = x_0 + r$ (*i.e. at the right-hand end point of the convergence interval, assuming that* $0 < r < +\infty$), *then its sum* $S(x)$ *possesses a limit from the left at the point* $x_0 + r$, *and this limit coincides with the sum of the series* (1) *for* $x = x_0 + r$. *A similar assertion is true for* $x = x_0 - r$. (Cf. Examples 15.4.1, 15.4.2.)

Theorem 5 (*Arithmetic Operations with Power Series*). *Let the series*

$$S_1(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \ldots, \tag{7}$$

$$S_2(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)^2 + \ldots \tag{8}$$

have radii of convergence r_1 *and* r_2, *respectively. Let* $r = \min(r_1, r_2)$. *Then in the interval* $(x_0 - r, x_0 + r)$ *we have*

$$S_1(x) \pm S_2(x) =$$

$$= (a_0 \pm b_0) + (a_1 \pm b_1)(x - x_0) + (a_2 \pm b_2)(x - x_0)^2 + \ldots, \tag{9}$$

$$S_1(x)\,S_2(x) =$$

$$= a_0b_0 + (a_0b_1 + a_1b_0)(x - x_0) + (a_0b_2 + a_1b_1 + a_2b_0)(x - x_0)^2 + \ldots, \tag{10}$$

i.e. series (7) *and* (8) *may be added, subtracted and multiplied within their common region of convergence* (cf. also Theorems 10.2.3 and 10.2.22). See also formula 27, §15.6. [The series (10) is known as the *Cauchy product* of the series (7), (8).]

Theorem 6 (*Inversion of a Series*). *Let the series*

$$S(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \ldots \tag{11}$$

have a non-zero radius of convergence and let $a_0 \neq 0$. *Then in a certain neighbourhood of the point* x_0 *the function* $1/S(x)$ *can be expanded in a power series*

$$\frac{1}{S(x)} = c_0 + c_1(x - x_0) + c_2(x - x_0)^2 + \dots . \tag{12}$$

REMARK 7. The coefficients c_n of series (12) may be found, for example, by the method of undetermined coefficients. Multiplying (12) by (11), we get

$$1 = [a_0 + a_1(x - x_0) + \dots][c_0 + c_1(x - x_0) + \dots] =$$
$$= a_0c_0 + (a_0c_1 + a_1c_0)(x - x_0) + \dots , \tag{13}$$

whence we have

$$1 = a_0c_0 , \quad 0 = a_0c_1 + a_1c_0 , \text{ etc.} \tag{14}$$

Since $a_0 \neq 0$, we can successively determine $c_0, c_1, \dots$ from these equations. Cf. also formula 28, § 15.6.

Example 2. We have (see formula 13, § 15.6)

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \quad (r = +\infty) . \tag{15}$$

Thus, in a neighbourhood of the point $x_0 = 0$,

$$\frac{1}{\cos x} = c_0 + c_1x + c_2x^2 + c_3x^3 + \dots . \tag{16}$$

By (13) we have

$$1 = \left(1 - \frac{x^2}{2} + \frac{x^4}{24} - \dots\right)(c_0 + c_1x + c_2x^2 + c_3x^3 + \dots) =$$
$$= c_0 + c_1x + \left(c_2 - \frac{c_0}{2}\right)x^2 + \left(c_3 - \frac{c_1}{2}\right)x^3 + \dots ,$$

so that

$$c_0 = 1, c_1 = 0, c_2 = \tfrac{1}{2}, c_3 = 0, \dots .$$

(Cf. also formula 28, § 15.6).

Theorem 7. *Let* (7) *and* (8) *be power series with* $r_1 > 0$, $r_2 > 0$, $b_0 \neq 0$. *Then in a certain neighbourhood of the point* x_0 *the function* $S_1(x)/S_2(x)$ *can be expanded in a power series*

$$\frac{S_1(x)}{S_2(x)} = d_0 + d_1(x - x_0) + d_2(x - x_0)^2 + \dots . \tag{17}$$

REMARK 8. The series (17) may be found, for example, in such a way that the power series for $1/S_2(x)$ is established first by Theorem 6 and Remark 7, and this series is then multiplied by the series for $S_1(x)$. Alternatively, we multiply equation (17) by the series $S_2(x)$ and then use the method of undetermined coefficients in the same way as in Remark 7.

Theorem 8 (*Substituting a Power Series into a Power Series*). *Let the series*

$$S(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \dots, \tag{18}$$

and

$$x - x_0 = g(t) = b_1(t - t_0) + b_2(t - t_0)^2 + \dots \tag{19}$$

have non-zero radii of convergence. Substituting formally the power series (19) *for* $x - x_0$ *into* (18) *and arranging the result by powers of* $t - t_0$, *we get a power series in* $t - t_0$. *This series converges in a certain neighbourhood of the point* t_0 *and, in this neighbourhood, its sum is equal to the function* $S(x_0 + g(t))$.

15.4. Theorems on Differentiation and Integration of Power Series. Power Series in Two or More Variables

Theorem 1. *Let the series*

$$a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \dots = S(x) \tag{1}$$

have radius of convergence r. Then the series

$$a_1 + 2a_2(x - x_0) + 3a_3(x - x_0)^2 + \dots, \tag{2}$$

obtained from (1) *by term-by-term differentiation, also has the same convergence radius r, and its sum in* $(x_0 - r, x_0 + r)$ *is equal to* $S'(x)$ (*i.e. to the derivative of the function* $S(x)$).

REMARK 1. Since the series (2) is again a power series with radius of convergence r, the series obtained from (2) by term-by-term differentiation again has radius of convergence r and defines the function $S''(x)$ in $(x_0 - r, x_0 + r)$, etc. *Hence, the function* $S(x)$ *defined by the series* (1) *with radius of convergence r possesses derivatives of all orders in the interval* $(x_0 - r, x_0 + r)$.

A function which may be expanded in a power series (1) in an interval $(x_0 - r, x_0 + r)$, is said to be *analytic* in this interval. Thus, every such function possesses derivatives of all orders in $(x_0 - r, x_0 + r)$.

Theorem 2. *Let the series*

$$a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \dots = S(x) \tag{3}$$

have radius of convergence r. Then the series

$$a_0(x - x_0) + \frac{a_1}{2}(x - x_0)^2 + \frac{a_2}{3}(x - x_0)^3 + \dots, \tag{4}$$

obtained from (3) *by term-by-term integration, also has the same radius of convergence r and defines, in* $(x_0 - r, x_0 + r)$, *the function*

$$\int_{x_0}^{x} S(t)\,\mathrm{d}t, \tag{5}$$

which is a primitive of $S(x)$.

REMARK 2. In (5) we choose x_0 as the lower limit of integration, since the series (4) has the sum zero for $x = x_0$, and consequently the function defined by this series must vanish for $x = x_0$. If we choose the primitive arbitrarily (i.e. not necessarily equal to zero at the point $x = x_0$), we have to add a constant of integration to the series (4), in general.

Example 1. Let us express the function

$$S(x) = \arctan x$$

by a power series in a neighbourhood of the origin (i.e. by a power series with the centre at the origin).

For the derivative of this function we have

$$\frac{1}{1 + x^2} = 1 - x^2 + x^4 - x^6 + \dots \quad (r = 1) \tag{6}$$

since the right-hand side is a geometric series with common ratio $-x^2$. By Theorem 2 (note that $\arctan 0 = 0$)

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots \quad (r = 1). \tag{7}$$

The series (7) converges even for $x = \pm 1$ (it is an alternating series, Theorem 10.2.16, p. 350) so that by Abel's theorem (Theorem 15.3.4) the equality (7) is valid also for those values of x. In particular, putting $x = 1$ we get the relationship

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

Example 2. Let us find the sum of the series

$$1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} + \dots. \tag{8}$$

We have

$$1 - x + x^2 - x^3 + \ldots = \frac{1}{1 + x} \quad (r = 1).$$

According to Theorem 2,

$$x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots = \ln(1 + x) \quad (r = 1). \tag{9}$$

The series (9) converges for $x = 1$ (it is an alternating series) so that by Abel's theorem (Theorem 15.3.4) we have

$$1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} + \ldots = \ln 2.$$

These methods are often used for finding sums of infinite series.

Theorem 3 (*Power Series in Two Variables*). *Let the double series*

$$S(x, y) = \sum_{m,n=0}^{\infty} a_{mn} x^m y^n \tag{10}$$

be absolutely convergent at a point (x_0, y_0) (cf. Remark 10.2.11, p. 351). *Then* (10) *converges absolutely at any point* (x, y), *whose coordinates satisfy the inequalities*

$$|x| \leqq |x_0|, \quad |y| \leqq |y_0|$$

(*i.e. in a rectangle* $\bar{O}$: $-|x_0| \leqq x \leqq |x_0|$, $-|y_0| \leqq y \leqq |y_0|$). *Moreover,* $S(x, y)$ *is continuous in* $\bar{O}$ *and possesses partial derivatives of all orders inside the rectangle in question. These derivatives may be obtained by term-by-term differentiation of the series* (10). For example,

$$\frac{\partial S}{\partial x} = \sum_{m,n=0}^{\infty} m a_{mn} x^{m-1} y^n = \sum_{m \geqq 1, n \geqq 0} m a_{mn} x^{m-1} y^n. \tag{11}$$

REMARK 3. An analogous theorem is true for power series

$$\sum_{m,n=0} a_{mn} (x - a)^m (y - b)^n, \tag{12}$$

which may, of course, be reduced to the form (10) by the substitution $x - a = \bar{x}$, $y - b = \bar{y}$.

REMARK 4. An analogous theorem is valid for power series in several variables.

REMARK 5. In contrast to power series in a single complex variable, the regions of *absolute* convergence of double series may be of different shapes.

REMARK 6. For double power series (or for power series in several variables) a theorem on addition, subtraction and multiplication of series, similar to Theorem 15.3.5, is true in the region of absolute convergence.

REMARK 7. In Theorem 3, the *absolute* convergence of the series (10) in the sense of Remark 10.2.11, p. 351, is under consideration. If the series (10) converges only in the usual sense (see (14)), then the assertion of Theorem 3 need not hold in general.

We draw the reader's attention to the fact that in technical literature, unless otherwise stated, by the convergence of the series

$$\sum_{m,n=0}^{\infty} a_{mn} x^m y^n \tag{13}$$

ordinary convergence is always understood in the sense of Remark 10.2.12, p. 352 i.e.: the series (13) converges at a point (x_0, y_0) and has the sum $S(x_0, y_0)$, if to each $\varepsilon > 0$ a positive integer P exists such that for any pair of integers M, N with $M > P$, $N > P$, we have

$$\left| \sum_{m=0}^{M} \sum_{n=0}^{N} a_{mn} x_0^m y_0^n - S(x_0, y_0) \right| < \varepsilon . \tag{14}$$

We interpret similarly the convergence of other double series than power series, e.g. of a Fourier series

$$\sum_{m,n=1}^{\infty} a_{mn} \sin mx \sin ny . \tag{15}$$

(For details as to (15) see Theorem 16.3.5.) The same is true for series in several variables.

15.5. Taylor's Series. The Binomial Series

REMARK 1. According to Taylor's theorem we have

$$f(x) = f(a) + \frac{f'(a)}{1!}(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \ldots + \frac{f^{(n)}(a)}{n!}(x - a)^n + R_{n+1}(x) ;$$

the most frequently used forms of the remainder R_{n+1} are given in Theorem 11.10.1, p. 396.

Theorem 1. *If $f(x)$ has derivatives of all orders in the interval $[a, x]$ (or $[x, a]$ provided $x < a$), then a necessary and sufficient condition for the series (the so-called Taylor series)*

$$f(a) + \frac{f'(a)}{1!}(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \ldots \tag{1}$$

to converge at given point x and to have the sum $f(x)$ is

$$\lim_{n \to \infty} R_{n+1}(x) = 0 . \tag{2}$$

REMARK 2. For $a = 0$, the series (1) is called the *Maclaurin series:*

$$f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \dots .$$

REMARK 3. It can happen that the series (1) converges at the point x, but that

$$\lim_{n\to\infty} R_{n+1}(x) \neq 0 ;$$

then, of course, its sum is not equal to $f(x)$.

Example 1. By Taylor's formula, we have

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \dots + \frac{x^n}{n!} + R_{n+1}(x) ,$$

where

$$R_{n+1}(x) = \frac{e^{\vartheta x}}{(n+1)!}x^{n+1} \quad (0 < \vartheta < 1) .$$

For any fixed x we have $|e^{\vartheta x}| < M$, where $M = \max(1, e^x)$ and

$$\lim_{n\to\infty} \frac{x^{n+1}}{(n+1)!} = 0 .$$

Consequently, by Theorem 1 we have, for every x,

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \dots \quad (r = +\infty) .$$

Theorem 2. *If $f(x)$ has derivatives of all orders in the interval $[a, c]$ (or $[c, a]$ when $c < a$) and if these derivatives are uniformly bounded (i.e. a constant M exists such that*

$$|f'(x)| \leqq M , \quad |f''(x)| \leqq M , \quad |f'''(x)| \leqq M , \quad \dots \text{in} \quad [a, c] \quad \text{or} \quad [c, a]) ,$$

then

$$f(x) = f(a) + \frac{f'(a)}{1!}(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \dots$$

in $[a, c]$ (or $[c, a]$).

Theorem 3 (*The Binomial Series*). *For any real n we have*

$$(1 + x)^n = 1 + \binom{n}{1}x + \binom{n}{2}x^2 + \binom{n}{3}x^3 + \dots \quad (r = 1) , \tag{3}$$

where

$$\binom{n}{k} = \frac{n(n-1)\dots(n-k+1)}{k!}.$$

For example

$$\frac{1}{\sqrt{(1-x^2)}} = (1-x^2)^{-1/2} = 1 + \frac{x^2}{2} + \frac{3x^4}{8} + \dots \quad (r=1), \tag{4}$$

since

$$\binom{-\frac{1}{2}}{1} = -\frac{1}{2}, \quad \binom{-\frac{1}{2}}{2} = \frac{(-\frac{1}{2})(-\frac{3}{2})}{2} = \frac{3}{8}, \text{ etc.}$$

Integrating the series (4) and making use of Theorem 15.4.2, we can get the power series for arcsin x (see formula 20, § 15.6).

REMARK 4. Formula (3) may also be used in cases where the first term of the binomial on the left-hand side is not necessarily equal to 1. For example, if $a > |b| > 0$, then

$$(a+b)^n = a^n\left(1 + \frac{b}{a}\right)^n = a^n(1+x)^n \quad (x = b/a).$$

15.6. Some Important Series, Particularly Power Series

(See also § 13.12).

1. $\dfrac{1}{1 \pm x} = 1 \mp x + x^2 \mp x^3 + x^4 \mp x^5 + \dots \qquad (-1 < x < 1).$

2. $\dfrac{1}{(1 \pm x)^2} = 1 \mp 2x + 3x^2 \mp 4x^3 + 5x^4 \mp 6x^5 + \dots \qquad (-1 < x < 1).$

3. $\sqrt{(1+x)} = 1 + \dfrac{1}{2}x - \dfrac{1}{2.4}x^2 + \dfrac{1.3}{2.4.6}x^3 - \dfrac{1.3.5}{2.4.6.8}x^4 + \dots$
$(-1 \leqq x \leqq 1).$

4. $\dfrac{1}{\sqrt{(1+x)}} = 1 - \dfrac{1}{2}x + \dfrac{1.3}{2.4}x^2 - \dfrac{1.3.5}{2.4.6}x^3 + \dfrac{1.3.5.7}{2.4.6.8}x^4 - \dots$
$(-1 < x \leqq 1).$

5. $e^x = 1 + \dfrac{x}{1!} + \dfrac{x^2}{2!} + \dfrac{x^3}{3!} + \dots \qquad (-\infty < x < +\infty).$

6. $e^{-x} = 1 - \frac{x}{1!} + \frac{x^2}{2!} - \frac{x^3}{3!} + \ldots$ $(-\infty < x < +\infty)$.

7. $a^x = 1 + \frac{\ln a}{1!}x + \frac{(\ln a)^2}{2!}x^2 + \frac{(\ln a)^3}{3!}x^3 + \ldots$ $(a > 0, -\infty < x < +\infty)$.

8. $\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots$ $(-1 < x \leqq 1)$.

9. $\ln(1 - x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \ldots$ $(-1 \leqq x < 1)$.

10. $\ln\frac{1 + x}{1 - x} = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \ldots\right)$ $(-1 < x < 1)$.

REMARK 1. This formula can be used for the computation of logarithms if the preceding two formulae fail to be applicable (i.e. if $|x|$ is not sufficiently small, so that convergence is slow). For example, for ln 2 we have

$$\frac{1 + x}{1 - x} = 2, \quad x = \tfrac{1}{3}, \quad \ln 2 = 2\left[\tfrac{1}{3} + \tfrac{1}{3}\cdot\left(\tfrac{1}{3}\right)^3 + \tfrac{1}{5}\cdot\left(\tfrac{1}{3}\right)^5 + \ldots\right].$$

For $|x| > 1$ we then have

11. $\ln\frac{x + 1}{x - 1} = 2\left(\frac{1}{x} + \frac{1}{3x^3} + \frac{1}{5x^5} + \ldots\right)$ $(|x| > 1)$.

REMARK 2. For $\ln x$, however, a power series with the centre at the origin does not exist, since $\ln x$ is not analytic in the neighbourhood of the origin. But we have

12. $\ln x = 2\left[\frac{x - 1}{x + 1} + \frac{1}{3}\left(\frac{x - 1}{x + 1}\right)^3 + \frac{1}{5}\left(\frac{x - 1}{x + 1}\right)^5 + \ldots\right]$ $(x > 0)$.

Furthermore:

13. $\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$, $\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots$
$(-\infty < x < +\infty)$.

14. $\sin^2 x = x^2\left(1 - \frac{1}{3}x^2 + \frac{2}{45}x^4 - \frac{1}{315}x^6 + \frac{2}{14,175}x^8 - \ldots\right)$ $(-\infty < x < +\infty)$.

15. $\cos^2 x = 1 - x^2 + \frac{1}{3}x^4 - \frac{2}{45}x^6 + \frac{1}{315}x^8 - \frac{2}{14,175}x^{10} + \ldots$
$(-\infty < x < +\infty)$.

16. $\sin^3 x = x^3\left(1 - \frac{1}{2}x^2 + \frac{13}{120}x^4 - \frac{41}{3,024}x^6 + \frac{671}{604,800}x^8 - \ldots\right)$
$(-\infty < x < +\infty)$.

17. $\cos^3 x = 1 - \frac{3}{2}x^2 + \frac{7}{8}x^4 - \frac{61}{240}x^6 + \frac{547}{13,440}x^8 - \frac{703}{172,800}x^{10} + \ldots$
$(-\infty < x < +\infty)$.

18. $\tan x = x + \frac{1}{3}x^3 + \frac{2}{15}x^5 + \frac{17}{315}x^7 + \ldots$ $(-\frac{1}{2}\pi < x < \frac{1}{2}\pi)$.

19. $\cot x = \frac{1}{x} - \frac{x}{3} - \frac{x^3}{45} - \frac{2}{945}x^5 - \ldots$ $(0 < |x| < \pi)$.

20. $\arcsin x = x + \frac{1}{2}\cdot\frac{x^3}{3} + \frac{1\cdot 3}{2^2\cdot 2!}\cdot\frac{x^5}{5} + \frac{1\cdot 3\cdot 5}{2^3\cdot 3!}\cdot\frac{x^7}{7} + \ldots$ $(-1 \leqq x \leqq 1)$.

21. $\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \ldots$ $(-1 \leqq x \leqq 1)$.

22. $\sinh x = \frac{x}{1!} + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \ldots$, $\cosh x = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \ldots$
$(-\infty < x < +\infty)$.

23. $\tanh x = x - \frac{1}{3}x^3 + \frac{2}{15}x^5 - \frac{17}{315}x^7 + \ldots$ $(-\frac{1}{2}\pi < x < \frac{1}{2}\pi)$.

24. $\sin x \sin y = xy[1 - \frac{1}{6}(x^2 + y^2) + \frac{1}{360}(3x^4 + 10x^2y^2 + 3y^4) -$
$- \frac{1}{5,040}(x^6 + 7x^4y^2 + 7x^2y^4 + y^6) +$
$+ \frac{1}{1,814,400}(5x^8 + 60x^6y^2 + 126x^4y^4 + 60x^2y^6 + 5y^8) - \ldots]$
$(-\infty < x, y < +\infty)$.

25. $\cos x \cos y = 1 - \frac{1}{2}(x^2 + y^2) + \frac{1}{24}(x^4 + 6x^2y^2 + y^4) -$
$- \frac{1}{720}(x^6 + 15x^4y^2 + 15x^2y^4 + y^6) +$
$+ \frac{1}{40,320}(x^8 + 28x^6y^2 + 70x^4y^4 + 28x^2y^6 + y^8) - \ldots$
$(-\infty < x,y < +\infty)$.

26. $\sin x \cos y = x[1 - \frac{1}{6}(x^2 + 3y^2) + \frac{1}{120}(x^4 + 10x^2y^2 + 5y^4) -$
$- \frac{1}{5,040}(x^6 + 21x^4x^2 + 35x^2y^4 + 7y^6) +$
$+ \frac{1}{362,880}(x^8 + 36x^6y^2 + 126x^4y^4 + 84x^2y^6 + 9y^8) + \ldots]$
$(-\infty < x, y < +\infty)$.

REMARK 3. The series for $\sin^2 x$, $\cos^2 x$, $\sin^3 x$, $\cos^3 x$, $\sin x \sin y$, $\cos x \cos y$, $\sin x \cos y$ given above may be obtained by multiplying the series corresponding to the individual functions (see Theorem 15.3.5).

In general, if

$$S(x) = a + bx + cx^2 + dx^3 + ex^4 + fx^5 + \ldots,$$

then:

27. $S^2(x) = a^2 + 2abx + (b^2 + 2ac)x^2 + 2(ad + bc)x^3 + \\ + (c^2 + 2ae + 2bd)x^4 + 2(af + be + cd)x^5 + \dots .$

28. $\frac{1}{S(x)} = \frac{1}{a}\left[1 - \frac{b}{a}x + \left(\frac{b^2}{a^2} - \frac{c}{a}\right)x^2 + \left(\frac{2bc}{a^2} - \frac{d}{a} - \frac{b^3}{a^3}\right)x^3 + \right.$

$\left. + \left(\frac{2bd}{a^2} + \frac{c^2}{a^2} - \frac{e}{a} - 3\frac{b^2c}{a^3} + \frac{b^4}{a^4}\right)x^4 + \dots\right] \qquad (a \neq 0).$

29. $\frac{1}{S^2(x)} = \frac{1}{a^2}\left[1 - 2\frac{b}{a}x + \left(3\frac{b^2}{a^2} - 2\frac{c}{a}\right)x^2 + \left(6\frac{bc}{a^2} - 2\frac{d}{a} - 4\frac{b^3}{a^3}\right)x^3 + \right.$

$\left. + \left(6\frac{bd}{a^2} + 3\frac{c^2}{a^2} - 2\frac{e}{a} - 12\frac{b^2c}{a^3} + 5\frac{b^4}{a^4}\right)x^4 + \dots\right] \qquad (a \neq 0).$

30. $\sqrt{[S(x)]} = \sqrt{(a)}\left[1 + \frac{1}{2}\frac{b}{a}x + \left(\frac{1}{2}\frac{c}{a} - \frac{1}{8}\frac{b^2}{a^2}\right)x^2 + \left(\frac{1}{2}\frac{d}{a} - \frac{1}{4}\frac{bc}{a^2} + \frac{1}{16}\frac{b^3}{a^3}\right)x^3 + \right.$

$\left. + \left(\frac{1}{2}\frac{e}{a} - \frac{1}{4}\frac{bd}{a^2} - \frac{1}{8}\frac{c^2}{a^2} + \frac{3}{16}\frac{b^2c}{a^3} - \frac{5}{128}\frac{b^4}{a^4}\right)x^4 + \dots\right] \quad (a > 0).$

31. $\frac{1}{\sqrt{[S(x)]}} = \frac{1}{\sqrt{(a)}}\left[1 - \frac{1}{2}\frac{b}{a}x + \left(\frac{3}{8}\frac{b^2}{a^2} - \frac{1}{2}\frac{c}{a}\right)x^2 + \left(\frac{3}{4}\frac{bc}{a^2} - \frac{1}{2}\frac{d}{a} - \frac{5}{16}\frac{b^3}{a^3}\right)x^3 + \right.$

$\left. + \left(\frac{3}{4}\frac{bd}{a^2} + \frac{3}{8}\frac{c^2}{a^2} - \frac{1}{2}\frac{e}{a} - \frac{15}{16}\frac{b^2c}{a^3} + \frac{35}{128}\frac{b^4}{a^4}\right)x^4 + \dots\right] \qquad (a > 0).$

32. $\frac{x}{e^x - 1} = D_0 + D_1\frac{x}{1!} + D_2\frac{x^2}{2!} + D_4\frac{x^4}{4!} + D_6\frac{x^6}{6!} + \dots \qquad (|x| < 2\pi).$

33. $x\cot x = D_0 - D_2\frac{(2x)^2}{2!} + D_4\frac{(2x)^4}{4!} - D_6\frac{(2x)^6}{6!} + \dots \qquad (|x| < \pi).$

For $x = 0$, the sums of the series in 32, 33 reduce to

$$D_0 = \lim_{x\to 0}\frac{x}{e^x - 1} = \lim_{x\to 0} x\cot x = 1.$$

The numbers D_n are defined by the recurrence formula

$$\binom{n+1}{1}D_n + \binom{n+1}{2}D_{n-1} + \dots + \binom{n+1}{n}D_1 + D_0 = 0, \quad D_0 = 1.$$

From this relation we have

$$D_1 = -\tfrac{1}{2},\ D_2 = \tfrac{1}{6},\ D_4 = -\tfrac{1}{30},\ D_6 = \tfrac{1}{42},\ D_8 = -\tfrac{1}{30},\ D_{10} = \tfrac{5}{66},\ \dots,$$

$$D_3 = D_5 = D_7 = \dots = 0.$$

15.7. Application of Series, Particularly of Power Series, to the Evaluation of Integrals. Asymptotic Expansions

On application to the solution of differential equations, see Chaps. 16, 17, 25.

REMARK 1. A further useful application of series is to the approximate evaluation of integrals. The integrated function (whose indefinite integral cannot be expressed, for example, by elementary functions) is expanded in a series and integrated term by term according to Theorem 15.4.2.

Example 1. Let us consider

$$\int_0^x e^{-t^2}\,dt\,. \tag{1}$$

By formula 6, § 15.6, we have

$$e^{-t^2} = 1 - \frac{t^2}{1!} + \frac{t^4}{2!} - \frac{t^6}{3!} + \ldots \quad (r = +\infty)\,. \tag{2}$$

According to Theorem 15.4.2, the series (2) can be integrated term by term over any interval (since $r = +\infty$). Thus,

$$\int_0^x e^{-t^2}\,dt = x - \frac{x^3}{3\,.\,1!} + \frac{x^5}{5\,.\,2!} - \frac{x^7}{7\,.\,3!} + \ldots \quad (r = +\infty)\,. \tag{3}$$

See also Remark 3 below.

Example 2. Let us consider

$$\int_0^{\pi/2} \frac{\sin x}{x}\,dx\,. \tag{4}$$

Referring to formula 13, § 15.6, we have

$$\frac{\sin x}{x} = 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \ldots \tag{5}$$

for every $x \neq 0$. If we define the function $\frac{\sin x}{x}$ as 1 at the point $x = 0$ (i.e. by its limit as $x \to 0$), then (5) will be true for all x. Since $r = +\infty$, the series (5) can by integrated term by term over any interval (Theorem 15.4.2). Hence,

$$\int_0^a \frac{\sin x}{x}\,dx = a - \frac{a^3}{3\,.\,3!} + \frac{a^5}{5\,.\,5!} - \ldots\,. \tag{6}$$

For $a = \pi/2$ it is sufficient to take the first five terms of (6) into account in order to ensure an accuracy of 5 decimal places (note that (6) is an alternating series, see Theorem 10.2.16, p. 350).

REMARK 2. If one of the limits of integration coincides with an end point of the interval of convergence, Abel's theorem can be used (Theorem 15.3.4; cf. also Example 15.4.2). See also §13.12 (application of series to the evaluation of elliptic integrals).

REMARK 3 (Application of Divergent Series; Asymptotic Expansions). The series (3) is suitable for evaluating the integral

$$\int_0^x e^{-t^2}\,dt \tag{7}$$

provided $|x|$ is sufficiently small. If $|x|$ is large, the series (3) is also convergent, but it is evidently not suitable for evaluating the integral (7). In such cases *asymptotic* expansions can advantageously be used, as we proceed to show using the integral (7) as an example. Let us assume $x > 0$. We have

$$\int_0^\infty e^{-t^2}\,dt = \frac{\sqrt{\pi}}{2}, \quad \int_0^x e^{-t^2}\,dt = \int_0^\infty e^{-t^2}\,dt - \int_x^\infty e^{-t^2}\,dt. \tag{8}$$

Integrating by parts ($e^{-t^2} = 2te^{-t^2}\cdot(\frac{1}{2}/t)$, so $u' = 2te^{-t^2}$, $v = \frac{1}{2}/t$) we get

$$\int_x^\infty e^{-t^2}\,dt = \frac{e^{-x^2}}{2x} - \frac{1}{2}\int_x^\infty \frac{e^{-t^2}}{t^2}dt. \tag{9}$$

A repeated integration by parts yields, after n steps,

$$\int_x^\infty e^{-t^2}\,dt = \frac{e^{-x^2}}{2x}\left[1 - \frac{1}{2x^2} + \frac{1\cdot 3}{(2x^2)^2} - \ldots + (-1)^{n-1}\frac{1\cdot 3\cdot 5\cdot\ldots\cdot(2n-3)}{(2x^2)^{n-1}}\right] + r_n, \tag{10}$$

where

$$r_n = (-1)^n \frac{1\cdot 3\cdot 5\cdot\ldots\cdot(2n-1)}{2^n}\int_x^\infty \frac{e^{-t^2}}{t^{2n}}\,dt. \tag{11}$$

The series in brackets in (10) diverges for every x. But obviously

$$\int_x^\infty \frac{e^{-t^2}}{t^{2n}}\,dt < \frac{1}{x^{2n}}\int_x^\infty e^{-t^2}\,dt$$

and

$$\int_x^\infty e^{-t^2}\,dt < \frac{e^{-x^2}}{2x}$$

by (9), so that we have

$$|r_n| < \frac{1\cdot 3\cdot 5\cdot\ldots\cdot(2n-1)}{2^{n+1}x^{2n+1}}e^{-x^2}. \tag{12}$$

If x is sufficiently large, then the remainder (12) can be made very small by a proper choice of n. (Divergent series with this property are sometimes called *semiconvergent*, but the more usual term is *asymptotic*.) For example, if $x = 5$, then for $n = 13$ we have $|r_n| < 10^{-20}$. Thus, taking $n = 13$ in (10), the integral

$$\int_5^\infty e^{-t^2}\,dt$$

can be evaluated with an accuracy of 20 decimal places. The former integral $\int_0^5 e^{-t^2}\,dt$ can then be evaluated by (8),

$$\int_0^5 e^{-t^2}\,dt = \frac{\sqrt{\pi}}{2} - \int_5^\infty e^{-t^2}\,dt\,.$$

REMARK 4. The logarithmic integral

$$\operatorname{li}(x) = \int_0^x \frac{dt}{\ln t} \quad (0 < x < 1)\,. \tag{13}$$

may be evaluated in a similar way. By the substitution $t = e^{-u}$, the integral (13) reduces to the integral

$$\operatorname{li}(x) = -\int_a^\infty \frac{e^{-u}}{u}\,du\,,$$

where

$$-\ln x = a \quad (a > 0)\,. \tag{14}$$

Integrating successively by parts, we get

$$\operatorname{li}(x) = -x\left[\frac{1}{a} - \frac{1!}{a^2} + \frac{2!}{a^3} - \ldots + (-1)^{n-1}\frac{(n-1)!}{a^n}\right] + r_n\,, \tag{15}$$

where

$$r_n = (-1)^{n+1}\,n!\int_a^\infty \frac{e^{-t}}{t^{n+1}}\,dt\,, \tag{16}$$

so that

$$|r_n| < n!\,\frac{e^{-a}}{a^{n+1}}\,. \tag{17}$$

Thus, if x is sufficiently small and, consequently, a sufficiently large, $|r_n|$ may be made small by a suitable choice of n, and the series (15) can be employed for calculating $\operatorname{li}(x)$.

It is important to note that as $n \to \infty$, $|r_n| \to \infty$; generally, the value of $|r_n|$ starts

by decreasing (as n increases) until it reaches a minimum value; thereafter, it increases and, indeed, becomes infinite. The best approximation is obtained, of course, by choosing n so that $|r_n|$ has its minimum value.

REMARK 5. A thorough treatment of asymptotic expansions may be found, e.g., in [75].

15.8. Survey of Some Important Formulae from Chapter 15

(See also §§ 15.6 and 15.7.)

1. $\lim\limits_{n\to\infty}\lim\limits_{x\to a+} f_n(x) = \lim\limits_{x\to a+}\lim\limits_{n\to\infty} f_n(x)$

under the assumptions of Theorem 15.1.4,

2. $\int_a^x \lim\limits_{n\to\infty} f_n(t)\,\mathrm{d}t = \lim\limits_{n\to\infty}\int_a^x f_n(t)\,\mathrm{d}t$ (Theorem 15.1.5),

3. $\frac{\mathrm{d}}{\mathrm{d}x}\left[\lim\limits_{n\to\infty} f_n(x)\right] = \lim\limits_{n\to\infty}\frac{\mathrm{d}f_n(x)}{\mathrm{d}x}$ (Theorem 15.1.6),

4. $\sum\limits_{n=1}^{\infty}\int_a^x f_n(t)\,\mathrm{d}t = \int_a^x \sum\limits_{n=1}^{\infty} f_n(t)\,\mathrm{d}t$ (Theorem 15.2.6),

5. $\frac{\mathrm{d}}{\mathrm{d}x}\sum\limits_{n=1}^{\infty} f_n(x) = \sum\limits_{n=1}^{\infty}\frac{\mathrm{d}f_n(x)}{\mathrm{d}x}$ (Theorem 15.2.7).

6. The radius of convergence of a power series is given by

$$r = \frac{1}{l}, \quad \text{where} \quad l = \lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| \quad \text{or} \quad l = \lim_{n\to\infty}\sqrt[n]{|a_n|}$$

(Theorem 15.3.2, Remark 15.3.5).

7. If

$$S(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \ldots \quad (|x - x_0| < r),$$

then

$$\int_{x_0}^x S(t)\,\mathrm{d}t = a_0(x - x_0) + \frac{a_1}{2}(x - x_0)^2 + \frac{a_2}{3}(x - x_0)^3 + \ldots \quad (|x - x_0| < r),$$

$$S'(x) = a_1 + 2a_2(x - x_0) + 3a_3(x - x_0)^2 + \ldots \quad (|x - x_0| < r)$$

(Theorems 15.4.1 and 15.4.2).

16. THE SPACE L_2. ORTHOGONAL SYSTEMS, FOURIER SERIES. SPECIAL FUNCTIONS (BESSEL FUNCTIONS, ETC.)

By KAREL REKTORYS

References: [3], [5], [7], [8], [26], [28], [32], [34], [38], [44], [49], [85], [86], [91], [92], [108], [112], [123], [127], [136], [148], [152], [168], [180], [183], [184], [189].

16.1. The Space L_2

In §13.14 the definition of functions square integrable on a set M has been given. In the present paragraph, this definition is specified for the case $M = [a, b]$, where $[a, b]$ is a bounded interval, and the so-called space $L_2(a, b)$ is introduced. At the end of the paragraph, the possibility is discussed how to define, in a similar way, the space $L_2(\Omega)$, where Ω is a bounded region in an N-dimensional Euclidean space.

For the properties of the space $L_2(a, b)$ (or $L_2(\Omega)$), introduced below, it is essential that the functions, we work with, are assumed to be integrable in the Lebesgue sense. In §13.14, a brief survey of the Lebesgue theory has been presented. *As far as the space L_2 is discussed in this chapter, integrability in the Lebesgue sense is always understood.* However, from the point of view of applications, no serious mistake can arise if the reader who is not familiar with the Lebesgue theory considers the integrals, he will meet, in the Riemann sense.

Definition 1. A real (Lebesgue) measureable function f is called *square integrable on* (or *in*) *a* (bounded) *interval* $[a, b]$, if the integral

$$\int_a^b f^2(x)\,\mathrm{d}x \tag{1}$$

is convergent (= finite). We write

$$f \in L_2(a, b)\ . \tag{2}$$

Every continuous or piecewise continuous function in $[a, b]$ is square integrable on this interval. As concerns unbounded functions, we have for example, for the functions $f(x) = x^{-1/3}$, $g(x) = x^{-1/2}$,

$$f \in L_2(0, 1)\ ,\quad g \notin L_2(0, 1)\ ,$$

because (Example 13.14.2, p. 565)

$$\int_0^1 f^2(x)\,\mathrm{d}x = \int_0^1 x^{-2/3}\,\mathrm{d}x = 3\ ,$$

$$\int_0^1 g^2(x)\,\mathrm{d}x = \int_0^1 x^{-1}\,\mathrm{d}x = +\infty\ .$$

Further (see Remark 13.14.9, p. 566)

(i) convergence of the integral (1) implies convergence of the integral

$$\int_a^b f(x)\,\mathrm{d}x\ ,$$

(ii)

$$f, g \in L_2(a,b) \Longrightarrow c_1 f + c_2 g \in L_2(a,b) \tag{3}$$

for arbitrary real numbers c_1, c_2.

(iii)

$$f, g \in L_2(a,b) \Longrightarrow \int_a^b f(x)\ g(x)\,\mathrm{d}x \text{ is convergent.} \tag{4}$$

From (3) it follows that the set of all square integrable functions on $[a,b]$ is a linear set, because with every couple of its elements it contains an arbitrary linear combination of them. The property (4) then admits to introduce the following definition:

Definition 2. Under a *scalar product of functions* $f, g \in L_2(a,b)$ we understand the number

$$(f,g) = \int_a^b f(x)\ g(x)\,\mathrm{d}x\ ; \tag{5}$$

the (nonnegative) number

$$\|f\| = \sqrt{(f,f)} \tag{6}$$

is called the *norm of the function* $f \in L_2(a,b)$, the (nonnegative) number

$$\rho(f,g) = \|f - g\| \tag{7}$$

is the so-called *distance of the functions* f and g.

The norm, or the distance can thus be obtained as (nonnegative) roots of

$$\|f\|^2 = \int_a^b f^2(x)\,\mathrm{d}x\ , \tag{8}$$

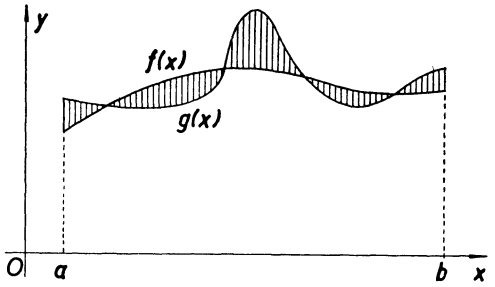


Fig. 16.1.

or

$$\rho^2(f,g) = \int_a^b (f-g)^2 \, \mathrm{d}x \ , \tag{9}$$

respectively.

Definition 3. The set of all square integrable functions in $[a,b]$ on which there are defined the scalar product (5), the norm (6) and the distance (we also say the *metric*) (7), is called the *metric space* $L_2(a,b)$, briefly the *space* $L_2(a,b)$.

(See Remark 2 and Theorem 1.)

REMARK 1. If the distance of two functions f,g is "small" in the space $L_2(a,b)$ (i.e. if (9) is "small"), then it does not follow that the difference of their values is "small" everywhere in $[a,b]$; it follows only that these functions do not differ "very much" in an integral sense as shown in Fig. 16.1. For this reason, one speaks also about the "*mean quadratic deviation*" instead of "distance" of the functions f and g.

REMARK 2 (*Equivalent Functions*). If two functions f and g are continuous in $[a,b]$ and if $\rho(f,g) = 0$ then, by (9),

$$f(x) = g(x) \qquad \text{for all } \ x \in [a,b] \ . \tag{10}$$

However, if they are *not* continuous in $[a,b]$, it need not be the case. The integral (9) is as well equal to zero if the functions f and g differ, for example, at a finite number of points or, more generally, on a set of measure zero. This fact is the motivation of the following definition:

Definition 4. Two functions $f, g \in L_2(a,b)$ are called *equivalent in the space* $L_2(a,b)$, if

$$\rho(f,g) = 0 \ . \tag{11}$$

We write

$$f = g \quad \text{in } L_2(a,b) \ . \tag{12}$$

Two equivalent functions are taken as equal in the space $L_2(a,b)$, they represent the *same element* of this space. While the meaning of (10) is that the functions f and g are equal in the whole interval $[a,b]$, the meaning of (12) is that f and g are equivalent functions in $L_2(a,b)$, which thus may be different on a set of measure zero. For example, if we write

$$f = 0 \quad \text{in } L_2(a,b) \tag{13}$$

it means that, in $[a,b]$, the function f is either equal to the function which is identically zero in this interval, or that it differs from that function on a set of points of measure zero, e.g. at a finite number of points (at which, may be, it need not be defined at all).

Theorem 1. *Scalar product, norm and distance have the following properties* (f, g, h *being arbitrary functions from* $L_2(a,b)$, c, c_1, c_2 *arbitrary real numbers):*

$$(f,g) = (g,f) \ , \tag{14}$$

$$(c_1 f_1 + c_2 f_2, g) = c_1(f_1,g) + c_2(f_2,g) \ , \tag{15}$$

$$(f,f) \geqq 0 \ , \tag{16}$$

$$(f,f) = 0 \Longleftrightarrow f = 0 \ \textit{in} \ L_2(a,b) \ \textit{(see (13))} \ , \tag{17}$$

$$\|f\| \geqq 0 \ , \tag{18}$$

$$\|f\| = 0 \Longleftrightarrow f = 0 \ \textit{in} \ L_2(a,b) \tag{19}$$

$$\|cf\| = |c| \, \|f\| \ , \tag{20}$$

$$\|f+g\| \leqq \|f\| + \|g\| \ \textit{(the triangle inequality)} \ , \tag{21}$$

$$|(f,g)| \leqq \|f\| \, \|g\| \ \textit{(the Schwarz inequality)} \ , \tag{22}$$

$$\rho(f,g) \geqq 0 \ , \tag{23}$$

$$\rho(f,g) = 0 \Longleftrightarrow f = g \ \textit{in} \ L_2(a,b) \ \textit{(see (12))} \ , \tag{24}$$

$$\rho(f,g) = \rho(g,f) \ , \tag{25}$$

$$\rho(f,h) \leqq \rho(f,g) + \rho(g,h) \quad (\textit{the triangle inequality}) \ . \tag{26}$$

REMARK 3. The just introduced properties of functions from $L_2(a,b)$ are very similar to those of geometric vectors, when replacing the length $|\mathbf{u}|$ of the vector $\mathbf{u}$ by the norm $\|f\|$ of the function f. In particular, the Schwarz inequality (22) corresponds to the well-known inequality $|(\mathbf{u},\mathbf{v})| \leqq |\mathbf{u}| \cdot |\mathbf{v}|$ for vectors.

Definition 5 (*Convergence in the Mean*). We say that a sequence $\{f_n\}$ of functions from $L_2(a,b)$ *converges in the space* $L_2(a,b)$ (or *in the mean*) to a function $f \in L_2(a,b)$ if

$$\lim_{n\to\infty} \rho(f,f_n) = 0 \ . \tag{27}$$

We write

$$\lim_{n\to\infty} f_n = f \ \text{ in } \ L_2(a,b) \ , \ \text{ or } \ f_n \to f \ \text{ in } \ L_2(a,b) \ . \tag{28}$$

Thus, if we have to establish that a sequence of functions $f_n \in L_2(a,b)$ converges in the space $L_2(a,b)$, i.e. in the mean, to a function $f \in L_2(a,b)$, we have to show that (27) is fulfilled, i.e. that

$$\lim_{n\to\infty} \rho^2(f,f_n) = 0 \ ,$$

or (see (9)) that

$$\lim_{n\to\infty} \int_a^b (f-f_n)^2 \, \mathrm{d}x = 0 \ . \tag{29}$$

REMARK 4. (i) It can be shown that the sequence $\{f_n\}$ may converge, in the mean, to a single function only (or to an equivalent function; *uniqueness of the limit*);

(ii) convergence in the mean does not imply ordinary (pointwise) convergence (Definition 15.1.1, p. 637) in the interval $[a,b]$, and vice versa:

Example 1. Let us show that the sequence of functions

$$f_n(x) = x^n \ , \quad n = 1,2,\dots \ , \tag{30}$$

i.e. of functions

$$f_1(x) = x \ , \ \ f_2(x) = x^2, \dots \ , \tag{31}$$

converges in the space $L_2(0,1)$ to the zero function.

According to (29) we have to show that

$$\lim_{n\to\infty} \int_0^1 (f-f_n)^2 \, \mathrm{d}x = 0 \ , \tag{32}$$

where f_n are the functions (30) and f is the function identically equal to zero, or an equivalent function. But

$$\lim_{n\to\infty} \int_0^1 (f - f_n)^2 \, dx = \lim_{n\to\infty} \int_0^1 (0 - x^n)^2 \, dx =$$

$$= \lim_{n\to\infty} \int_0^1 x^{2n} \, dx = \lim_{n\to\infty} \left[\frac{x^{2n+1}}{2n+1}\right]_0^1 = \lim_{n\to\infty} \frac{1}{2n+1} = 0 \, ,$$

by which (32) is established. (As concerns the ordinary (pointwise) convergence, the sequence (30) converges, in the interval $[0, 1]$, to the function

$$f(x) = \begin{cases} 0 \text{ if } x \in [0,1) \, , \\ \\ 1 \text{ if } x = 1 \, .) \end{cases}$$

Example 2. Let

$$f(x) = \begin{cases} n \text{ if } 0 < x < \dfrac{1}{n} \, , \\ \\ 0 \text{ otherwise in } [0,1] \, , \end{cases}$$

$n = 1, 2, \ldots$. For every fixed $x \in [0, 1]$ we have obviously

$$\lim_{n\to\infty} f_n(x) = 0 \, ,$$

so that the sequence of functions f_n converges pointwise to the function $f(x) \equiv 0$ in $[0, 1]$. However, it does *not* converge to it in the mean, i.e. we have not

$$\lim_{n\to\infty} \int_0^1 (f - f_n)^2 \, dx = 0 \, ,$$

but

$$\lim_{n\to\infty} \int_0^1 (f - f_n)^2 \, dx = \lim_{n\to\infty} \int_0^1 f_n^2 \, dx =$$

$$= \lim_{n\to\infty} \int_0^{1/n} n^2 \, dx = \lim_{n\to\infty} n = +\infty \, .$$

Example 3. The trigonometric Fourier series corresponding to a function $f \in L_2(-\pi, \pi)$ converges to this function in the mean (Remark 16.2.13). As well known, it need not converge pointwise to that function everywhere in $[-\pi, \pi]$.

REMARK 5. On base of the concept of convergence (in the mean) of a sequence, convergence of an infinite series is defined: We say that a *series*

$$\sum_{n=1}^{\infty} f_n \, , \quad f_n \in L_2(a, b) \, , \tag{33}$$

is convergent in the space $L_2(a,b)$ (or *in the mean*) and *has the sum* $s \in L_2(a,b)$, we write

$$\sum_{n=1}^{\infty} f_n = s \quad \text{in} \quad L_2(a,b) \ , \tag{34}$$

if the sequence of its partial sums,

$$s_k = \sum_{n=1}^{k} f_n \ , \tag{35}$$

converges in the space $L_2(a,b)$ (in the mean) to the function s, i.e. (see (29)) if

$$\lim_{k\to\infty} \int_a^b (s - \sum_{n=1}^{k} f_n)^2 \, \mathrm{d}x = 0 \ . \tag{36}$$

REMARK 6 (*The Complex Space* $L_2(a,b)$). In applications (namely in electrotechnics), we often meet complex functions of a real variable, thus functions of the form

$$f(x) = f_1(x) + \mathrm{i} f_2(x) \ , \tag{37}$$

where f_1, f_2 are real functions. We then say that a complex function f is *square integrable* in the interval $[a,b]$, if the functions f_1 and f_2 are measureable in that interval and if the integral

$$\int_a^b |f(x)|^2 \, \mathrm{d}x \tag{38}$$

is convergent (= finite) (what takes place exactly if both integrals

$$\int_a^b f_1^2 \, \mathrm{d}x \ , \quad \int_a^b f_2^2 \, \mathrm{d}x$$

are convergent). The *scalar product* on the set of these functions is defined by

$$(f,g) = \int_a^b f(x) \, \overline{g(x)} \, \mathrm{d}x \ , \tag{39}$$

where $\overline{g(x)}$ is the complex conjugate of $g(x)$. For the scalar product we then have

$$(f,g) = \overline{(g,f)} \ ; \tag{40}$$

the relation

$$(cf,g) = c(f,g)$$

remains true, but by (39)

$$(f,cg) = \bar{c}(f,g)$$

(if c is a complex number). All definitions and results presented in this paragraph remain unchanged except that in (8), or (9), or (29), or (36) we have to write

$$|f|^2 \ , \quad \text{or} \quad |f-g|^2 \ , \quad \text{or} \quad |f-f_n|^2 \ , \quad \text{or} \quad |s-\sum_{n=1}^{k} f_n|^2$$

instead of

$$f^2 \ , \quad \text{or} \quad (f-g)^2 \ , \quad \text{or} \quad (f-f_n)^2 \ , \quad \text{or} \quad (s-\sum_{n=1}^{k} f_n)^2 \ ,$$

respectively.

REMARK 7 (*The Space* $L_2(\Omega)$). Let Ω be a bounded region in E_N. Similarly as in §13.14 we define: A real function $f(x_1,\dots,x_N)$ is called *square integrable on* (or *in*) Ω, we write

$$f \in L_2(\Omega) \ ,$$

if it is (Lebesgue) measureable in Ω (Remark 13.14.10, p. 566) and if the integral

$$\int \dots \int_{\Omega} f^2(x_1,\dots,x_N)\,\mathrm{d}x_1 \dots \mathrm{d}x_N$$

is convergent (= of finite value). On the set of these functions then scalar product, norm and distance are introduced quite similarly as in Definition 2, and so the space $L_2(\Omega)$ is obtained. In this space convergence is defined in a quite similar way as in the space $L_2(a,b)$.

Similarly as in Remark 6 then *complex space* $L_2(\Omega)$ can be defined with the scalar product

$$(f,g) = \int \dots \int_{\Omega} f(x_1,\dots,x_N)\,\overline{g(x_1,\dots,x_N)}\,\mathrm{d}x_1 \dots \mathrm{d}x_N \ .$$

16.2. Orthogonal Systems, Fourier Series

Definition 1. We say that two functions $f,g \in L_2(a,b)$ are *orthogonal in the space* $L_2(a,b)$ (or *in*, or *on the interval* $[a,b]$), if

$$(f,g) = 0 \ , \quad \text{i.e. if} \quad \int_a^b f(x)\,g(x)\,\mathrm{d}x = 0 \ . \tag{1}$$

Definition 2. A function f is called *normed* (or *normalized*) *in* $L_2(a,b)$ (or *in*, or *on the interval* $[a,b]$), if its norm is equal to unity, i.e. if

$$\|f\| = 1 \ . \tag{2}$$

Definition 3. Let a finite or countable (Definition 22.1.11) system of functions $f_n \in L_2(a,b)$ be given. This system is called *orthogonal on* $[a,b]$ (or *in the space* $L_2(a,b)$), if for every two mutually different functions f_i, f_k we have

$$(f_i, f_k) = 0 \quad (i \neq k) \ . \tag{3}$$

If, moreover, every function f_n is normed, the system is called *orthonormal*.

REMARK 1. Thus, in this case, the functions from this system fulfill

$$(f_i, f_k) = \delta_{ik} \ , \tag{4}$$

where δ_{ik} is the so-called *Kronecker delta symbol*, i.e.

$$\delta_{ik} = \begin{cases} 0 \ \text{ if } \ i \neq k \ , \\ \\ 1 \ \text{ if } \ i = k \ . \end{cases} \tag{5}$$

REMARK 2. From an orthogonal system $\{f_n\}$ an orthonormal system $\{\varphi_n\}$ is obtained when putting, for every n,

$$\varphi_n(x) = \frac{f_n(x)}{\|f_n\|}$$

(provided $\|f_n\| \neq 0$, of course).

Example 1. Direct integration shows (see also §13.10) that the system of functions

$$1, \ \cos x, \ \sin x, \ \cos 2x, \ \sin 2x, \ldots \tag{6}$$

is orthogonal in the interval $[-\pi, \pi]$ (even in every interval $[a, a + 2\pi]$). However, this system is not orthonormal. Because

$$\int_{-\pi}^{\pi} 1^2 \, \mathrm{d}x = 2\pi, \ \int_{-\pi}^{\pi} \cos^2 kx \, \mathrm{d}x = \pi, \ \int_{-\pi}^{\pi} \sin^2 kx \, \mathrm{d}x = \pi$$

for every k ($k = 1, 2, \ldots$), the corresponding orthonormal system is, by Remark 2,

$$\frac{1}{\sqrt{(2\pi)}}, \ \frac{\cos x}{\sqrt{\pi}}, \ \frac{\sin x}{\sqrt{\pi}}, \ \frac{\cos 2x}{\sqrt{\pi}}, \ \frac{\sin 2x}{\sqrt{\pi}}, \ldots \ . \tag{7}$$

In $[-\pi, \pi]$ also the systems

$$\sin x,\ \sin 2x,\ \sin 3x, \dots, \tag{8}$$

$$1,\ \cos x,\ \cos 2x,\ \cos 3x, \dots \tag{9}$$

are orthogonal, as follows from the fact that each of them is a subsystem of the system (6).

The system (6) is not orthogonal in the interval $[0, \pi]$, because, e.g.,

$$\int_0^\pi 1 \times \sin x \, \mathrm{d}x = 2 \neq 0 \ .$$

On the other hand, the systems (8) and (9) are orthogonal in $[0, \pi]$.

REMARK 3. Definitions 1–3 can be easily generalized for a more-dimensional case (cf. Remark 16.1.7):

Definition 4. Two functions $f, g \in L_2(\Omega)$ are called *orthogonal in that space*, or *on (in) the region* Ω, if

$$(f, g) = \int \dots \int_\Omega f(x_1, \dots, x_N)\, g(x_1, \dots, x_N)\, \mathrm{d}x_1 \dots \mathrm{d}x_N = 0 \ . \tag{10}$$

If

$$\|f\| = \sqrt{(f, f)} = 1 \ , \tag{11}$$

the function f is called *normed* (or *normalized*) in $L_2(\Omega)$. A system of functions $f_n \in L_2(\Omega)$ is called *orthogonal on the region* Ω (or *in the space* $L_2(\Omega)$), if

$$(f_i, f_k) = 0 \ \text{ for } \ i \neq k \ . \tag{12}$$

If, moreover, all the functions of this system are normed, the system is called *orthonormal.*

An example of an orthogonal system in $L_2(\Omega)$, where Ω is the square $(0, \pi) \times (0, \pi)$, is the system of functions

$$\sin mx\ \sin ny \ , \quad m = 1, 2, 3, \dots, \quad n = 1, 2, 3, \dots \ , \tag{13}$$

thus the system of functions

$$f_1 = \sin x\ \sin y \ , \quad f_2 = \sin 2x\ \sin y \ , \quad f_3 = \sin x\ \sin 2y \ ,$$

$$f_4 = \sin 3x\ \sin y \ , \ f_5 = \sin 2x\ \sin 2y \ , \ f_6 = \sin x\ \sin 3y \ , \dots \ .$$

The corresponding orthonormal system is

$$\frac{2}{\pi} \sin mx \sin ny , \quad m = 1,2,3,\ldots, \quad n = 1,2,3,\ldots . \tag{14}$$

In a similar way further definitions of this paragraph can be extended, almost literally, to the more-dimensional case.

REMARK 4. Similarly, these definitions can be extended to the comlex space L_2 (see Remark 16.1.6). For example, two functions f, g from a complex space $L_2(a,b)$ are called *orthogonal in that space* (or *in*, or *on the interval* $[a,b]$) if

$$(f,g) = \int_a^b f(x)\,\overline{g(x)}\,\mathrm{d}x = 0 . \tag{15}$$

Example 2. The system of functions

$$f_n(x) = \mathrm{e}^{\mathrm{i}nx} = \cos nx + \mathrm{i} \sin nx \quad (n \text{ an integer})$$

is orthogonal in the interval $[-\pi,\pi]$: In fact,

$$(f_m,f_n) = \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}mx}\,\overline{\mathrm{e}^{\mathrm{i}nx}}\,\mathrm{d}x = \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}mx}\,\mathrm{e}^{-\mathrm{i}nx}\,\mathrm{d}x =$$

$$= \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}(m-n)x}\,\mathrm{d}x = \left[\frac{1}{\mathrm{i}(m-n)}\,\mathrm{e}^{\mathrm{i}(m-n)x}\right]_{-\pi}^{\pi} = 0 \quad \text{for} \quad m \neq n .$$

To make the system orthonormal, we compute the norm of the functions f_n. We have

$$\|f\|^2 = (f_n,f_n) = \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}nx}\,\overline{\mathrm{e}^{\mathrm{i}nx}}\,\mathrm{d}x =$$

$$= \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}nx}\,\mathrm{e}^{-\mathrm{i}nx}\,\mathrm{d}x = \int_{-\pi}^{\pi} 1 \times \mathrm{d}x = 2\pi .$$

So the corresponding orthonormal system is

$$\varphi_n(x) = \frac{\mathrm{e}^{\mathrm{i}nx}}{\sqrt{(2\pi)}} \quad (n \text{ an integer}) .$$

Definition 5. A system of (real) functions $f_n(x)$ is called *orthogonal with a weight function* $\rho(x)$ ($\rho(x) \geqq 0$, $\rho \not\equiv 0$ in $L_2(a,b)$) *in the interval* $[a,b]$, if for every pair of functions $f_i(x)$, $f_k(x)$ we have

$$\int_a^b \rho(x)\, f_i(x)\, f_k(x)\,\mathrm{d}x = 0 , \quad \text{whenever} \quad i \neq k . \tag{16}$$

If, moreover, every function f_n is normed with the weight function $\rho(x)$, i.e. if

$$\int_a^b \rho(x)\, f_n^2(x)\,\mathrm{d}x = 1\ , \quad n = 1, 2, \dots\ ,$$

we say that this system is *orthonormal with the weight function* $\rho(x)$ *in* $[a, b]$. Hence

$$\int_a^b \rho(x)\, f_i(x)\, f_k(x)\,\mathrm{d}x = \delta_{ik}$$

(see Remark 1, p. 670; for examples of such systems see §§16.5, 16.6; see also Theorem 16.4.7, p. 697).

REMARK 5. From a system of functions $f_n(x)$ which are orthogonal in $[a, b]$ with a weight function $\rho(x)$ an orthonormal system of functions $\varphi_n(x)$ (with this weight function) is obtained when putting

$$\varphi(x) = \frac{f_n(x)}{\sqrt{\int_a^b \rho(x)\, f_n^2(x)\,\mathrm{d}x}}\ .$$

REMARK 6. A rich source of orthogonal systems (or orthogonal systems with weight functions) are eigenvalue problems in differential equations. See e.g., §17.17.

Definition 6. Let in $[a, b]$, a countable orthonormal system (thus an orthonormal sequence) of functions

$$\varphi_1(x),\ \varphi_2(x),\ \varphi_3(x), \dots,\ \varphi_n \in L_2(a, b)\ , \tag{17}$$

be given. Let $f \in L_2(a, b)$. The series

$$c_1\varphi_1(x) + c_2\varphi_2(x) + c_3\varphi_3(x) + \dots\ , \tag{18}$$

where

$$c_k = (f, \varphi_k) = \int_a^b f(x)\, \varphi_k(x)\,\mathrm{d}x\ , \quad k = 1, 2, 3, \dots\ , \tag{19}$$

is called a *(generalized) Fourier series corresponding to the function* $f(x)$. The numbers c_k are called the *Fourier coefficients of the function* $f(x)$ *with respect to the system* (17).

Example 3. The Fourier coefficients of a function $f \in L_2(-\pi, \pi)$ with respect to the system (7) are

$$c_1 = \int_{-\pi}^{\pi} \frac{f(x)}{\sqrt{(2\pi)}}\,\mathrm{d}x\ , \quad c_2 = \int_{-\pi}^{\pi} \frac{f(x)\,\cos x}{\sqrt{\pi}}\,\mathrm{d}x\ , \quad c_3 = \int_{-\pi}^{\pi} \frac{f(x)\,\sin x}{\sqrt{\pi}}\,\mathrm{d}x, \dots\ .$$

REMARK 7. About pointwise convergence of the series (18) to the function $f(x)$ nothing can be said, in general. On the convergence in the mean see below.

Theorem 1. *Let the orthonormal system* (17), *a function* $f \in L_2(a,b)$ *and a positive integer* n *be given. Then from among all functions of the form*

$$k_1\varphi_1 + k_2\varphi_2 + \ldots + k_n\varphi_n \ , \tag{20}$$

exactly the function

$$c_1\varphi_1 + c_2\varphi_2 + \ldots + c_n\varphi_n \ , \tag{21}$$

with c_k *given by* (19), *has, in* $L_2(a,b)$, *the least distance from the function* f, *i.e. the least quadratic deviation.*

REMARK 8. Thus for every system of numbers $k_1, k_2, \ldots, k_n$ we have

$$\int_a^b \{f(x) - [c_1\varphi_1(x) + c_2\varphi_2(x) + \ldots + c_n\varphi_n(x)]\}^2 \, \mathrm{d}x \leqq$$

$$\leqq \int_a^b \{f(x) - [k_1\varphi_1(x) + k_2\varphi_2(x) + \ldots + k_n\varphi_n(x)]\}^2 \, \mathrm{d}x \ .$$

Theorem 2. *The Fourier coefficients* (19) *of every function* $f \in L_2(a,b)$ *fulfil the so-called Bessel inequality*

$$\sum_{n=1}^{\infty} c_n^2 \leqq \|f\|^2 \ . \tag{22}$$

REMARK 9. Let us remind that all theorems introduced here are valid for orthonormal systems.

REMARK 10. Convergence of the series (22) implies that the Fourier coefficients c_n (with respect to the orthonormal system (17)) of an arbitrary function $f \in L_2(a,b)$ fulfill

$$\lim_{n\to\infty} c_n = 0 \ .$$

Theorem 3. *Let* (17) *be an orthonormal system in* $L_2(a,b)$. *Let* $b_1, b_2, b_3, \ldots$ *be such numbers that*

$$\sum_{n=1}^{\infty} b_n^2 < \infty \ .$$

Then the series

$$\sum_{n=1}^{\infty} b_n \varphi_n(x)$$

is convergent in $L_2(a,b)$ (thus in the mean, see (16.1.34)). *If* $s(x)$ *is its sum, then* b_n *are Fourier coefficients of this function* $s(x)$ *with respect to the system* (17).

From Theorems 2 and 3, it follows:

Theorem 4. *Let* (17) *be an orthonormal system in* $L_2(a,b)$, $f \in L_2(a,b)$. *Then the Fourier series* (18), *corresponding to this function, converges in the mean to a function* $s \in L_2(a,b)$.

However, (18) need not converge (in the mean) to the function f (or to an equivalent function), in general.

Definition 7. The orthonormal system (17) is called *complete in* $L_2(a,b)$, if the Fourier series (18) of every function $f \in L_2(a,b)$ converges in the mean to that function, i.e. if for every $f \in L_2(a,b)$ we have

$$\lim_{k\to\infty} \|f - \sum_{n=1}^{k} c_n\varphi_n\| = 0 \ .$$

REMARK 11. For conditions on completeness see Theorems 5, 6. See also Remark 13.

Theorem 5. *A necessary and sufficient condition for the system* (17) *to be complete in* $L_2(a,b)$ *is that for every function* $f \in L_2(a,b)$ *the Bessel inequality* (22) *becomes equality (the so-called Parseval equality),*

$$\sum_{n=1}^{\infty} c_n^2 = \|f\|^2 \ . \tag{23}$$

Definition 8. The system (17) is called *closed in* $L_2(a,b)$, if no nonzero function (see, of course, Remark 16.1.2) $g \in L_2(a,b)$ exists orthogonal to every function of this system.

Theorem 6. *The system* (17) *is complete in* $L_2(a,b)$ *if and only if it is closed.*

REMARK 12. In particular: If the system (17) is complete, then the relations

$$g \in L_2(a,b),\ (g,\varphi_n) = 0 \ , \quad n = 1,2,3,\ldots$$

imply that $g = 0$ in $L_2(a,b)$ (in the sense of Remark 16.1.2).

Theorem 7 *(on Uniqueness). If two functions $f, g \in L_2(a,b)$ have the same Fourier coefficients with respect to a complete system* (17), *then $f = g$ in $L_2(a,b)$ (in the sense of Remark* 16.1.2*).*

Theorem 8. *If a Fourier series* (18) *corresponding to a function $f \in L_2(a,b)$ converges to that function in the mean and, moreover, if it is uniformly convergent in the interval $[a,b]$, then it converges pointwise in $[a,b]$ to that function (or to an equivalent function).*

In particular:

Theorem 9. *If the system* (17) *is complete and if the Fourier series corresponding to that function is uniformly convergent in $[a,b]$, then it converges uniformly in $[a,b]$ to that function (or to an equivalent function).*

REMARK 13. It is not easy, in general, to prove completeness of a given orthonormal system. Often completeness follows from Theorems of §17.17. Typical examples of complete orthonormal systems in L_2 are trigonometric systems:

$$\text{Interval } [-\pi,\pi]: \quad \frac{1}{\sqrt{(2\pi)}}, \ \frac{\cos x}{\sqrt{\pi}}, \ \frac{\sin x}{\sqrt{\pi}}, \ \frac{\cos 2x}{\sqrt{\pi}}, \ \frac{\sin 2x}{\sqrt{\pi}}, \dots, \tag{24}$$

$$\text{interval } [0,\pi]: \quad \sqrt{\left(\frac{2}{\pi}\right)} \sin x, \ \sqrt{\left(\frac{2}{\pi}\right)} \sin 2x, \ \sqrt{\left(\frac{2}{\pi}\right)} \sin 3x, \dots, \tag{25}$$

$$\text{interval } [0,\pi]: \quad \frac{1}{\sqrt{\pi}}, \ \sqrt{\left(\frac{2}{\pi}\right)} \cos x, \ \sqrt{\left(\frac{2}{\pi}\right)} \cos 2x, \ \sqrt{\left(\frac{2}{\pi}\right)} \cos 3x, \dots. \tag{26}$$

Thus the Fourier series of an arbitrary function $f \in L_2(-\pi,\pi)$, or $g \in L_2(0,\pi)$, converges (in the corresponding interval) in the mean to the function f, or g, respectively. The same remains valid for Fourier series transformed onto the interval $[-l,l]$, or $[0,l]$, or $[a,b]$ (§16.3).

See also Remark 16; a typical example of a complete orthonormal system of functions of two variables in $L_2(\Omega)$ is the system (14) on the square $\Omega = (0,\pi) \times (0,\pi)$.

REMARK 14. Instead of "complete system" the term "complete sequence" is frequently used. This term is also often used in the following, more general sense:

Let $f_n \in L_2(a,b)$ be a sequence of functions which are linearly independent in $[a,b]$. (This means that every finite number of terms of the sequence constitutes a linearly independent system; the functions f_n need not be pairwise orthogonal.) The sequence $f_n(x)$ is called *complete in* $L_2(a,b)$, if any function $f \in L_2(a,b)$ can be approximated in the mean by linear combinations of functions f_n with any accuracy. In other words: Let a function $f \in L_2(a,b)$ and an $\varepsilon > 0$ be given. Then an integer m and constants $k_1, \ldots, k_m$ can be found such that

$$||f - (k_1 f_1 + k_2 f_2 + \ldots + k_m f_m)|| < \varepsilon ,$$

i.e.

$$\int_a^b \{f(x) - [k_1 f_1(x) + k_2 f_2(x) + \ldots + k_m f_m(x)]^2\} \, \mathrm{d}x < \varepsilon^2 .$$

For example, for any interval $[a,b]$, the sequence

$$1, \; x, \; x^2, \; x^3, \ldots \tag{27}$$

is complete in $L_2(a,b)$.

REMARK 15 (*The Schmidt Orthogonalization Process*). Let $f_n \in L_2(a,b)$, $n = 1,2,\ldots$, be a sequence of functions which are linearly independent in $[a,b]$ (see Remark 14) but not necessarily orthogonal. From this sequence a sequence $g_n = L_2(a,b)$ orthogonal in $[a,b]$ can be constructed as follows:

First, put $g_1(x) = f_1(x)$.

The function $g_2(x)$ is first sought in the form

$$g_2(x) = f_2(x) + k_1 g_1(x) .$$

We determine the constant k_1 so that $g_2(x)$ be orthogonal to $g_1(x)$, i.e.

$$(g_2, g_1) = (f_2 + k_1 g_1, g_1) = 0 ,$$

thus

$$(f_2, g_1) + k_1 (g_1, g_1) = 0 . \tag{28}$$

As $g_1(x) = f_1(x)$, $(f_1, f_1) > 0$ (f_n are linearly independent), the constant k_1 is uniquely determined by (28). Obviously, $g_2(x)$ is a non-zero function, since $f_1(x)$ and $f_2(x)$ are linearly independent. The function $g_3(x)$ is now sought in the form

$$g_3(x) = f_3(x) + c_2 g_2(x) + c_1 g_1(x) ,$$

choosing the constants so that $(g_3, g_1) = 0$, $(g_3, g_2) = 0$. Expanding the products we get

$$(f_3, g_1) + c_2 (g_2, g_1) + c_1 (g_1, g_1) = 0 ,$$

$$(f_3, g_2) + c_2 (g_2, g_2) + c_1 (g_1, g_2) = 0 .$$

As $(g_1, g_2) = 0$ and $(g_1, g_1) > 0$, c_1 is uniquely determined by the first equation. For the same reason, c_2 is uniquely determined by the second one. Obviously, $g_3(x)$ is again a non-zero function.

In this manner we obtain a system of non-zero orthogonal functions $g_n \in L_2(a, b)$ from the functions $f_n \in L_2(a, b)$. Finally, by normalizing each of the functions $g_n(x)$ we get an orthonormal system. If the original system $\{f_n\}$ is complete, so is the obtained system $\{g_n\}$.

Example 4. The orthogonalization of the system (27) in $[-1, 1]$ yields the (normalized) Legendre polynomials

$$(\tfrac{1}{2})^{1/2},\ (\tfrac{3}{2})^{1/2}\,x,\ (\tfrac{5}{2})^{1/2} \times \tfrac{1}{2}(3x^2 - 1),\ (\tfrac{7}{2})^{1/2} \times \tfrac{1}{2}(5x^3 - 3x), \ldots$$

(§16.5).

REMARK 16. Everything said above about functions of one variable may be extended to functions of two or more variables. Instead of an interval (a, b) a region Ω in a plane or in space, etc., is then considered. Cf. Definition 4. See also §22.4.

16.3. Trigonometric Fourier Series. Fourier Series in Two and Several Variables. Fourier Integral

REMARK 1. According to Remark 16.2.13, the orthonormal system of trigonometric functions

$$\frac{1}{\sqrt{(2\pi)}},\ \frac{\cos x}{\sqrt{\pi}},\ \frac{\sin x}{\sqrt{\pi}},\ \frac{\cos 2x}{\sqrt{\pi}},\ \frac{\sin 2x}{\sqrt{\pi}}, \ldots$$

is complete in $[-\pi, \pi]$. Thus, taking any function $f \in L_2(-\pi, \pi)$, the corresponding Fourier series converges in the mean to f.

An analogous statement is true for the systems (16.2.25), (16.2.26) in the interval $[0, \pi]$.

As far as the *pointwise* convergence of a Fourier series is concerned, we have:

Theorem 1. *Let $f(x)$ be a periodic function with period 2π (i.e. $f(x + 2\pi) = f(x)$ for every x) and let $f(x)$ and $f'(x)$ be piecewise continuous in $[-\pi, \pi]$. Define constants a_n, b_n by*

$$\begin{aligned} a_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)\, \cos nx\, \mathrm{d}x \quad (n = 0, 1, 2, \ldots)\,, \\ b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)\, \sin nx\, \mathrm{d}x \quad (n = 1, 2, 3, \ldots)\,. \end{aligned} \tag{1}$$

Then at each point x, where $f(x)$ is continuous, we have

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) = f(x) , \tag{2}$$

while at each point of discontinuity,

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) = \tfrac{1}{2}(f(x+0) + f(x-0)) . \tag{3}$$

(The symbols $f(x+0)$, $f(x-0)$ denote the limits from the right and the left, respectively, of the function $f(x)$ at the point x.)

REMARK 2. Equations (2) and (3) hold under far more general assumptions. They are true if: (i) $f(x)$ is measureable in $[-\pi, \pi]$; (ii) the integral

$$\int_{-\pi}^{\pi} |f(x)| \, \mathrm{d}x$$

is convergent ($f(x)$ may be unbounded) and (iii) the point x is interior to an interval, in which $f(x)$ has a bounded variation (Definition 11.3.7). Moreover, in the case of convergence of the above integral, the series (2) converges uniformly in each interval $[a, b]$, which is interior to an interval on which $f(x)$ is continuous and is of bounded variation.

REMARK 3. Brackets in the sum on the left-hand side of equation (2) or (3) cannot in general be omitted.

REMARK 4. For a periodic function with period 2π (Theorem 1) we may take not only the interval $[-\pi, \pi]$ as "basic" interval, but also any interval $[a, a+2\pi]$ (for example, the interval $[0, 2\pi]$). The lower and upper limits in integrals (1), of course, are then a and $a + 2\pi$, respectively.

Example 1. Let $f(x)$ be a periodic function with period 2π and let $f(x) = x$ on the basic interval $[0, 2\pi)$ (thus at points $x = 2k\pi$, where k in an integer, we have $f(x) = 0$, see Fig. 16.2).

Integrating by parts we get

$$a_n = \frac{1}{\pi} \int_0^{2\pi} x \cos nx \, \mathrm{d}x = \frac{1}{\pi} \left[x \frac{\sin nx}{n} \right]_0^{2\pi} - \frac{1}{n\pi} \int_0^{2\pi} \sin nx \, \mathrm{d}x = 0$$

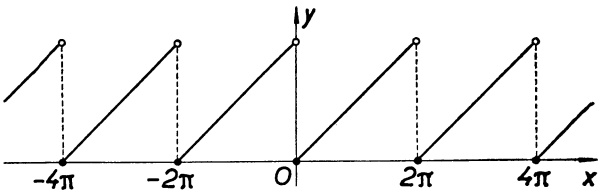


Fig.16.2.

for $n = 1, 2, 3, \ldots$,

$$a_0 = \frac{1}{\pi} \int_0^{2\pi} x \, dx = 2\pi ,$$

$$b_n = \frac{1}{\pi} \int_0^{2\pi} x \sin nx \, dx = -\frac{1}{\pi} \left[x \frac{\cos nx}{n} \right]_0^{2\pi} + \frac{1}{n\pi} \int_0^{2\pi} \cos nx \, dx =$$

$$= -\frac{2}{n} \; (n = 1, 2, 3, \ldots) .$$

Thus by (2) we have

$$f(x) = \pi - 2 \left(\frac{\sin x}{1} + \frac{\sin 2x}{2} + \frac{\sin 3x}{3} + \ldots \right) \tag{4}$$

at any point x different from an integral multiple of 2π. If $x = 2k\pi$ (where k is an integer), $f(x)$ is discontinuous and the sum on the right-hand side of equation (4) equals, in view of (3),

$$\tfrac{1}{2} [f(x + 0) + f(x - 0)] = \tfrac{1}{2} (0 + 2\pi) = \pi .$$

This can, of course, readily be verified from (4). In any closed interval which does not contain the point $x = 2k\pi$, the series (4) converges uniformly.

REMARK 5. For a function $f(x)$ periodic with period $2l$ a theorem analogous to Theorem 1 is true (and remarks similar to Remarks 2, 3 and 4 also hold). Here the formula reads:

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{l} + b_n \sin \frac{n\pi x}{l} \right) = \begin{cases} f(x) \text{ at a point of continuity of } f(x), \\ \tfrac{1}{2} (f(x + 0) + f(x - 0)) \text{ at a point of} \\ \text{discontinuity of } f(x), \end{cases} \tag{5}$$

where

$$a_n = \frac{1}{l}\int_{-l}^{l} f(x)\ \cos\frac{n\pi x}{l}\,dx\ ,\quad b_n = \frac{1}{l}\int_{-l}^{l} f(x)\ \sin\frac{n\pi x}{l}\,dx\ . \tag{6}$$

If f is a periodic function of time t, with period T and basic interval $[0,T]$, then putting $2\pi/T = \omega$ and writing t instead of x, we have

$$\frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n\ \cos n\omega t + b_n\ \sin n\omega t) = \begin{cases} f(t)\ , \\ \\ \frac{1}{2}(f(t+0) + f(t-0))\ , \end{cases} \tag{5'}$$

where

$$a_n = \frac{2}{T}\int_0^T f(t)\ \cos n\omega t\,dt\ ,\quad b_n = \frac{2}{T}\int_0^T f(t)\ \sin n\omega t\,dt\ . \tag{6'}$$

REMARK 6. If $f(x)$ is an odd function of x (i.e. if $f(-x) \equiv -f(x)$), then, by (1), or (6), $a_n = 0$ for $n = 0, 1, \ldots$ and the series involves only sine terms. If $f(x)$ is even in x (i.e. if $f(-x) \equiv f(x)$), then $b_n = 0$ for $n = 1, 2, \ldots$ and the series involves only the constant term $a_0/2$ and cosine terms.

If $f(x)$ is defined in the interval $[0,\pi]$, it can be expressed in this interval by either a sine series or a cosine series. Defining $f(x)$ in $(-\pi, 0)$ by $f(-x) = -f(x)$ (and as a periodic function with period 2π for remaining x), we get $f(x)$ expressed by a sine series. Defining $f(-x) = f(x)$ in $(-\pi, 0)$, we get $f(x)$ expressed by a cosine series. In the first case,

$$\sum_{n=1}^{\infty} b_n\ \sin nx = \begin{cases} f(x) \\ \\ \frac{1}{2}\,(f(x+0) + f(x-0)) \end{cases} \quad (\text{for } x \in [0,\pi]), \tag{7}$$

where

$$b_n = \frac{2}{\pi}\int_0^{\pi} f(x)\ \sin nx\,dx\ . \tag{8}$$

In the second case,

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nx = \begin{cases} f(x)\,, \\ \frac{1}{2}(f(x+0) + f(x-0)) \end{cases} \quad (\text{for } x \in [0,\pi]) \tag{9}$$

with

$$a_n = \frac{2}{\pi}\int_0^{\pi} f(x)\cos nx\,dx\ . \tag{10}$$

If the interval $[0, \pi]$ is replaced by the interval $[0, l]$, formulae (7) and (8) assume the form

$$\sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{l} = \begin{cases} f(x), \\ \frac{1}{2}(f(x+0)+f(x-0)) \end{cases} \quad (\text{for } x \in [0, l]),$$

$$b_n = \frac{2}{l}\int_0^l f(x)\sin\frac{n\pi x}{l}\,dx$$

and formulae (9), (10) the form

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{l} = \begin{cases} f(x), \\ \frac{1}{2}(f(x+0)+f(x-0)) \end{cases} \quad (\text{for } x \in [0, l]),$$

$$a_n = \frac{2}{l}\int_0^l f(x)\cos\frac{n\pi x}{l}\,dx.$$

Example 2. For the function $f(x) = x$ in $[0, \pi)$, $f(\pi) = 0$ we get in the first case

$$f(x) = 2\left(\frac{\sin x}{1} - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \ldots\right) \tag{11}$$

(the continuation of the function over all real x is plotted in Fig. 16.3).

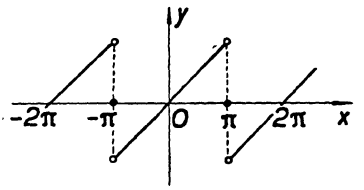


Fig. 16.3.

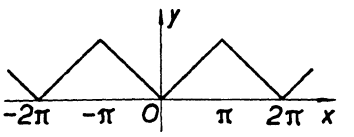


Fig. 16.4.

In the second case (we define $f(\pi) = \pi$; now the continuation of the function for all x is as plotted in Fig. 16.4),

$$f(x) = \frac{\pi}{2} - \frac{4}{\pi}\left(\frac{\cos x}{1^2} + \frac{\cos 3x}{3^2} + \frac{\cos 5x}{5^2} + \ldots\right). \tag{12}$$

REMARK 7 (*Fourier Expansions of Some Important Functions*).

1. $f(x) = |\sin x|$ (Fig. 16.5):

$$f(x) = \frac{2}{\pi} - \frac{4}{\pi}\left(\frac{\cos 2x}{1 \cdot 3} + \frac{\cos 4x}{3 \cdot 5} + \frac{\cos 6x}{5 \cdot 7} + \ldots\right).$$

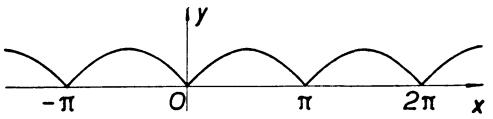


Fig. 16.5.

2. $f(x) = \frac{h}{p}x$ for $0 \leqq x \leqq p$, $f(x) = h$ for $p \leqq x \leqq \pi - p$,

$f(x) = -\frac{h}{p}(x - \pi)$ for $\pi - p \leqq x \leqq \pi$, $f(x + \pi) = -f(x)$ for every x

(Fig. 16.6):

$$f(x) = \frac{4}{\pi}\frac{h}{p}\left(\frac{1}{1^2}\sin p \sin x + \frac{1}{3^2}\sin 3p \sin 3x + \frac{1}{5^2}\sin 5p \sin 5x + \ldots\right).$$

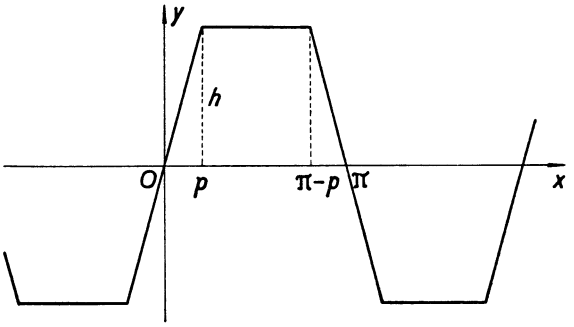


Fig. 16.6.

3. Particularly, for $p = \frac{1}{2}\pi$ (Fig. 16.7):

$$f(x) = \frac{8h}{\pi^2}\left(\frac{\sin x}{1^2} - \frac{\sin 3x}{3^2} + \frac{\sin 5x}{5^2} - \ldots\right).$$

4. $f(x) = h$ for $0 < x < \pi$, $f(0) = 0$, $f(x + \pi) = -f(x)$ for every x (Fig. 16.8):

$$f(x) = \frac{4}{\pi}h\left(\frac{\sin x}{1} + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \ldots\right).$$

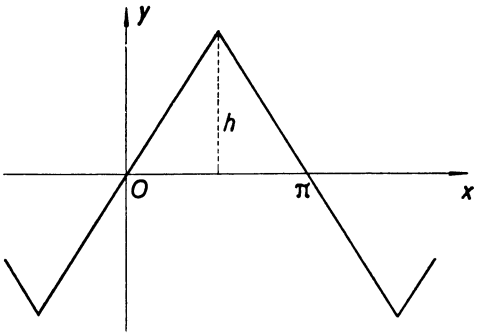


Fig. 16.7.

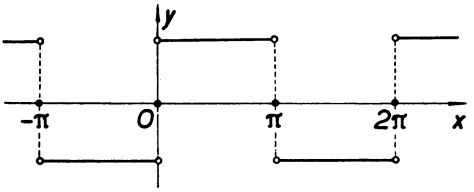


Fig. 16.8.

5. $f(x) = 0$ for $0 \leqq x < p$ and for $\pi - p < x \leqq \pi$,
$f(x) = a$ for $p < x < \pi - p$, $f(x + \pi) = -f(x)$ for every x (Fig. 16.9):

$$f(x) = \frac{4a}{\pi}\left(\cos p \sin x + \tfrac{1}{3}\cos 3p \sin 3x + \tfrac{1}{5}\cos 5p \sin 5x + \ldots\right).$$

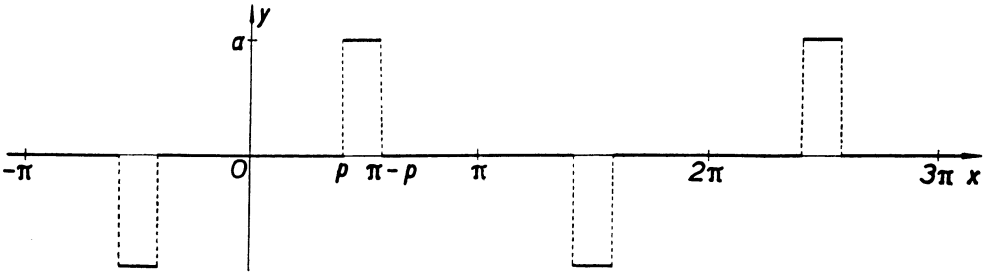


Fig. 16.9.

6. $f(x) = x^2$ for $-\pi \leqq x \leqq \pi$, $f(x + 2\pi) = f(x)$ for every x (Fig. 16.10):

$$f(x) = \frac{\pi^2}{3} - 4\left(\frac{\cos x}{1^2} - \frac{\cos 2x}{2^2} + \frac{\cos 3x}{3^2} - \ldots\right).$$

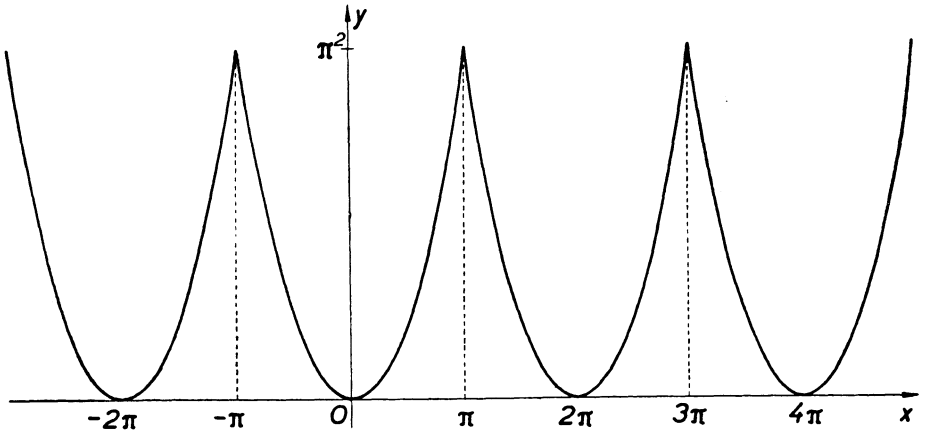


Fig. 16.10.

7. $f(x) = x(\pi - x)$ for $0 \leqq x \leqq \pi$, $f(x + \pi) = f(x)$ for every x (Fig. 16.11):

$$f(x) = \frac{\pi^2}{6} - \left(\frac{\cos 2x}{1^2} + \frac{\cos 4x}{2^2} + \frac{\cos 6x}{3^2} + \ldots\right).$$

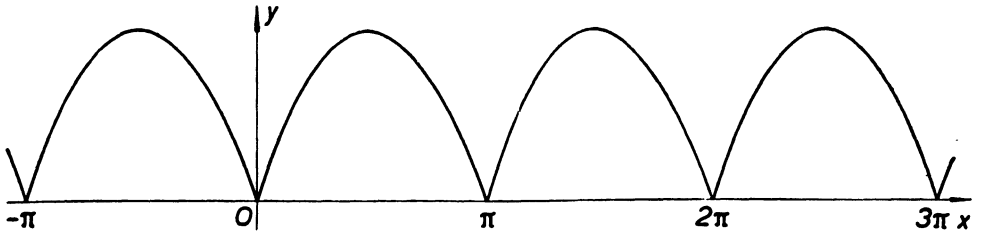


Fig. 16.11.

8. $f(x) = x(\pi - x)$ for $0 \leqq x \leqq \pi$, $f(x + \pi) = -f(x)$ for every x (Fig. 16.12):

$$f(x) = \frac{8}{\pi}\left(\frac{1}{1^3}\sin x + \frac{1}{3^3}\sin 3x + \frac{1}{5^3}\sin 5x + \ldots\right).$$

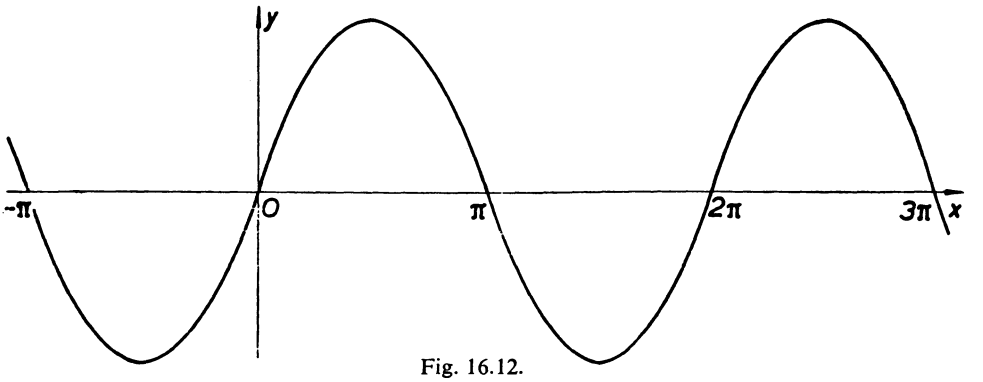


Fig. 16.12.

9. $f(x) = \cos x$ for $0 < x < \pi$, $f(0) = 0$, $f(x + \pi) = f(x)$ for every x (Fig. 16.13):

$$f(x) = \frac{4}{\pi}\left(\frac{2 \sin 2x}{1 \cdot 3} + \frac{4 \sin 4x}{3 \cdot 5} + \frac{6 \sin 6x}{5 \cdot 7} + \ldots\right).$$

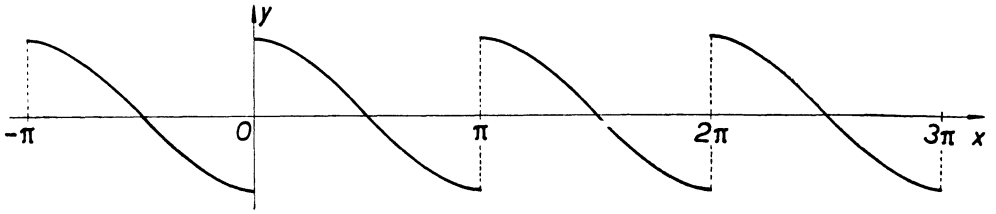


Fig. 16.13.

10. $f(x) = \sin x$ for $0 \leqq x \leqq \pi$, $f(x) = 0$ for $\pi \leqq x \leqq 2\pi$, $f(x + 2\pi) = f(x)$ for every x (Fig. 16.14):

$$f(x) = \frac{1}{\pi} + \frac{1}{2}\sin x - \frac{2}{\pi}\left(\frac{\cos 2x}{1 \cdot 3} + \frac{\cos 4x}{3 \cdot 5} + \frac{\cos 6x}{5 \cdot 7} + \ldots\right).$$

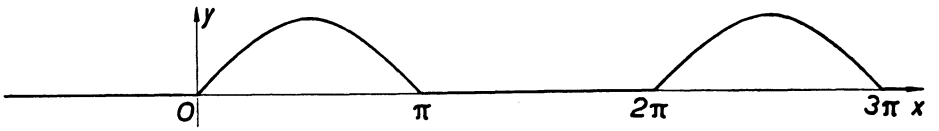


Fig. 16.14.

11. $f(x) = \cos ux$ for $-\pi \leqq x \leqq \pi$, u arbitrary, different from any integer:

$$f(x) = \frac{2u \sin u\pi}{\pi}\left(\frac{1}{2u^2} - \frac{\cos x}{u^2 - 1} + \frac{\cos 2x}{u^2 - 4} - \frac{\cos 3x}{u^2 - 9} + \ldots\right).$$

12. $f(x) = \sin ux$ for $-\pi < x < \pi$, u arbitrary, different from any integer:

$$f(x) = \frac{2 \sin u\pi}{\pi}\left(\frac{\sin x}{1 - u^2} - \frac{2 \sin 2x}{4 - u^2} + \frac{3 \sin 3x}{9 - u^2} - \ldots\right).$$

13. $f(x) = x \cos x$ for $-\pi < x < \pi$:

$$f(x) = -\tfrac{1}{2}\sin x + \frac{4 \sin 2x}{1 \cdot 3} - \frac{6 \sin 3x}{3 \cdot 5} + \frac{8 \sin 4x}{5 \cdot 7} - \ldots.$$

14. $f(x) = -\ln(2\sin\frac{1}{2}x)$ for $0 < x \leqq \pi$:

$$f(x) = \cos x + \tfrac{1}{2}\cos 2x + \tfrac{1}{3}\cos 3x + \ldots .$$

15. $f(x) = \ln(2\cos\frac{1}{2}x)$ for $0 \leqq x < \pi$:

$$f(x) = \cos x - \tfrac{1}{2}\cos 2x + \tfrac{1}{3}\cos 3x - \ldots .$$

16. $f(x) = \frac{1}{2}\ln\cot\frac{1}{2}x$ for $0 < x < \pi$:

$$f(x) = \cos x + \tfrac{1}{3}\cos 3x + \tfrac{1}{5}\cos 5x + \ldots .$$

REMARK 8 (*Fourier Series in Complex Form*). For a periodic function (satisfying the assumptions of Theorem 1 or Remark 2) with basic interval $[0, 2\pi]$ we have

$$\sum_{n=-\infty}^{\infty} c_n e^{inx} = \begin{cases} f(x), \\ \frac{1}{2}(f(x+0) + f(x-0)), \end{cases} \quad c_n = \frac{1}{2\pi}\int_0^{2\pi} f(x)\, e^{-inx}\, dx .$$

By a summation from $-\infty$ to ∞ we mean the summation

$$c_0 + \sum_{n=1}^{\infty}(c_n e^{inx} + c_{-n}e^{-inx}).$$

For a periodic function with basic interval $[0, l]$ we have

$$\sum_{n=-\infty}^{\infty} c_n e^{2\pi inx/l} = \begin{cases} f(x), \\ \frac{1}{2}(f(x+0) + f(x-0)), \end{cases} \quad c_n = \frac{1}{l}\int_0^{l} f(x)\, e^{-2\pi inx/l}\, dx .$$

Theorem 2 (*Differentiation and Integration of Fourier Series*). *Let the Fourier series corresponding to a function $f(x)$ be given*:

$$\frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos nx + b_n \sin nx)$$

$$\left(a_n = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos nx\, dx, \quad b_n = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin nx\, dx\right).$$

If $f(x)$ is continuous in $[-\pi, \pi]$, $f(-\pi) = f(\pi)$ and $f'(x)$ is piecewise continuous in $[-\pi, \pi]$, then at each point where $f'(x)$ has a derivative (and consequently is continuous) we have

$$f'(x) = \sum_{n=1}^{\infty} n(-a_n \sin nx + b_n \cos nx)$$

(*term-by-term differentiation*).

If $f(x)$ is piecewise continuous in $[-\pi, \pi]$, then (no matter whether (3) is true or not) *we have*

$$\int_{-\pi}^{x} f(t)\,dt = \tfrac{1}{2}a_0(x+\pi) + \sum_{n=1}^{\infty} \frac{1}{n}\left[a_n \sin nx - b_n(\cos nx - \cos n\pi)\right] \quad (-\pi \leqq x \leqq \pi)$$

(term-by-term integration).

Similar theorems hold for an interval $[-l, l]$, etc.

Theorem 3 *(Riemann-Lebesgue Theorem). Let $f(x)$ be a function such that the integral $\int_{-\pi}^{\pi} |f(x)|\,dx$ is convergent. Then $a_n \to 0$, $b_n \to 0$, as $n \to \infty$, i.e.*

$$\lim_{n\to\infty} \int_{-\pi}^{\pi} f(x) \cos nx\,dx = 0\,, \quad \lim_{n\to\infty} \int_{-\pi}^{\pi} f(x) \sin nx\,dx = 0\,.$$

(A similar result is true for an interval $[-l, l]$, etc.)

Theorem 4 *(A theorem frequently used, for example in solving partial differential equations by the Fourier method). If $f(x)$, $f'(x)$, $f''(x)$, $f'''(x)$ are continuous in $[0, l]$ and vanish for $x = 0$ and $x = l$, and if $f^{(4)}(x)$ is piecewise continuous in $[0, l]$, then as $n \to \infty$, the Fourier coefficients*

$$b_n = \frac{2}{l} \int_0^l f(x) \sin \frac{n\pi x}{l}\,dx$$

converge to zero as rapidly as $1/n^4$, i.e. $b_n = O(1/n^4)$ when $n \to \infty$. (In other words $n^4 b_n$ is bounded for all n.)

The above theorem, given for $k = 4$, may be stated for any positive integer k.

REMARK 9. In applications, Fourier series in two variables are often encountered. In this case the convergence tests are rather complicated (see e.g. [44]). Let us present a simple criterion:

Theorem 5. *Let the function $f(x, y)$ have continuous derivatives*

$$\frac{\partial f}{\partial x}, \quad \frac{\partial f}{\partial y}, \quad \frac{\partial^2 f}{\partial x\,\partial y}$$

in the rectangle $R(-l \leqq x \leqq l,\ -h \leqq y \leqq h)$. *Then at each interior point of* R *we have*

$$f(x,y) = \sum_{m,n=0}^{\infty} \lambda_{mn} \left[a_{mn} \cos \frac{m\pi x}{l} \cos \frac{n\pi y}{h} + b_{mn} \sin \frac{m\pi x}{l} \cos \frac{n\pi y}{h} + \right. \\ \left. + c_{mn} \cos \frac{m\pi x}{l} \sin \frac{n\pi y}{h} + d_{mn} \sin \frac{m\pi x}{l} \sin \frac{n\pi y}{h} \right], \tag{13}$$

where

$$a_{mn} = \frac{1}{lh} \iint_R f(x,y) \cos \frac{m\pi x}{l} \cos \frac{n\pi y}{h} \, dx\, dy , \tag{14}$$

$$b_{mn} = \frac{1}{lh} \iint_R f(x,y) \sin \frac{m\pi x}{l} \cos \frac{n\pi y}{h} \, dx\, dy , \tag{15}$$

$$c_{mn} = \frac{1}{lh} \iint_R f(x,y) \cos \frac{m\pi x}{l} \sin \frac{n\pi y}{h} \, dx\, dy , \tag{16}$$

$$d_{mn} = \frac{1}{lh} \iint_R f(x,y) \sin \frac{m\pi x}{l} \sin \frac{n\pi y}{h} \, dx\, dy , \tag{17}$$

$$\lambda_{mn} = \begin{cases} \frac{1}{4} & \text{if } m = n = 0 , \\ \frac{1}{2} & \text{if either } n = 0,\ m > 0 \text{ or } m = 0,\ n > 0 \\ 1 & \text{if } m > 0,\ n > 0 . \end{cases} \tag{18}$$

If, in addition, the function $f(x,y)$ has continuous derivatives $\partial^2 f/\partial x^2$, $\partial^3 f/\partial x^2 \partial y$ in R and $f(-l,y) = 0$, $f(l,y) = 0$, then the series (13) can be differentiated term-by-term with respect to x and a series with sum $\partial f/\partial x$ is obtained. A similar assertion is true for differentiation with respect to y.

REMARK 10. The exact meaning of the term "(pointwise) convergence of the series (13) to the function $f(x,y)$" is the following: For any chosen $\varepsilon > 0$ an integer N can be found such that for any pair of integers $p > N$, $q > N$ we have

$$|f(x,y) - s_{pq}(x,y)| < \varepsilon ,$$

where s_{pq} denotes a partial sum of series (13) with m assuming all the values $0, 1, \ldots, p$ and n assuming all the values $0, 1, \ldots, q$ (see Remark 10.2.12).

Example 3. For the function $f(x,y) = xy$ on the square $K(-\pi \leqq x \leqq \pi,\ -\pi \leqq y \leqq \pi)$ we get, by formulae (14) to (17),

$$a_{mn} = b_{mn} = c_{mn} = 0,\ m,\ n \text{ non-negative integers},$$

$$d_{00} = d_{01} = d_{10} = 0,\ d_{mn} = (-1)^{m+n}\frac{4}{mn},\ m,\ n \text{ positive integers}.$$

Thus, by Theorem 5, at any interior point of K we have

$$xy = 4\sum_{m,n=1}^{\infty}(-1)^{m+n}\frac{\sin mx\ \sin ny}{mn}.$$

REMARK 11 (*The Fourier Integral*). For $l \to +\infty$, the Fourier series (5) becomes the Fourier integral:

Theorem 6. *If the integral*

$$\int_{-\infty}^{\infty}|f(x)|\,dx$$

is convergent and if $f(x)$ and $f'(x)$ are piecewise continuous in any finite interval, then

$$\frac{1}{\pi}\int_0^{\infty}du\int_{-\infty}^{\infty}f(t)\cos\{u(t-x)\}\,dt = \begin{cases} f(x) \text{ at each point of continuity} \\ \text{of the function } f(x), \\ \frac{1}{2}\{f(x+0)+f(x-0)\} \text{ at} \\ \text{each point of discontinuity} \\ \text{of the function } f(x). \end{cases} \tag{19}$$

REMARK 12. Expanding the term $\cos\{u(t-x)\}$ in (19), we get

$$\frac{1}{\pi}\int_0^{\infty}\cos ux\,du\int_{-\infty}^{\infty}f(t)\cos ut\,dt + \frac{1}{\pi}\int_0^{\infty}\sin ux\,du\int_{-\infty}^{\infty}f(t)\sin ut\,dt =$$

$$= \begin{cases} f(x) \\ \\ \frac{1}{2}(f(x+0)+f(x-0)). \end{cases}$$

If $f(x)$ is an even function, i.e. if $f(-x) \equiv f(x)$, then

$$\frac{2}{\pi}\int_0^{\infty}\cos ux\,du\int_0^{\infty}f(t)\cos ut\,dt = \begin{cases} f(x), \\ \\ \frac{1}{2}(f(x+0)+f(x-0)). \end{cases}$$

If $f(x)$ is odd, i.e. if $f(-x) \equiv -f(x)$, then

$$\frac{2}{\pi}\int_0^{\infty}\sin ux\,du\int_0^{\infty}f(t)\sin ut\,dt = \begin{cases} f(x), \\ \\ \frac{1}{2}(f(x+0)+f(x-0)). \end{cases}$$

Using complex form, we have:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixu}\,du\int_{-\infty}^{\infty} f(t)\,e^{iut}\,dt = \begin{cases} f(x)\,, \\ \frac{1}{2}\,(f(x+0)+f(x-0))\,, \end{cases}$$

where the first improper integral is to be taken in the sense of the Cauchy principal value, i.e. as the limit of $\int_{-a}^{a}$ for $a \to +\infty$.

REMARK 13 (*Harmonic Analysis*). Consider a real function f with period b and let us expand it in $[0,b]$ into the corresponding Fourier series

$$\frac{a_0}{2}+\sum_{n=1}^{\infty}\left(a_n\,\cos\frac{2\pi nx}{b}+b_n\,\sin\frac{2\pi nx}{b}\right), \tag{20}$$

where

$$a_n=\frac{2}{b}\int_0^b f(x)\,\cos\frac{2\pi nx}{b}\,dx\,,\quad b_n=\frac{2}{b}\int_0^b f(x)\,\sin\frac{2\pi nx}{b}\,dx \tag{21}$$

(see Remark 5, formulae (5'), (6'), with $T=b$, $\omega=2\pi/b$ and $t=x$).

Often, the values of the function f are known only at some discrete points of the interval $[0,b]$ (when obtained by measurement, for example). Then to the calculation of a_n, b_n, quadrature formulae must be used, as usual. According to Remark 13.13.1 (the integrated functions being periodic), the trapezoidal rule is of a particularly convenient use in this case. Writing in (13.13.5) $h=b/m$, $a_k=kb/m$ and taking into account that the values of the integrated functions are the same at the points $x=0$, $x=b$, we obtain

$$\begin{aligned} a_n &\approx \frac{2}{m}\sum_{k=0}^{m-1} f(a_k)\,\cos\frac{2\pi nk}{m}\,, \\ b_n &\approx \frac{2}{m}\sum_{k=0}^{m-1} f(a_k)\,\sin\frac{2\pi nk}{m}\,. \end{aligned} \tag{22}$$

For computing these sums, the so-called *fast Fourier transformation* is very profitable to be used. This effective algorithm, considerably significant also in other fields of numerical analysis, serves for numerical computation of sums of the form

$$\sum_{k=0}^{m-1} e^{\frac{2\pi ink}{m}}\,f_k\,,\quad n=0,1,\ldots,m-1 \tag{23}$$

into which the sums in (22) may obviously be easily transformated. We describe it for the case $m = 2^r$:

Computation of m sums of the form (23) is equivalent to determination of m values of the polynomial

$$p_{m-1}(z) = f_{m-1}\, z^{m-1} + \ldots + f_0$$

of degree $m-1$ in m points $e^{\frac{2\pi i n}{m}}$, $m = 0, 1, \ldots, m-1$. However, this polynomial can be rewritten as

$$p_{m-1}(z) = z p_{\frac{m}{2}-1}(z^2) + q_{\frac{m}{2}-1}(z^2)\,, \tag{24}$$

where

$$p_{\frac{m}{2}-1}(z) = \sum_{k=0}^{\frac{m}{2}-1} f_{2k+1}\, z^k$$

and

$$q_{\frac{m}{2}-1}(z) = \sum_{k=0}^{\frac{m}{2}-1} f_{2k}\, z^k\,.$$

(Here, the index denotes the degree of the corresponding polynomials.) Using (24), we thus have replaced computation of m values of a polynomial of degree $m-1$ by computation of $m/2$ values of two polynomials of degree $\frac{m}{2}-1$, since we have

$$\left(e^{\frac{2\pi i n}{m}}\right)^2 = \left(e^{\frac{2\pi i (n-\frac{m}{2})}{m}}\right)^2$$

for $n = \frac{m}{2}, \frac{m}{2}+1, \ldots, m-1$.

Any of the polynomials $p_{\frac{m}{2}-1}$, $q_{\frac{m}{2}-1}$ can again be replaced, similarly as in (24), by two polynomials of degree $m/2^2 - 1$, and we thus may compute the values of these polynomials at $m/2^2$ points. Going on in this way, we reduce, successively, the degrees of polynomials, the values of which we had to compute, till we come to 2^{r-1} polynomials of the first degree the values of which we compute at the points $+1$ and -1.

The number of operations is not greater than $2m \log_2 m$, while the direct use of (23) leads to a number of operations which is proportional to m^2. Thus the difference is significant even for a relatively small m. For example, for $m = 2^7 = 128$, $m^2 = 16\,384$, while $2m \log_2 m = 2 \times 2^7 \times 7 = 1\,792$. In this case, the algorithm of the fast Fourier transform is approximately ten times faster than the straightforward procedure.

16.4. Bessel Functions

In this paragraph, ν, or n denotes a real number, or an integer, respectively.

Definition 1. *Bessel functions (cylindrical functions) of the first kind and index (order)* ν are defined by

$$J_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,\Gamma(\nu+k+1)} \left(\frac{x}{2}\right)^{2k}\,. \tag{1}$$

The infinite series in (1) is convergent for every real ν (see, however, Remark 1) and every real, or complex x.
About the meaning of $(x/2)^\nu$ if x is complex, see in Remark 20.6.4.

REMARK 1. For $k = 0$ we have $k! = 1$. On the function Γ see in §13.11. If $\nu = n$ is a nonnegative integer, then $\Gamma(n + k + 1) = (n + k)!$ (see (13.11.6)). If $\nu + k + 1 = 0, -1, -2, \ldots$, we define $1/\Gamma(\nu + k + 1) = 0$. In this way, (1) has sense also if ν is a negative integer.

As functions of a complex variable x, (1) are analytic functions, with singularities at the point $x = 0$ and $x = \infty$, in general. Definition 1 of Bessel functions can be extended for complex ν: It suffices, for example, to define the function Γ for the case of a complex variable by the relation (13.11.10), see e.g. [183].

Example 1.

$$J_2(it) = \frac{1}{2!}\left(\frac{it}{2}\right)^2 - \frac{1}{1!\,3!}\left(\frac{it}{2}\right)^4 + \frac{1}{2!\,4!}\left(\frac{it}{2}\right)^6 - \ldots =$$

$$= -\left(\frac{t^2}{8} + \frac{t^4}{96} + \frac{t^6}{3072} + \ldots\right)$$

(for every t).

Theorem 1. *If n is an integer, then*

$$J_{-n}(x) = (-1)^n J_n(x) . \tag{2}$$

REMARK 2. Thus if n is a negative integer, it is not necessary to use (1) for computing $J_n(x)$, but it suffices to put $n = -m$ (where thus m is a positive integer) and to compute $J_m(x)$.

Theorem 2. *The function $J_\nu(x)$ satisfies the so-called Bessel differential equation*

$$x^2y'' + xy' + (x^2 - \nu^2)\,y = 0 \tag{3}$$

(see §17.15 *and* §17.21, *equation* 117*).*

REMARK 3. If ν is *not* an integer, then the functions $J_\nu(x)$ and $J_{-\nu}(x)$ can be shown to be linearly independent and

$$y = c_1 J_\nu(x) + c_2 J_{-\nu}(x) \tag{4}$$

is the general integral of equation (3). If $\nu = n$ is an integer, then (2) implies that the functions $J_n(x)$ and $J_{-n}(x)$ are not linearly independent and that (4) is not a general integral of that equation. The general integral is then

$$y = c_1 J_n(x) + c_2 Y_n(x) , \tag{5}$$

see Remark 16.

Theorem 3. *The Bessel functions of integral indices can be taken for coefficients of the Laurent expansion (§20.4) of the so-called generating function: For every x and $t \neq 0$ we have*

$$e^{\frac{x}{2}(t-\frac{1}{t})} = \sum_{n=-\infty}^{\infty} J_n(x)\, t^n =$$

$$= J_0(x) + (J_1(x)\, t + J_2(x)\, t^2 + \ldots) + (J_{-1}(x)\, t^{-1} + J_{-2}(x)\, t^{-2} + \ldots) .$$

Theorem 4 *(the Integral Form). For $n = 0, 1, 2, \ldots$ we have*

$$J_n(x) = \frac{1}{\pi} \int_0^{\pi} \cos(x \sin \theta - n\theta)\, d\theta . \tag{6}$$

REMARK 4. For applications, the most important Bessel Functions are the functions

$$J_0(x) = 1 - \frac{x^2}{2^2} + \frac{x^4}{(2 \times 4)^2} - \frac{x^6}{(2 \times 4 \times 6)^2} + \ldots , \tag{7}$$

$$J_1(x) = \frac{x}{2} \left(1 - \frac{x^2}{2 \times 4} + \frac{x^4}{2 \times 4^2 \times 6} - \right.$$

$$\left. - \frac{x^6}{2 \times (4 \times 6)^2 \times 8} + \frac{x^8}{2 \times (4 \times 6 \times 8)^2 \times 10} + \ldots \right) . \tag{8}$$

Their graphs are sketched in Fig. 16.15. In Table 16.1 their values for $x = 0$, $x = 0.5, \ldots$ are given. Of frequent application are also Tables 16.2, 16.3, giving positive, or nonnegative roots of the equation $J_n(x) = 0$, or $J_n'(x) = 0$, respectively.

Theorem 5 *(Recursion Formulae):*

$$2\nu J_\nu(x) = x J_{\nu-1}(x) + x J_{\nu+1}(x) , \tag{9}$$

$$2J_\nu'(x) = J_{\nu-1}(x) - J_{\nu+1}(x) , \tag{10}$$

$$x J_\nu'(x) = \nu J_\nu(x) - x J_{\nu+1}(x) , \tag{11}$$

$$x J_\nu'(x) = -\nu J_\nu(x) + x J_{\nu-1}(x) . \tag{12}$$

TABLE 16.1

Bessel functions of orders zero and one

x	$J_0(x)$	$J_1(x)$	x	$J_0(x)$	$J_1(x)$	x	$J_0(x)$	$J_1(x)$
0.0	1.000 0	0.000 0	5.5	−0.006 8	− 0.341 4	11.0	− 0.171 2	− 0.176 8
0.5	0.938 5	0.242 3	6.0	0.150 6	− 0.276 7	11.5	− 0.067 7	− 0.228 4
1.0	0.765 2	0.440 1	6.5	0.260 1	− 0.153 8	12.0	0.047 7	− 0.223 4
1.5	0.511 8	0.557 9	7.0	0.300 1	− 0.004 7	12.5	0.146 9	− 0.165 5
2.0	0.223 9	0.576 7	7.5	0.266 3	0.135 2	13.0	0.206 9	− 0.070 3
2.5	− 0.048 4	0.497 1	8.0	0.171 7	0.234 6	13.5	0.215 0	0.038 0
3.0	− 0.260 1	0.339 1	8.5	0.041 9	0.273 1	14.0	0.171 1	0.133 4
3.5	− 0.380 1	0.137 4	9.0	− 0.090 3	0.245 3	14.5	− 0.087 5	0.193 4
4.0	− 0.397 1	− 0.066 0	9.5	− 0.193 9	0.161 3	15.0	− 0.014 2	0.205 1
4.5	− 0.320 5	− 0.231 1	10.0	− 0.245 9	0.043 5	15.5	− 0.109 2	0.167 2
5.0	− 0.177 6	− 0.327 6	10.5	− 0.236 6	− 0.078 9	16.0	− 0.174 9	0.090 4

TABLE 16.2

Positive Roots of the Equation $J_n(x) = 0$

	Order of the Root							
Index n	1	2	3	4	5	6	7	8
0	2.404 83	5.520 08	8.653 73	11.791 53	14.930 92	18.071 06	21.211 64	24.352 47
1	3.831 71	7.015 59	10.173 47	13.323 69	16.470 63	19.615 86	22.760 08	
2	5.135 62	8.417 24	11.619 84	14.795 95	17.959 82	21.117 00	24.271 12	
3	6.380 16	9.761 02	13.015 20	16.223 47	19.409 42	22.582 73		
4	7.588 34	11.064 71	14.372 54	17.616 0	20.826 9	24.199 0		
5	8.771 42	12.338 60	15.700 17	18.980 1	22.217 8			
6	9.936 11	13.589 29	17.003 8	20.320 8	23.586 1			
7	11.086 37	14.821 27	18.287 6	21.641 6	24.934 9			

TABLE 16.3

Nonnegative Roots of the Equation $J_n'(x) = 0$

	Order of the Root								
Index n	1	2	3	4	5	6	7	8	9
0	0.000 0	3.831 7	7.015 6	10.173 5	13.323 7	16.470 6	19.615 9	22.760 1	25.903 7
1	1.841 2	5.331 4	8.536 3	11.706 0	14.863 6	18.015 5	21.164 4	24.311 3	
2	3.054 2	6.706 1	9.969 5	13.170 4	16.347 5	19.512 9	22.672 1		
3	4.201 2	8.015 2	11.345 9	14.585 9	17.788 8	20.972 4	24.146 9		
4	5.317 5	9.282 4	12.681 9	15.964 1	19.196 0	22.401 0			
5	6.415 6	10.519 9	13.987 2	17.312 8	20.575 5	23.803 3			
6	7.501 3	11.734 9	15.268 2	18.637 4	21.931 8				
7	8.577 8	12.932 4	16.529 4	19.941 9	23.268 1				

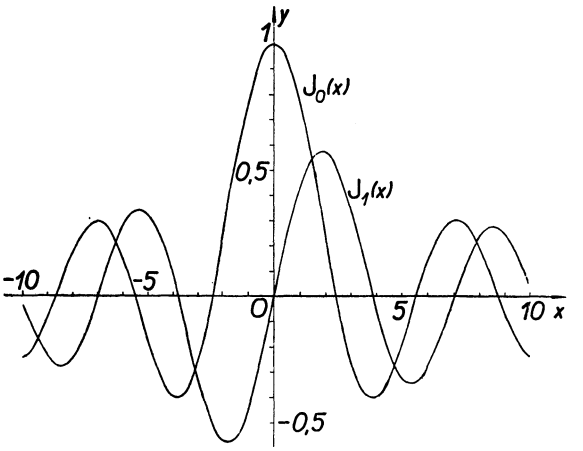


Fig. 16.15.

In particular, (11) yields, for $\nu = 0$,

$$J_0'(x) = -J_1(x) \ . \tag{13}$$

From (6) we then obtain, for $\nu = 0, 1, 2, \ldots$,

$$|J_n(x)| \leqq 1, \ |J_n^{(k)}(x)| < 1 \quad (k = 1, 2, 3, \ldots)$$

for all real x; further,

$$\lim_{x \to \infty} J_n(x) = 0$$

(see also (16)).

REMARK 5 (the Functions $J_\nu(x)$ for $\nu = n + \frac{1}{2}$ or $\nu = -n - \frac{1}{2}$, $n = 0, 1, 2, \ldots$):

$$J_{n+\frac{1}{2}}(x) = (-1)^n \sqrt{\left(\frac{2}{\pi}\right)} \, x^{n+\frac{1}{2}} \left(\frac{1}{x} \frac{\mathrm{d}}{\mathrm{d}x}\right)^n \frac{\sin x}{x} \ , \tag{14}$$

$$J_{-n-\frac{1}{2}}(x) = \sqrt{\left(\frac{2}{\pi}\right)} \, x^{n+\frac{1}{2}} \left(\frac{1}{x} \frac{\mathrm{d}}{\mathrm{d}x}\right)^n \frac{\cos x}{x} \ . \tag{15}$$

In particular,

$$J_{1/2}(x) = \sqrt{\left(\frac{2}{\pi x}\right)} \sin x \ , \quad J_{3/2}(x) = \sqrt{\left(\frac{2}{\pi x}\right)} \left(\frac{\sin x}{x} - \cos x\right) \ ,$$

$$J_{-1/2}(x) = \sqrt{\left(\frac{2}{\pi x}\right)}\ \cos x\ ,\quad J_{-3/2}(x) = -\sqrt{\left(\frac{2}{\pi x}\right)}\ \left(\frac{\cos x}{x} + \sin x\right)\ .$$

REMARK 6 ("Limit Form" of Bessel Functions). If the numbers ν and x are real, then for $x \gg 1$ we have

$$J_\nu(x) \approx \sqrt{\left(\frac{2}{\pi x}\right)}\ \cos\varphi \tag{16}$$

with

$$\varphi = x - (\nu + \tfrac{1}{2})\ \frac{\pi}{2}\ .$$

Theorem 6. *For every real ν, the function $J_\nu(x)$ has a countable set of positive zeros $x_1 < x_2 < x_3 < \ldots$. At the same time, the relation*

$$x_n \to +\infty \quad \textit{for} \quad n \to \infty$$

holds.

Theorem 7. *Let $\lambda_1 < \lambda_2 < \lambda_3 < \ldots$ be positive roots of the equation*

$$J_\nu(\lambda c) = 0 \quad (c > 0 \textit{ fixed, } \nu \textit{ fixed, real, nonnegative})\ . \tag{17}$$

Then the functions $J_\nu(\lambda_1 x)$, $J_\nu(\lambda_2 x)$, $J_\nu(\lambda_3 x), \ldots$ form an orthogonal system with the weight function x (Definition 16.2.5) in the interval $[0, c]$, i.e. we have

$$\int_0^c x J_\nu(\lambda_i x)\ J_\nu(\lambda_k x)\,dx = 0 \quad \text{for} \quad i \neq k\ . \tag{18}$$

REMARK 7. The same is valid if $\lambda_1 < \lambda_2 < \lambda_3 < \ldots$ are nonnegative roots of the equation

$$\lambda c\, J'_\nu(\lambda c) = -h J_\nu(\lambda c)\ , \tag{19}$$

where h is a real constant, not necessarily different from zero.

Theorem 8. *If $\nu \geqq 0$, $h \geqq 0$, then the equation* (17), *or* (19) *has only real roots.*

REMARK 8. From among them only nonnegative roots are taken into account when constructing orthogonal systems mentioned in Theorem 7 and Remark 7. At the same time, the root $\lambda = 0$ is used only in the case of equation (19) with $\nu = 0$ and $h = 0$.

REMARK 9. On how to come to the equation (17), or (19), see in §26.4.

REMARK 10. (Complete) orthonormal systems

$$\varphi_{\nu 1}(x)\ ,\quad \varphi_{\nu 2}(x)\ ,\quad \varphi_{\nu 3}(x)\ ,\dots$$

with the weight function x, for which thus the relations

$$\int_0^c x\,\varphi_{\nu i}(x)\,\varphi_{\nu k}(x)\,\mathrm{d}x = \begin{cases} 1 & \text{for } i = k\ , \\ 0 & \text{for } i \neq k\ , \end{cases}$$

hold, are obtained from orthogonal systems, mentioned in Theorem 7 and Remark 7, in the following way:

In the case of the equation (17):

$$\varphi_{\nu k}(x) = \frac{\mathrm{J}_\nu(\lambda_k x)}{\sqrt{\left[\frac{c^2}{2}\,\mathrm{J}^2_{\nu+1}(\lambda_k c)\right]}}\ ;$$

in the case of the equation (19):

$$\varphi_{\nu k}(x) = \frac{\mathrm{J}_\nu(\lambda_k x)}{\sqrt{\left[\frac{\lambda_k^2 c^2 + h^2 - \nu^2}{2\lambda_k^2}\right]\,\mathrm{J}^2_\nu(\lambda_k c)}}\ .$$

For $\nu = 0,\ h = 0,\ \lambda = 0$ we take

$$\varphi_{\nu 1}(x) = \sqrt{(2)}/c\ .$$

Let us note that the λ_k's from Theorem 7 are different, in general, from those from Remark 7; thus also the corresponding functions $\mathrm{J}_\nu(\lambda_k x)$ are different in both cases, in general.

Theorem 9 *(The Fourier-Bessel Expansion). Let*

$$\mathrm{J}_\nu(\lambda_1 x)\ ,\quad \mathrm{J}_\nu(\lambda_2 x)\ ,\quad \mathrm{J}_\nu(\lambda_3 x),\dots$$

be functions from Theorem 7, or Remark 7, respectively, $\nu \geqq 0$, $h \geqq 0$. (It is assumed that $\lambda_1 < \lambda_2 < \lambda_3 < \dots$ are all positive, or nonnegative roots of the equation (17), *or* (19), *respectively, while $\lambda = 0$ is considered only in the case of equation* (19) *with*

$\nu = 0$, $h = 0$; ν, h are fixed). Let $f(x)$ and $f'(x)$ be piecewise continuous functions in the interval $[0, c]$. Then

$$\sum_{k=1}^{\infty} a_k \, \mathrm{J}_\nu(\lambda_k x) =$$

$$= \begin{cases} f(x) & \text{at every point } x \text{ at which } f(x) \text{ is continuous} , \\ \frac{1}{2}\,[f(x+) + f(x-)] & \text{at every point } x \text{ at which } f(x) \text{ is discontinuous} \end{cases} \tag{20}$$

$(0 < x < c$; $f(x+)$, or $f(x-)$ means the right-hand, or left-hand limit of the function $f(x)$ at the point x, respectively). The convergence of the series (20) *is uniform in every closed interval lying in the interior of an interval in which $f(x)$ is continuous. Here,*

$$a_k = \frac{2}{c^2 \mathrm{J}_{\nu+1}^2(\lambda_k c)} \int_0^c x \, \mathrm{J}_\nu(\lambda_k x) \, f(x) \, \mathrm{d}x , \quad k = 1, 2, 3, \ldots \tag{21}$$

if λ_k are positive roots of equation (17) *and*

$$a_k = \frac{2\lambda_k^2}{(\lambda_k^2 c^2 + h^2 - \nu^2) \, \mathrm{J}_\nu^2(\lambda_k c)} \int_0^c x \, \mathrm{J}_\nu(\lambda_k x) \, f(x) \, \mathrm{d}x , \quad k = 1, 2, 3, \ldots \tag{22}$$

if λ_k are nonnegative roots of equation (19).

REMARK 11. If $\nu = 0$, $h = 0$, then $\lambda_1 = 0$ in (22) and

$$a_1 = \frac{2}{c^2} \int_0^c x f(x) \, \mathrm{d}x .$$

REMARK 12. The validity of (20) is ensured if only $f(x)$ is integrable on every closed interval $[\varepsilon, c]$ with $\varepsilon > 0$ arbitrarily small, if integral

$$\int_0^c \sqrt{(x)} \, |f(x)| \, \mathrm{d}x$$

is convergent and the point x is an interior point of an interval where $f(x)$ is of bounded variation. The convergence is then uniform, again, in every closed interval which lies in the interior of an interval where $f(x)$ is continuous.

REMARK 13. Theorem 9 can be generalized for the case $\nu \geqq -\frac{1}{2}$ (see, e.g., [183], where Bessel functions are thoroughly treated).

REMARK 14. Theorem 9 is of fundamental significance in applications (see also §26.4).

REMARK 15 *(Bessel Functions of the Second Kind, the Weber (Neumann) Functions). Bessel functions of the second kind* are defined, if ν is not an integer, by

$$Y_\nu(x) = \frac{J_\nu(x)\ \cos \nu\pi - J_{-\nu}(x)}{\sin \nu\pi}\ , \quad \nu \neq n\ , \tag{23}$$

if ν is an integer, by

$$Y_n(x) = \lim_{\nu \to n} \frac{J_\nu(x)\ \cos \nu\pi - J_{-\nu}(x)}{\sin \nu\pi}\ . \tag{24}$$

The functions $Y_n(x)$ are called the *Weber* (or *Neumann*) functions. They are often denoted by $N_n(x)$. Let us note that under the Neumann functions sometimes also the functions

$$Y_\nu(x) = \tfrac{1}{2}\pi\ Y_\nu(x) + (\ln 2 - C)\ J_\nu(x)$$

are understood in the literature; here

$$C = 0.577\ 215\ 664\ 9\ldots \tag{25}$$

is the Euler constant.

Theorem 10. *For $n \geqq 0$, n an integer, we have*

$$Y_n(x) = \frac{2}{\pi} J_n(x)\ (\ln \frac{x}{2} + C) - \frac{1}{\pi} \left(\frac{x}{2}\right)^{-n} \sum_{k=0}^{n-1} \frac{(n-k-1)!}{k!} \left(\frac{x}{2}\right)^{2k} -$$

$$- \frac{1}{\pi} \left(\frac{x}{2}\right)^{n} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(n-k)!} \left(\frac{x}{2}\right)^{2k} \left[\sum_{l=1}^{k} \frac{1}{l} + \sum_{l=1}^{k+n} \frac{1}{l}\right]\ . \tag{26}$$

Here, we have to put

$$\sum_{l=1}^{k} \frac{1}{l} + \sum_{l=1}^{k+n} \frac{1}{l} = 1 + \ldots + \frac{1}{n} \quad \text{if} \quad k = 0$$

and

$$\sum_{l=1}^{k} \frac{1}{l} + \sum_{l=1}^{k+n} \frac{1}{l} = 0 \quad \text{if} \quad k = 0 \quad \text{and} \quad n = 0\ .$$

The infinite series in (26) is convergent for every x (complex, in general). As functions of a complex variable, the functions (26) are analytic functions (multivalued with a cut along the negative real half-axis, due to the presence of the logarithmic term in (26)).

A table is given below for the functions $Y_0(x)$, $Y_1(x)$, most often encountered in applications (Tab. 16.4).

TABLE 16.4

The Weber (Neumann) Functions $Y_0(x)$, $Y_1(x)$

x	$Y_0(x)$	$Y_1(x)$	x	$Y_0(x)$	$Y_1(x)$
0.0	–	–	5.5	−0.339 5	−0.023 8
0.5	−0.444 5	−1.471	6.0	−0.288 2	−0.175 0
1.0	0.088 3	−0.781 2	6.5	−0.173 2	−0.274 1
1.5	0.382 4	−0.412 3	7.0	−0.025 9	−0.302 7
2.0	0.510 4	−0.107 0	7.5	0.117 3	−0.259 1
2.5	0.498 1	0.145 9	8.0	0.223 5	−0.158 1
3.0	0.376 9	0.324 7	8.5	0.270 2	−0.026 2
3.5	0.189 0	0.410 2	9.0	0.249 9	0.104 3
4.0	−0.016 9	0.397 9	9.5	0.171 2	0.203 2
4.5	−0.194 7	0.301 0	10.0	0.055 7	0.249 0
5.0	−0.308 5	0.147 9			

REMARK 16. Similarly as the functions $J_\nu(x)$, also the functions $Y_\nu(x)$ are solutions of equation (3). If $\nu = n$ is an integer, then we have

$$Y_{-n}(x) = (-1)^n\, Y(x)\ ,$$

and the functions of the form

$$c_1 Y_n(x) + c_2 Y_{-n}(x)$$

do not represent the general integral of that equation. Its general solution is then of the form

$$c_1 J_n(x) + c_2 Y_n(x)\ ,$$

cf. Remark 3.

REMARK 17. Relations (9)–(12) (and thus also the relation (13)) hold as well for the functions Y_ν. Further, if $\nu = n + \frac{1}{2}$, or $\nu = -n - \frac{1}{2}$, $n = 0, 1, 2, \ldots$, we have

$$Y_{n+\frac{1}{2}}(x) = (-1)^{n+1} J_{-n-\frac{1}{2}}(x)\ ,$$

$$Y_{-n-\frac{1}{2}}(x) = (-1)^n J_{n+\frac{1}{2}}(x)\ .$$

For ν, x real, $x \gg 1$, we obtain

$$Y_\nu(x) \approx \sqrt{\left(\frac{2}{\pi x}\right)}\ \sin\varphi\ ,$$

where $\varphi = x - (\nu + \frac{1}{2})\,\frac{\pi}{2}$.

REMARK 18. The *Bessel functions of the third kind* (the *Hankel functions*) are defined by

$$\mathrm{H}_\nu^{(1)}(x) = \mathrm{J}_\nu(x) + \mathrm{i}\mathrm{Y}_\nu(x) \ ,$$

$$\mathrm{H}_\nu^{(2)}(x) = \mathrm{J}_\nu(x) - \mathrm{i}\mathrm{Y}_\nu(x) \ ,$$

where i is the imaginary unit.

Also for them the relations (9)–(13) are valid.

REMARK 19 (*Modified Bessel Functions of the First and Second Kinds*). The *modified Bessel function of the first kind* $\mathrm{I}_\nu(x)$ is defined by

$$\mathrm{I}_\nu(x) = \mathrm{i}^{-\nu}\,\mathrm{J}_\nu(\mathrm{i}x) \ , \tag{27}$$

where i is the imaginary unit. We have

$$\mathrm{I}_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{1}{k!\,\Gamma(\nu + k + 1)} \left(\frac{x}{2}\right)^{2k} \ . \tag{28}$$

The infinite series in (28) is convergent for every x (complex, in general; see also Remark 1).

For integral n we have

$$\mathrm{I}_{-n}(x) = \mathrm{I}_n(x) \ . \tag{29}$$

Recursion formulae:

$$2\nu\mathrm{I}(x) = x\mathrm{I}_{\nu-1}(x) - x\mathrm{I}_{\nu+1}(x) \ ,$$

$$2\mathrm{I}'_\nu(x) = \mathrm{I}_{\nu-1}(x) + \mathrm{I}_{\nu+1}(x) \ ,$$

$$x\mathrm{I}'_\nu(x) = \nu\mathrm{I}_\nu(x) + x\mathrm{I}_{\nu+1}(x) \ ,$$

$$x\mathrm{I}'_\nu(x) = -\nu\mathrm{I}_\nu(x) + x\mathrm{I}_{\nu-1}(x) \ .$$

Especially, the third of these formulae yields

$$\mathrm{I}'_0(x) = \mathrm{I}_1(x) \ . \tag{30}$$

The functions $\mathrm{I}_\nu(x)$ satisfy the differential equation

$$x^2y'' + xy' - (x^2 + \nu^2)\,y = 0 \ . \tag{31}$$

If ν is no integer, then

$$y = c_1 \mathrm{I}_\nu(x) + c_2 \mathrm{I}_{-\nu}(x)$$

is the general integral of equation (31). If $\nu = n$ is an integer, then the functions $\mathrm{I}_n(x)$ and $\mathrm{I}_{-n}(x)$ are linearly dependent (see (30)) and

$$y = c_1 \mathrm{I}_n(x) + c_2 \mathrm{I}_{-n}(x)$$

is no general integral of that equation. Its general integral is then

$$y = c_1 \mathrm{I}_n(x) + c_2 \mathrm{K}_n(x) \ ,$$

where $\mathrm{K}_n(x)$ is the so-called *modified Bessel function of the second kind* (called also the *Mac Donald function*):

If ν is no integer, we define

$$\mathrm{K}_\nu(x) = \frac{\pi}{2} \frac{\mathrm{I}_{-\nu}(x) - \mathrm{I}_\nu(x)}{\sin \nu\pi} \ ,$$

and if $\nu = n$ is an integer, then

$$\mathrm{K}_n(x) = \lim_{\nu \to n} \frac{\pi}{2} \frac{\mathrm{I}_{-\nu}(x) - \mathrm{I}_\nu(x)}{\sin \nu\pi} \ .$$

We have

$$\mathrm{K}_\nu(x) = \frac{\pi}{2} \mathrm{i}^{\nu+1} \mathrm{H}_\nu^{(1)}(\mathrm{i}x) \ .$$

Recursion Formulae:

$$2\nu \mathrm{K}_\nu(x) = x\mathrm{K}_{\nu+1}(x) - x\mathrm{K}_{\nu-1}(x) \ ,$$

$$2\mathrm{K}'_\nu(x) = -\mathrm{K}_{\nu+1}(x) - \mathrm{K}_{\nu-1}(x) \ ,$$

$$x\mathrm{K}'_\nu(x) = \nu \mathrm{K}_\nu(x) - x\mathrm{K}_{\nu+1}(x) \ ,$$

$$x\mathrm{K}'_\nu(x) = -\nu \mathrm{K}_\nu(x) - x\mathrm{K}_{\nu-1}(x) \ .$$

In particular, from the third of these formulae, it follows that

$$\mathrm{K}'_0(x) = -\mathrm{K}_1(x) \ .$$

REMARK 20 (The *Kelvin Functions*). From the point of view of applications, the differential equation

$$x^2 y'' + xy' - \mathrm{i}x^2 y = 0$$

is of interest. Its general integral is

$$y = c_1 \mathrm{J}_0(\mathrm{i}^{\frac{2}{3}} x) + c_2 \mathrm{K}_0(\mathrm{i}^{\frac{1}{2}} x) \ .$$

Expanding the function $J_0(i^{\frac{2}{3}}x)$ in a power series, we obtain

$$J_0(i^{\frac{3}{2}}x) = 1 + i\,\frac{\left(\frac{x}{2}\right)^2}{(1!)^2} - \frac{\left(\frac{x}{2}\right)^4}{(2!)^2} - i\,\frac{\left(\frac{x}{2}\right)^6}{(3!)^2} + \dots .$$

Thus we can write

$$J_0(i^{\frac{3}{2}}x) = \operatorname{ber} x + i \operatorname{bei} x ,$$

where the functions

$$\operatorname{ber} x = 1 - \frac{\left(\frac{x}{2}\right)^4}{(2!)^2} + \frac{\left(\frac{x}{2}\right)^8}{(4!)^2} - \frac{\left(\frac{x}{2}\right)^{12}}{(6!)^2} + \dots ,$$

$$\operatorname{bei} x = \frac{\left(\frac{x}{2}\right)^2}{(1!)^2} - \frac{\left(\frac{x}{2}\right)^6}{(3!)^2} + \frac{\left(\frac{x}{2}\right)^{10}}{(5!)^2} - \dots$$

are the so-called *Kelvin functions* (the "*Bessel real*" and the "*Bessel imaginary*" *function*, respectively), corresponding to the function J_0.

It can be easily shown that

$$J_0(i^{-\frac{3}{2}}x) = \operatorname{ber} x - i \operatorname{bei} x .$$

Similarly, by the relations

$$K_0(i^{\frac{1}{2}}x) = \ker x + i \operatorname{kei} x , \quad K_0(i^{-\frac{1}{2}}x) = \ker x - i \operatorname{kei} x$$

the *Kelvin functions* $\ker x$ and $\operatorname{kei} x$ are defined. We have

$$\ker x = -(\ln \frac{x}{2} + C) \operatorname{ber} x + \frac{\pi}{4} \operatorname{bei} x -$$

$$- \frac{\left(\frac{x}{2}\right)^4}{(2!)^2} (1 + \tfrac{1}{2}) + \frac{\left(\frac{x}{2}\right)^8}{(4!)^2} (1 + \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{4}) - \dots ,$$

$$\operatorname{kei} x = -(\ln \frac{x}{2} + C) \operatorname{bei} x - \frac{\pi}{4} \operatorname{ber} x +$$

$$+ \frac{\left(\frac{x}{2}\right)^2}{(1!)^2} - \frac{\left(\frac{x}{2}\right)^6}{(3!)^2} (1 + \tfrac{1}{2} + \tfrac{1}{3}) + \dots ,$$

where $C = 0.577\ 215\ 664\ 9$ is the Euler constant.

More generally, by the relations

$$J_\nu(i^{\frac{3}{2}}x) = \operatorname{ber}_\nu x + i \operatorname{bei}_\nu x ,$$

$$J_\nu(i^{-\frac{3}{2}}x) = \operatorname{ber}_\nu x - i \operatorname{bei}_\nu x$$

$$i^{-\nu}\mathrm{K}_\nu(\mathrm{i}^{\frac{1}{2}}x) = \mathrm{ker}_\nu x + \mathrm{i}\,\mathrm{kei}_\nu x\,,$$

$$i^{\nu}\mathrm{K}_\nu(\mathrm{i}^{-\frac{1}{2}}x) = \mathrm{ker}_\nu x - \mathrm{i}\,\mathrm{kei}_\nu(x)$$

the so-called *Kelvin functions of order* ν can be defined.

16.5. Legendre Polynomials. Spherical Harmonics

Definition 1. The polynomial

$$\mathrm{P}_n(x) = \frac{1\,.\,3\,.\,5\,.\,\ldots\,.\,(2n-1)}{n!}\left[x^n - \frac{n(n-1)}{2(2n-1)}\,x^{n-2} + \right.$$

$$\left. + \frac{n(n-1)\,(n-2)\,(n-3)}{2\,.\,4(2n-1)\,(2n-3)}\,x^{n-4} - \ldots\right] \tag{1}$$

is called the *Legendre polynomial of degree* n.

Theorem 1. *We have*

$$\mathrm{P}_n(x) = \frac{1}{2^n n!}\,\frac{\mathrm{d}^n}{\mathrm{d}x^n}\,(x^2-1)^n \tag{2}$$

(*the Rodrigues formula*).

Example 1. For $n = 2$ we have by (2):

$$\mathrm{P}_2(x) = \frac{1}{2^2\,.\,2!}\,\frac{\mathrm{d}^2}{\mathrm{d}x^2}\,(x^4 - 2x^2 + 1) = \tfrac{3}{2}x^2 - \tfrac{1}{2}\,.$$

This result, of course, may also be obtained from (1).

Theorem 2. *The function* $y = \mathrm{P}_n(x)$ *satisfies the* (*Legendre*) *differential equation* (see also §17.21, equation 129 and further)

$$(1-x^2)\,y'' - 2xy' + n(n+1)\,y = 0\,. \tag{3}$$

REMARK 1. The polynomial $\mathrm{P}_n(\cos\vartheta)$ satisfies the equation

$$\frac{1}{\sin\vartheta}\,\frac{\mathrm{d}}{\mathrm{d}\vartheta}\left(\sin\vartheta\,\frac{\mathrm{d}y}{\mathrm{d}\vartheta}\right) + n(n+1)\,y = 0\,, \tag{4}$$

which can be obtained from (3) by the substitution $x = \cos\vartheta$.

Theorem 3. *The first five Legendre polynomials* (*in the variables* x *and* ϑ) *are*:

$\mathrm{P}_0(x) = 1\,,$

$\mathrm{P}_1(x) = x = \cos\vartheta\,,$

$\mathrm{P}_2(x) = \tfrac{3}{2}x^2 - \tfrac{1}{2} = \tfrac{1}{2}(3\cos^2\vartheta - 1) = \tfrac{1}{4}(3\cos 2\vartheta + 1)\,,$

$$P_3(x) = \tfrac{5}{2}x^3 - \tfrac{3}{2}x = \tfrac{1}{2}(5\cos^3\vartheta - 3\cos\vartheta) = \tfrac{1}{8}(5\cos 3\vartheta + 3\cos\vartheta),$$

$$P_4(x) = \tfrac{35}{8}x^4 - \tfrac{15}{4}x^2 + \tfrac{3}{8} = \tfrac{1}{8}(35\cos^4\vartheta - 30\cos^2\vartheta + 3) =$$

$$= \tfrac{1}{64}(35\cos 4\vartheta + 20\cos 2\vartheta + 9).$$

The graphs and some numerical values of these functions are shown in Figs 16.16, 16.17 and Table 16.5.

Theorem 4 (*Fundamental Properties*).

1. $P_n(-x) = (-1)^n P_n(x)$.
2. $P_n(1) = 1$.

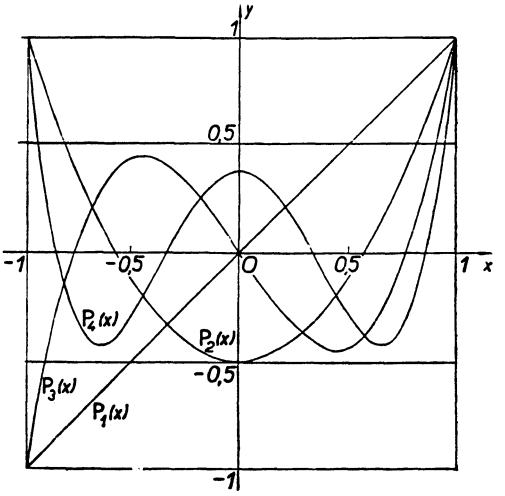


Fig. 16.16.

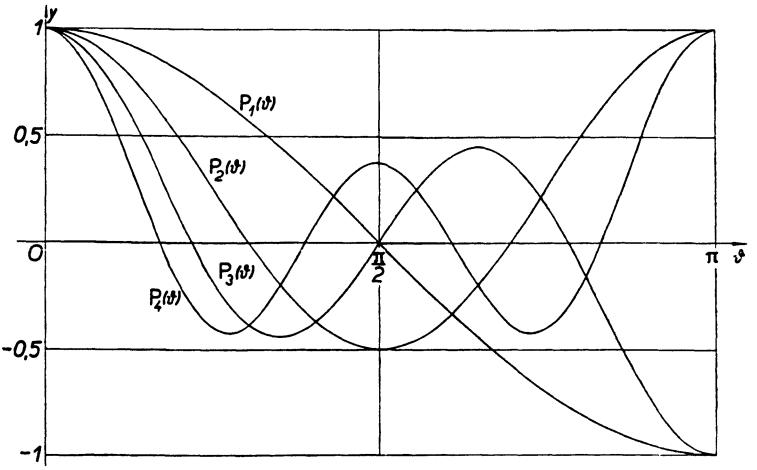


Fig. 16.17.

TABLE 16.5

Legendre polynomials

x	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	ϑ	$P_1(\vartheta)$	$P_2(\vartheta)$	$P_3(\vartheta)$	$P_4(\vartheta)$
0·0	0·000	−0·500	0·000	0·375	0°	1·000	1·000	1·000	1·000
0·1	0·100	−0·485	−0·148	0·338	10°	0·985	0·955	0·911	0·853
0·2	0·200	−0·440	−0·280	0·232	20°	0·940	0·825	0·665	0·475
0·3	0·300	−0·365	−0·383	0·073	30°	0·866	0·625	0·325	0·023
0·4	0·400	−0·260	−0·440	−0·113	40°	0·766	0·380	−0·025	−0·319
0·5	0·500	−0·125	−0·438	−0·289	50°	0·643	0·120	−0·300	−0·428
0·6	0·600	0·040	−0·360	−0·408	60°	0·500	−0·125	−0·438	−0·289
0·7	0·700	0·235	−0·193	−0·412	70°	0·342	−0·325	−0·413	−0·004
0·8	0·800	0·460	0·080	−0·233	80°	0·174	−0·455	−0·247	0·266
0·9	0·900	0·715	0·473	0·208	90°	0·000	−0·500	0·000	0·375
1·0	1·000	1·000	1·000	1·000					

3. *For* $|x| \leqq 1$ *we have* $|P_n(x)| \leqq 1$, $\frac{1}{n^2}|P_n'(x)| \leqq 1$, $\frac{1}{n^4}|P_n''(x)| \leqq 1, \ldots,$

$$\frac{1}{n^{2k}}|P_n^{(k)}(x)| \leqq 1 .$$

4. *The roots of the equation* $P_n(x) = 0$ $(n = 1, 2, \ldots)$ *lie in the interval* $(-1, 1)$.

5. $$P_n(x) = \frac{2n-1}{n} x P_{n-1}(x) - \frac{n-1}{n} P_{n-2}(x) \quad (n \geqq 2) .$$

Theorem 5 (*The Generating Function*). *For* $|x| \leqq 1$, $|t| < 1$ *we have*

$$\frac{1}{\sqrt{(1 - 2xt + t^2)}} = P_0(x) + P_1(x)\, t + P_2(x)\, t^2 + \ldots . \tag{5}$$

By this relationship the Legendre polynomials are uniquely determined.

Theorem 6. *The functions*

$$\varphi_n(x) = \sqrt{(n + \tfrac{1}{2})}\, P_n(x)$$

constitute a complete orthonormal system in $L_2(-1, 1)$ (§ 16.2).

REMARK 2. This means that:

1. $$\int_{-1}^{1} \varphi_i(x)\, \varphi_k(x)\, dx = \begin{cases} 1 \text{ for } i = k , \\ 0 \text{ for } i \neq k . \end{cases}$$

2. The Fourier series (Definition 16.2.6)

$$\sum_{k=0}^{\infty} c_k \varphi_k(x)$$

of any function $f \in L_2(-1, 1)$ converges in the mean to $f(x)$. For pointwise convergence we have:

Theorem 7. *Let $f(x)$ be bounded and integrable in $[-1, 1]$. Let*

$$a_n = \frac{2n+1}{2} \int_{-1}^{1} f(x) \, \mathrm{P}_n(x) \, \mathrm{d}x \quad (n = 0, 1, 2, \ldots) .$$

Then at each point x $(-1 < x < 1)$ interior to an interval in which $f(x)$ is of bounded variation, we have

$$\sum_{n=0}^{\infty} a_n \mathrm{P}_n(x) = \begin{cases} f(x), & \text{provided } f(x) \text{ is continuous at } x, \\ \frac{1}{2}(f(x+0) - f(x-0)) & \text{if } f(x) \text{ is discontinuous at } x \end{cases}$$

($f(x + 0)$ and $f(x - 0)$ denote the limits of the function $f(x)$ from the right and the left, respectively, at the point x).

REMARK 3. Thus, in particular, for every function with a continuous derivative in $[-1, 1]$ we have in $(-1, 1)$,

$$f(x) = \sum_{n=0}^{\infty} a_n \mathrm{P}_n(x) .$$

REMARK 4. The application of Theorem 7 is similar to that of Theorem 16.4.9.

REMARK 5 (*Spherical Harmonics*). Denote*

$$\sqrt{[(1-x^2)^m]} \, \frac{\mathrm{d}^m \mathrm{P}_n(x)}{\mathrm{d}x^m} = \mathrm{P}_n^m(x) . \tag{6}$$

(also the notation $\mathrm{P}_{n,m}(x)$ is usual), where $\mathrm{P}_n(x)$ is the Legendre polynomial of degree n. The function (6) is often called the *associated Legendre function.* For example, if $m = 2$, $n = 2$, we have

$$\mathrm{P}_2^2(x) = (1-x^2) \, \frac{\mathrm{d}^2 \mathrm{P}_2(x)}{\mathrm{d}x^2} = 3(1-x^2) . \tag{7}$$

In what follows, the notation $\mathrm{P}_n(\cos\theta)$, or $\mathrm{P}_n^m(\cos\theta)$ is used, by which the function $\mathrm{P}_n(x)$, or $\mathrm{P}_n^m(x)$ is to be understood, respectively, with x replaced by $\cos\theta$. For example we have, by (7),

$$\mathrm{P}_2^2(\cos\theta) = 3 \sin^2\theta . \tag{8}$$

The functions

$$\mathrm{Y}^m_{n(c)}(\theta,\varphi) = \mathrm{P}^m_n(\cos\theta)\ \cos m\varphi\ . \tag{9}$$

$$\mathrm{Y}^m_{n(s)}(\theta,\varphi) = \mathrm{P}^m_n(\cos\theta)\ \sin m\varphi\ , \tag{10}$$

are called *spherical harmonics.* (Note that this term if often used also for Legendre polynomials themselves, in the literature). For example (see (8)),

$$\mathrm{Y}^2_{2(c)}(\theta,\varphi) = 3\sin^2\theta\ \cos 2\varphi\ . \tag{11}$$

Theorem 8. *The spherical harmonics are orthogonal on the unit spherical surface S, i.e.*

$$\iint_S \mathrm{Y}^m_{n(c)}(\theta,\varphi)\ \mathrm{Y}^{m'}_{n'(c)}(\theta,\varphi)\,\mathrm{d}S =$$

$$= \int_0^\pi \int_0^{2\pi} \mathrm{Y}^m_{n(c)}(\theta,\varphi)\ \mathrm{Y}^{m'}_{n'(c)}(\theta,\varphi)\ \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi = 0$$

for $m \neq m'$, or $n \neq n'$, and similarly

$$\iint_S \mathrm{Y}^m_{n(s)}(\theta,\varphi)\ \mathrm{Y}^{m'}_{n'(s)}(\theta,\varphi)\,\mathrm{d}S = 0\ ,$$

$$\iint_S \mathrm{Y}^m_{n(c)}(\theta,\varphi)\ \mathrm{Y}^{m'}_{n'(s)}(\theta,\varphi)\,\mathrm{d}S = 0$$

for $m \neq m'$, or $n \neq n'$. The corresponding orthonormal system is complete (§16.2).

Theorem 9. *Let $f(\theta,\varphi)$ be square integrable on S, i.e. let $f \in L_2(S)$. Denote*

$$a^m_{n(c)} = \frac{(2n+1)\ (n-m)!}{2\pi(n+m)!} \iint_S f(\theta,\varphi)\ \mathrm{Y}^m_{n(c)}(\theta,\varphi)\,\mathrm{d}S \quad (m \neq 0)\ ,$$

$$a^m_{n(s)} = \frac{(2n+1)\ (n-m)!}{2\pi(n+m)!} \iint_S f(\theta,\varphi)\ \mathrm{Y}^m_{n(s)}(\theta,\varphi)\,\mathrm{d}S \quad (m \neq 0)\ ,$$

$$a^0_{n(c)} = \frac{2n+1}{4\pi} \iint_S f(\theta,\varphi)\ \mathrm{Y}^0_{n(c)}(\theta)\,\mathrm{d}S = \frac{2n+1}{4\pi} \iint_S f(\theta,\varphi)\ \mathrm{P}_n(\theta)\,\mathrm{d}S\ ,$$

$$a^0_{n(s)} = 0\ .$$

Then

$$\begin{aligned} f(\theta,\varphi) &= \sum_{n=0}^{\infty} \sum_{m=0}^{n} [a^m_{n(c)}\ \mathrm{Y}^m_{n(c)}(\theta,\varphi) + a^m_{n(s)}\ \mathrm{Y}^m_{n(s)}(\theta,\varphi)] = \\ &= \sum_{n=0}^{\infty} \sum_{m=0}^{n} [a^m_{n(c)}\ \cos m\varphi + a^m_{n(s)}\ \sin m\varphi]\ \mathrm{P}^m_n(\cos\theta) \end{aligned} \tag{12}$$

in $L_2(S)$ (thus the series (12) converges, on S, to the function f in the mean). *If, moreover,* f *and its partial derivatives of the first and second order are continuous on* S, (12) *converges to* f *uniformly on* S.

Theorem 10. *The functions*

$$U^m_{n(c)}(x,y,z) = r^n Y^m_{n(c)}(\theta,\varphi)\ ,\quad U^m_{n(s)}(x,y,z) = r^n Y^m_{n(s)}(\theta,\varphi)$$

are harmonic in E_3, *i.e. they satisfy the equation*

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} = 0\,.$$

(*Here, the relationship between* x, y, z *and* r, φ, ϑ *is given by the equations* $x = r\sin\vartheta\cos\varphi$, $y = r\sin\vartheta\sin\varphi$, $z = r\cos\vartheta$,

$$r = \sqrt{(x^2 + y^2 + z^2)}\,.)$$

16.6. Some Further Important Functions (Hypergeometric Functions, Jacobi Polynomials, Chebyshev Polynomials, Laguerre Polynomials, Hermite Polynomials)

Legendre polynomials appear as a particular case of the so-called *hypergeometric function*, namely*

$$F(\alpha,\beta,\gamma,x) = 1 + \frac{\alpha\beta}{1\,.\,\gamma}x + \frac{\alpha(\alpha+1)\,\beta(\beta+1)}{1\,.\,2\,.\,\gamma(\gamma+1)}x^2 + \\ + \frac{\alpha(\alpha+1)\,(\alpha+2)\,\beta(\beta+1)\,(\beta+2)}{1\,.\,2\,.\,3\,.\gamma(\gamma+1)\,(\gamma+2)}x^3 + \ldots \qquad (1)$$

(the series converging for $|x| < 1$ and for γ different from any non-positive integer) which is a solution of the so-called *hypergeometric* (*Gauss*) *equation*

$$x(1-x)\,y'' + [\gamma - (\alpha+\beta+1)\,x]\,y' - \alpha\beta y = 0 \qquad (2)$$

(α, β, γ being constants) (see also § 17.21, equation 140 and further).

If $\alpha = 1$, $\beta = \gamma$, the series (1) reduces to a geometric series

$$1 + x + x^2 + \ldots = \frac{1}{1-x}$$

* The slightly different notation $F(\alpha,\beta;\gamma;x)$ is often used.

(a solution of equation (2) for $\alpha = 1$, $\beta = \gamma$).

For $\alpha = -n$, $\beta = n + p$, $\gamma = q$ $(q > 0, p - q > -1)$ we get the so-called *Jacobi polynomials*

$$J_n(p, q, x) = F(-n, n + p, q, x) =$$

$$= 1 + \sum_{k=1}^{n} (-1)^k \binom{n}{k} \frac{(n + p)(n + p + 1) \dots (n + p + k - 1)}{q(q + 1) \dots (q + k - 1)} x^k .$$

(It should be noted that the term *Jacobi polynomial* is applied also to a different polynomial in English literature.)

As their particular case we get Legendre polynomials for $p = 1$, $q = 1$ (setting $\frac{1}{2}(1 - x)$ for x),

$$P_n(x) = J_n\left(1, 1, \frac{1 - x}{2}\right) = F\left(-n, n + 1, 1, \frac{1 - x}{2}\right). \tag{3}$$

Example 1. For $n = 2$ we have from (1) and (3),

$$P_2(x) = F\left(-2, 3, 1, \frac{1 - x}{2}\right) = 1 + \frac{(-2) \cdot 3}{1 \cdot 1} \cdot \frac{1 - x}{2} + \frac{(-2)(-1) \cdot 3 \cdot 4}{1 \cdot 2 \cdot 1 \cdot 2} \left(\frac{1 - x}{2}\right)^2 + 0$$

(since $\alpha + 2 = 0$ in the fourth term); collecting the powers of x we get

$$\tfrac{3}{2}x^2 - \tfrac{1}{2}$$

in agreement with Theorem 16.5.3.

For $p = 0$, $q = \frac{1}{2}$ the Jacobi polynomials (again setting $\frac{1}{2}(1 - x)$ for x) yield the so-called *Chebyshev polynomials*,

$$T_n(x) = \frac{1}{2^{n-1}} J_n(0, \tfrac{1}{2}, \tfrac{1}{2}(1 - x)) .$$

(It should be noted that in English literature the factor $1/2^{n-1}$ is omitted in the definition of $T_n(x)$.)

It can be shown that

$$T_n(x) = \frac{1}{2^{n-1}} \cos (n \arccos x) \quad (|x| \leqq 1)$$

and that $T_n(x)$ constitute an orthogonal system in the interval $[-1, 1]$ for the weight function $1/\sqrt{(1 - x^2)}$, i.e.

$$\int_{-1}^{1} \frac{1}{\sqrt{(1-x^2)}} T_i(x)\, T_k(x)\, dx = 0 \quad \text{for} \quad i \neq k .$$

Moreover, $T_n(x)$ is a solution of the so-called *Chebyshev equation with index n*,

$$(1-x^2)\, y'' - xy' + n^2 y = 0 .$$

Laguerre polynomials are defined by the equation

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

(where the coefficient of x^n is taken to be $(-1)^n$) or by the equation

$$L_n(x) = e^x \frac{d^n}{dx^n}(x^n e^{-x}) .$$

They are solutions of the so-called *Laguerre equation*

$$xy'' + (1-x)\, y' + ny = 0 ,$$

and are orthogonal in the interval $[0, \infty)$ for the weight function e^{-x}, i.e.

$$\int_0^{\infty} e^{-x} L_i(x)\, L_k(x)\, dx = 0 \quad \text{for} \quad i \neq k .$$

Hermite polynomials are defined by the equation

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

where $h = \frac{1}{2}n$ for n even, $h = \frac{1}{2}(n-1)$ for n odd, or by the equation

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2} .$$

They satisfy the so-called *Hermite equation*

$$y'' - 2xy' + 2ny = 0 ,$$

and are orthogonal in the interval $(-\infty, \infty)$ for the weight function e^{-x^2}, i.e.

$$\int_{-\infty}^{\infty} e^{-x^2} H_i(x)\, H_k(x)\, dx = 0 \quad \text{for} \quad i \neq k .$$

16.7. Special Functions and Group Representation

In foregoing paragraphs, the most current special functions have been treated, often called *special functions of mathematical physics.* The whole complex of the above presented theorems, formulae, integral representations, etc., is rather unserveyable and looks a little chaotic. However, it can be shown that a certain systematic theory can be built up on the base of representation of groups. Here, we have no possibility to show the whole theory, even in a brief survey. So we try to outline, at least, its basic idea, and refer the reader to remarkable monographs [7] and [180].

Let us note, first, that many operators of mathematical physics are invariant with respect to transformation of coordinates. For example, the Laplace operator is invariant with respect to translation in E_n, the wave equation with respect to the Lorentz transformation, etc. Also eigenvalues and eigenfunctions of these operators are then invariant. For example, the problem

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \lambda u = 0 \quad \text{in} \quad \Omega = (0,\pi) \times (0,\pi) \,, \tag{1}$$

$$u = 0 \quad \text{on the boundary } S \text{ of the square } \Omega \,, \tag{2}$$

has – as well known – eigenvalues

$$\lambda_{mn} = m^2 + n^2 \,,$$

and corresponding eigenfunctions

$$\sin mx \, \sin ny \quad (m = 1,2,\ldots,\ n = 1,2,\ldots) \,.$$

By the translation

$$x' = x - a \,, \quad y' = y - b$$

the problem (1), (2) is transformed into

$$\frac{\partial^2 u}{\partial x'^2} + \frac{\partial^2 u}{\partial y'^2} + \lambda' u = 0 \quad \text{in} \quad \Omega' \,, \quad u = 0 \quad \text{on} \quad S' \,, \tag{3}$$

on the translated square Ω' with the boundary S', having the same eigenvalues

$$\lambda'_{mn} = m^2 + n^2$$

and the eigenfunctions

$$\sin mx' \, \sin ny' = \sin m(x-a) \, \sin n(y-b) \,.$$

Let us consider a topologic group G with elements g. (On the concept of a group see §1.21; a group is called *topologic (continuous)*, if convergence is defined there so

that basic group operations are continuous: If $g_n \to g$, $h_n \to h$, then $g_n h_n \to gh$ and $g_n^{-1} \to g^{-1}$ $(g \neq 0)$. In the just mentioned applications, we meet usually very simple topologic groups, most often groups of transformations in E_2 and E_3.

Let a linear space L be given (a space of functions, as a rule). By a *representation of the group* G (with respect to the space L) a mapping is understood assigning to every element $g \in G$ an operator $T(g)$ such that the following three requirements are fulfilled:

(i) for every fixed element g of the group G, T is a bounded linear operator on L;

(ii) T is continuous on G: If $g_n \to g$ (in the given sense), then also $T(g_n) \to Tg$ in a certain (before defined) sense;

(iii) $T(g_1 g_2) = T(g_1)\, T(g_2)$.

Example 1. Let k be the circumference of the unit circle in E_2 with centre at the origin, let L be the linear space $C(k)$ of functions $f(\psi)$ continuous on k. Let G be the group of all rotations of k, with individual rotations combined in the usual way: If g_α is a rotation by an angle α, then $g_\alpha g_\beta = g_{\alpha+\beta}$. Then (if convergence is suitably chosen on G), the mapping given by

$$T(g_\alpha)\, f(\psi) = f(\psi + \alpha)$$

is a representation of the group G. Because for every fixed g, the operator T is a bounded linear operator on $C(k)$, continuous if convergence is suitably defined on G, and

$$T(g_\alpha g_\beta) = T(g_{\alpha+\beta}) = T(g_\alpha)\, T(g_\beta)\ ,$$

since for every function $f(\psi)$ we have

$$T(g_{\alpha+\beta})\, f(\psi) = f(\psi + \alpha + \beta) =$$

$$= T(g_\alpha)\, f(\psi + \beta) = T(g_\alpha)\, T(g_\beta)\, f(\psi)\ .$$

Now, it can be shown that for simple groups of transformations, or for their subgroups (groups of translations, or rotations) and for a suitable choice of the space L (for example of the space of eigenfunctions of an operator which is invariant with respect to the group considered) and of a base in this space, the operator T can be expressed in a simple matrix form, where elements of such a matrix are just special functions of mathematical physics – up to a certain coefficient, or an exponential function, may be. Which of these functions they are, this depends on the choice of the group G, or its subgroup G' and on the choice of the space L and a base in it.

Example 2. Let us investigate the group G of motions (of transformations) in E_2, given by the equations

$$x' = x\cos\alpha - y\sin\alpha + a\ ,$$

$$y' = x\sin\alpha + y\cos\alpha + b\ .$$

Let the elements of this group be denoted by $g(a,b,\alpha)$. Because the point $[a,b]$ can be expressed in polar coordinates in the form

$$a = \rho\cos\varphi\ ,\quad b = \rho\sin\varphi\ ,\quad \rho \geqq 0\ ,\quad 0 \leqq \varphi < 2\pi\ ,$$

the notation $g(\rho,\varphi,\alpha)$ is also used.

Let us consider, in E_2 the circumference k of the unit circle with its center in the origin (as above) and the linear space L of functions $f(\psi)$ (complex, in general), square integrable on the interval $[0,2\pi)$. It can be shown that the mapping $g \to T(g)$, where the operator T is defined by,

$$T(g)\ f(\psi) = \mathrm{e}^{\mathrm{i}\rho\cos(\psi-\varphi)}\ f(\psi-\alpha)\ , \tag{4}$$

is a representation of the group G with respect to the space L.

In particular, for the subgroup G' of all translations in the direction of the x-axis we obtain the representation (we have then $\varphi = 0$, $\alpha = 0$ in (4), so that we can write $g(\rho)$ only instead of $g(\rho,0,0)$)

$$T(g(\rho))\ f(\psi) = \mathrm{e}^{\mathrm{i}\rho\cos\psi}\ f(\psi)\ . \tag{5}$$

On the space L, let a scalar product be defined by

$$(f_1,f_2) = \frac{1}{2\pi}\int_0^{2\pi} f_1(\psi)\,\overline{f_2(\psi)}\,\mathrm{d}\psi\ , \tag{6}$$

which thus differs only by the coefficient $1/(2\pi)$ from the scalar product in the space $L_2(0,2\pi)$. In the so obtained Hilbert space – let the notation L be preserved for this space – let us choose the base

$$\{\mathrm{e}^{\mathrm{i}n\psi}\}\ ,\quad n \text{ integral}\ ,\quad -\infty < n < \infty\ .$$

The elements of the matrix by which the operator (5) can then be expressed, will be of the form

$$a_{mn} = (T(g(\rho))\ \mathrm{e}^{\mathrm{i}m\psi},\ \mathrm{e}^{\mathrm{i}n\psi})\ ,$$

or

$$a_{mn} = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{e}^{\mathrm{i}\rho\cos\psi}\,\mathrm{e}^{\mathrm{i}n\psi}\,\overline{\mathrm{e}^{\mathrm{i}m\psi}}\,\mathrm{d}\psi =$$

$$= \frac{1}{2\pi}\int_0^{2\pi} \mathrm{e}^{\mathrm{i}[\rho\cos\psi+(n-m)\psi]}\,\mathrm{d}\psi\ . \tag{7}$$

If we denote, as usual, by the symbol $J_n(x)$ the Bessel function of the first kind and integral index n and take its integral representation into account, then (7) yields – after the substitution $\psi = \pi/2 - \theta$ and some small rearrangements –

$$a_{mn} = i^{n-m} J_{n-m}(\rho) . \tag{8}$$

Thus the elements of the matrix, corresponding to the considered representation of the group of translations in the direction of the x-axis, differ only by the coefficient i^{n-m} from the Bessel functions of index $n - m$ and argument ρ.

The aim of Example 2 was only to outline, to the reader, in a brief way, fundamental ideas of the theory. A lot of details was omitted here, that play an important role when an exact theory should be built up.

The connection between group representation and special functions made it possible to find a relatively simple way – uniform and general enough, at the same time – how to derive properties of these functions. For example, a very simple operation

$$Qf(\psi) = f(-\psi)$$

in Example 2 yields almost immediately the well-known property

$$J_n(x) = (-1)^n J_{-n}(x)$$

of Bessel functions with an integral index. Further operations, which are similar for all current special functions, in essential, lead to well-known recursive formulae, etc. Of course, applications of group representation discover also new (unknown) relations for known special functions, or lead to new special functions. The whole theory is extensive, with numerous applications. The reader is referred to the already quoted monographs [7], [180].

17. ORDINARY DIFFERENTIAL EQUATIONS

By KAREL REKTORYS

References: [2], [12], [17], [33], [36], [44], [45], [53], [61], [69], [77], [87], [88], [89], [99], [112], [129], [133], [144], [155], [168], [184], [189], [191], [211], [250], [254], [264], [277], [282], [286], [289], [290], [318], [327], [341], [370], [375], [388], [401], [403], [409], [430], [432], [444], [450], [465], [470], [471], [476], [483], [485], [489], [490], [494], [495].

INTRODUCTORY REMARK. The problematics of ordinary differential equations is very broad. Therefore, it was not possible to cover the whole field, in our Survey. For example, the reader will not find topics here, concerning theory of transformations of differential equations, optimal control, abstract differential equations, etc. The book being written for a very wide circle of readers, we tried, as far as possible, to work with classical concepts and methods (we do not consider solutions in the Kurzweil sense, for example). Systems of equations are not treated in the vector form exclusively, because for many readers the classical treating in components is more usual. For similar reasons it was not possible, when considering nonlinear equations, to avoid the concepts of general, or singular integrals which are quite current in technical literature. At the same time, we are well aware of the fact that these concepts are rather vague and thus not often used in modern mathematical literature. The encyclopedical character of the book made in necessary to include, in §§ 17.5 and 17.7, methods for solving equations of some special types. However, the reader's attention should be drawn to the fact that numerical methods, discussed in Chap. 25, represent often a more effective tool of their solution.

A particular attention has been paid to *boundary value problems* in ordinary differential equations and to *eigenvalue problems* (§ 17.17). In § 17.21 a table of solved differential equations has been given. The main source for it is the book [250]. Let us note here that in that remarkable book a lot of classical methods and results can be found, other than those introduced here, as well as a lot of useful references.

Unless the contrary is stated, all functions and numbers considered in this chapter are assumed to be real.

17.1. Classification of Differential Equations. Ordinary and Partial Differential Equations. Order of a Differential Equation. Systems of Differential Equations

A differential equation is a relation (given in a certain domain) between the unknown function and its derivatives. If the function to be found is a function of one variable only, we speak of an *ordinary differential equation*, if it is a function of several variables (so that the equation contains partial derivatives), we speak of a *partial differential equation.*

If m differential equations are given for n unknown functions (not necessarily with $m = n$), we speak of a *system of differential equations.*

The *order* of a given differential equation is that of the derivative of the highest order occurring in it. Similarly by the *order of a system* we understand (as usual) that of the derivative of the highest order occurring in the system.

Example 1.

$$y'' + xy'^3 - e^y = 0$$

is an ordinary differential equation of the second order.

Example 2.

$$\frac{\partial^2 u}{\partial x \partial y} - \frac{\partial^3 u}{\partial y^3} - u = 0$$

is a partial differential equation of the third order for the unknown function $u(x, y)$.

Example 3.

$$\frac{\partial u_1}{\partial x} - \frac{\partial^2 u_2}{\partial y^2} = 0, \qquad \frac{\partial^3 u_1}{\partial y^3} - \frac{\partial u_2}{\partial x} = 0$$

is a system of partial differential equations of the third order for two unknown functions $u_1(x, y)$, $u_2(x, y)$.

17.2. Basic Concepts. Solution (Integral) of a Differential Equation. Theorems Relating to Existence and Uniqueness of Solution. General Integral, Particular Integral, Singular Integral

Definition 1. By an *ordinary differential equation of the first order*, we mean an equation of the form

$$F(x, y, y') = 0 \tag{1}$$

or, in the special case where the equation is solved with respect to y', an equation of the form

$$y' = f(x, y). \tag{2}$$

Definition 2. By an *ordinary differential equation of order n* we mean an equation of the form

$$F\left(x, y, y', y'', \ldots, y^{(n)}\right) = 0 \tag{3}$$

or, if the equation is solved with respect to the derivative of the highest order, the equation of the form

$$y^{(n)} = f\left(x, y, y', y'', \ldots, y^{(n-1)}\right). \tag{4}$$

Definition 3. By a *solution* or *integral* (or *particular integral* or *integral curve*, if speaking geometrically) *of equation* (3), we mean any function $y = g(x)$ which has derivatives up to the n-th order and satisfies equation (3) identically in the domain considered. (See also Remark 3.)

REMARK 1. Most frequently, the domain considered is an interval I. Thus the function $y = g(x)$ is the solution of equation (3) in the interval I, if it has derivatives up to the order n in I (the right-hand, or left-hand derivatives at the point a, or b, respectively, if the closed interval $[a, b]$ is in question) and if on substituting $g(x)$ for y, $g'(x)$ for y', etc., in (3), equation (3) is satisfied for all x in I.

Example 1. The function $y = \sin x$ is a (particular) integral of the equation $y'' + y = 0$ in the whole interval $(-\infty, \infty)$.

REMARK 2. In Definition 3, definitions of the solution of equations (1), (2) and (4) as special cases are included.

REMARK 3. The solution of equation (3) need not be given in the explicit form $y = g(x)$, but possibly in an implicit form by an equation $h(x, y) = 0$. Then the derivatives $y', y'', \ldots$ may be found according to the theorem on implicit functions (Theorem 12.9.1, in a region in which the corresponding conditions are fulfilled, of course); $h(x, y) = 0$ is then called a solution of equation (3), if on substituting for y', $y'', \ldots$ in (3), the differential equation is identically satisfied (in the variables x and y) at all the points of the curve $h(x, y) = 0$ (see also Remark 16 and Example 7).

Example 2. The equation $x^2 + y^2 = 4$ constitutes an integral of the equation

$$y' = -\frac{x}{y} \tag{5}$$

in an implicit form, since at every point of the circle $x^2 + y^2 - 4 = 0$, the relation $2x + 2yy' = 0$, holds, i.e. (5) is valid. The points $(-2, 0)$ and $(2, 0)$, where $y = 0$, are exceptions; for this case see Remark 16 and Example 7.

REMARK 4. Geometrical interpretation of equation (2): Let $f(x, y)$ be defined in a region Ω. Then by virtue of equation (2), to every point $(x, y) \in \Omega$ there

corresponds a line element (*directional element*) determined by the point (x, y) and by the slope y'. Thus in Ω a field of directions is given, the so-called *directional field*. The problem of finding solutions of equation (2) can thus be given the following geometrical interpretation: to find curves in Ω, the tangents to which coincide at every point with the corresponding line element (i.e. the slope y' of such a curve at the point considered is prescribed by (2)).

In the case of an equation of the second order

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

there is assigned at every point $(x, y) \in \Omega$ and for each slope y' the value of the second derivative (and thus the curvature of the integral curve which is to be found). Equations of higher orders have no longer such a simple geometrical interpretation.

Definition 4. By a *solution (integral) of a system of equations of the first order*,

$$\begin{aligned} y_1' &= f_1(x, y_1, y_2, \ldots, y_n), \\ y_2' &= f_2(x, y_1, y_2, \ldots, y_n), \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n' &= f_n(x, y_1, y_2, \ldots, y_n) \end{aligned} \tag{6}$$

we mean a system of functions

$$y_1 = g_1(x), y_2 = g_2(x), \ldots, y_n = g_n(x) \tag{7}$$

such that if the functions (7) and their derivatives are substituted into (6), all the equations (6) are identically satisfied (in the domain considered). (On more general systems of equations see also § 17.18.)

In this case it is also possible to speak about a solution in implicit form (see Remark 3).

REMARK 5. In the case where $n = 2$, let us denote the functions to be found by y, z instead of y_1, y_2. Then the functions (7)

$$y = g_1(x), \qquad z = g_2(x)$$

represent geometrically a *curve* (in three-dimensional space). For this reason we often call the system of functions (7) an *integral curve* of the system (6).

System (6) is often written in a vector form: If we use the notation

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}, \quad \boldsymbol{y}' = \begin{bmatrix} y_1' \\ y_2' \\ \cdots \\ y_n' \end{bmatrix}, \quad \boldsymbol{f}(x, \boldsymbol{y}) = \begin{bmatrix} f_1(x, y_1, y_2, \ldots, y_n) \\ f_2(x, y_1, y_2, \ldots, y_n) \\ \cdots\cdots\cdots\cdots\cdots \\ f_n(x, y_1, y_2, \ldots, y_n) \end{bmatrix},$$

we can write (6) in the form

$$\boldsymbol{y}' = f(x, \boldsymbol{y}) \tag{6'}$$

and its solution (7) as

$$\boldsymbol{y} = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \dots \\ g_n(x) \end{bmatrix}. \tag{7'}$$

This simple surveyable vector form is current in modern mathematical literature. Also Theorem 1 below can be easily "translated" into this "vector language".

REMARK 6. In general, the number of equations and the number of unknown functions in the system (6) need not be equal. The definition of the solution remains unchanged. In this case, however, the theorem on existence and uniqueness of the solution (see below) need no longer be valid, in general.

Theorem 1. *Let the system* (6) *and a point*

$$P(a, b_1, b_2, \dots, b_n) \tag{8}$$

be given. Let the functions $f_1, f_2, \dots, f_n$ *of the system* (6) *be continuous* (*as functions of the* $n + 1$ *variables* $x, y_1, y_2, \dots, y_n$) *in a neighbourhood* U *of the point* P *and let them have, in* U, *continuous partial derivatives with respect to the variables* $y_1, y_2, \dots, y_n$. *Then in a certain neighbourhood of the point* a *there exists exactly one system of functions* (7), *which is a solution of the system* (6) *and satisfies the conditions* (the so-called *initial conditions*)

$$g_1(a) = b_1, g_2(a) = b_2, \dots, g_n(a) = b_n. \tag{9}$$

REMARK 7. In the terminology of Remark 5: *If the functions* $f_1, f_2, \dots, f_n$ *satisfy the above conditions in* U, *then, in a certain neighbourhood of the point* a (or *locally*, in brief), *there exists precisely one integral curve of the system* (6) *passing through the point* P.

In particular, if the functions

$$f(x, y), \qquad \frac{\partial f}{\partial y}(x, y) \tag{10}$$

are continuous in a neighbourhood U of the point $P(a, b)$, then there exists, locally, just one integral curve of the equation

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

passing through the point P.

Using the Lipschitz condition (see below), a more sharp theorem, similar to Theorem 2, can be formulated for the system (6).

REMARK 8. We say that the function $f(x, y)$ satisfies a *Lipschitz condition with respect to the variable* y in the rectangle $\overline{R}(A \leqq x \leqq B,\ C \leqq y \leqq D)$, if a constant K can be found such that for every $x \in [A, B]$ and for any numbers y_1, y_2 from $[C, D]$ the inequality

$$|f(x, y_1) - f(x, y_2)| \leqq K\,|y_1 - y_2| \tag{11}$$

holds. (In particular, condition (11) is satisfied, if $f(x, y)$ has a derivative with respect to y in $\overline{R}$ and

$$\left|\frac{\partial f}{\partial y}\right| \leqq K \tag{12}$$

holds in $\overline{R}$.)

Example 3. The function $f(x, y) = |y|$ does not possess partial derivative $\partial f/\partial y$ at points of the x-axis, but it nevertheless satisfies the Lipschitz condition in the whole plane xy. It is sufficient to choose $K = 1$, since

$$\big||y_1| - |y_2|\big| \leqq |y_1 - y_2|.$$

Theorem 2. *Let* $f(x, y)$ *be continuous in the rectangle*

$$\overline{Q}\,(a - h \leqq x \leqq a + h,\ b - k \leqq y \leqq b + k)\,.$$

(*Then* (see Theorem 12.1.5) *there exists a constant* $M > 0$ *such that* $|f(x, y)| \leqq$ $\leqq M$ *in* $\overline{Q}$.) *Let* $f(x, y)$ *satisfy condition* (11) *in* $\overline{Q}$. *Let us write*

$$d = \min\left(h, \frac{k}{M}\right).$$

Then, in the interval $[a - d, a + d]$, *there exists precisely one solution* $y = g(x)$ *of the equation*

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

passing through the point (a, b).

REMARK 9 (*extension of solution, maximal solution*). The just discussed solution $y = g(x)$ can often be extended onto a larger interval than is the interval $[a-d, a+d]$ specified in Theorem 2. If we start at the point (a, b) and proceed "to the right" along the integral curve $y = g(x)$, we come, in the sense of the theorem, at the point $(a + d, g(a + d))$. Let us denote that point briefly by (a_1, b_1) (Fig. 17.1).

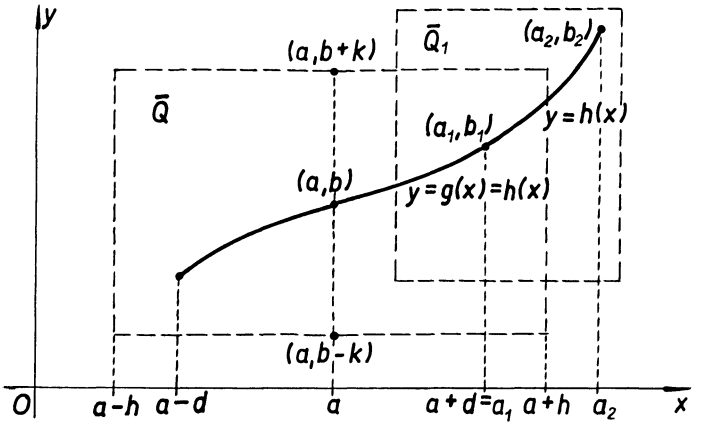


Fig. 17.1.

If the right-hand side $f(x, y)$ of the equation $y' = f(x, y)$ satisfies, in a certain neighbourhood of the point (a_1, b_1), conditions similar to those of Theorem 2, then again a unique solution $y = h(x)$ of the given equation, passing through the point (a_1, b_1), will exist in a sufficiently small neighbourhood $[a_1 - d_1, a_1 + d_1]$ of the point a_1. Because of uniqueness, this solution will be identical with the solution $y = g(x)$ in the left-hand neighbourhood of the point a_1. Now, if we start at the point (a_1, b_1) and proceed to the right along the integral curve $y = h(x)$, we come to another point — denote it by (a_2, b_2). Let us define a new function $y = \overline{g}(x)$ in the interval $[a - d, a_2]$ by

$$\overline{g}(x) = \begin{cases} g(x) & \text{for} \quad x \in [a - d, a_1], \\ h(x) & \text{for} \quad x \in [a_1, a_2]. \end{cases}$$

The function $y = \overline{g}(x)$ which is the (unique) solution of the equation $y' = f(x, y)$, passing through the point (a, b) and is defined in the whole interval $[a - d, a_2]$, is called the *extension* (*continuation*) of the original solution (from the interval $[a - d, a + d]$ onto the interval $[a - d, a_2]$). In this way, we can go on further. Similarly, we can extend the solution $y = g(x)$, defined originally on the interval $[a - d, a + d]$, "to the left".

A solution of the given equation (passing through the given point) is called *maximal* if it is no longer possible to extend it to the right as well as to the left.

Let us draw the reader's attention to the fact that even if the function $f(x, y)$ has derivatives of all orders in the whole plane xy, the maximal solution of the equation $y' = f(x, y)$ passing through the point (a, b) need not be defined for *all* x. For example, the (only) solution of the equation

$$y' = y^2,$$

passing through the point (0, 1), is

$$y = \frac{1}{1-x}.$$

This solution is defined in the interval $(-\infty, 1)$ and cannot be extended "to the right" over the point $x = 1$, obviously.

REMARK 10. The Lipschitz condition and a theorem similar to Theorem 2 can be formulated for the system (6) as well.

To prove existence of a solution of the equation $y' = f(x, y)$, the only assumption of continuity of the function $f(x, y)$ is sufficient. However, continuity itself, without further assumptions (boundedness of the derivative with respect to y, the Lipschitz condition, etc.) is not sufficient to guarantee uniqueness:

Example 4. The functions

$$y \equiv 0 \qquad \text{and} \qquad y = \tfrac{1}{27}(x-a)^3$$

are two different solutions of the equation

$$y' = \sqrt[3]{y^2}$$

passing through the point $(a, 0)$.

A similar remark holds for solutions of the system (6).

Theorem 3. *Let the equation*

$$y^{(n)} = f(x, y, y', \ldots, y^{(n-1)}) \tag{13}$$

of the n-th order $(n \geqq 1)$ *and the point* $P(a, b_1, b_2, \ldots, b_n)$ *be given. Let the functions*

$$f, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial y'}, \ldots, \frac{\partial f}{\partial y^{(n-1)}}$$

be continuous (as functions of the $n+1$ *variables) in a neighbourhood of the point* P. *Then, in a certain neighbourhood of the point* a, *there exists precisely one solution* $y = g(x)$ *of equation* (13), *which satisfies the initial conditions*

$$g(a) = b_1, g'(a) = b_2, \ldots, g^{(n-1)}(a) = b_n \tag{14}$$

(i.e. through the point $T(a, b_1)$ *there passes one and only one integral curve of equation* (13) *satisfying the conditions*

$$g'(a) = b_2, \ldots, g^{(n-1)}(a) = b_n).$$

REMARK 11. By employing a Lipschitz condition, it is possible to formulate a theorem relating to equation (13) similar to Theorem 2.

Theorem 3 has a *local* character, i.e. it guarantees existence and uniqueness of the solution in a certain neighbourhood of the point a. In the case of a *linear* equation

$$y^{(n)} + a_{n-1}(x)y^{(n-1)} + \ldots + a_1(x)y' + a_0(x)y = b(x) \tag{15}$$

existence and uniqueness of the solution (satisfying the prescribed initial conditions (14)) is guaranteed in the entire interval I in which the functions $a_{n-1}(x)$, ..., $a_1(x)$, $a_0(x)$, $b(x)$ are continuous and which contains the point a. The interval I may even be infinite. A similar statement holds for the so-called normal systems of linear equations (Theorem 17.18.1).

REMARK 12. For equation (13) let the initial conditions (14) be prescribed. We say briefly that an $(n+1)$-*dimensional point* $P(a, b_1, b_2, \ldots, b_n)$ (cf. Theorem 3) *is given.* Let Q be an $(n+1)$-dimensional region constituted of such points P, for which equation (13) has precisely one solution in the sense of Theorem 3. We define:

Definition 5. By the *general integral* (or *general solution*, or *general form of solution*) of equation (13) in the region Q we mean a function $y = g(x, C_1, C_2, \ldots, C_n)$ which depend, in addition to x, on n parametrs C_1, C_2, ..., C_n and which is, as the function of x only, a solution of equation (13) for arbitrary values of these parameters. At the same time, the parameters C_1, C_2, ..., C_n are independent in the following sense: If we choose an arbitrary point $P \in Q$, there is one and only one set of numerical values of these parameters for which the solution satisfies the conditions (14).

REMARK 13. Roughly speaking, these parameters are independent if it is not possible to replace any of them by others, i.e. if none of them is "superfluous".

Example 5. The function

$$y = C_1 \mathrm{e}^{2x} + C_2 \mathrm{e}^{-x}$$

is the general integral of the equation

$$y'' - y' - 2y = 0,$$

even for the case that Q is the whole three-dimensional space. Because, first, choosing C_1, C_2 arbitrarily, this function is a solution of the given equation. And, secondly, the point $P(a, b_1, b_2)$ being chosen arbitrarily, C_1 and C_2 are uniquely determined by

$$\begin{aligned} y(a) &\equiv C_1 \mathrm{e}^{2a} + C_2 \mathrm{e}^{-a} = b_1, \\ y'(a) &\equiv 2C_1 \mathrm{e}^{2a} - C_2 \mathrm{e}^{-a} = b_2. \end{aligned}$$

The function

$$C_1 \, e^{x+C_2}$$

is *not* a general integral of the equation

$$y'' - y = 0$$

since, evidently,

$$C_1 \, e^{x+C_2} = C_1 \, e^x \, . \, e^{C_2} = K \, e^x \, .$$

REMARK 14. The concept of a general integral has its historical origin in the theory of *linear* homogeneous equations, i.e. of equations of the form

$$y^{(n)} + a_{n-1}(x)y^{(n-1)} + \ldots + a_0(x)y = 0.$$

As well known (see § 17.11), if the functions $a_i(x)$ $(i = 0, 1, \ldots, n-1)$ are continuous in an interval I and if $y_1(x), \ldots, y_n(x)$ are linearly independent solutions of the given equation in I, then *every* solution of this equation in that interval is of the form

$$C_1 y_1(x) + \ldots + C_n y_n(x).$$

Thus *every* solution is contained in this general form for a proper choice of the coefficients $C_1, \ldots, C_n$.

In the case of *nonlinear* equations, it is not the case, in general. For example, in the "upper half-plane" $y > 0$, as well as in the "lower half-plane" $y < 0$, the function

$$y = \tfrac{1}{27} (x - C)^3$$

is the general integral of the equation

$$y' = \sqrt[3]{y^2} \, ,$$

by Definition 5. However, it does *not* contain all solutions of this equation: For example, the solution $y \equiv 0$ is not contained in it for any value of the constant C. Thus in the case of nonlinear equations, the concept of the general integral is not "fitting". Therefore (but also for other reasons), this concept is not often used in modern mathematical literature. However, in technical literature it is so customary that it was not possible to omit it here.

In the case of the equation

$$F(x, y, y', \ldots, y^{(n)}) = 0 \tag{16}$$

it is not possible to speak about the general integral exactly in the sense of Remark 12 and Definition 5, since uniqueness of the solution might be in question: For example, given the point (x_0, y_0), the equation

$$F(x, y, y') = 0 \tag{17}$$

can be satisfied by several values of y'_0, in general, even if F is a very "reasonable" function. For example, the equation

$$y'^2 - x^4y^2 = 0 \tag{18}$$

is satisfied by both y' given by

$$y' = x^2y, \qquad y' = -x^2y.$$

Evidently, through every point (x_0, y_0), for which $x_0^2y_0 \neq 0$, there pass two integral curves of equation (18). Often we succeed in finding a function or relation

$$y = g(x, C_1, C_2, \ldots, C_n) \quad \text{or} \quad h(x, y, C_1, C_2, \ldots, C_n) = 0 \tag{19}$$

such that by a proper choice of the parameters C_1, C_2, ..., C_n (dependent, for example, on the possible choice of the derivative of the highest order in (16) as was the case in (18)), it is possible to satisfy — in a certain region — any initial condition. In such a case, we can also speak of a general integral of equation (16).

Definition 6. By a *singular integral* (*singular solution*) of equation $y' = f(x, y)$ we mean an integral curve of this equation such that at every its point uniqueness is broken (i.e. through every point of this curve there also passes another integral curve of the equation).

Example 6. The integral curve $y \equiv 0$ is a singular integral of the equation

$$y' = \sqrt[3]{y^2}$$

(see Example 4), since through every point $(a, 0)$ of this curve there passes another integral curve

$$y = \tfrac{1}{27}(x - a)^3$$

of the given equation.

REMARK 15. The general integral of the equation $y' = f(x, y)$ constitutes a one-parameter system of curves. It easily follows that the envelope of this system (if it exists) is a singular integral of the given equation.

In the case of the equation $F(x, y, y') = 0$, we define the singular integral in the same way as in Definition 6, with the supplementary condition (see the question of uniqueness of the solution mentioned in Remark 14) that through every point of the singular integral curve there passes also another integral curve *with the same tangent.*

REMARK 16. For $f(x, y) \neq 0$, the equation

$$\frac{dy}{dx} = \frac{1}{f(x, y)} \tag{20}$$

is equivalent to the equation

$$\frac{dx}{dy} = f(x, y). \tag{21}$$

At points where $f(x, y) = 0$, equation (20) has no meaning, but equation (21) has. Consequently, we often "add" equation (21) to equation (20) and then understand by an integral curve of equation (20) an integral curve satisfying either equation (20) or equation (21).

Example 7. In this sense, the circle

$$x^2 + y^2 - 4 = 0 \tag{22}$$

is an integral curve of the equation

$$y' = -\frac{x}{y}$$

(Example 2) even at the points $(-2, 0)$, $(2, 0)$, since in the neighbourhood of these points it satisfies the equation

$$\frac{dx}{dy} = -\frac{y}{x}.$$

17.3. Elementary Methods of Integration of Equations of the First Order. Separation of Variables. Homogeneous Equations. Linear Equations. Bernoulli's Equation. Riccati's Equation

The subsections IV, V of this section are of little significance. They have been introduced here for completeness only.

I. The equation

$$y' = f(x)$$

(where the function $f(x)$ is supposed to be continuous in the interval I considered) has the general integral

$$y = \int f(x)\,dx.$$

(The indefinite integral already contains an arbitrary constant.) The integral, for which $y(x_0) = y_0$ is then

$$y = y_0 + \int_{x_0}^{x} f(t)\,dt.$$

II. The equation

$$y' = f(y)$$

(where $f(y)$ is continuous in the region considered) can be written, for $f(y) \neq 0$, in the form (Remark 17.2.16)

$$\frac{dx}{dy} = \frac{1}{f(y)}.$$

The general integral (in the region where $f(y) \neq 0$) is

$$x = \int \frac{dy}{f(y)},$$

and the integral curve passing through the point (x_0, y_0) is of the form

$$x - x_0 = \int_{y_0}^{y} \frac{dt}{f(t)}.$$

If $f(y_0) = 0$, then the curve

$$y \equiv y_0$$

satisfies the given equation and the initial condition.

Example 1. Let us consider the equation

$$y' = y^2$$

or (for $y \neq 0$)

$$\frac{dx}{dy} = \frac{1}{y^2}.$$

The general integral is

$$x = -\frac{1}{y} + C$$

or

$$y = \frac{1}{C - x}.$$

The integral curve passing through the point $(0, 1)$ is the curve

$$y = \frac{1}{1 - x}$$

(because, putting $x = 0$, $y = 1$, it follows that $C = 1$). The integral curve passing through the point $(3, 0)$ is

$$y \equiv 0$$

(for in this case $f(y_0) = y_0^2 = 0$). In both cases there is only one solution, in accordance with Theorem 17.2.1 and Remark 17.2.7.

III (*Separation of Variables*).

Theorem 1. *Let* $f(x)$ *be continuous in* $[a, b]$, $g(y)$ *continuous in* $[c, d]$ *and let* $g(y) \neq 0$ *in* $[c, d]$. *Then the equation*

$$y' = \frac{f(x)}{g(y)}$$

has in the rectangle $R(a < x < b,\ c < y < d)$ *the general integral*

$$\int f(x)\,\mathrm{d}x = \int g(y)\,\mathrm{d}y.$$

The integral curve passing through the point $(x_0, y_0) \in R$ *has the equation*

$$\int_{x_0}^{x} f(x)\,\mathrm{d}x = \int_{y_0}^{y} g(y)\,\mathrm{d}y$$

(*and it is, in a neighbourhood of the point* (x_0, y_0), *the only integral curve of the given equation which passes through this point*).

REMARK 1. An analogous theorem holds for the equation

$$y' = f(x)g(y).$$

For $g(y_0) \neq 0$, the general integral is

$$\int \frac{\mathrm{d}y}{g(y)} = \int f(x)\,\mathrm{d}x.$$

The integral curve passing through the point (x_0, y_0) is

$$\int_{y_0}^{y} \frac{\mathrm{d}y}{g(y)} = \int_{x_0}^{x} f(x)\,\mathrm{d}x.$$

If $g(y_0) = 0$, then the equation considered and the initial condition are satisfied by the function

$$y \equiv y_0.$$

Example 2. Let us solve the equation

$$y' = xy^3 \sin x$$

subject to the condition that $y(0) = 1$.

By separation of variables we get

$$\int \frac{\mathrm{d}y}{y^3} = \int x \sin x \, \mathrm{d}x.$$

The general integral is then

$$\frac{1}{2y^2} = x \cos x - \sin x + C.$$

From the condition $y(0) = 1$ it follows that $C = \frac{1}{2}$, so the solution is

$$y = \frac{1}{\sqrt{(2x \cos x - 2 \sin x + 1)}}.$$

(The positive sign before the square root follows from the condition $y(0) = +1$.)

The solution of the same equation, but with the condition $y(0) = 0$, is $y \equiv 0$.

IV. An equation of the form

$$y' = f\left(\frac{y}{x}\right) \qquad (x \neq 0) \tag{1}$$

is often called *homogeneous*. (The right-hand side of equation (1) is a homogeneous function of degree zero (Definition 12.6.1).)

An example of such an equation is the equation

$$y' = \frac{x^2 + y^2}{xy} \qquad \text{or} \qquad y' = \frac{1 + \left(\frac{y}{x}\right)^2}{\frac{y}{x}}.$$

Instead of looking for the solution $y(x)$ of equation (1), we try to find a new unknown function $z(x)$, related to $y(x)$ by the equation

$$z(x) = \frac{y(x)}{x} \qquad \text{or} \qquad y(x) = xz(x).$$

From the second equation it follows that $y' = z + xz'$, so that after substituting into (1) we find that z satisfies the equation

$$z + xz' = f(z) \qquad \text{or} \qquad z' = \frac{f(z) - z}{x}$$

which is an equation with separable variables (see III). If we find its solution $z(x)$, then the solution of equation (1) is $y(x) = xz(x)$ (see Example 3 below).

V. Equations of the form

$$y' = \frac{a_1 x + b_1 y + c_1}{a_2 x + b_2 y + c_2}, \tag{2}$$

where

$$\begin{vmatrix} a_1, & b_1 \\ a_2, & b_2 \end{vmatrix} \neq 0, \tag{3}$$

can be transformed into homogeneous equations by making a substitution of the form

$$x = u + A, \quad y = v + B, \quad \text{so that} \quad \mathrm{d}x = \mathrm{d}u, \quad \mathrm{d}y = \mathrm{d}v, \quad \mathrm{d}y/\mathrm{d}x = \mathrm{d}v/\mathrm{d}u.$$

Equation (2) becomes

$$\frac{\mathrm{d}v}{\mathrm{d}u} = \frac{a_1 u + b_1 v + a_1 A + b_1 B + c_1}{a_2 u + b_2 v + a_2 A + b_2 B + c_2}. \tag{4}$$

Now we choose the constants A, B such that

$$\begin{aligned} a_1 A + b_1 B + c_1 &= 0, \\ a_2 A + b_2 B + c_2 &= 0, \end{aligned} \tag{5}$$

which is possible by virtue of (3). Then equation (4) becomes

$$\frac{\mathrm{d}v}{\mathrm{d}u} = \frac{a_1 u + b_1 v}{a_2 u + b_2 v} = \frac{a_1 + b_1 \dfrac{v}{u}}{a_2 + b_2 \dfrac{v}{u}} \tag{6}$$

and this is a homogeneous equation. Such an equation can be solved by the substitution $v(u) = uz(u)$. After solving this equation we substitute back $z = v/u$ and $v = y - B$, $u = x - A$.

If the determinant (3) is zero, i.e. if $a_2 x + b_2 y = k(a_1 x + b_1 y)$, equation (2) can be easily solved by introducing a new unknown function $z(x)$ given by

$$a_1 x + b_1 y = z, \quad \text{so that} \quad a_2 x + b_2 y = kz,$$

and

$$a_1 + b_1 y' = z', \quad \text{hence} \quad y' = \frac{z' - a_1}{b_1}.$$

We thus obtain an equation for z of the form

$$\frac{z'}{b_1} = \frac{a_1}{b_1} + \frac{z + c_1}{kz + c_2} \quad \text{or} \quad z' = f(z)$$

whose solution is described in II above.

If in (2) $c_1 = c_2 = 0$, then case V becomes case IV.

Example 3.

$$y' = \frac{2x - y + 9}{x - 3y + 2}. \tag{7}$$

By solving equations (5) we obtain $A = -5$, $B = -1$. Substituting $x = u - 5$, $y = v - 1$ we get (compare (6))

$$\frac{\mathrm{d}v}{\mathrm{d}u} = \frac{2u - v}{u - 3v} = \frac{2 - \frac{v}{u}}{1 - 3\frac{v}{u}}.$$

By a new substitution

$$z = \frac{v}{u}, \quad \text{or} \quad uz = v$$

we have, as shown in IV,

$$u\frac{\mathrm{d}z}{\mathrm{d}u} + z = \frac{2 - z}{1 - 3z} \quad \text{or} \quad u\frac{\mathrm{d}z}{\mathrm{d}u} = \frac{2 - 2z + 3z^2}{1 - 3z}.$$

Then, as in III, we get

$$\int \frac{1 - 3z}{2 - 2z + 3z^2}\,\mathrm{d}z = \int \frac{\mathrm{d}u}{u},$$

$$-\tfrac{1}{2}\ln\left(2 - 2z + 3z^2\right) = \ln ku \quad (ku > 0),$$

$$\ln\left(2 - 2z + 3z^2\right) = -2\ln(ku) = \ln\frac{1}{k^2u^2} = \ln\frac{C}{u^2} \quad (C = 1/k^2 > 0),$$

$$2 - 2z + 3z^2 = \frac{C}{u^2}$$

and after substituting back the original variables $z = v/u$, $u = x + 5$, $v = y + 1$, we obtain, subsequently,

$$2u^2 - 2uv + 3v^2 = C,$$

$$2(x + 5)^2 - 2(x + 5)(y + 1) + 3(y + 1)^2 = C.$$

VI. *Linear equations* are equations of the form

$$y' + a(x)y = b(x). \tag{8}$$

The equation

$$y' + a(x)y = 0 \tag{9}$$

is the so-called *homogeneous linear equation corresponding to equation* (8).

The term "homogeneous" here has a quite different meaning from that in IV. However, this common terminology is normally used in the literature.

If the functions $a(x)$, $b(x)$ are continuous in the interval I (which can also be infinite), then according to Remark 17.2.11, existence and uniqueness of solution of equation (8) or (9) with prescribed initial condition is guaranteed in the whole interval I.

Equation (9) can be solved by separation of variables:

$$\begin{aligned} \int \frac{\mathrm{d}y}{y} &= -\int a(x)\,\mathrm{d}x, \\ \ln(ky) &= -\int a(x)\,\mathrm{d}x \quad (ky > 0), \\ ky &= \mathrm{e}^{-\int a(x)\,\mathrm{d}x}, \\ y &= C\,\mathrm{e}^{-\int a(x)\,\mathrm{d}x} \quad (C = 1/k), \end{aligned} \tag{10}$$

which is the general integral of equation (9).

The general integral of equation (8) can be obtained from (10) by the so-called *method of variation of the parameter* (or *variation of the constant*). Let us assume that the integral of equation (8) is of the form (10), where C is now a function of the variable x,

$$y = C(x)\,\mathrm{e}^{-\int a(x)\,\mathrm{d}x}. \tag{11}$$

Differentiating (11),

$$y' = C'\,\mathrm{e}^{-\int a\,\mathrm{d}x} - Ca\,\mathrm{e}^{-\int a\,\mathrm{d}x}, \tag{12}$$

and substituting (11) and (12) into (8), we have

$$C'\,\mathrm{e}^{-\int a\,\mathrm{d}x} - Ca\,\mathrm{e}^{-\int a\,\mathrm{d}x} + a\,C\,\mathrm{e}^{-\int a\,\mathrm{d}x} = b, \tag{13}$$

wherefrom

$$C'\,\mathrm{e}^{-\int a\,\mathrm{d}x} = b$$

and

$$C(x) = \int b(x)\,\mathrm{e}^{\int a(x)\,\mathrm{d}x}\,\mathrm{d}x.$$

Substituting this result into (11), we obtain the general integral of equation (8) in the form

$$y = \mathrm{e}^{-\int a(x)\,\mathrm{d}x} \cdot \int b(x)\,\mathrm{e}^{\int a(x)\,\mathrm{d}x}\,\mathrm{d}x. \tag{14}$$

Example 4.

$$y' + 2xy = x^3. \tag{15}$$

First, we solve the corresponding homogeneous equation

$$y' + 2xy = 0:$$

$$\int \frac{dy}{y} = -\int 2x\,dx,$$

$$\ln(ky) = -x^2 \quad (ky > 0),$$

$$ky = e^{-x^2},$$

$$y = Ce^{-x^2} \quad (C = 1/k).$$

The general integral of equation (15) can now be found, according to (11), in the form

$$y = C(x)e^{-x^2}. \tag{16}$$

Then

$$y' = C'e^{-x^2} - C.2xe^{-x^2}. \tag{17}$$

Substituting (16) and (17) into (15), we get

$$C'e^{-x^2} - 2xCe^{-x^2} + 2xCe^{-x^2} = x^3,$$

$$C' = x^3e^{x^2}, \quad C = \int x^3e^{x^2}\,dx = \tfrac{1}{2}e^{x^2}(x^2 - 1) + c.$$

Substituting this result into (16), we obtain the general integral of equation (15),

$$y = \tfrac{1}{2}(x^2 - 1) + ce^{-x^2}. \tag{18}$$

REMARK 2. When solving equation (9) by separation of variables we divide by the unknown function y. Nevertheless, (10) or (14) gives the unique solution even in the case where the integral curve has to pass through the point (x_0, y_0), where $y_0 = 0$. For example, the solution (and indeed the only solution) of equation (15) passing through the point (0, 0) is according to (18)

$$y = \tfrac{1}{2}(x^2 - 1) + \tfrac{1}{2}e^{-x^2}.$$

REMARK 3. Instead of using the method of variation of the parameter it is possible to proceed as follows: The integral of equation (8) is assumed to be of the form

$$y(x) = u(x)v(x). \tag{19}$$

Substituting into (8), we get

$$u'v + uv' + auv = b. \tag{20}$$

Let the function $u(x)$ be chosen so as to satisfy the equation

$$u' + au = 0, \tag{21}$$

i.e.

$$u = \mathrm{e}^{-\int a(x)\,\mathrm{d}x}. \tag{22}$$

Since it is possible to write (20) in the form

$$(u' + au)v + uv' = b,$$

it is sufficient for v, in view of condition (21), to satisfy the equation

$$uv' = b$$

or (see (22))

$$v = \int b\,\mathrm{e}^{\int a\,\mathrm{d}x}\,\mathrm{d}x. \tag{23}$$

Substituting (22) and (23) into (19), we get the previous result (14).

VII. *Bernoulli's equation* is an equation of the form

$$y' + a(x)y = b(x)y^n \tag{24}$$

($a(x)$, $b(x)$ being continuous functions in an interval I). For $n = 0$, or $n = 1$, we get a linear equation of the form (8) or (9), respectively.

In the following text, let n be an arbitrary real number different from 0 and 1.

Dividing by y^n, we get from (24) (assuming $y \neq 0$)

$$\frac{y'}{y^n} + \frac{a}{y^{n-1}} = b. \tag{25}$$

Substituting

$$z = y^{-n+1}, \quad \text{from which} \quad z' = (-n+1)y^{-n}y' \quad \text{or} \quad \frac{y'}{y^n} = \frac{z'}{-n+1},$$

we obtain

$$\frac{z'}{-n+1} + az = b,$$

i.e. we get a linear equation for the function $z(x)$.

Example 5.

$$y' + xy = xy^3. \tag{26}$$

Dividing by y^3 and substituting

$$z = \frac{1}{y^2}, \quad \text{so that} \quad z' = -\frac{2y'}{y^3}$$

we have

$$-\frac{z'}{2} + xz = x,$$

$$z' - 2xz = -2x.$$

According to VI the solution of this equation is

$$z = 1 + C\,\mathrm{e}^{x^2}.$$

Hence, the general integral of equation (26) is

$$y^2 = \frac{1}{1 + C\,\mathrm{e}^{x^2}}.$$

VIII. *Riccati's equation* is an equation of the form

$$y' = a(x)y^2 + b(x)y + c(x). \tag{27}$$

For $a(x) \equiv 0$ we get a linear equation and for $c(x) \equiv 0$ we get Bernoulli's equation.

It is not possible, in general, to solve Riccati's equation by *quadratures*, i.e. it is not possible, in general, to reduce its solution to a mere integration, as was the case in the preceding types. However, if one solution of equation (27) is known, the general integral can be obtained by means of quadratures as follows: Let $y_1(x)$ be a solution of equation (27), i.e.

$$y_1' = ay_1^2 + by_1 + c. \tag{28}$$

Introducing a new unknown function $z(x)$ by the relation

$$y = y_1 + z \tag{29}$$

and substituting into (27), we have

$$y_1' + z' = ay_1^2 + 2ay_1z + az^2 + by_1 + bz + c;$$

in consequence of (28), the function z then satisfies the equation

$$z' = az^2 + (2ay_1 + b)z,$$

which is Bernoulli's equation, thus integrable by quadratures.

A particular solution y_1 of equation (27) can often be easily found:

Example 6.

$$xy' - 3y + y^2 = 4x^2 - 4x. \tag{30}$$

Let us assume the particular solution to be of the form

$$y_1 = Ax + B.$$

(In guessing the form of the particular solution, some experience is necessary.) Substituting y_1 in (30) and comparing coefficients of the same powers of x, we get

$$y_1 = 2x.$$

Using the substitution (29), $y = y_1 + z = 2x + z$, we get the equation

$$xz' + (4x - 3)z + z^2 = 0,$$

which is of Bernoulli's type.

Two following special cases of Riccati's equation are easy to solve:

A. $$y' = ay^2 + by + c,$$

where a, b, c are constants; this equation is solved by II above.

B. $$y' = ay^2 + \frac{b}{x^2},$$

where a, b are constants; this equation may be solved by substituting

$$y = \frac{1}{z},$$

giving

$$y' = -\frac{z'}{z^2},$$

$$-\frac{z'}{z^2} = \frac{a}{z^2} + \frac{b}{x^2}$$

and

$$z' = -a - b\left(\frac{z}{x}\right)^2$$

which is a homogeneous equation to be solved by IV.

17.4. Exact Differential Equations. The Integrating Factor. Singular Points

REMARK 1. The equation $y' = \varphi(x, y)$ is often of the form

$$y' + \frac{f(x, y)}{g(x, y)} = 0, \tag{1}$$

where f and g and their derivatives of the first order are continuous functions in a region Ω and $g(x, y) \neq 0$ in Ω. According to Theorem 17.2.1 and Remark 17.2.7, an initial condition being given, (local) existence and uniqueness of the solution is guaranteed in Ω. Written in the so-called *differential form*, equation (1) becomes

$$f(x, y)\,dx + g(x, y)\,dy = 0. \tag{2}$$

Theorem 1. *Let the left-hand side of equation* (2) *be the total differential of a function* $F(x, y)$ *in* Ω. (In this case equation (2) is said to be *exact*. If Ω is a simply connected region, then the necessary and sufficient condition for this is that

$$\frac{\partial f}{\partial y} = \frac{\partial g}{\partial x} \quad \text{in } \Omega.) \tag{3}$$

Then the general integral of (2) *is*

$$F(x, y) = C. \tag{4}$$

REMARK 2. The function $F(x, y)$ may be obtained in the same way as in Example 14.7.2.

Example 1.

$$\left(x^2 - y^2\right) dx + \left(y^3 - 2xy\right) dy = 0.$$

The validity of (3) may be easily verified, because

$$\frac{\partial}{\partial y}\left(x^2 - y^2\right) = \frac{\partial}{\partial x}\left(y^3 - 2xy\right) = -2y. \tag{5}$$

Thus, in every simply connected region, the given equation is exact. By Example 14.7.2 we find that

$$F(x, y) = \frac{x^3}{3} - xy^2 + \frac{y^4}{4} + k,$$

so that the general integral of the given equation is

$$\frac{x^3}{3} - xy^2 + \frac{y^4}{4} = C$$

(obviously in the whole xy-plane), the constant k being incorporated into the constant C.

Theorem 2. *If the equation*

$$f(x, y)\,\mathrm{d}x + g(x, y)\,\mathrm{d}y = 0$$

is exact and if the functions f, g are continuous and homogeneous (Definition 12.6.1) *of the same degree n ($n \neq -1$), then the general integral of the given equation is*

$$x\,f(x, y) + y\,g(x, y) = C.$$

Example 2. The general integral of the equation

$$\frac{2x}{y}\,\mathrm{d}x + \left(1 - \frac{x^2}{y^2}\right)\,\mathrm{d}y = 0$$

is

$$x\frac{2x}{y} + y\left(1 - \frac{x^2}{y^2}\right) = C.$$

REMARK 3. If equation (2) is not exact, we try to find a function $m(x, y) \neq 0$ such that the new equation

$$m(x, y)f(x, y)\,\mathrm{d}x + m(x, y)g(x, y)\,\mathrm{d}y = 0$$

is an exact equation, i.e. that (under some assumptions on the smoothness of the function m)

$$\frac{\partial(mf)}{\partial y} = \frac{\partial(mg)}{\partial x} \tag{6}$$

holds. The function $m(x, y)$ is called an *integrating factor* of equation (2).

REMARK 4. The existence of an integrating factor can be proved under quite general assumptions.

REMARK 5. From (6) it follows that the integrating factor $m(x, y)$ satisfies a partial differential equation

$$g(x, y)\frac{\partial m}{\partial x} - f(x, y)\frac{\partial m}{\partial y} = \left(\frac{\partial f}{\partial y} - \frac{\partial g}{\partial x}\right)m. \tag{7}$$

In general, it is more difficult to integrate this equation than to integrate the given equation (2).

In many cases m can be found as a function of the variable x only. From equation (7) it follows that a necessary and sufficient condition for this is that

$$\frac{\partial f/\partial y - \partial g/\partial x}{g}$$

be a function of the variable x only; then $m(x)$ can be obtained from the equation (assuming that $m(x) > 0$)

$$\frac{\mathrm{d} \ln m}{\mathrm{d}x} = \frac{\partial f/\partial y - \partial g/\partial x}{g}. \tag{8}$$

A similar statement is valid for the case where m is only a function of the variable y. The result is:

$$\frac{\mathrm{d} \ln m}{\mathrm{d}y} = \frac{\partial f/\partial y - \partial g/\partial x}{-f}. \tag{9}$$

Example 3. The equation

$$\left(2xy + x^2 y + \tfrac{1}{3}y^3\right) \mathrm{d}x + \left(x^2 + y^2\right) \mathrm{d}y = 0$$

is not exact (the condition (3) is not satisfied). But we note that

$$\frac{\partial f/\partial y - \partial g/\partial x}{g} = \frac{2x + x^2 + y^2 - 2x}{x^2 + y^2} = 1.$$

According to (8)

$$\frac{\mathrm{d} \ln m(x)}{\mathrm{d}x} = 1, \quad \text{hence} \quad m(x) = \mathrm{e}^x .$$

(It is not necessary to write $m(x) = C\,\mathrm{e}^x$ with a constant C, because we are only trying to find *a* solution of equation (7).)

The equation

$$\mathrm{e}^x \left(2xy + x^2 y + \tfrac{1}{3}y^3\right) \mathrm{d}x + \mathrm{e}^x \left(x^2 + y^2\right) \mathrm{d}y = 0$$

is thus exact, as may also be easily verified by (3).

REMARK 6 (*Singular Points of Equation* (1)). From Theorem 17.2.1 and Remark 17.2.7 it follows: Let f, g, $\partial f/\partial y$, $\partial g/\partial y$ be continuous in the neighbourhood of the point (x_0, y_0) and let $g(x_0, y_0) \neq 0$. Then exactly one integral curve of equation (1) exists (locally) which passes through the point (x_0, y_0). If $g(x_0, y_0) = 0$,

but $f(x_0, y_0) \neq 0$, then (see Remark 17.2.16) equation (1) may be written in the form

$$\frac{\mathrm{d}x}{\mathrm{d}y} = -\frac{g(x, y)}{f(x, y)}, \tag{10}$$

and under the supplementary assumption of continuity of $\partial f/\partial x$, $\partial g/\partial x$ again exactly one integral curve passes through the point (x_0, y_0). If the relations

$$f(x_0, y_0) = 0 \quad \text{and} \quad g(x_0, y_0) = 0$$

hold simultaneously, then the point (x_0, y_0) is called a *singular point of equation* (1). In the neighbourhood of a singular point various possibilities can occur as to existence and uniqueness of the solution as well as to the form of the integral curves of the given equation. Let us give some simple examples:

The equation

$$y' = \frac{y}{x} \tag{11}$$

has one and only one singular point, namely (0, 0). By separation of variables, the general integral

$$y = Cx \tag{12}$$

can easily be obtained in each of the half-planes $x > 0$, $x < 0$. Thus integral curves are the half-lines (12) (Fig. 17.2). It is usual, in the literature, to complete them

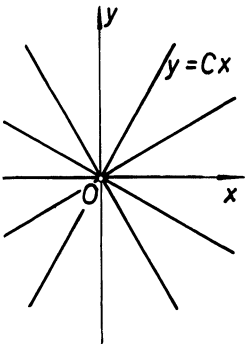


Fig. 17.2.

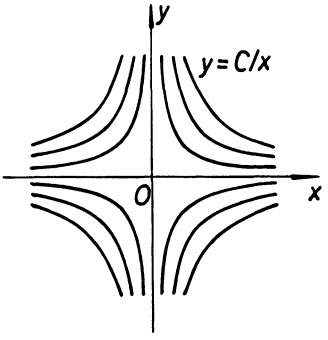


Fig. 17.3.

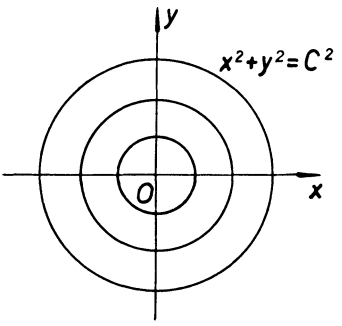


Fig. 17.4.

by the point (0, 0). In this sense, integral curves are straight lines passing through the singular point considered. A singular point of this type is called a *node*.

The (only) singular point of the equation

$$y' = -\frac{y}{x} \tag{13}$$

is also the origin. The general integral is a system of hyperbolae

$$y = \frac{C}{x} \tag{14}$$

with asymptotes

$$y = 0, \quad x = 0$$

(Fig. 17.3). This type of singular point is called a *saddle point.*

The equation

$$y' = -\frac{x}{y} \tag{15}$$

has also only one singular point, the origin. The general integral is a system of circles

$$x^2 + y^2 = C$$

(Example 17.2.7), thus of closed curves round that singular point (Fig. 17.4). This type of singular point is called a *centre.*

The reader will find a more detailed analysis in [87] or in [44] where he will also find some results concerning questions on differential equations in a complex domain.

17.5. Equations of the First Order not Solved with Respect to the Derivative. Lagrange's Equation. Clairaut's Equation. Singular Solutions

Equations of the first order not solved with respect to the derivative are equations of the form

$$F(x, y, y') = 0, \tag{1}$$

or, if the common notation $y' = p$ is used, of the form

$$F(x, y, p) = 0. \tag{2}$$

I. In some cases (see Theorem 12.9.2 on Implicit Functions) it is possible to solve (1) with respect to y'.

Example 1.

$$y'^2 = 1 + x^2 + y^2. \tag{3}$$

On solving equation (1), we get two equations

$$y' = \sqrt{(1 + x^2 + y^2)}, \quad y' = -\sqrt{(1 + x^2 + y^2)}.$$

The general integral of each of these equations represents a one-parameter system of curves. The curves of both the first and the second system are integral curves of equation (3). Thus the general integral of equation (3) consists of two systems of curves.

REMARK 1. Equation (1) is not always so easy to solve with respect to y'. In the domain of real functions, this equation need not have a solution at all, as is to be seen from the simple example of the equation

$$y'^2 + y^2 + 1 = 0.$$

In the general case, the problem of solving equation (1), both from a theoretical and practical point of view, is more complicated than in the case of the equation $y' = f(x, y)$. In this paragraph we shall deal only with the most important results and methods of solution.

II. *Method of two parameters.* Under rather simple assumptions about the function F, equation (2) represents a surface. The coordinates of the point (x, y, p) of this surface can be expressed as functions of two parameters u, v:

$$x = f(u, v), \quad y = g(u, v), \quad p = h(u, v). \tag{4}$$

From the relation $\mathrm{d}y = p\,\mathrm{d}x$ it follows that

$$\frac{\partial g}{\partial u}\,\mathrm{d}u + \frac{\partial g}{\partial v}\,\mathrm{d}v = h(u, v)\left[\frac{\partial f}{\partial u}\,\mathrm{d}u + \frac{\partial f}{\partial v}\,\mathrm{d}v\right]$$

or

$$\frac{\mathrm{d}v}{\mathrm{d}u} = \frac{h\dfrac{\partial f}{\partial u} - \dfrac{\partial g}{\partial u}}{\dfrac{\partial g}{\partial v} - h\dfrac{\partial f}{\partial v}}. \tag{5}$$

This is a differential equation for the unknown function $v(u)$, solved with respect to the derivative $\mathrm{d}v/\mathrm{d}u$. Finding its general integral $v = t(u, C)$ and substituting into the first two equations (4), we obtain

$$x = f\,(u, t(u, C))\,, \quad y = g\,(u, t(u, C))\,.$$

Eliminating u from these two equations, we have

$$\varphi(x, y, C) = 0,$$

which, as a rule, gives the general integral of equation (1).

REMARK 2. We have used the phrase "as a rule". The above-mentioned procedure is formal; in particular, the concept of elimination of a variable or parameter is very uncertain. Consequently, the result so obtained must be analysed, especially to verify whether it is really the solution of the given equation. *This remark applies throughout the whole paragraph* 17.5.

It is particularly simple to express the "surface" (2) by means of equations (4) in these two cases:

III. Let equation (2) be of the form

$$y = f(x, p). \tag{6}$$

We choose x and p as parameters. Then, from the equation $\mathrm{d}y = p\,\mathrm{d}x$, it follows that

$$\frac{\partial f}{\partial x}\,\mathrm{d}x + \frac{\partial f}{\partial p}\,\mathrm{d}p = p\,\mathrm{d}x \quad \text{or} \quad \frac{\partial f}{\partial x} + \frac{\partial f}{\partial p}\frac{\mathrm{d}p}{\mathrm{d}x} = p. \tag{7}$$

(It is possible to obtain (7) directly by differentiating equation (6) with respect to x, if p is taken as a function of x and p is substituted for y'.) If we find the general integral of (7),

$$p = t(x, C) \tag{8}$$

and substitute for p into (6), we get, as a rule (see Remark 2), the general integral of equation (6). (It would not be correct to substitute y' for p into (8) and to try to find the general integral of equation (6) by integrating equation (8) thus obtained. When integrating, a second constant would arise. It is now easy to show that by this procedure we would get the general integral of the equation of the second order,

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y'}y'' = y',$$

which results from (7) by substituting y' for p.)

IV. Let equation (2) be of the form

$$x = f(y, p). \tag{9}$$

If we consider p as a function of y (by virtue of a relation between x and y) and differentiate (9) with respect to y, using

$$\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{1}{\mathrm{d}y/\mathrm{d}x} = \frac{1}{p},$$

we get

$$\frac{1}{p} = \frac{\partial f}{\partial y} + \frac{\partial f}{\partial p}\frac{\mathrm{d}p}{\mathrm{d}y}. \tag{10}$$

This is an equation for the unknown function $p(y)$. If we find its general integral

$$p = k(y, C),$$

and then substitute for p into (9) we get the general integral of equation (9). (See, however, Remark 2.)

REMARK 3. Using the above-mentioned procedure, it is naturally possible to integrate special cases of III and IV,

$$y = f(p), \quad \text{or} \quad x = f(p).$$

If, instead of the equation $x = f(p)$, a rather more general equation $g(x, p) = 0$ is given, we can use a method similar to method II. This case becomes simpler, because the equation $g(x, p) = 0$ represents a curve, geometrically. Let its parametric equations be of the form

$$x = \psi(t), \quad p = \chi(t).$$

Since

$$\frac{\mathrm{d}y}{\mathrm{d}x} = p = \chi(t),$$

then

$$\mathrm{d}y = p\,\mathrm{d}x = \chi(t)\psi'(t)\,\mathrm{d}t.$$

Integrating, we get

$$y = \int \chi(t)\psi'(t)\,\mathrm{d}t = \mu(t) + C,$$

say. Then the equations

$$x = \psi(t), \quad y = \mu(t) + C$$

give the general solution of the given equation.

In a similar way, it is possible to deal with the equation $g(y, p) = 0$ (see Example 2).

Example 2.

$$y = y'^2 + 2. \tag{11}$$

Following III, we differentiate with respect to x and write p instead of y'. Then

$$p = 2p\frac{\mathrm{d}p}{\mathrm{d}x}.$$

If $p \neq 0$ (for $p = 0$ see below), then

$$\frac{\mathrm{d}p}{\mathrm{d}x} = \frac{1}{2}$$

so that

$$p = \tfrac{1}{2}x + C.$$

Substituting into (11), we have the general integral of equation (11):

$$y = \left(\tfrac{1}{2}x + C\right)^2 + 2. \tag{12}$$

It may be easily verified that if we solve (11) to give y', then the general integral of the two equations so obtained is included in (12).

The case $p = 0$ gives the solution $y = 2$ of the given equation.

If we had used the method of Remark 3, we would have got:

$$p = t, \quad y = t^2 + 2,$$

$$\frac{dy}{dx} = p, \quad dx = \frac{dy}{p} = \frac{2t\,dt}{t} = 2\,dt, \quad x = 2t + k.$$

Thus, the general solution in a parametric form is

$$x = 2t + k, \quad y = t^2 + 2$$

and, by eliminating the parameter t,

$$y = \left(\frac{x}{2} - \frac{k}{2}\right)^2 + 2$$

in accordance with (12).

V. *Lagrange's equation*

$$y = \varphi(y')x + \psi(y') \tag{13}$$

or

$$y = \varphi(p)x + \psi(p) \tag{14}$$

is a special case of type III; it is always possible to integrate it by means of quadratures. Differentiating (14) with respect to x and substituting p for y', we obtain

$$p = \varphi(p) + x\varphi'(p)\frac{dp}{dx} + \psi'(p)\frac{dp}{dx}. \tag{15}$$

If we consider x as a function of the variable p, then we obtain from (15) (assuming $\varphi(p) \neq p$; on the case $\varphi(p) \equiv p$ see VI, Clairaut's equation)

$$\frac{dx}{dp} = \frac{\varphi'(p)}{p - \varphi(p)}x + \frac{\psi'(p)}{p - \varphi(p)} \quad (\varphi(p) \neq p). \tag{16}$$

Equation (16) is a linear equation for the function $x(p)$. If we substitute its general integral

$$x = t(p, C) \tag{17}$$

into (14), we get

$$y = \varphi(p)t(p, C) + \psi(p); \tag{18}$$

(17) and (18) give a parametric form of the general integral of (14). Eliminating the parameter p from (17) and (18), we get the general integral in the form $G(x, y, C) = 0$. (See, however, Remark 2.)

Example 3.

$$y = 2xy' + y'^2 \quad \text{or} \quad y = 2px + p^2. \tag{19}$$

Differentiating with respect to x, we have

$$p = 2p + 2(x + p)\frac{\mathrm{d}p}{\mathrm{d}x},$$

that is

$$\frac{\mathrm{d}x}{\mathrm{d}p} = -\frac{2x}{p} - 2 \quad (p \neq 0). \tag{20}$$

(20) is a linear equation for the function $x(p)$. Its solution is (cf. Example 17.3.4)

$$x = \frac{C}{p^2} - \frac{2p}{3}. \tag{21}$$

Substituting into the second of equations (19), we get

$$y = \frac{2C}{p} - \frac{p^2}{3} \tag{22}$$

which together with (21) constitute the parametric equations of the general integral of (19).

For $p = 0$ we get, on substituting into (19), the solution $y \equiv 0$.

VI. *Clairaut's equation*

$$y = xy' + \psi(y') \quad \text{or} \quad y = px + \psi(p) \tag{23}$$

(where $\psi(p)$ is a given differentiable function) is a particular case of Lagrange's equation (for the case $\varphi(p) \equiv p$). Differentiating with respect to x, we get

$$p = p + [x + \psi'(p)]\frac{\mathrm{d}p}{\mathrm{d}x} \quad \text{or} \quad \frac{\mathrm{d}p}{\mathrm{d}x}[x + \psi'(p)] = 0. \tag{24}$$

From the equation

$$\frac{dp}{dx} = 0$$

we get $p = C$, which substituted into (23) gives the general integral of equation (23):

$$y = Cx + \psi(C). \tag{25}$$

This is a one-parameter system of straight lines.

If the second term on the left-hand side of the second equation (24) is equal to zero,

$$x + \psi'(p) = 0, \tag{26}$$

then (26) determines p as a function of x,

$$p = t(x) \tag{27}$$

(if, for example, $\psi''(p) \neq 0$). Substituting into (23), we get

$$y = xt(x) + \psi\left(t(x)\right). \tag{28}$$

This solution may be proved to be the envelope of the system (25). It is a singular integral of equation (23).

REMARK 4. When investigating equations of the first order, it is often important to find the *singular integral* of the given equation (Definition 17.2.6). The solution of this problem in its full generality is rather difficult. We introduce here only two special theorems:

Definition 1. By a *discriminant curve* of the equation

$$F(x, y, y') = 0 \tag{29}$$

we mean the curve $G(x, y) = 0$ the points (x, y) of which satisfy (for a certain range of values of α) the equations

$$F(x, y, \alpha) = 0, \quad \frac{\partial F}{\partial \alpha}(x, y, \alpha) = 0. \tag{30}$$

REMARK 5. It may happen that the curve $G(x, y) = 0$ is not a real curve. For example, for the equation

$$y'^3 + (y^2 + 2)y' - xy = 0 \tag{31}$$

we get

$$\frac{\partial F}{\partial \alpha} = 3\alpha^2 + y^2 + 2 \tag{32}$$

(since $F(x, y, \alpha) = \alpha^3 + (y^2+2)\alpha - xy$), and this cannot be zero for any real values of the variables x, y, α, so that $G(x, y) = 0$ does not represent a real curve.

Theorem 1. *Let a discriminant curve of equation* (29) *exist. Then the singular integral of this equation* (*if it exists*) *is contained in this discriminant curve.*

REMARK 6. From the theorem on implicit functions it follows that validity of the relation

$$\frac{\partial F}{\partial y'}(x, y, y') = 0 \tag{33}$$

at each point of the integral curve concerned is a *necessary* (but not sufficient) condition for that curve to be a singular integral of the given equation. If the discriminant curve of equation (29) is an integral curve of this equation, then it need not be the singular integral of this equation.

Theorem 2. *Let*

$$f(x, y, C) = 0 \tag{34}$$

be the general integral of equation (29). *If the curve given by the equation*

$$H(x, y) = 0 \tag{35}$$

which we obtain by eliminating C from the equations

$$f(x, y, C) = 0, \quad \frac{\partial f}{\partial C}(x, y, C) = 0, \tag{36}$$

is the envelope (§ 9.7) *of the one-parameter system* (34), *then*

1. *it is an integral curve of equation* (29);
2. *it is the singular integral of this equation.*

Example 4. Let the equation

$$y = x - \tfrac{4}{9}p^2 + \tfrac{8}{27}p^3 \quad (p = y') \tag{37}$$

be solved by the method of differentiation with respect to x (see V). We get (see (16)):

$$\frac{dx}{dp} = \frac{-\frac{8}{9}p + \frac{8}{9}p^2}{p-1} = \tfrac{8}{9}p, \quad \text{so that} \quad x = \tfrac{4}{9}p^2 + C. \tag{38}$$

Substituting into (37), we get

$$y = C + \tfrac{8}{27}p^3 \quad \text{or} \quad (y-C)^2 = \left(\tfrac{8}{27}\right)^2 \left(\tfrac{9}{4}\right)^3 (x-C)^3$$

(using the second equation (38)), i.e.

$$(y-C)^2 - (x-C)^3 = 0. \tag{39}$$

The second of equations (36) gives

$$2(y - C) = 3(x - C)^2. \tag{40}$$

If we substitute from here for $y - C$ into the first equation (36), we get

$$\tfrac{9}{4}(x - C)^4 = (x - C)^3, \tag{41}$$

so that either

$$x - C = 0 \quad \text{and then also} \quad y - C = 0, \quad \text{consequently} \quad y - x = 0, \tag{42}$$

or $x - C \neq 0$ and then from (41) we have

$$x - C = \tfrac{4}{9}, \quad \text{so that according to (39)} \quad y - C = \tfrac{8}{27} \quad \text{or} \quad y - x = -\tfrac{4}{27}. \tag{43}$$

The straight line $y - x = -\frac{4}{27}$ is (see Example 9.7.3) the envelope of the system (39) and according to Theorem 2 it is the singular integral of equation (37). The straight line $y - x = 0$ is not the envelope of the system (39) (Example 9.7.3). Substituting into (37), we can easily verify that it is not an integral curve of this equation.

We could reach the same result using Theorem 1 (i.e. without integrating equation (37)): We have

$$\frac{\partial F}{\partial y'} = \frac{8}{9}y' - \frac{8}{9}y'^2,$$

so that the second of equations (30) (i.e. equation (33)) is

$$y'(1 - y') = 0. \tag{44}$$

Thus, either $y' = 0$, or $y' = 1$. Substituting into (37), we get

$$y - x = 0, \quad \text{and} \quad y - x = -\tfrac{4}{27},$$

respectively, and these are the lines (42), (43).

17.6. Orthogonal and Isogonal (Oblique) Trajectories

Let a one-parameter system of curves be given:

$$F(x, y, C) = 0 \tag{1}$$

and let

$$y' = f(x, y) \tag{2}$$

be the differential equation of this system.

Theorem 1. *The differential equation of a system of such curves* $G(x, y, k) = 0$, *each of which intersects every curve of the system* (1) *at right angles* (*the system of so-called orthogonal trajectories*), *is*

$$y' = -\frac{1}{f(x, y)}. \tag{3}$$

Example 1. Find orthogonal trajectories of the system of parabolas

$$y = Cx^2. \tag{4}$$

Differentiating (4) with respect to x we get $y' = 2Cx$ and eliminating C from both equations we obtain the differential equation of the system (4):

$$y' = \frac{2y}{x}. \tag{5}$$

According to (3), the differential equation of the orthogonal trajectories is

$$\frac{dy}{dx} = -\frac{1}{2y/x} = -\frac{x}{2y}.$$

Integrating this by separation of variables (§ 17.3), we obtain a system of ellipses

$$y^2 + \tfrac{1}{2}x^2 = k. \tag{6}$$

REMARK 1. The differential equation of the system $H(x, y, C) = 0$, every curve of which intersects every curve of the system (1) at an angle $\alpha \neq \pi/2$ (the so-called *isogonal trajectories* or *oblique trajectories*), is given by the equation

$$\frac{dy}{dx} = \frac{k + f(x, y)}{1 - kf(x, y)},$$

where $k = \tan\alpha$ and $y' = f(x, y)$ is the differential equation of the system (1).

17.7. Differential Equations of Order n. Simple Types of Equations of Order n. The Method of a Parameter

For existence and uniqueness of the solution see Theorem 17.2.3.

Throughout this paragraph, the continuity of the functions considered and of the respective derivatives is assumed.

I. The equation

$$y'' = f(x) \tag{1}$$

is easy to solve:

$$y' = \int f(x)\,dx, \quad y = \int y'(x)\,dx. \tag{2}$$

There are two arbitrary constants in the result, because two indefinite integrals are involved. In a similar way we solve the equation

$$y^{(n)} = f(x).$$

II. The solution of the equation

$$y^{(n)} = f(x), \tag{3}$$

that satisfies the initial conditions

$$y(x_0) = y'(x_0) = \ldots = y^{(n-1)}(x_0) = 0$$

is given by the *Cauchy-Dirichlet Formula*:

$$y(x) = \frac{1}{(n-1)!}\int_{x_0}^{x} (x-z)^{n-1} f(z)\,dz. \tag{4}$$

By adding to (4) the expression

$$y_0^{(n-1)}\frac{(x-x_0)^{n-1}}{(n-1)!} + y_0^{(n-2)}\frac{(x-x_0)^{n-2}}{(n-2)!} + \ldots + y_0'(x-x_0) + y_0, \tag{5}$$

we get the integral of equation (3) that satisfies the initial conditions

$$y(x_0) = y_0,\ y'(x_0) = y_0', \ldots, y^{(n-1)}(x_0) = y_0^{(n-1)}.$$

Example 1. Solve the equation $y''' = \ln x$ under the initial conditions $y(1) = y_0$, $y'(1) = y_0'$, $y''(1) = y_0''$.

We get the solution as the sum of expressions (4) and (5), $y = y_1 + y_2$. Integrating (4) by parts, we get

$$y_1 = \tfrac{1}{2}\int_1^x (x-z)^2 \ln z\,dz = \tfrac{1}{6}x^3 \ln x - \tfrac{11}{36}x^3 + \tfrac{1}{2}x^2 - \tfrac{1}{4}x + \tfrac{1}{18}. \tag{6}$$

According to (5)

$$y_2 = y_0 + y_0'(x-1) + y_0''\frac{(x-1)^2}{2}. \tag{7}$$

The sum of (6) and (7) gives the solution required.

III. The equation of the form

$$y'' = f(x, y'), \quad \text{or} \quad F(x, y', y'') = 0, \tag{8}$$

in which y does not occur explicitly, can be transformed, using the substitution $y'(x) = z(x)$ into an equation of the first order for the function $z(x)$. More generally, the equation of the form

$$y^{(n)} = f(x, y^{(n-1)}), \quad \text{or} \quad F(x, y^{(n)}, y^{(n-1)}) = 0 \tag{9}$$

can be transformed by the substitution $y^{(n-1)}(x) = z(x)$ into an equation of the first order for $z(x)$. Solving it, we get, by repeated integration (see I and II), the function $y(x)$.

REMARK 1 (*Method of a Parameter*). If a particular case of the second of equations (9),

$$F\big(y^{(n)}, y^{(n-1)}\big) = 0, \tag{10}$$

is to be solved and if this equation cannot be easily transformed into the form $y^{(n)} = f\big(y^{(n-1)}\big)$, we can often use the method of a parameter with success. Let us consider, for example, the equation

$$F(y''', y'') = 0. \tag{11}$$

First, we express y''' and y'' parametrically,

$$y''' = \varphi(t), \quad y'' = \psi(t),$$

so that equation (11) is identically satisfied in t, i.e. $F(\varphi(t), \psi(t)) \equiv 0$. Further, $dy'' = y''' dx$ holds, so that (under obvious assumptions)

$$dx = \frac{dy''}{y'''} = \frac{\psi'(t)\,dt}{\varphi(t)}, \quad \text{whence} \quad x(t) = \int \frac{\psi'(t)\,dt}{\varphi(t)} = \mu(t) + C_1; \tag{12}$$

then

$$dy' = y''\,dx = \frac{\psi(t)\psi'(t)}{\varphi(t)}dt, \quad \text{whence} \quad y'(t) = \int \frac{\psi(t)\psi'(t)\,dt}{\varphi(t)} = \chi(t) + C_2, \tag{13}$$

$$dy = y'\,dx = [\chi(t) + C_2]\frac{\psi'(t)}{\varphi(t)}dt, \quad \text{whence}$$

$$y(t) = \int [\chi(t) + C_2]\frac{\psi'(t)}{\varphi(t)}dt = \kappa(t, C_2) + C_3. \tag{14}$$

Thus we get the general integral of equation (11) in parametric form

$$x = \mu(t) + C_1, \quad y = \kappa(t, C_2) + C_3. \tag{15}$$

REMARK 2. In this paragraph, we often use a formal procedure in the same way as in § 17.5, so that also here a remark similar to Remark 17.5.2 must be taken into account.

Example 2.

$$y'''^2 + y''^2 - 4 = 0.$$

Let us express y''' and y'' parametrically in the form

$$y''' = 2\cos t, \quad y'' = 2\sin t.$$

Following (12), (13), (14) we get

$$x = \int \frac{\psi'(t)\,\mathrm{d}t}{\varphi(t)} = \int \mathrm{d}t = t + C_1,$$

$$y' = \int 2\sin t\,\mathrm{d}t = -2\cos t + C_2,$$

$$y = \int \left(-2\cos t + C_2\right)\mathrm{d}t = -2\sin t + C_2 t + C_3.$$

Eliminating t from the equations for x and y, we get the general integral of the given equation:

$$y = -2\sin(x - C_1) + C_2 x + K \qquad (K = -C_1 C_2 + C_3).$$

REMARK 3. We can obviously proceed as follows: Substituting $y''(x) = z(x)$, we first transform equation (11) into an equation of the form

$$F(z, z') = 0$$

and then integrate this equation using the method of a parameter. From the parametric equations for x and z so obtained, we then eliminate t, in order to get $y(x)$ by a further integration (which, however, need not be a simple matter, in general).

Example 3. Taking the equation of Example 2, we put $y'' = z$ and choose

$$z' = 2\cos u, \quad z = 2\sin u.$$

From the relation $\mathrm{d}z = z'\,\mathrm{d}x$, or $2\cos u\,\mathrm{d}u = 2\cos u\,\mathrm{d}x$, it follows that $\mathrm{d}x = \mathrm{d}u$, $x = u + C$. Hence

$$z = 2\sin(x - C) = y'',$$

and by integrating twice we get the same result as in Example 2.

IV. The equation

$$y'' = f(y) \tag{16}$$

multiplied by $y'(x)$ becomes

$$y'y'' = f(y)y', \quad \text{hence} \quad \tfrac{1}{2}y'^2 = \int f(y)\,dy = F(y) + C.$$

In this way, the original equation is transformed into an equation of the first order. A similar procedure may be used when solving an equation of the form

$$y^{(n)} = f\left(y^{(n-2)}\right) \tag{17}$$

(after the initial substitution $y^{(n-2)}(x) = z(x)$, if wanted).

REMARK 4. If equation (16) or (17) is given in implicit form,

$$F(y'', y) = 0, \quad \text{or} \quad F\left(y^{(n)}, y^{(n-2)}\right) = 0, \tag{18}$$

we can use the method of a parameter in a similar way as in Remark 1.

For example, considering the equation

$$F(y''', y') = 0, \tag{19}$$

we express, first of all, y''' and y' parametrically:

$$y''' = \varphi(t), \quad y' = \psi(t)$$

(so that $F\left(\varphi(t), \psi(t)\right) \equiv 0$). Further, by eliminating dx from the equations

$$dy'' = y'''\,dx, \quad dy' = y''\,dx,$$

we get

$$y''\,dy'' = y'''\,dy' = \varphi(t)\psi'(t)\,dt, \quad \text{whence} \quad \tfrac{1}{2}\left[y''(t)\right]^2 = \int \varphi(t)\psi'(t)\,dt.$$

Once $y''(t)$ known, we proceed further according to Remark 1

$$\left(dx = \frac{dy''}{y'''}, \quad dy' = y''dx, \text{ etc.}\right).$$

REMARK 5. A method similar to case IV may also be used for equations of the form

$$y'' = ay'^2 + f(y). \tag{20}$$

Introducing a new unknown function $u(x)$ by the relation $u(x) = y'^2(x)$, we get

$$du = 2y'\,dy' = 2y''dy \tag{21}$$

using the relations

$$y'\,dx = dy, \quad dy' = y''\,dx.$$

Hence, from (20), (21)

$$du = 2\big[au + f(y)\big]\,dy,$$

which is a *linear* equation for the function $u(y)$. Having solved it, we come to an equation of the first order

$$y'^2 = u(y).$$

17.8. First Integral of a Differential Equation of the Second Order. Reduction of the Order of a Differential Equation. Equations, the Left-hand Sides of which are Exact Derivatives

REMARK 1. In physics, we often meet the concept of the first integral of a differential equation of the second order. Its meaning is as follows:

Definition 1. We say that the equation

$$g(x, y, y') = C \tag{1}$$

gives (is, or represents) the *first integral* of the equation

$$F(x, y, y', y'') = 0,$$

if the function $g(x, y, y')$ is constant along each integral curve of the given equation (and, at the same time, is not identically equal to a constant in the variables x, y, y').

Example 1. The equation

$$y'^2 + 4y^3 = C \tag{2}$$

gives the first integral of the equation

$$y'' + 6y^2 = 0 \tag{3}$$

since

$$\frac{d}{dx}(y'^2 + 4y^3) = 2y'y'' + 12y^2y' = 2y'(y'' + 6y^2),$$

which, because of (3), is zero along each integral curve of equation (3).

REMARK 2. Similarly, it is possible to define the k-th integral of a given differential equation of order n. In this way, differential equations of lower orders can be obtained which are (at least theoretically) easier to integrate.

Some typical examples of *reduction of orders of differential equations*:

I. The equation

$$F\left(x, y^{(m)}, y^{(m+1)}, \ldots, y^{(n)}\right) = 0 \tag{4}$$

is easily transformed by the substitution $y^{(m)}(x) = z(x)$ into the equation

$$F\left(x, z, z', \ldots, z^{(n-m)}\right) = 0, \tag{5}$$

the order of which is $n - m$.

II. In the equation

$$F\left(y, y', y'', \ldots, y^{(n)}\right) = 0, \tag{6}$$

let us write $y' = p$ and let p be considered as a function of the independent variable y, i.e. $p = p(y)$. Then

$$y'' = \frac{\mathrm{d}y'}{\mathrm{d}x} = \frac{\mathrm{d}p}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}x} = p\frac{\mathrm{d}p}{\mathrm{d}y}, \tag{7}$$

$$y''' = \frac{\mathrm{d}y''}{\mathrm{d}x} = \frac{\mathrm{d}\left(p\,\dfrac{\mathrm{d}p}{\mathrm{d}y}\right)}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}x} = p^2\frac{\mathrm{d}^2p}{\mathrm{d}y^2} + p\left(\frac{\mathrm{d}p}{\mathrm{d}y}\right)^2, \tag{8}$$

.....

So we get from (6) an equation of order $n - 1$

$$G\left(y, p, \frac{\mathrm{d}p}{\mathrm{d}y}, \ldots, \frac{\mathrm{d}^{n-1}(p)}{\mathrm{d}y^{n-1}}\right) = 0.$$

III. Let the left-hand side of the equation $F(x, y, y', \ldots, y^n) = 0$ be a homogeneous function of degree m in the arguments $y, y', y'', \ldots, y^{(n)}$, i.e. let

$$F\left(x, ty, ty', \ldots, ty^{(n)}\right) = t^m F\left(x, y, y', \ldots, y^{(n)}\right) \tag{9}$$

be satisfied for all t in a certain neighbourhood of the point $t = 1$. Then the order of that equation may be reduced by introducing a new unknown function $z(x)$, given by the relation

$$y(x) = \mathrm{e}^{\int z(x)\,\mathrm{d}x}. \tag{10}$$

Example 2. The equation

$$x^2yy'' - (y - xy')^2 = 0 \tag{11}$$

is homogeneous (of the second degree) with respect to y, y', y''. Following (10), we get

$$y = \mathrm{e}^{\int z\,dx}, \quad y' = z\,\mathrm{e}^{\int z\,dx}, \quad y'' = (z' + z^2)\,\mathrm{e}^{\int z\,dx}. \tag{12}$$

Substituting (12) into (11) and dividing by $\mathrm{e}^{2\int z\,dx}$, we get the equation

$$x^2(z' + z^2) - (1 - xz)^2 = 0,$$

or

$$x^2z' + 2xz - 1 = 0.$$

The solution of this linear equation (see VI, § 17.3) is

$$z = \frac{1}{x} + \frac{C_1}{x^2},$$

so that

$$y = \mathrm{e}^{\int z\,dx} = \mathrm{e}^{\int (1/x + C_1/x^2)\,dx} = C_2x\,\mathrm{e}^{-C_1/x}.$$

REMARK 3. The homogeneous linear equation

$$a_n(x)y^{(n)} + a_{n-1}(x)y^{(n-1)} + \ldots + a_1(x)y' + a_0(x)y = 0 \tag{13}$$

is also of type III. The substitution (10), however, is not suitable in this case, because it leads to a non-linear equation, in general. If a (non-zero) particular integral $y = u(x)$ of equation (13) is known, we can reduce the order of the equation by introducing a new unknown function $z(x)$ by the relation

$$y(x) = u(x)z(x). \tag{14}$$

Example 3. One of the solutions of the equation

$$xy''' - y'' + xy' - y = 0 \tag{15}$$

is obviously $u(x) = x$. Using the substitution (14), we get

$$y' = xz' + z, \quad y'' = xz'' + 2z', \quad y''' = xz''' + 3z''.$$

Thus, equation (15) is transformed into the equation

$$x^2z''' + 2xz'' + (x^2 - 2)z' = 0 \quad \text{or} \quad x^2v'' + 2xv' + (x^2 - 2)v = 0 \tag{16}$$

(putting $z' = v$), and this is an equation of the second order. From (14) and from $v = z'$ it follows that $v = (y/u)'$. Further, if we knew another integral $U(x)$ of equation (15), we could obtain another integral $v = (U/u)'$ of equation (16), so that we could again reduce the order of equation (16). In general: If we know k (*independent*) integrals of equation (13), we can reduce its order by k.

REMARK 4. It is possible to prove, by the method mentioned in Remark 3, that if $y_1(x)$ is a non-zero particular integral of the equation

$$y'' + a_1(x)y' + a_0(x)y = 0, \tag{17}$$

then the second function of the fundamental system of solutions (Definition 17.11.2) is given by the function

$$y_2 = y_1 \int \frac{1}{y_1^2} \mathrm{e}^{-\int a_1 \,\mathrm{d}x} \,\mathrm{d}x. \tag{18}$$

Example 4. A particular integral of the equation

$$y'' - y = 0$$

is $y_1 = \mathrm{e}^x$. According to (18), the second function of the fundamental system is the function

$$y_2 = \mathrm{e}^x \int \frac{1}{\mathrm{e}^{2x}} \mathrm{e}^{-\int 0 \,\mathrm{d}x} \,\mathrm{d}x = k\,\mathrm{e}^{-x}.$$

REMARK 5. Another way of reducing the order of a differential equation is to transform the left-hand side of the given equation into a form which is the complete derivative (with respect to x) of an expression of lower order. We shall demonstrate this procedure by an example.

Example 5. Let us consider the equation

$$yy'' - y'^2 = 0. \tag{19}$$

Dividing by y^2, we get

$$\frac{yy'' - y'^2}{y^2} = 0 \quad (y \neq 0). \tag{20}$$

The left-hand side of equation (20) is obviously a complete derivative:

$$\frac{yy'' - y'^2}{y^2} = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{y'}{y}\right). \tag{21}$$

Thus the first integral of equation (20) (or (19)) is

$$\frac{y'}{y} = C_1,$$

hence

$$y = C_2 \, e^{C_1 x}.$$

17.9. Dependence of Solutions on Parameters of the Differential Equation and on Initial Conditions

Theorem 1. *Let an equation of the form*

$$y^{(n)} = f\left(x, y, y', \ldots, y^{(n-1)}\right) \tag{1}$$

be given. Let us denote by A the point with coordinates $a, b_1, b_2, \ldots, b_n$ (cf. Remark 17.2.12). *Let the function f (as a function of $n+1$ variables) be continuous in an $(n+1)$-dimensional region Ω and let for any point $A \in \Omega$ there exist precisely one solution of equation* (1),

$$y = \varphi(x, a, b_1, b_2, \ldots, b_n),$$

satisfying the conditions

$$y(a) = b_1, \; y'(a) = b_2, \ldots, y^{(n-1)}(a) = b_n.$$

Then the function φ is a continuous function of all $n+2$ variables.

Moreover, if the function f has in Ω continuous partial derivatives of order r with respect to all the variables, then the function φ has continuous partial derivatives of order r (and therefore of all lower orders) with respect to all the variables.

REMARK 1. If the function f depends also on parameters $\lambda_1, \lambda_2, \ldots, \lambda_k$, then, of course, the function φ also depends on these parameters. If f is continuous in $\lambda_1, \lambda_2, \ldots, \lambda_k$, then the same holds for φ. A similar statement holds for derivatives with respect to $\lambda_1, \lambda_2, \ldots, \lambda_k$.

REMARK 2. The statement of Theorem 1 (or Remark 1) is very natural: If the function f is "well behaved", then, in the neighbourhood of the point a, integral curves of equation (1) change only slightly with small changes of initial conditions or of the parameters involved.

On questions concerning the stability of the solution see § 17.19.

17.10. Asymptotic Behaviour of Integrals of Differential Equations (for $x \to +\infty$). Oscillatory Solutions. Periodic Solutions

REMARK 1. The problem of asymptotic behaviour of integrals has been most thoroughly studied in the case of linear differential equations

$$y^{(n)} + f_{n-1}(x)y^{(n-1)} + \ldots + f_1(x)y' + f_0(x)y = f(x), \tag{1}$$

where existence of the solution is guaranteed in the whole interval I in which $f_k(x)$ and $f(x)$ are continuous functions.

Theorem 1. *Let $f_k(x)$ $(k = 0, 1, \ldots, n-1)$ and $f(x)$ be (real) continuous functions for $x > x_0$ and let, for $x \to +\infty$,*

$$f_k(x) \to a_k \quad (a_0 \neq 0), \qquad f(x) \to a. \tag{2}$$

Let the equation

$$\varrho^n + a_{n-1}\varrho^{n-1} + \ldots + a_1\varrho + a_0 = 0, \tag{3}$$

the coefficients of which are given by the limits (2), have only real, distinct roots. Then equation (1) has in the interval $(x_0, +\infty)$ at least one solution $y(x)$, which satisfies

$$y(x) \to \frac{a}{a_0}, \quad y^{(m)}(x) \to 0 \quad \text{for} \quad x \to +\infty \quad (m = 1, 2, \ldots, n). \tag{4}$$

If, in addition, all roots of equation (3) are negative, then (4) holds for all solutions of equation (1).

Example 1. Let us consider the equation

$$y'' - y = 3.$$

Here, $a_1 = 0$, $a_0 = -1$, $a = 3$. Equation (3),

$$\varrho^2 - 1 = 0,$$

has the roots $+1$ and -1. Thus, in the interval $(-\infty, +\infty)$, there exists at least one solution for which

$$y(x) \to -3, \quad y'(x) \to 0, \quad y''(x) \to 0 \quad \text{if} \quad x \to +\infty. \tag{5}$$

(In this example, this solution is obviously the function $y = -3 + \mathrm{e}^{-x}$; at the same time, another solution exists that does not have the property (5), namely $y = -3 + \mathrm{e}^{x}$.)

REMARK 2 (*Oscillatory Solutions of Linear Equations of the Second Order*). We shall consider only equations of the form

$$y'' + Q(x)y = 0, \tag{6}$$

where $Q(x)$ is a continuous function in an interval I, since any equation of the form

$$y'' + p(x)y' + q(x)y = 0$$

can be transformed (see Theorem 17.15.3) by the substitution

$$y = \mathrm{e}^{-\frac{1}{2}\int p\,\mathrm{d}x}\, z$$

into an equation of the form (6).

REMARK 3. *In the following text, the identically zero solution of equation* (6) *will be excluded from our considerations.*

Definition 1. The solution of equation (6), which has, in the given interval I, at most one zero point, is said to be *non-oscillatory* (*non-oscillating*) *in the interval* I; if it has at least two zero-points, it is said to be *oscillatory* (*oscillating*)/ or to *oscillate* in this interval.

Theorem 2. *If at all points of the interval* I *the inequality* $Q(x) \leqq 0$ *holds, then every solution of equation* (6) *is non-oscillatory in* I.

REMARK 4. *In the following text, we shall consider only the case where* $Q(x) > 0$ *in* I.

Theorem 3 (*Sturm's Theorem*). *Let* x_0 *and* x_1 *be two* (*distinct*) *consecutive zeros of a non-zero solution* $y_1(x)$ *of equation* (6). *Then every other solution* $y_2(x)$ *of this equation, which is not a multiple of the solution* $y_1(x)$, *has exactly one zero between the points* x_0, x_1.

REMARK 5. Roughly speaking: The zeros of two linear independent solutions of equation (6) are mutually alternating. The solutions $y_1 = \cos x$, $y_2 = \sin x$ of the equation $y'' + y = 0$ may be mentioned as a typical example. (The zeros of such pairs of functions are sometimes said to *interlace*.)

REMARK 6. If one solution of equation (6) has more than two (distinct) zeros in the interval I, then all solutions of equation (6) are oscillatory in I.

Theorem 4 (*The Comparison Theorem*). *Let us consider two equations of the form*

$$y'' + Q_1(x)y = 0, \tag{7}$$

$$z'' + Q_2(x)z = 0. \tag{8}$$

If the relation $Q_2(x) > Q_1(x) > 0$ *holds in* I, *then between any two (distinct) zeros* x_0, x_1 *of an arbitrary solution* $y(x)$ *of equation* (7), *there is at least one zero of every solution* $z(x)$ *of equation* (8). (*The conclusion of the theorem remains valid if in the interval* (x_0, x_1) *only the inequality* $Q_2(x) \geqq Q_1(x) > 0$ *holds and if at least in a subinterval of this interval the relation* $Q_2(x) > Q_1(x)$ *holds.*)

Theorem 5. *Let* x_0, x_1 $(x_0 < x_1)$ *be two consecutive zeros of the solution* $y(x)$ *of equation* (7). *Let the assumptions of Theorem* 4 *be satisfied (in the interval* (x_0, x_1) *it is sufficient if the conditions in brackets at the end of that theorem are satisfied). If, at the same time,* x_0 *is a zero point of the solution* $z(x)$ *of equation* (8), *then the next zero* x_2 *of this solution* $(x_2 > x_0)$ *lies to the left of the point* x_1 (*i.e.* $x_2 < x_1$).

REMARK 7. Theorems 4 and 5 express — roughly speaking — the fact that, if $Q_2(x) > Q_1(x) > 0$ in I, then the solutions of equation (8) oscillate more rapidly in I than the solutions of equation (7). The simplest example of this statement are the solutions of the equations

$$y'' + ay = 0, \quad z'' + bz = 0 \quad (0 < a < b).$$

Theorem 6. *If in the interval* I *the relation*

$$0 < m \leqq Q(x) \leqq M$$

holds for equation (6), *then the distance between two neighbouring zeros of the solution is not greater than* $\pi/\sqrt{m}$ *and not smaller than* $\pi/\sqrt{M}$.

Theorem 7 (*Kneser's Theorem*). *Let* $I = [x_0, \infty)$, $x_0 > 0$. *If in* I *the relation*

$$0 < Q(x) \leqq \frac{1}{4x^2}$$

holds, then the solution of equation (6) *cannot have an infinite number of zeros in* I. *If in* I *the relation*

$$Q(x) > \frac{1+\alpha}{4x^2}, \quad \textit{where} \quad \alpha > 0$$

holds, then the solution of equation (6) *has an infinite number of zeros in* I.

Theorem 8 (*Späthe's Theorem*). *Let the relation*

$$Q(x) = O\left(\frac{1}{x^{k+2}}\right) \quad (k > 0)$$

hold in the interval $I = [x_0, \infty)$ $(x_0 > 0)$. (This means (see Definition 11.4.6) that the expression $x^{k+2}|Q(x)|$ is bounded for all $x \in I$.) *Then equation* (6) *has a fundamental system of solutions* y_1, y_2 (Definition 17.11.2) *such that*

$$y_1(x) - 1 = O\left(\frac{1}{x^k}\right) \quad \textit{and} \quad y_2(x) - x = \begin{cases} O\left(\dfrac{1}{x^{k-1}}\right) & \textit{for} \quad k \neq 1, \\ O(\ln x) & \textit{for} \quad k = 1. \end{cases}$$

Example 2. Let us consider the equation

$$y'' + \frac{1}{x^4}y = 0.$$

According to Theorem 8, for large x the functions

$$y \equiv 1, \quad y \equiv x$$

constitute "nearly" the fundamental system of the given equation.

REMARK 8. For the equation of the n-th order,

$$y^{(n)} + g(x)y = 0, \tag{9}$$

the following assertion is true: If the function $g(x) > 0$ is continuous in $I = [x_0, \infty)$ and if

$$\int_{x_0}^{\infty} g(x)\,\mathrm{d}x \quad \text{is divergent,}$$

then 1° for any even n, every solution of equation (9) has an infinite number of zero points in I; 2° for any odd n, the solution has either an infinite number of zero points or it has zero as its limit as $x \to +\infty$.

REMARK 9. Remarkable results in the above-mentioned theory of oscillatory solutions, as well as in other fields concerning global properties of differential equations, have been achieved by applying the theory of transformations of these equations, developed by O. Borůvka and his school ([53]).

Periodic Solutions of Homogeneous Linear Equations

Theorem 9 (*Floquet's Theorem*). *Let us consider the equation*

$$y^{(n)} + f_{n-1}(x)y^{(n-1)} + \ldots + f_1(x)y' + f_0(x)y = 0 \tag{10}$$

in which $f_k(x)$ are holomorphic functions (of the complex variable x, Definition 20.1.9*) on the whole complex plane, and are periodic with a common period ω. Then, there exists at least one non-zero solution $\varphi(x)$ of equation* (10) *such that for a properly chosen constant s (complex, in general) the relation*

$$\varphi(x + \omega) = s\varphi(x) \tag{11}$$

holds. (The function $\varphi(x)$ with the property (11) is called *periodic of the second kind* or *pseudo-periodic.*) *The number α, determined by the equation*

$$s = \mathrm{e}^{\alpha\omega},$$

is called the characteristic exponent. The function

$$\psi(x) = \mathrm{e}^{-\alpha x}\,\varphi(x)$$

is then periodic with period ω.

REMARK 10. Finding the characteristic exponent is rather difficult, in general ([250], volume I). When solving an equation of the form

$$y'' = f(x)y \tag{12}$$

we can proceed in the following way: Let $y_1(x)$, $y_2(x)$ denote the *normal* (*standard*) fundamental system of equation (12) at the point $x = 0$, i.e. such for which the relations

$$y_1(0) = 1, \quad y_1'(0) = 0, \quad y_2(0) = 0, \quad y_2'(0) = 1 \tag{13}$$

are satisfied. The number s (see (11)) is then given by the solution of the equation

$$s^2 - \left[y_1(\omega) + y_2'(\omega)\right]s + 1 = 0.$$

(Since it may be difficult to construct the above-mentioned normal fundamental system even if the function $f(x)$ is relatively simple, the constants $y_1(\omega)$ and $y_2'(\omega)$ are frequently evaluated approximately by numerical integration starting from the known values (13).)

REMARK 11. In the case of linear *non-homogeneous* equations, existence of periodic solutions follows in some simple cases immediately from Theorem 17.14.1. For the case

$$y'' + a^2 y = g(x)$$

see e.g. [250], volume I.

17.11. Linear Equations of the n-th Order

Linear equations of the n-th order are equations of the form

$$y^{(n)} + a_{n-1}(x)y^{(n-1)} + \ldots + a_1(x)y' + a_0(x)y = f(x). \tag{1}$$

Throughout § 17.11, *the functions* $a_0(x), \ldots, a_{n-1}(x)$, $f(x)$ *are assumed to be continuous in an interval* I (which can be infinite). According to Remark 17.2.11 we know that if we choose an arbitrary number $x_0 \in I$ and n arbitrary numbers

$$y_0, y_0', \ldots, y_0^{(n-1)}, \tag{2}$$

then there exists exactly one solution of equation (1) satisfying the initial conditions

$$y(x_0) = y_0,\ y'(x_0) = y_0', \ldots, y^{(n-1)}(x_0) = y_0^{(n-1)}. \tag{3}$$

This solution is defined in the whole interval I.

REMARK 1. We draw the reader's attention to the fact that this conclusion is not true in general, when the linear equation is of the form

$$a_n(x)y^{(n)} + a_{n-1}(x)y^{(n-1)} + \ldots + a_1(x)y' + a_0(x)y = f(x)$$

and $a_n(x_1) = 0$ for some $x_1 \in I$. (As an example of an equation of this form, the equation $x^3y'' + y = 0$ in the interval $(-1, 1)$ may be mentioned.) However, if a_n is continuous and $a_n(x) \neq 0$ in the whole interval I, then, dividing through by $a_n(x)$, we get the previous case.

Definition 1. The equation

$$y^{(n)} + a_{n-1}(x)y^{(n-1)} + \ldots + a_1(x)y' + a_0(x)y = 0 \tag{4}$$

is called the *linear homogeneous equation corresponding to the non-homogeneous equation* (1).

Theorem 1. *If*

$$y_1(x), y_2(x), \ldots, y_k(x)$$

are solutions of equation (4), *then an arbitrary linear combination of them,*

$$y = c_1y_1 + c_2y_2 + \ldots + c_ky_k \tag{5}$$

(*where* $c_1, c_2, \ldots, c_k$ *are arbitrary numbers*), *is also a solution of equation* (4). *In particular,* $y_1 + y_2$ *and* $y_1 - y_2$ *are also solutions of the given equation.*

Theorem 2. *Let the functions*

$$f_1(x), f_2(x), \ldots, f_k(x) \tag{6}$$

have $k-1$ *continuous derivatives in the interval* I *and let they be linearly dependent in* I (§ 12.8). *Then the determinant*

$$W(x) = \begin{vmatrix} f_1(x), & f_2(x), & \ldots, & f_k(x) \\ f_1'(x), & f_2'(x), & \ldots, & f_k'(x) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ f_1^{(k-1)}(x), & f_2^{(k-1)}(x), & \ldots, & f_k^{(k-1)}(x) \end{vmatrix} \tag{7}$$

(the so-called *Wronskian determinant* (or briefly *Wronskian*) *of the functions* (6)) *is identically zero in* I. Often we write $W(f_1, f_2, \ldots, f_k)$.

Theorem 3. *Let the functions*

$$y_1(x), y_2(x), \ldots, y_n(x) \tag{8}$$

be solutions of equation (4), *linearly independent in the interval* I, *in which the coefficients* $a_0(x)$, $a_1(x)$, $\ldots$, $a_{n-1}(x)$ *of equation* (4) *are continuous functions of* x. *Then their Wronskian*

$$W(x) = \begin{vmatrix} y_1, & y_2, & \ldots, & y_n \\ y_1', & y_2', & \ldots, & y_n' \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ y_1^{(n-1)}, & y_2^{(n-1)}, & \ldots, & y_n^{(n-1)} \end{vmatrix} \tag{9}$$

is different from zero in the whole interval I.

REMARK 2. Thus, if we are investigating linear dependence or independence of functions (8), that are *solutions of equation* (4), it is sufficient to evaluate the determinant (9) at one point $x_0 \in I$ only. If $W(x_0) = 0$, the solutions (8) are linearly dependent in I, if $W(x_0) \neq 0$, they are linearly independent.

Example 1. The functions

$$y_1 = \mathrm{e}^x, \quad y_2 = \mathrm{e}^{-x} \tag{10}$$

are, as can be easily verified, solutions of the equation

$$y'' - y = 0 \tag{11}$$

the coefficients of which are constants. (As the interval I, we may therefore take the interval $(-\infty, \infty)$.) The Wronskian of the functions (10) is

$$W(x) = \begin{vmatrix} \mathrm{e}^x, & \mathrm{e}^{-x} \\ \mathrm{e}^x, & -\mathrm{e}^{-x} \end{vmatrix} = -2. \tag{12}$$

Thus, the functions (10) are linearly independent in the interval $(-\infty, \infty)$.

REMARK 3. Here, the evaluation of the determinant was easy, so that we could easily calculate its value for any x. However, according to Remark 2 it is sufficient to evaluate $W(x)$ at one point $x_0 \in I$ only, say, at the point $x_0 = 0$ in this case.

Theorem 3 and Remark 2 follow from the so-called *Liouville's formula* (or *Abel identity*): If the coefficients of equation (4) are continuous in I, $y_1(x)$, $y_2(x)$, $\ldots$ $\ldots$, $y_n(x)$ are arbitrary solutions of equation (4) and x_0 is an arbitrary point of I, then the relation

$$W(x) = W(x_0)\, \mathrm{e}^{-\int_{x_0}^{x} a_{n-1}(t)\, \mathrm{d}t} \tag{13}$$

holds. Since $e^z \neq 0$ for arbitrary z, $W(x)$ is zero or different from zero in I according to whether $W(x_0) = 0$ or $W(x_0) \neq 0$, respectively.

Example 2. The functions (10) are solutions of equation (11) in the interval $(-\infty, \infty)$. In (11) $a_{n-1}(x) \equiv a_1(x) = 0$, thus

$$W(x) = W(x_0)\, e^{-\int_{x_0}^{x} 0.dt} = W(x_0) = \text{const.},$$

as may also be seen from (12).

Definition 2. A system of solutions $y_1(x), y_2(x), \ldots, y_n(x)$ of equation (4) which are linearly independent in the interval I, is called a *fundamental system of solutions* (briefly a *fundamental system*, or a *basis*) of equation (4) (in that interval).

For example, the functions (10) form a fundamental system of solutions of equation (11).

Theorem 4. *A fundamental system*

$$y_1(x), y_2(x), \ldots, y_n(x)$$

of equation (4) *in the interval* I *having been found, every solution* $y(x)$ *of this equation in this interval can be written in the form*

$$y(x) = c_1 y_1(x) + c_2 y_2(x) + \ldots + c_n y_n(x), \tag{14}$$

where $c_1, c_2, \ldots, c_n$ *are suitable constants.*

REMARK 4. In details: If we find n functions $y_1(x), y_2(x), \ldots, y_n(x)$ which

(i) are solutions of the homogeneous linear equation (4) of the n-th order in the interval I,

(ii) are linearly independent in I,

then every solution of this equation in this interval can be expressed as a linear combination of these functions. We also say that *all solutions of equation* (4) (*in* I) *form an* n*-dimensional linear space*, the basis of which is formed by arbitrary n linearly independent solutions of this equation.

In accordance with Definition 17.2.5 and Remark 17.2.14 we call (14) the *general integral of equation* (4).

Example 3. The functions

$$y_1 = \cos x, \quad y_2 = \sin x \tag{15}$$

are solutions of the homogeneous equation of the second order

$$y'' + y = 0 \tag{16}$$

in the interval $(-\infty, \infty)$.

We easily verify that (15) is the fundamental system of equation (16), because

$$W(x) = \begin{vmatrix} \cos x, & \sin x \\ -\sin x, & \cos x \end{vmatrix} = 1 \neq 0;$$

thus the functions (15) are linearly independent in the interval $(-\infty, \infty)$. The general integral of equation (16) is then

$$y = c_1 \cos x + c_2 \sin x. \tag{17}$$

REMARK 5. In applications we frequently have to solve the following problem: Given a differential equation of the form (4), to find a solution $y(x)$ of this equation satisfying, at a given point $x_0 \in I$, the initial conditions

$$y(x_0) = y_0,\ y'(x_0) = y_0',\ \ldots,\ y^{(n-1)}(x_0) = y_0^{(n-1)} \tag{18}$$

($y_0, y_0', \ldots, y_0^{(n-1)}$ being prescribed numbers). If the general integral of equation (4) is known, then the required solution $y(x)$ is of the form (14), where the constants $c_1, c_2, \ldots, c_n$ are uniquely determined by the initial conditions (18).

Example 4. Let us find the solution of the equation

$$y'' + y = 0 \tag{19}$$

satisfying the initial conditions

$$y(0) = 1, \quad y'(0) = -2. \tag{20}$$

The solution will be of the form (17) (see Example 3). Substituting $x = 0$ into (17) and into the equation $y' = -c_1 \sin x + c_2 \cos x$, which arises by differentiating (17) with respect to x, we obtain from conditions (20)

$$c_1 \,.\, 1 + c_2 \,.\, 0 = 1, \quad -c_1 \,.\, 0 + c_2 \,.\, 1 = -2;$$

hence $c_1 = 1$, $c_2 = -2$, and the solution is

$$y = \cos x - 2 \sin x.$$

(In the general case, the determination of the constants $c_1, c_2, \ldots, c_n$ leads to the solution of a system of n linear (algebraic) equations with a non-zero determinant by Theorem 3.)

Theorem 5. *For every equation* (4) *with continuous coefficients in* I, *there exists at least one fundamental system in* I. (*In fact there is an infinite number of them.*)

Definition 3. The fundamental system $y_1(x), y_2(x), \ldots, y_n(x)$ for which the relations

$$\begin{aligned}
&y_1(x_0) = 1, \quad y_1'(x_0) = 0, \quad y_1''(x_0) = 0, \quad \ldots, \quad y_1^{(n-1)}(x_0) = 0,\\
&y_2(x_0) = 0, \quad y_2'(x_0) = 1, \quad y_2''(x_0) = 0, \quad \ldots, \quad y_2^{(n-1)}(x_0) = 0,\\
&y_3(x_0) = 0, \quad y_3'(x_0) = 0, \quad y_3''(x_0) = 1, \quad \ldots, \quad y_3^{(n-1)}(x_0) = 0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&y_n(x_0) = 0, \quad y_n'(x_0) = 0, \quad y_n''(x_0) = 0, \quad \ldots, \quad y_n^{(n-1)}(x_0) = 1,
\end{aligned} \tag{21}$$

hold, is called the *normal* (or *standard*) *fundamental system of equation* (4) *at the point* x_0 (or *with respect to the point* x_0).

The normal fundamental system of the given equation being known, it is easy to find the solution of this equation satisfying conditions (18):

Theorem 6. *If* $y_1(x), y_2(x), \ldots, y_n(x)$ *is the normal fundamental system of equation* (4) *at the point* x_0, *then the function*

$$y = y_0 y_1(x) + y_0' y_2(x) + y_0'' y_3(x) + \ldots + y_0^{(n-1)} y_n(x)$$

is the solution of equation (4) *that satisfies the initial conditions*

$$y(x_0) = y_0, \; y'(x_0) = y_0', \ldots, y^{(n-1)}(x_0) = y_0^{(n-1)}.$$

REMARK 6. In the general case, it need not be easy to find a fundamental system of equation (4). However, in frequently occurring case where $a_0, a_1, \ldots, a_{n-1}$ are constants, the problem is simple (see § 17.13).

17.12. Non-homogeneous Linear Equations. The Method of Variation of Parameters

Theorem 1. *If a fundamental system* $y_1(x), y_2(x), \ldots, y_n(x)$ *of the equation*

$$y^{(n)} + a_{n-1}(x) y^{(n-1)} + \ldots + a_1(x) y' + a_0(x) y = 0 \tag{1}$$

is known, then the general integral of the corresponding non-homogeneous equation

$$y^{(n)} + a_{n-1}(x) y^{(n-1)} + \ldots + a_1(x) y' + a_0(x) y = g(x) \tag{2}$$

is of the form

$$y = y_p + c_1 y_1 + c_2 y_2 + \ldots + c_n y_n, \tag{3}$$

where $c_1, c_2, \ldots, c_n$ are (arbitrary) parameters and $y_p(x)$ is any function satisfying equation (2).

Theorem 2. *The function $y_p(x)$ can be obtained by quadratures (by the so-called method of variation of parameters or variation of constants) when assumed to be of the form*

$$y_p(x) = c_1(x)y_1(x) + c_2(x)y_2(x) + \ldots + c_n(x)y_n(x), \tag{4}$$

where

$$y_1(x), y_2(x), \ldots, y_n(x) \tag{5}$$

is the fundamental system of equation (1).

REMARK 1. Thus, $y_p(x)$ has the form of the general integral of equation (1), where however, instead of constants $c_1, c_2, \ldots, c_n$, we have functions which at the moment are unknown. We have to find these functions so that the function (4) satisfies equation (2).

Theorem 3. *If the functions $c_1(x), c_2(x), \ldots, c_n(x)$ satisfy the equations*

$$\begin{aligned}
&c_1'y_1 + c_2'y_2 + \ldots + c_n'y_n = 0,\\
&c_1'y_1' + c_2'y_2' + \ldots + c_n'y_n' = 0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&c_1'y_1^{(n-2)} + c_2'y_2^{(n-2)} + \ldots + c_n'y_n^{(n-2)} = 0,\\
&c_1'y_1^{(n-1)} + c_2'y_2^{(n-1)} + \ldots + c_n'y_n^{(n-1)} = g(x),
\end{aligned} \tag{6}$$

then the function y_p, given by (4), *satisfies equation* (2).

REMARK 2. The system (6) for the unknown functions $c_1'(x), c_2'(x), \ldots, c_n'(x)$ is uniquely solvable because its determinant is the Wronskian of the fundamental system (5), and is thus different form zero in the interval considered. Integrating $c_1'(x), c_2'(x), \ldots, c_n'(x)$, we get the functions $c_1(x), c_2(x), \ldots, c_n(x)$. (We can write them without constants of integration, because these constants appear in the remaining terms of (3).)

Example 1. Let us find the general integral of the equation

$$y'' + y = x^2. \tag{7}$$

The fundamental system of the equation

$$y'' + y = 0 \tag{8}$$

is

$$y_1 = \cos x, \quad y_2 = \sin x \tag{9}$$

(see Example 17.11.3). The general integral of equation (7) will be of the form (3),

$$y = y_p + c_1 \cos x + c_2 \sin x, \tag{10}$$

where y_p may be found in the form (4),

$$y_p = c_1(x)y_1(x) + c_2(x)y_2(x). \tag{11}$$

The system (6) is

$$c_1' \cos x + c_2' \sin x = 0,$$
$$-c_1' \sin x + c_2' \cos x = x^2,$$

whence

$$c_1' = -x^2 \sin x, \quad c_2' = x^2 \cos x.$$

Integration by parts yields

$$c_1(x) = (x^2 - 2)\cos x - 2x \sin x, \quad c_2(x) = (x^2 - 2)\sin x + 2x \cos x. \tag{12}$$

Substituting (12) into (11), we get

$$y_p(x) = (x^2 - 2)\cos^2 x - 2x \sin x \cos x + (x^2 - 2)\sin^2 x + 2x \cos x \sin x = x^2 - 2.$$

The general integral of equation (7) is then, according to (10),

$$y = x^2 - 2 + c_1 \cos x + c_2 \sin x.$$

REMARK 3. The right-hand side of equation (2) is often of a special form. For example, $g(x)$ may be a polynomial, as it was in the above example. If the left-hand side of equation (2) has constant coefficients, then, for some special forms of the right-hand side of equation (2), y_p can be found in a much simpler way than by the method of variation of parameters (see § 17.14).

17.13. Homogeneous Linear Equations with Constant Coefficients. Euler's Equation

Let us consider the equation

$$y^{(n)} + a_{n-1}y^{(n-1)} + \ldots + a_1 y' + a_0 y = 0, \tag{1}$$

where $a_0, a_1, \ldots, a_{n-1}$ are constants (complex, in general). Assuming the solution of equation (1) in the form

$$y = e^{\alpha x},$$

we obtain for the number α (after substituting for $y, y', \ldots, y^{(n)}$ into (1) and dividing the whole equation by $e^{\alpha x}$) the so-called *characteristic* (or *auxiliary*) *equation*

$$\alpha^n + a_{n-1}\alpha^{n-1} + \ldots + a_1\alpha + a_0 = 0. \tag{2}$$

I. If the characteristic equation has distinct roots

$$\alpha_1, \alpha_2, \ldots, \alpha_n,$$

then the fundamental system of equation (1) is given by the functions (complex, in general)

$$y_1 = e^{\alpha_1 x}, y_2 = e^{\alpha_2 x}, \ldots, y_n = e^{\alpha_n x}. \tag{3}$$

II. If α_i is a root of multiplicity r, then r functions correspond to it in the fundamental system. These r functions are

$$y_1 = e^{\alpha_i x}, y_2 = x e^{\alpha_i x}, \ldots, y_r = x^{r-1} e^{\alpha_i x}. \tag{4}$$

Example 1. Let us consider the equation

$$y''' - 3y' + 2y = 0. \tag{5}$$

The characteristic equation

$$\alpha^3 - 3\alpha + 2 = 0 \tag{6}$$

has obviously the root $\alpha = 1$. After dividing by $\alpha - 1$, we get

$$\alpha^2 + \alpha - 2 = 0$$

with the roots 1, -2. Thus, equation (6) has a simple root $\alpha_1 = -2$ and a double root $\alpha_2 = 1$. To α_1 there corresponds the function $y_1 = e^{-2x}$, to α_2 there correspond two functions (because $r = 2$), $y_2 = e^x$, $y_3 = x e^x$. Thus the fundamental system of equation (5) is

$$y_1 = e^{-2x}, \quad y_2 = e^x, \quad y_3 = x e^x.$$

REMARK 1. The roots of the characteristic equation need not always be real. Let the coefficients of equation (1) (and thus also of equation (2)) be *real*; then, as is well known from algebra, if the characteristic equation has a complex root $a + ib$ (a, b real), it also has the complex conjugate root $a - ib$, and both these roots are

of the same multiplicity. (If, for example, $a + ib$ is a double root, $a - ib$ is also a double root, etc.)

Let the root $a + ib$ (and hence also $a - ib$) be a simple root of equation (2). Then to the two roots

$$a + ib, \quad a - ib$$

there correspond, in the fundamental system, two (real) functions

$$e^{ax} \cos bx, \quad e^{ax} \sin bx. \tag{7}$$

Let $a + ib$ (and hence also $a - ib$) be a double root of equation (2). Then four functions correspond to these roots in the fundamental system:

$$e^{ax} \cos bx, \quad e^{ax} \sin bx, \quad x e^{ax} \cos bx, \quad x e^{ax} \sin bx; \tag{8}$$

if $a + ib$ is an r-fold root, then $2r$ functions correspond to the two roots $a + ib$, $a - ib$:

$$\begin{aligned} &e^{ax} \cos bx, \quad x e^{ax} \cos bx, \quad \ldots, \quad x^{r-1} e^{ax} \cos bx, \\ &e^{ax} \sin bx, \quad x e^{ax} \sin bx, \quad \ldots, \quad x^{r-1} e^{ax} \sin bx. \end{aligned} \tag{9}$$

Example 2. Let us consider the equation

$$y^{(V)} - y^{(IV)} + 2y''' - 2y'' + y' - y = 0.$$

The characteristic equation

$$\alpha^5 - \alpha^4 + 2\alpha^3 - 2\alpha^2 + \alpha - 1 = 0$$

has obviously the root $\alpha_1 = 1$. Dividing by the factor $\alpha - 1$, we get

$$\alpha^4 + 2\alpha^2 + 1 = 0$$

or

$$(\alpha^2 + 1)^2 = 0$$

with double roots $\alpha_{2,3} = \mathrm{i}$, $\alpha_{4,5} = -\mathrm{i}$. The fundamental system, written in the complex form, is (according to (4))

$$y_1 = e^x, \quad y_2 = e^{ix}, \quad y_3 = x e^{ix}, \quad y_4 = e^{-ix}, \quad y_5 = x e^{-ix},$$

written in the real form (according to (8)); $a = 0$, $b = 1$)

$$y_1 = e^x, \quad y_2 = \cos x, \quad y_3 = \sin x, \quad y_4 = x \cos x, \quad y_5 = x \sin x.$$

The general integral (written in the real form) is

$$y = c_1 e^x + c_2 \cos x + c_3 \sin x + c_4 x \cos x + c_5 x \sin x.$$

REMARK 2. Solving the characteristic equation involves, in general, solving an algebraic equation of the n-th degree. For methods of solution of these equations see Chap. 31. We often succeed in finding one root in a simple way (e.g. we look to see if one root of the characteristic equation may perhaps be an integer, as in the case of Example 1 or of Example 2, where a negative reciprocal equation was to be solved). Then dividing by the corresponding linear factor, we reduce the order of the characteristic equation.

REMARK 3. By substituting $x = e^t$, the so-called *Euler equation*

$$y^{(n)} + \frac{a_{n-1}}{x} y^{(n-1)} + \frac{a_{n-2}}{x^2} y^{(n-2)} + \ldots + \frac{a_1}{x^{n-1}} y' + \frac{a_0}{x^n} y = 0 \tag{10}$$

(where $a_0, a_1, \ldots, a_{n-1}$ are constants) can be transformed into an equation with constant coefficients (for $x > 0$; for negative x, we use the substitution $x = -e^t$). The procedure will be shown in the following example.

Example 3. Let us consider the equation

$$x^2 y'' + 3xy' + y = 0. \tag{11}$$

Then (cf. § 12.11)

$$\frac{dy}{dx} = \frac{dy}{dt}\frac{dt}{dx} = \frac{dy}{dt}\frac{1}{dx/dt} = \frac{dy}{dt}\frac{1}{e^t} = e^{-t}\frac{dy}{dt},$$

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d}{dt}\left(\frac{dy}{dx}\right)\frac{dt}{dx} = \frac{d}{dt}\left(e^{-t}\frac{dy}{dt}\right)e^{-t} = e^{-2t}\left(\frac{d^2y}{dt^2} - \frac{dy}{dt}\right).$$

Substituting these results into (11), we get

$$e^{2t}e^{-2t}\left(\frac{d^2y}{dt^2} - \frac{dy}{dt}\right) + 3e^t e^{-t}\frac{dy}{dt} + y = 0$$

or

$$\frac{d^2y}{dt^2} + 2\frac{dy}{dt} + y = 0 \tag{12}$$

which is an equation with constant coefficients for the unknown function $y(t)$. The characteristic equation

$$\alpha^2 + 2\alpha + 1 = 0$$

has exactly one double root $\alpha = -1$, so, according to (4), the general integral is

$$y = c_1 \mathrm{e}^{-t} + c_2 t \mathrm{e}^{-t} = \mathrm{e}^{-t}(c_1 + c_2 t).$$

From the substitution $x = \mathrm{e}^t$ if follows that $t = \ln x$, thus the general integral of equation (11) is

$$y = \frac{1}{x}(c_1 + c_2 \ln x). \tag{13}$$

In the case of Euler equations of higher order we proceed in a similar way. Instead of transforming the equation into the form (12), we can directly assume the solution of equation (10) in the form

$$y = x^\alpha. \tag{14}$$

The determination of α leads to the solution of a characteristic equation for α. If the roots $\alpha_1, \alpha_2, \ldots, \alpha_n$ are simple, then the fundamental system of equation (10) is

$$y_1 = x^{\alpha_1}, y_2 = x^{\alpha_2}, \ldots, y_n = x^{\alpha_n}.$$

If one of the roots is of multiplicity r, then the corresponding r functions of the fundamental system are

$$x^\alpha, x^\alpha \ln x, x^\alpha \ln^2 x, \ldots, x^\alpha \ln^{r-1} x. \tag{15}$$

The complex functions

$$x^{a+\mathrm{i}b}, \; x^{a-\mathrm{i}b} \quad \text{or} \quad x^{a+\mathrm{i}b} \ln^k x, \; x^{a-\mathrm{i}b} \ln^k x$$

can be replaced by real functions

$$x^a \cos(b \ln x), \; x^a \sin(b \ln x) \quad \text{or} \quad x^a \cos(b \ln x) \ln^k x, \; x^a \sin(b \ln x) \ln^k x.$$

Example 4. Substituting (14) into (11), we get

$$x^2 \,.\, \alpha(\alpha - 1)x^{\alpha-2} + 3x\alpha x^{\alpha-1} + x^\alpha = 0.$$

Dividing through by $x^\alpha \neq 0$, we get the characteristic equation

$$\alpha^2 + 2\alpha + 1 = 0$$

with a double root $\alpha_{1,2} = -1$. According to (15), the fundamental system is

$$y_1 = x^{-1} = \frac{1}{x}, \quad y_2 = x^{-1} \ln x = \frac{1}{x} \ln x,$$

in agreement with (13).

17.14. Non-homogeneous Linear Equations with Constant Coefficients and a Special Right-hand Side

Let us consider the equation

$$y^{(n)} + a_{n-1}y^{(n-1)} + \ldots + a_1y' + a_0y = \mathrm{e}^{ax}\left[P(x)\cos bx + Q(x)\sin bx\right], \qquad (1)$$

where a_0, a_1, ..., a_{n-1}, a, b are real constants, $P(x)$ is a polynomial of the p-th degree and $Q(x)$ is a polynomial of the q-th degree, both with real coefficients. (The special cases $P(x) \equiv 0$ or $Q(x) \equiv 0$ are not excluded.)

Equation (1) covers the great majority of differential equations which occur in applications. For example, the equation

$$y'' + y = \mathrm{e}^x \sin x$$

is of type (1). Here $a_1 = 0$, $a_0 = 1$, $a = 1$, $b = 1$, $P(x) \equiv 0$, $Q(x) \equiv 1$ (therefore $q = 0$). The equation

$$y'' - 3y' + 2y = x^2$$

is also of the type (1). Here $a_1 = -3$, $a_0 = 2$, $a = 0$, $b = 0$, $P(x) = x^2$ (therefore $p = 2$); $Q(x)$ can be considered as a zero polynomial, $Q(x) \equiv 0$.

The general integral of equation (1) is of the form (cf. Theorem 17.12.1)

$$y = y_p + c_1y_1 + c_2y_2 + \ldots + c_ny_n,$$

where y_1, y_2, ..., y_n is the fundamental system of the homogeneous equation

$$y^{(n)} + a_{n-1}y^{(n-1)} + \ldots + a_1y' + a_0y = 0 \qquad (2)$$

(which can be found by the method mentioned in the previous paragraph) and y_p is an arbitrary solution of equation (1). If the right-hand side of the equation is of the form indicated in (1), it is possible to assume the function y_p to be of a special form:

The characteristic equation corresponding to equation (2) is

$$\alpha^n + a_{n-1}\alpha^{n-1} + \ldots + a_1\alpha + a_0 = 0. \qquad (3)$$

Let us denote by s the greater of the two numbers p, q, where p is the degree of the polynomial $P(x)$ and q is the degree of the polynomial $Q(x)$ in (1). (If $p = q$, then obviously $s = p = q$. If e.g. $Q(x) \equiv 0$, we shall consider $s = p$.)

Theorem 1. *Suppose* $a + ib$ *is not a root of equation* (3) (*so that neither is* $a - ib$ *a root of equation* (3)). *Then* y_p *can be assumed to be of the form*

$$y_p = e^{ax}\left[R(x)\cos bx + S(x)\sin bx\right], \tag{4}$$

where $R(x)$ *and* $S(x)$ *are polynomials of the* s*-th degree.*

If $a + ib$ *is an* r*-fold root of the characteristic equation, then* y_p *can be assumed to be of the form*

$$y_p = x^r e^{ax}\left[R(x)\cos bx + S(x)\sin bx\right], \tag{5}$$

where again $R(x)$ *and* $S(x)$ *are polynomials of the* s*-th degree.*

REMARK 1. The coefficients of these polynomials can be determined by the method of undetermined coefficients, as will be clear from the examples given below.

REMARK 2. If the right-hand side of equation (1) is a sum of terms of the form given in (1), then y_p is a sum of functions of the form (4) or (5). (It is often advantageous to find the integral corresponding to each term of the right-hand side separately; their sum then gives y_p.)

For example, consider the equation

$$y'' + 4y = 2\sin x + \cos 3x.$$

The right-hand side of this equation cannot be expressed in the form

$$e^{ax}\left[P(x)\cos bx + Q(x)\sin bx\right],$$

because for each term of the right-hand side of this equation the value of b is different. So we find first the particular integral y_{p_1} of the equation

$$y'' + 4y = 2\sin x$$

and then the particular integral y_{p_2} of the equation

$$y'' + 4y = \cos 3x.$$

Their sum $y_{p_1} + y_{p_2}$ is the required particular integral y_p.

Example 1. Let us consider the equation

$$y'' + y = x^2. \tag{6}$$

If we use the notation of (1), then $a = 0$, $b = 0$ (whence $a + ib = 0$), $P(x) = x^2$, $Q(x) \equiv 0$ (therefore $s = 2$); $a + ib$ is not a root of the characteristic equation

$\alpha^2 + 1 = 0$ (because this has the roots $\alpha_1 = \mathrm{i}$, $\alpha_2 = -\mathrm{i}$ while $a + ib = 0$), so according to (4) we may assume y_p to be of the form

$$y_p = \mathrm{e}^{0.x}\left[(Ax^2 + Bx + C)\cos(0 \,.\, x) + (Dx^2 + Ex + F)\sin(0 \,.\, x)\right] = \\ = Ax^2 + Bx + C.$$

Hence $y_p'' = 2A$, $y_p = Ax^2 + Bx + C$; substituting into (6), we get

$$2A + Ax^2 + Bx + C = x^2.$$

Comparing the coefficients of equal powers of x, we get $A = 1$, $B = 0$, $2A + C = 0$, so that

$$y_p = x^2 - 2$$

in accordance with example 17.12.1; the general integral of equation (6) is therefore

$$y = x^2 - 2 + c_1 \cos x + c_2 \sin x.$$

REMARK 3. It may be easily verified that the equation $y'' + y' = x^2$ has no particular solution of the form $Ax^2 + Bx + C$; the function y_p is to be obtained in the form $x(Ax^2 + Bx + C)$, in accordance with (5). The difference between this case and Example 1 lies in the fact that here $a + ib = 0$ *is* a simple root of the characteristic equation $\alpha^2 + \alpha = 0$.

Example 2. Let us consider the equation

$$y'' - 3y' + 2y = x\,\mathrm{e}^x\,. \tag{7}$$

Here $a = 1$, $b = 0$ (thus $a + ib = 1$), $P(x) = x$ (therefore $p = 1$), $Q(x) \equiv 0$ (thus $s = 1$). The characteristic equation

$$\alpha^2 - 3\alpha + 2 = 0$$

has the roots $\alpha_1 = 1$, $\alpha_2 = 2$, thus $a + ib = 1$ is a simple root of this equation (i.e. in (5), we have $r = 1$). Following (5), we assume y_p in the form

$$y_p = x\,\mathrm{e}^x\left[(Ax + B)\cos(0 \,.\, x) + (Cx + D)\sin(0 \,.\, x)\right] = x\,\mathrm{e}^x(Ax + B) = \\ = \mathrm{e}^x(Ax^2 + Bx).$$

Substituting for y_p, y_p', y_p'' into (7), we get

$$\mathrm{e}^x\left[x^2(2A - 3A + A) + x(B + 4A - 3B - 6A + 2B) + (2B + 2A - 3B)\right] = \mathrm{e}^x \,.\, x.$$

Dividing by e^x ($\neq 0$) and comparing the coefficients of equal powers of x we get

$$-2A = 1, \quad 2A - B = 0 \quad \text{whence} \quad A = -\tfrac{1}{2}, \quad B = -1,$$

thus

$$y_p = e^x\left(-\tfrac{1}{2}x^2 - x\right),$$

and the general integral of equation (7) is

$$y = e^x\left(-\tfrac{1}{2}x^2 - x\right) + c_1 e^x + c_2 e^{-2x} = e^x\left(-\tfrac{1}{2}x^2 - x + c_1\right) + c_2 e^{-2x}.$$

Example 3. Let us consider the equation

$$y'' + y = 2\sin x. \tag{8}$$

Here $a = 0$, $b = 1$ (thus $a + ib = i$), $P(x) \equiv 0$, $Q(x) = 2$ (so $q = 0$ and $s = 0$); $a + ib = i$ is a *simple* root of the characteristic equation

$$\alpha^2 + 1 = 0.$$

According to (5) we have

$$y_p = x e^{0 \cdot x}\left[A\cos x + B\sin x\right] = Ax\cos x + Bx\sin x.$$

Substituting for y_p, y_p'' into (8), we get

$$-2A\sin x - Ax\cos x + 2B\cos x - Bx\sin x + Ax\cos x + Bx\sin x = 2\sin x,$$
$$-2A\sin x + 2B\cos x = 2\sin x.$$

Comparing coefficients of the linearly independent functions $\sin x$, $\cos x$ we get

$$-2A = 2, \quad 2B = 0, \quad \text{i.e.} \quad A = -1, \quad B = 0;$$

consequently

$$y_p = -x\cos x,$$

and the general integral of equation (8) is

$$y = -x\cos x + c_1\cos x + c_2\sin x = (-x + c_1)\cos x + c_2\sin x.$$

From this example it is clear that even in the case when only a sine term appears on the right-hand side of equation (1) (in this case, of equation (8)), when searching for y_p we must write the expression (4) (or (5)) in the complete form containing both sine and cosine terms.

On harmonic, damped, undamped and forced oscillations see in § 4.13 and § 17.21 (eqs. 90, 91, 92, 97, 98, 99).

17.15. Linear Equations of the Second Order with Variable Coefficients. Transformation into Self-adjoint Form, into Normal Form. Invariant. Equations with Regular Singularities (Equations of the Fuchsian Type). Some Special Equations (Bessel's Equation etc.)

In this paragraph equations of the type

$$p_0(x)y'' + p_1(x)y' + p_2(x)y = 0 \tag{1}$$

are considered, where $p_0(x)$, $p_1(x)$, $p_2(x)$ are continuous functions of the variable x on an interval I (which can be infinite). Unlike § 17.11 we *do not assume* that $p_0(x) \equiv 1$, which leads in some cases to different results from those given in the paragraph mentioned.

Theorem 1. *In any interval in which* $p_0(x) \neq 0$, *equation* (1) *can be transformed, by multiplying by the function*

$$\mu(x) = \frac{1}{p_0(x)} e^{\int [p_1(x)/p_0(x)]\,dx}, \tag{2}$$

into the so-called self-adjoint form

$$\frac{d}{dx}[p(x)y'] + q(x)y = 0, \tag{3}$$

where

$$p(x) = e^{\int (p_1/p_0)\,dx}, \quad q(x) = \frac{p_2}{p_0} e^{\int (p_1/p_0)\,dx}. \tag{4}$$

Example 1. For the Bessel equation

$$x^2y'' + xy' + (x^2 - \nu^2)y = 0, \tag{5}$$

$p_0 = x^2$, $p_1 = x$, $p_2 = x^2 - \nu^2$, thus (carrying through the transformation for $x > 0$)

$$\mu = \frac{1}{x^2} e^{\int (x/x^2)\,dx} = \frac{1}{x^2} e^{\ln x} = \frac{1}{x}, \quad p = e^{\int (x/x^2)\,dx} = x,$$

$$q = \frac{x^2 - \nu^2}{x^2} e^{\int (x/x^2)\,dx} = \frac{x^2 - \nu^2}{x};$$

thus equation (5), multiplied by the function $1/x$, is transformed into the self-adjoint equation

$$(xy')' + \frac{x^2 - \nu^2}{x} y = 0. \tag{6}$$

Theorem 2. *Introducing a new variable*

$$u(x) = \int e^{-\int (p_1/p_0)\,dx}\,dx, \tag{7}$$

equation (1) *can be transformed, in any interval in which* $p_0(x) \neq 0$, *into the form*

$$\frac{d^2y}{du^2} + Q(u)y = 0, \tag{8}$$

where

$$Q = \frac{p_2}{p_0} e^{2\int (p_1/p_0)\,dx} \tag{9}$$

and where, on the right-hand side of (9), u *is to be substituted for* x *according to the relation* (7).

Example 2. In the equation

$$xy'' + \tfrac{1}{2}y' - y = 0, \tag{10}$$

$p_0 = x$, $p_1 = \frac{1}{2}$, $p_2 = -1$. Then (for $x > 0$)

$$u = \int e^{-\int (p_1/p_0)\,dx}\,dx = \int e^{-\frac{1}{2}\ln x}\,dx = \int \frac{1}{\sqrt{x}}\,dx = 2\sqrt{x}\,, \tag{11}$$

$$Q = -\frac{1}{x} e^{2\int (1/(2x))\,dx} = -\frac{1}{x} e^{\ln x} = -1.$$

(Here Q is a constant. In the general case, we should substitute $x = \frac{1}{4}u^2$ in accordance with (11).) Consequently, equation (8) becomes

$$\frac{d^2y}{du^2} - y = 0. \tag{12}$$

Theorem 3. *By a transformation of the form*

$$y = u(x)z, \tag{13}$$

where

$$u(x) = e^{-\frac{1}{2}\int (p_1/p_0)\,dx}, \tag{14}$$

and by dividing by $p_0(x)$ *we can transform equation* (1) (*in any interval in which* $p_0 \neq 0$) *into the form*

$$z'' + I(x)z = 0, \tag{15}$$

where

$$I(x) = \frac{p_2}{p_0} - \frac{1}{4}\left(\frac{p_1}{p_0}\right)^2 - \frac{1}{2}\left(\frac{p_1}{p_0}\right)'. \tag{16}$$

Example 3. For the equation

$$y'' + \frac{2}{x}y' + y = 0 \tag{17}$$

we have $p_0 = 1$, $p_1 = 2/x$, $p_2 = 1$. Further

$$u(x) = \mathrm{e}^{-\frac{1}{2}\int (2/x)\,\mathrm{d}x} = \mathrm{e}^{-\ln x} = \frac{1}{x}, \tag{18}$$

giving

$$I(x) = 1 - \frac{1}{4}\left(\frac{2}{x}\right)^2 - \frac{1}{2}\left(\frac{2}{x}\right)' = 1 - \frac{1}{x^2} + \frac{1}{x^2} = 1. \tag{19}$$

Consequently, equation (15) becomes

$$z'' + z = 0. \tag{20}$$

REMARK 1. Forms (8) and (15) are often called *normal* (or *normalized*) forms of the differential equation of the second order. The expression $I(x)$ is the so-called *invariant* of the given equation. By transforming an equation into the normal form, we often obtain the solution in a simple way, as may be seen from Examples 2 and 3: Equation (12) has the general integral

$$y = c_1\,\mathrm{e}^{u} + c_2\,\mathrm{e}^{-u}$$

whence, applying (11), we get the general integral of equation (10):

$$y = c_1\,\mathrm{e}^{2\sqrt{x}} + c_2\,\mathrm{e}^{-2\sqrt{x}}.$$

The general integral of equation (20) is

$$z = c_1 \cos x + c_2 \sin x,$$

whence, applying (13) and (18), we get the general integral of equation (17):

$$y = c_1\frac{\cos x}{x} + c_2\frac{\sin x}{x}.$$

Definition 1. Equations of the form

$$(x-a)^2y'' + (x-a)P_1(x)y' + P_2(x)y = 0, \tag{21}$$

where $P_1(x)$, $P_2(x)$ are functions which can be expanded into power series in a neighbourhood of the point a, are called *equations with a regular singularity at the point* a.

REMARK 2. The name "*equations of Fuchsian type*", is also used although this name is more often used in a rather different sense (see e.g. [44]).

REMARK 3. Equation (21) is often studied in the *complex domain* (y being a function of the complex variable x). In this case we require that $P_1(x)$, $P_2(x)$ be holomorphic in a neighbourhood of the point a (Definition 20.1.9).

REMARK 4. In applications $P_1(x)$ and $P_2(x)$ are very often polynomials (if $P_1(x)$ and $P_2(x)$ are constants, we have (by substitution $x - a = u$) Euler's equation (Remark 17.13.3)). In the general case, we have

$$P_1(x) = \sum_{k=0}^{\infty} \alpha_k (x-a)^k, \quad P_2(x) = \sum_{k=0}^{\infty} \beta_k (x-a)^k. \tag{22}$$

Theorem 4. *Equation* (21) *has always at least one solution of the form*

$$y = (x-a)^{\varrho} \sum_{k=0}^{\infty} c_k (x-a)^k. \tag{23}$$

REMARK 5. ϱ need not be an integer (it need not even be real). Substituting (23) into (21) and comparing coefficients of equal powers of $x - a$, we get equations for the evaluation of ϱ and the unknown constants $c_0, c_1, c_2, \ldots$:

$$\begin{aligned} &c_0 f_0(\varrho) = 0, \\ &c_1 f_0(\varrho+1) + c_0 f_1(\varrho+1) = 0, \\ &c_2 f_0(\varrho+2) + c_1 f_1(\varrho+2) + c_0 f_2(\varrho+2) = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned} \tag{24}$$

Here we have used the notation

$$\begin{aligned} f_0(u) &= u(u-1) + u\alpha_0 + \beta_0, \\ f_k(u) &= u\alpha_k + \beta_k, \quad k = 1, 2, 3, \ldots, \end{aligned} \tag{25}$$

where α_k, β_k are coefficients of the series (22).

Let us determine ϱ so that

$$f_0(\varrho) \equiv \varrho(\varrho - 1) + \varrho\alpha_0 + \beta_0 = 0. \tag{26}$$

Equation (26) (the so-called *indicial equation* or *fundamental equation* of the given equation (21) at the singularity $x = a$) gives, in general, two roots (not necessarily real) ϱ_1, ϱ_2. In the following text, let us assume that

$$\operatorname{Re} \varrho_1 \geqq \operatorname{Re} \varrho_2, \tag{27}$$

i.e. that the real part of the root ϱ_1 is greater than (or equal to) the real part of the root ϱ_2. Let us consider first the root ϱ_1 and let us choose $c_0 \neq 0$ arbitrarily (e.g. $c_0 = 1$). In consequence of (27) we get $f_0(\varrho_1 + 1) \neq 0$, $f_0(\varrho_1 + 2) \neq 0$, etc.; thus, the values of c_1, c_2, c_3, ... can be uniquely determined from (24).

Theorem 5. *The series*

$$\sum_{k=0}^{\infty} c_k(x - a)^k \tag{28}$$

is convergent in a neighbourhood of the point a. Its radius of convergence is equal at least to the smaller of the radii of convergence of the series (22). (*In particular, if P_1 and P_2 are polynomials, the radius of convergence is infinite.*)

REMARK 6. By this procedure, we obtain the first solution of equation (21),

$$y_1 = (x - a)^{\varrho_1} \sum_{k=0}^{\infty} c_k(x - a)^k. \tag{29}$$

(ϱ_1 need not, of course, be a natural number. Therefore (29) need not be defined in the whole neighbourhood of the point a. More precisely, we should speak about an analytic element of the solution.)

Example 4. For the Bessel equation of order $\frac{1}{2}$, namely

$$y'' + \frac{1}{x}y' + \frac{x^2 - \frac{1}{4}}{x^2}y = 0,$$

we have

$$P_1(x) \equiv 1, \quad P_2(x) = x^2 - \frac{1}{4},$$

thus

$$\alpha_0 = 1, \quad \alpha_1 = \alpha_2 = \alpha_3 = \ldots = 0,$$
$$\beta_0 = -\tfrac{1}{4}, \quad \beta_1 = 0, \quad \beta_2 = 1, \quad \beta_3 = \beta_4 = \beta_5 = \ldots = 0,$$

so that by (25)

$$f_0(\varrho) = \varrho(\varrho - 1) + \varrho - \tfrac{1}{4} = \varrho^2 - \tfrac{1}{4},$$
$$f_1(\varrho) \equiv 0, f_2(\varrho) \equiv 1, f_3(\varrho) \equiv 0, f_4(\varrho) \equiv 0, \ldots$$

The fundamental equation

$$\varrho^2 - \frac{1}{4} = 0$$

has the solutions $\varrho_1 = \frac{1}{2}$, $\varrho_2 = -\frac{1}{2}$. For $\varrho_1 = \frac{1}{2}$ we have, according to (24),

$$\begin{aligned} &c_1\left[\left(\tfrac{1}{2}+1\right)^2 - \tfrac{1}{4}\right] + 0 = 0, \\ &c_2\left[\left(\tfrac{1}{2}+2\right)^2 - \tfrac{1}{4}\right] + 0 + c_0 = 0, \\ &c_3\left[\left(\tfrac{1}{2}+3\right)^2 - \tfrac{1}{4}\right] + 0 + c_1 + 0 = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned} \tag{30}$$

If we choose $c_0 = \sqrt{(2/\pi)}$, we have

$$y_1 = x^{1/2}\sqrt{\left(\frac{2}{\pi}\right)}\left(1 - \frac{1}{6}x^2 + \frac{1}{120}x^4 - \ldots\right).$$

By constructing a recurrence formula for the coefficients,

$$c_{2k+2} = \frac{-c_{2k}}{\left(\frac{1}{2}+2k+2\right)^2 - \frac{1}{4}}, \qquad c_{2k+1} = 0 \quad (k = 0, 1, 2, \ldots)$$

on the basis of (30), we would obtain the result (see Remark 16.4.5)

$$y_1 = \sqrt{\left(\frac{2}{\pi x}\right)}\sin x.$$

REMARK 7. If $\varrho_1 - \varrho_2$ is *not* an integer, we try to find a second solution of equation (21) in the form

$$y_2 = (x-a)^{\varrho_2}\sum_{k=0}^{\infty} d_k(x-a)^k. \tag{31}$$

For the coefficients d_k, we get the same system (24) as for the c_k but, naturally, ϱ must be given the value ϱ_2. If we choose d_0 arbitrarily ($d_0 \neq 0$), then all other d_k are uniquely determined (because $f_0(\varrho_2 + 1) \neq 0$, $f_0(\varrho_2 + 2) \neq 0$, etc.). Theorem 5 holds true for the series thus obtained. Moreover, the following theorem is true:

Theorem 6. *The functions* (29) *and* (31) *constitute a fundamental system* (Definition 17.11.2) *of the equation* (21).

REMARK 8. If $\varrho_1 - \varrho_2$ is an integer, we cannot use the above procedure because we find that, for some m, $f_0(\varrho_2 + m) = 0$. We can, of course, use the formula

$$y_2 = y_1 \int \frac{1}{y_1^2} e^{-\int P_1/(x-a)\,dx}\,dx \tag{32}$$

(see (17.8.18)). Since, however, we know the solution y_1 in a form of an infinite series, it is often more convenient to use the following theorem:

Theorem 7. *Let* $\varrho_1 - \varrho_2$ *be an integer and let* y_1 *be given by equation* (29). *Then a second solution, completing the fundamental system, is of the form*

$$y_2 = (x-a)^{\varrho}\ .A\ln(x-a)\sum_{k=0}^{\infty} c_k(x-a)^k + (x-a)^{\varrho_2}\sum_{k=0}^{\infty}\gamma_k(x-a)^k. \tag{33}$$

The radius of convergence of both infinite series is at least equal to the smaller of the radii of convergence of the series (22).

REMARK 9. In (33), c_k is known (see (28)). We get the conditions for the unknown constants A, γ_k by substituting (33) into (21). If $\varrho_1 - \varrho_2$ is not an integer, the first term in (33) vanishes ($A = 0$). If $\varrho_1 - \varrho_2$ is an integer, then, in general, $A \neq 0$ although in exceptional cases we may have $A = 0$. This happens, for instance, if $\alpha_1 = \beta_1 = \beta_2 = 0$, which occurs quite frequently in practice. If $\varrho_1 = \varrho_2$, however, then invariably $A \neq 0$.

SOME SPECIAL EQUATIONS WITH VARIABLE COEFFICIENTS

I. *The Bessel equation* (see also § 17.21, equation 117 and others):

$$x^2y'' + xy' + (x^2 - \nu^2)y = 0. \tag{34}$$

A solution (for any x and any real ν) is:

$$\mathrm{J}_\nu(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\Gamma(\nu + k + 1)}\left(\frac{x}{2}\right)^{\nu+2k} \tag{35}$$

(*the Bessel function of the first kind of order* ν). If ν is a negative integer, (35) still has a meaning if we define

$$\frac{1}{\Gamma(t)} = 0 \quad \text{for} \quad t = 0, -1, -2, \ldots$$

($\Gamma(t)$ is the gamma function, see Definition 13.11.1).

REMARK 10. For a more detailed treatment see § 16.4.

REMARK 11. We get the series (35) by using the method explained above (calculating c_k from (24), see also Example 4). In Example 4 we had $\varrho_1 - \varrho_2 = 1$, thus $f_0(\varrho_2 + 1) = 0$. When solving the system (30) for $\varrho_2 = -\frac{1}{2}$, this gives the possibility of a free choice of c_1. If we choose $c_1 = 0$ (then all other c_k will equal zero for odd k), we get a result identical with (35). Similarly, for $\varrho_2 = -\frac{3}{2}$, $\varrho_2 = -\frac{5}{2}$, etc., we choose all odd coefficients equal to zero.

Theorem 8. *If ν is not an integer (even if $\varrho_1 - \varrho_2 = 1$, etc.), then*

$$y_1(x) = \mathrm{J}_\nu(x) \quad \textit{and} \quad y_2(x) = \mathrm{J}_{-\nu}(x) \tag{36}$$

form a fundamental system of solutions of equation (34).

REMARK 12. If $\nu = n$ is an integer, then (36) is not a fundamental system because (cf. (16.4.2)) $\mathrm{J}_{-n}(x) = (-1)^n \mathrm{J}_n(x)$. Let $n > 0$ be an integer. If we determine $y_1(x) = \mathrm{J}_n(x)$, then we get the second solution of the fundamental system by applying (33),

$$\begin{aligned} y_2(x) = \mathrm{Y}_n(x) = {} & \frac{2}{\pi} \mathrm{J}_n(x) \left(\ln \frac{x}{2} + C \right) - \\ & - \frac{1}{\pi} \left(\frac{x}{2} \right)^{-n} \sum_{k=0}^{n-1} \frac{(n-k-1)!}{k!} \left(\frac{x}{2} \right)^{2k} - \\ & - \frac{1}{\pi} \left(\frac{x}{2} \right)^{n} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(n+k)!} \left(\frac{x}{2} \right)^{2k} \left[\sum_{l=1}^{k} \frac{1}{l} + \sum_{l=1}^{k+n} \frac{1}{l} \right], \end{aligned} \tag{37}$$

where

$$C = 0{\cdot}5772156649$$

is the Euler constant. Here, we put, for $k = 0$,

$$\sum_{l=1}^{k} \frac{1}{l} + \sum_{l=1}^{k+n} \frac{1}{l} = 1 + \ldots + \frac{1}{n}.$$

For $n = 0$, we have

$$\mathrm{Y}_0(x) = \frac{2}{\pi} \mathrm{J}_0(x) \left(\ln \frac{x}{2} + C \right) - \frac{2}{\pi} \sum_{k=1}^{\infty} \frac{(-1)^k}{(k!)^2} \left(\frac{x}{2} \right)^{2k} \sum_{l=1}^{k} \frac{1}{l}.$$

$\mathrm{Y}_n(x)$ is the so-called *Bessel function of the second kind.*

II. *The Gauss (or hypergeometric) equation* (see also § 17.21, equation 140 and further)

$$x(1-x)y'' + [\gamma - (\alpha+\beta+1)x]y' - \alpha\beta y = 0 \tag{38}$$

has if γ is not an integer and if $x \in (0, 1)$ the general integral

$$y = C_1 F(\alpha, \beta, \gamma, x) + C_2 x^{1-\gamma} F(\alpha+1-\gamma, \beta+1-\gamma, 2-\gamma, x), \tag{39}$$

where

$$F(\alpha, \beta, \gamma, x) = 1 + \frac{\alpha\beta}{1\,.\,\gamma}x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{1\,.\,2\,.\,\gamma(\gamma+1)}x^2 + \\ + \frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{1\,.\,2\,.\,3\,.\,\gamma(\gamma+1)(\gamma+2)}x^3 + \ldots \tag{40}$$

is the so-called *hypergeometric series.*

III. *The Legendre differential equation* (see also § 17.21, equation 129 and further)

$$(1-x^2)y'' - 2xy' + n(n+1)y = 0 \tag{41}$$

(n is a non-negative integer) arises from the hypergeometric equation ($\alpha = -n$, $\beta = n+1$, $\gamma = 1$) by the substitution $x = \frac{1}{2}(1-z)$. The function (40) then becomes

$$F\left(-n, n+1, 1, \frac{1-x}{2}\right) = \mathrm{P}_n(x) = \frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n \tag{42}$$

(*which is known as the Legendre polynomial of degree* n).

IV. *The Laguerre differential equation*

$$xy'' + (1-x)y' + ny = 0$$

(n being a non-negative integer) has a solution

$$\mathrm{L}_n(x) = \mathrm{e}^x \frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^n \mathrm{e}^{-x})$$

(*the Laguerre polynomial of degree* n).

V. *The Hermite differential equation*

$$y'' - 2xy' + 2ny = 0$$

(n being a non-negative integer) has a solution

$$\mathrm{H}_n(x) = (-1)^n \mathrm{e}^{x^2} \frac{\mathrm{d}^n}{\mathrm{d}x^n} \mathrm{e}^{-x^2}$$

(*the Hermite polynomial of degree n*).

For more details on Bessel functions, Legendre polynomials, etc., see Chap. 16.

17.16. Discontinuous Solutions of Linear Equations

In applications, we encounter cases where the integral of a given equation or its derivative has a prescribed "jump" at various points. (For example, we come across this situation when finding the elastic deflection of a bar which is supporting concentrated loads, etc.)

Definition 1. We say that $f(x)$ is *piecewise n times smooth in* $[a, b]$ if a finite number of points

$$a = x_0 < x_1 < x_2 < \ldots < x_m < x_{m+1} = b$$

exist so that in any interval (x_k, x_{k+1}) $(k = 0, 1, 2, \ldots, m)$ $f(x)$ and its derivatives up to the n-th order inclusive are continuous and have finite limits from the right and from the left at the points $x_0, x_1, x_2, \ldots, x_{m+1}$ (at the point $x_0 = a$ from the right and at the point $x_{m+1} = b$ from the left).

REMARK 1. The difference of these limits is called the *jump* of the function $f(x)$, or of its derivative $f^{(i)}(x)$ at the point x_k $(k = 1, 2, \ldots, m)$, and is denoted by ${}^kS_0$, or ${}^kS_i$, respectively. Thus

$${}^kS_i = \lim_{x \to x_k +} f^{(i)}(x) - \lim_{x \to x_k -} f^{(i)}(x). \tag{1}$$

(We consider the function $f(x)$ itself as its zero-th derivative.)

REMARK 2. If, in Definition 1, $n = 0$, then $f(x)$ is called a *piecewise continuous function in* $[a, b]$.

REMARK 3. Let an equation of the form

$$y^{(n)} + a_{n-1}(x)y^{(n-1)} + \ldots + a_1(x)y' + a_0(x)y = f(x) \tag{2}$$

be given, where $a_j(x)$, $f(x)$ are continuous functions (in the domain considered); let us denote by

$${}^ky_0(x), {}^ky_1(x), \ldots, {}^ky_{n-1}(x) \tag{3}$$

the normal (standard) fundamental system of the corresponding homogeneous equation at the point x_k, i.e. the fundamental system, for which

$$\begin{gathered}
{}^ky_0(x_k) = 1, {}^ky_1(x_k) = 0, {}^ky_2(x_k) = 0, \ldots, {}^ky_{n-1}(x_k) = 0, \\
{}^ky_0'(x_k) = 0, {}^ky_1'(x_k) = 1, {}^ky_2'(x_k) = 0, \ldots, {}^ky_{n-1}'(x_k) = 0, \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{gathered}$$

(cf. Definition 17.11.3).

Let us define the function ${}^kY_i(x)$ such that

$$ {}^kY_i(x) = \begin{cases} 0 & \text{for} \quad x < x_k, \\ {}^ky_i(x) & \text{for} \quad x \geqq x_k. \end{cases} \tag{4} $$

Theorem 1. *Let $y(x)$ be piecewise n-times smooth in the interval $[a, b]$. Further, in each of intervals (x_k, x_{k+1}) $(k = 0, 1, 2, \ldots, m$,* cf. Definition 1*) let it satisfy equation* (2) *and at the points x_k $(k = 1, 2, \ldots, m)$ let $y(x)$ and its derivatives have prescribed jumps ${}^kS_i$* (see (1)). *Then $y(x)$ is of the form*

$$ y = y_p + c_1y_1 + c_2y_2 + \ldots + c_ny_n + \\ + \sum_{k=1}^{m} {}^kS_0\,{}^kY_0(x) + \sum_{k=1}^{m} {}^kS_1\,{}^kY_1(x) + \ldots + \sum_{k=1}^{m} {}^kS_{n-1}\,{}^kY_{n-1}(x) \tag{5} $$

where y_p is a particular solution of equation (2), *$y_1, y_2, \ldots, y_n$ is the fundamental system of the corresponding homogeneous equation and ${}^kS_i$, ${}^kY_i$ are defined by equations* (1), (4).

Example 1. Let us find the solution of the equation

$$ y'' + y = 3x \tag{6} $$

with initial conditions

$$ y(1) = 1, \quad y'(1) = -1 \tag{7} $$

such that at the point $x_1 = 2$, $y(x)$ has the jump ${}^1S_0 = -1$ and $y'(x)$ has the jump ${}^1S_1 = 1$.

In this case we thus have $m = 1$.

The function $y_p = 3x$ is obviously a particular solution of equation (6), while the general integral of the corresponding homogeneous equation may with advantage be written, in view of the condition (7), in the form

$$ c_1 \cos(x - 1) + c_2 \sin(x - 1). $$

The normal fundamental system with respect to the point $x_1 = 2$ may be written in the form

$$ {}^1y_0(x) = \cos(x - 2), \quad {}^1y_1(x) = \sin(x - 2). $$

Thus

$$ {}^1Y_0(x) = \begin{cases} 0, \\ \cos(x - 2), \end{cases} \qquad {}^1Y_1(x) = \begin{cases} 0 & \text{for} \quad x < 2, \\ \sin(x - 2) & \text{for} \quad x \geqq 2. \end{cases} $$

Applying (5), the solution will be

$$\begin{aligned} &y = 3x + c_1 \cos(x-1) + c_2 \sin(x-1), \quad \text{for} \quad x < 2, \\ &y = 3x + c_1 \cos(x-1) + c_2 \sin(x-1) - \cos(x-2) + \sin(x-2) \quad \text{for} \quad x \geqq 2. \end{aligned} \tag{8}$$

From (8) and (7) it easily follows that $c_1 = -2$, $c_2 = -4$, thus giving the required solution.

REMARK 4. In applications we often come across cases where discontinuities occur not in the required solution, but in the coefficients of the given equation:

Theorem 2. *In equation* (2) *let the functions* $a_j(x)$, $f(x)$ *be piecewise continuous in the interval* $[a, b]$. *Let us denote the points of discontinuity by* x_k ($k = 1, 2, \ldots, m$). *With the exception of these points all the functions* $a_j(x)$, $f(x)$ *are then continuous in* $[a, b]$.) *If we prescribe at a point* $c \in [a, b]$ ($c \neq x_k$) *initial conditions for* $y^{(i)}(x)$ ($i = 0, 1, \ldots, n-1$), *then one and only one solution of equation* (2) *exists in* $[a, b]$ *that has* $n-1$ *continuous derivatives and a piecewise continuous* n*-th derivative in* $[a, b]$ *with points of discontinuity at* x_k *and which at the same time satisfies the prescribed initial conditions.*

REMARK 5. The procedure of finding the solution may be seen from an example:

Example 2. Let the equation

$$y'' + ay = 0 \tag{9}$$

be given, where $a = 1$ for $x \leqq 0$, $a = 4$ for $x > 0$. Let us find a solution which satisfies the following initial conditions:

$$y\left(-\frac{\pi}{2}\right) = 0, \quad y'\left(-\frac{\pi}{2}\right) = 1. \tag{10}$$

For $x \leqq 0$, the general integral of equation (9) is

$$y = c_1 \cos x + c_2 \sin x.$$

From conditions (10) we get

$$y = \cos x \quad (x \leqq 0). \tag{11}$$

Thus

$$y(0) = 1, \quad y'(0) = 0. \tag{12}$$

For $x > 0$ the general integral of equation (9) is

$$y = k_1 \cos 2x + k_2 \sin 2x. \tag{13}$$

Since we are trying to find the solution of equation (9) which is continuous, together with its first derivative, in the interval $(-\infty, \infty)$ (we note that, in the notation of Theorem 2, $n = 2$ and the interval may be taken as infinite), the function (13) must satisfy conditions (12) for $x = 0$. We easily find that

$$y = \cos 2x \quad (x > 0). \tag{14}$$

By (11) and (14), the solution is defined in the whole interval $(-\infty, \infty)$. At the point $x = 0$ the second derivative has a jump equal to -3.

17.17. Boundary Value Problems. Eigenvalue Problems. Expansion Theorem. Green's Function

INTRODUCTORY REMARK. It is well known that for the problem

$$y'' + \lambda y = 0, \quad y(0) = 0, \quad y(1) = 0$$

there exist certain values of λ ($\lambda_1 = \pi^2$, $\lambda_2 = 4\pi^2$, $\lambda_3 = 9\pi^2$, ...), the so-called *eigenvalues* of this problem, for which the given problem has non-zero solutions ($y_1 = \sin \pi x$, $y_2 = \sin 2\pi x$, $y_3 = \sin 3\pi x$, ...), the so-called *eigenfunctions* of this problem (besides the identically zero solution $y = 0$ which exists for all values of λ). Any function $f(x)$, which has a continuous derivative in the interval $[0, 1]$ and fulfils the conditions $f(0) = f(1) = 0$ can be expanded in an absolutely and uniformly convergent Fourier series with respect to these eigenfunctions.

In this paragraph similar questions are treated for more general problems which are important for applications (stability problems for stressed bars, solution of partial differential equations by Fourier's method, etc.). A sufficiently general theory of eigenvalue problems can be developed in Sobolev spaces. The reader who is interested in that theory, based on some results of functional analysis, is referred to Chap. 22 of this book (§ 22.6) and, in particular, to the book [389]. Here, we are going to follow the classical "Collatz theory" ([88]), based on the application of the Green function and giving, for the case of ordinary differential equations, relatively very natural results.

Notation: We shall write the linear equation

$$f_n(x)y^{(n)} + f_{n-1}(x)y^{(n-1)} + \ldots + f_1(x)y' + f_0(x)y = f(x) \tag{1}$$

briefly in the form

$$L(y) = f(x). \tag{2}$$

L is thus a *linear differential operator of the n-th order*. Whenever we shall write $L(u)$, we shall automatically assume that $u(x)$ has (in the domain considered, often

in some interval I) n continuous derivatives. The functions $f_k(x)$ are assumed to be (real) continuous functions in the domain considered, and $f_n(x) \neq 0$. If necessary, we shall assume that these functions have continuous derivatives of the required order.

Definition 1. The differential expression

$$K(y) \equiv (-1)^n[f_n(x)y]^{(n)} + (-1)^{n-1}[f_{n-1}(x)y]^{(n-1)} + \ldots \\ \ldots + [f_2(x)y]'' - [f_1(x)y]' + f_0(x)y \tag{3}$$

is called the *adjoint expression to the expression* $L(y)$. The equation $K(y) = 0$ is called the *adjoint equation to the equation* $L(y) = 0$, and K is called the *adjoint* (or rather *formaly adjoint*) *differential operator to the operator* L.

Theorem 1. *A necessary and sufficient condition for the expression* $u(x)L(y)$ *to be a complete derivative of a differential expression* $V(y)$ *of the* $(n-1)$*-th order (i.e. that* $uL(y) = \mathrm{d}V(y)/\mathrm{d}x$ *for an arbitrary choice of a sufficiently smooth function* $y(x)$*) is that* $u(x)$ *should be a solution of the adjoint equation* $K(u) = 0$.

Definition 2. If $L(y) = K(y)$ for each n-times differentiable function $y(x)$, the expression $L(y)$ (and the differential operator L) is called *self-adjoint.*

Example 1.

$$L(y) = f_2(x)y'' + f_1(x)y' + f_0(x)y.$$

Then

$$K(y) = (f_2y)'' - (f_1y)' + f_0y = f_2y'' + (2f_2' - f_1)y' + (f_2'' - f_1' + f_0)y.$$

If (and only if) $f_1 = f_2'$, i.e. if

$$L(y) = (f_2y')' + f_0y,$$

then L is a self-adjoint differential operator. So any self-adjoint expression of the second order can be put into the form

$$L(y) = [p(x)y']' + q(x)y. \tag{4}$$

REMARK 1. In applications, self-adjoint expressions of the $2n$-th order

$$L(y) = \sum_{k=0}^{n}(-1)^k[f_k(x)y^{(k)}]^{(k)} = (-1)^n[f_n(x)y^{(n)}]^{(n)} + \\ +(-1)^{n-1}[f_{n-1}(x)y^{(n-1)}]^{(n-1)} + \ldots + f_0(x)y$$

play an important role.

Let an equation of the form

$$M(y) - \lambda N(y) = f(x) \tag{5}$$

be given, where $M(y)$, $N(y)$ are self-adjoint expressions of the $2m$-th and $2n$-th order, respectively,

$$M(y) = \sum_{k=0}^{m} (-1)^k [f_k(x) y^{(k)}]^{(k)}, \quad N(y) = \sum_{k=0}^{n} (-1)^k [g_k(x) y^{(k)}]^{(k)}, \quad m > n. \tag{6}$$

In the interval $[a, b]$ let the (real) functions $f_k(x)$, $g_k(x)$ have k continuous derivatives (for $k = 0$ this means the continuity of the functions f_0, g_0 themselves), $f_m(x) \neq 0$, $g_n(x) \neq 0$ in $[a, b]$, and $f(x)$ be continuous in $[a, b]$. In addition, let $2m$ linear homogeneous *boundary conditions* of the form

$$\begin{gathered} \alpha_{i0} y(a) + \beta_{i0} y(b) + \alpha_{i1} y'(a) + \beta_{i1} y'(b) + \dots \\ \dots + \alpha_{i,2m-1} y^{(2m-1)}(a) + \beta_{i,2m-1} y^{(2m-1)}(b) = 0, \quad i = 1, 2, \dots, 2m, \end{gathered} \tag{7}$$

be prescribed, where $\alpha_{i0}, \beta_{i0}, \dots$ are real constants, not simultaneously equal to zero in anyone of the equations (7). Conditions (7) (of which there are $2m$) are supposed to be linearly independent (roughly speaking, none of them is a consequence of the others).

Definition 3. Let the (real) number λ in (5) be *given*. The problem of finding a solution $y(x)$ of equation (5) (with $2m$ continuous derivatives in $[a, b]$) satisfying conditions (7), is called a *boundary value problem* for equation (5).

REMARK 2. According to our assumptions, the operator M is of a higher order than the operator N. Often $N(y) = g_0(x)y$, i.e. $n = 0$.

REMARK 3. The number of boundary conditions (7) is equal to the order of the differential equation (5), i.e. equal to $2m$. Derivatives up to the $(2m - 1)$-th order of the required solution can occur in (7).

REMARK 4. Boundary value problems for linear ordinary differential equations can be defined more generally. We do not, however, do this, because our definition involves practically all boundary value problems occurring in applications.

REMARK 5. If $f(x) \equiv 0$ in (5), i.e. if this equation is of the form

$$M(y) - \lambda N(y) = 0, \tag{8}$$

the problem (5), (7), i.e. the problem (8), (7), is called *homogeneous*. For an arbitrary λ, this problem has obviously a solution $y(x) \equiv 0$. This solution is, naturally, of no interest. Thus *it is excluded from our considerations in what follows.*

Definition 4. Such a λ, for which the problem (8), (7) has, a non-zero solution $y(x) \not\equiv 0$, is called an *eigenvalue* (or a *characteristic value*, or *proper value*) of this problem. The corresponding non-zero solution (i.e. not identically equal to zero), is called an *eigenfunction* (a *characteristic function*, a *proper function*) corresponding to that eigenvalue λ.

The problem (8), (7) (or in more details, the problem of finding all eigenvalues and eigenfunctions of (8), (7)) is called an *eigenvalue problem.*

REMARK 6. It follows from homogeneity of the problem (8), (7) that if $y(x)$ is an eigenfunction corresponding to an eigenvalue λ, then $cy(x)$, with arbitrary $c \neq 0$, is also an eigenfunction, corresponding to the same λ.

Example 2. An example of an eigenvalue problem, according to our definition, is the following:

$$ay'' - \lambda y = 0 \quad (a \text{ a non-zero constant}), \tag{9}$$

$$y(0) = 0, \quad y'(1) - 2y(1) = 0. \tag{10}$$

$M(y) = ay''$ is a self-adjoint expression of the second order, because $M(y) = (ay')'$; $N(y) = y$. Conditions (10) are linear and homogeneous (they are obviously satisfied for $y \equiv 0$) and contain no derivative of a higher order than of the first.

Now suppose that a is not a constant, but $a = a(x) \not\equiv$ const.; then $M(y) = a(x)y''$ is not a self-adjoint expression. If, however, $a(x) \not\equiv 0$ in $[0, 1]$, we can divide equation (9) by this function and change it to the form

$$y'' - \lambda \frac{y}{a(x)} = 0. \tag{11}$$

Here $M(y) = y'' = (y')'$ is a self-adjoint expression, while

$$N(y) = \frac{1}{a(x)} y = b(x)y$$

is also a self-adjoint expression. From this example it can be seen that a simple rearrangement of the given equation may be all that is needed to put it into the form (8).

Example 3. The eigenfunctions of the problem

$$-y'' - \lambda y = 0, \quad y(0) = 0, \quad y(\pi) = 0 \tag{12}$$

are the functions $y = \sin px$, while the corresponding eigenvalues are $\lambda = p^2$, where p ranges over all positive integers.

In applications, we encounter, most often, the so-called symmetric and positive problems. Let us introduce corresponding definitions:

Definition 5. The (real) function $y(x)$ is called a *comparison function* (or a *test function* or a *trial function*) of the eigenvalue problem (8), (7), if it has $2m$ continuous derivatives in $[a, b]$ and satisfies the boundary conditions (7).

REMARK 7. A comparison function *need not* be a solution of the differential equation (8). For example, the function $y = x(\pi - x)$ is a comparison function of the problem (12), but obviously it is not an eigenfunction of that problem for any λ. Evidently, any eigenfunction is also a comparison function.

Definition 6. The eigenvalue problem (8), (7) is called *symmetric*, if for any *comparison* functions $u(x)$, $v(x)$ the relations

$$\int_a^b [uM(v) - vM(u)]\,dx = 0, \quad \int_a^b [uN(v) - vN(u)]\,dx = 0 \tag{13}$$

hold.

Definition 7. The eigenvalue problem is called *positive* if for any non-zero comparison function the relations

$$\int_a^b uM(u)\,dx > 0, \quad \int_a^b uN(u)\,dx > 0 \tag{14}$$

hold.

Example 4. The problem

$$-y'' - \lambda c(x)y = 0 \quad (c(x) > 0 \text{ in } [a, b]), \tag{15}$$

$$y(a) = 0, \quad y(b) = 0 \tag{16}$$

is (as we shall show) symmetric and positive. We have

$$M(y) = -y'', \quad N(y) = c(x)y.$$

Then, integrating by parts and using the fact that any comparison function satisfies conditions (16), we have

$$\int_a^b uM(v)\,dx = -\int_a^b uv''\,dx = -[uv']_a^b + \int_a^b u'v'\,dx = \int_a^b u'v'\,dx. \tag{17}$$

If we change the roles of u and v, we get similarly

$$\int_a^b vM(u)\,dx = -\int_a^b u'v'\,dx, \quad \text{thus} \quad \int_a^b [uM(v) - vM(u)]\,dx = 0.$$

Moreover

$$\int_a^b u \,.\, cv \,\mathrm{d}x = \int_a^b v \,.\, cu \,\mathrm{d}x, \quad \text{thus} \quad \int_a^b [uN(v) - vN(u)] \,\mathrm{d}x = 0. \tag{18}$$

The problem is thus symmetric. Further (see (17))

$$\int_a^b uM(u) \,\mathrm{d}x = \int_a^b u'^2 \,\mathrm{d}x, \quad \int_a^b uN(u) \,\mathrm{d}x = \int_a^b cu^2 \,\mathrm{d}x,$$

thus (since we are given that $c(x) > 0$) for any non-zero comparison function $u(x)$, (14) is satisfied. The problem is thus positive.

REMARK 8. Similarly the so-called *Sturm-Liouville problem*

$$-(py')' + qy - \lambda ry = 0, \quad y(a) = 0, \quad y(b) = 0$$

can be shown to be symmetric and positive provided that, in $[a, b]$, $p(x) > 0$, $r(x) > 0$, $q(x) \geqq 0$.

Example 5. In the same way as in Example 4 the problem

$$-y'' - \lambda y = 0, \quad y'(a) = 0, \quad y'(b) = 0$$

can be shown to be symmetric. It is not, however, positive because, for example, for every comparison function $u(x) = \text{const.} \neq 0$ we have

$$\int_a^b uM(u) \,\mathrm{d}x = 0.$$

REMARK 9. From Example 4 it can be seen that by using the method of integration by parts we can often decide easily whether a given problem is symmetric and positive or not. In other cases, we can make use of the so-called *Green's formula* (*Dirichlet's formula*)

$$\int_a^b [uM(v) - vM(u)] \,\mathrm{d}x =$$

$$= \Big[\sum_{k=1}^{m} \sum_{l=0}^{k-1} (-1)^{k+l} \{u^{(l)}[f_k v^{(k)}]^{(k-l-1)} - v^{(l)}[f_k u^{(k)}]^{(k-l-1)}\}\Big]_a^b, \tag{19}$$

where as usual $[F]_a^b$ denotes $F(b) - F(a)$. For a given k, it is necessary for l in the sum to range through all the integers from 0 to $k - 1$. If $k = 1$, only one value of l, $l = 0$, is concerned.

A similar formula holds for the operator N (instead of f_k, or m, we have here g_k, or n, respectively). Obviously, the eigenvalue problem is symmetric if and only if the boundary conditions are such that the right-hand side of (19) is equal to zero for every comparison function, for both the operators M and N.

REMARK 10. We very often meet such problems where the self-adjoint operator N has only one term,

$$N(y) = (-1)^n[g_n(x)y^{(n)}]^{(n)} \tag{20}$$

and the boundary conditions are such that for every two comparison functions u, v the relation

$$\int_a^b uN(v)\,dx = \int_a^b g_n u^{(n)} v^{(n)}\,dx \tag{21}$$

holds. Then the eigenvalue problem is called *regular*. (The terminology is not uniform in the literature.) An example of this is the case where the operator N is of zero order, i.e.

$$N(y) = g_0(x)y$$

(see Example 4). In this very simple case, (21) is obviously satisfied, the boundary conditions being arbitrary.

Theorem 2. *If the eigenvalue problem is symmetric, then the eigenfunctions $y_s(x)$, $y_t(x)$, corresponding to different eigenvalues λ_s, λ_t are orthogonal in the so-called generalized sense, i.e.*

$$\int_a^b y_s N(y_t)\,dx = 0 \quad \textit{for} \quad \lambda_s \neq \lambda_t. \tag{22}$$

REMARK 11. In the special case where $N(y) = g_0(x)y$, equation (22) gives orthogonality with a weight function $g_0(x)$ (Definition 16.2.5), i.e.

$$\int_a^b g_0 y_s y_t\,dx = 0.$$

Theorem 3. *If the eigenvalue problem is positive, then it can have only positive eigenvalues.*

Theorem 4 (*Solvability of the problem* (8), (7)). *If the eigenvalue problem is symmetric and positive, then there exists a countable set of positive mutually different eigenvalues of this problem* (see also Theorem 5 below).

REMARK 12. To every eigenvalue λ, there corresponds in this case either one or, in general, a finite (not infinite) number p of *linearly independent* (Definition 12.8.3) eigenfunctions. We say then that λ is a *simple* eigenvalue, or that its multiplicity is

p, respectively. (In almost all technical problems λ is simple, i.e. only one linearly independent eigenfunction corresponding to this λ exists, all others being multiples of it.) Linearly independent eigenfunctions corresponding to a given λ (for $p > 1$) can be orthogonalized in the generalized sense similarly as in Remark 16.2.15. Because, according to Theorem 2, the eigenfunctions corresponding to *different* λ are orthogonal in this sense, we can thus associate with the set of all eigenvalues λ a system of linearly independent eigenfunctions mutually orthogonal in the sense (22). Let us number the eigenvalues according to their magnitude,

$$\lambda_1 \leqq \lambda_2 \leqq \lambda_3 \leqq \dots \tag{23}$$

and at the same time in such a way that the correspondence between the orthogonal system of eigenfunctions and the system of corresponding eigenvalues be one-to-one. Thus, every eigenvalue λ will occur in the system (23) the same number of times as the number of functions of the orthogonal system corresponding to it. If, e.g., three functions of the orthogonal system correspond to the smallest eigenvalue, then in (23) three numbers $\lambda_1 = \lambda_2 = \lambda_3$ will correspond to them; if to a further number λ two functions of the orthogonal system correspond, then in (23) two numbers $\lambda_4 = \lambda_5$ will correspond to them, etc. For this reason it was necessary to allow the possibility of equalities in the ordering (23).

A typical example of a symmetric positive problem is the problem (12). In this case we have

$$\lambda_1 = 1,\ \lambda_2 = 4,\ \lambda_3 = 9,\ \lambda_4 = 16,\ \dots.$$

Every eigenvalue is simple here. The system of eigenfunctions is the system

$$y_1 = \sin x,\ y_2 = \sin 2x,\ y_3 = \sin 3x,\ y_4 = \sin 4x,\ \dots.$$

These functions are orthogonal in the interval $[0, \pi]$ in the usual sense, i.e. with the weight function 1.

Theorem 5. *Let the eigenvalue problem be symmetric and positive. Then*

I.

$$\lambda_n \to +\infty \quad \textit{for} \quad n \to \infty;$$

the point $+\infty$ *is the only point of accumulation of the sequence* λ_n.

II. *Let us define the so-called Rayleigh quotient* $R(u)$ *by the relation*

$$R(u) = \frac{\int_a^b uM(u)\,\mathrm{d}x}{\int_a^b uN(u)\,\mathrm{d}x}. \tag{24}$$

Then

$$\lambda_1 = \min R(u),$$

if $u(x)$ *runs through all comparison functions of the given problem.*

III. *More generally*:

$$\lambda_{k+1} = \min R(u),$$

where $u(x)$ *runs through those comparison functions which are orthogonal in the generalized sense to the first* k *eigenfunctions* $\varphi_1(x), \ldots, \varphi_k(x)$, *i.e. for which*

$$\int_a^b uN(\varphi_i)\,\mathrm{d}x = 0 \quad (i = 1, 2, \ldots, k). \tag{25}$$

REMARK 13. It can be shown that $R(u)$ assumes the minimal value λ_1 exactly for the eigenfunction $\varphi_1(x)$ corresponding to λ_1. If we consider an arbitrary comparison function $u(x)$, which is not an eigenfunction, then $R(u) > \lambda_1$. Thus, $R(u)$ gives an estimate of λ_1 from above.

Considering, for example, the problem (12), we have $M(u) = -u''$, $N(u) = u$. Let us consider the comparison function $u = x(\pi - x)$. According to (24) we get

$$R\big(x(\pi - x)\big) = \frac{\displaystyle\int_0^\pi 2x(\pi - x)\,\mathrm{d}x}{\displaystyle\int_0^\pi x^2(\pi - x)^2\,\mathrm{d}x} = \frac{\dfrac{\pi^3}{3}}{\dfrac{\pi^5}{30}} = \frac{10}{\pi^2} \doteq 1{\cdot}0132,$$

which is a good estimate of the first eigenvalue $\lambda_1 = 1$ "from above".

REMARK 14. Part II of Theorem 5 provides the possibility of applying variational methods for finding λ_1 or $\varphi_1(x)$, respectively. This is true for part III, as well. See Chaps. 24 and 25.

The application of variational methods in part III is more difficult, on account of condition (25). The following theorem is then often useful:

Theorem 6 (*Courant's Maximum – Minimum Principle*). *Let the eigenvalue problem be symmetric and positive. Let* $w_1(x), \ldots, w_k(x)$ *be an arbitrary system of linearly independent integrable functions. Let us denote by* $m(w_1, \ldots, w_k)$ *the minimum* (*or infimum*) *of the Rayleigh quotient, if* $u(x)$ *runs through all comparison functions which are orthogonal to all functions* $w_i(x)$, i.e. *for which*

$$\int_a^b u(x)w_i(x)\,\mathrm{d}x = 0, \quad i = 1, 2, \ldots, k.$$

Then λ_{k+1} *is equal to the maximum of* $m(w_1, \ldots, w_k)$ *if all systems of* (*linearly independent and integrable*) *functions* $w_i(x)$ *are considered.*

Theorem 7 (*Comparison Theorem*). *Let us consider two symmetric positive eigenvalue problems*

$$M_1(y) = \lambda N_1(y), \quad M_2(y) = \lambda^* N_2(y) \tag{26}$$

with the same boundary conditions. For any comparison function $u(x)$ *let the relations*

$$\int_a^b uM_1(u)\,\mathrm{d}x \leqq \int_a^b uM_2(u)\,\mathrm{d}x, \quad \int_a^b uN_1(u)\,\mathrm{d}x \geqq \int_a^b uN_2(u)\,\mathrm{d}x \tag{27}$$

be satisfied. Then (*supposing the eigenvalues of both problems to be arranged according to their magnitude,* see Remark 12) *we have*

$$\lambda_k \leqq \lambda_k^* \quad (k = 1, 2, 3, \ldots). \tag{28}$$

Example 6. If, for example, in the problems

$$-(p_1y')' + q_1y - \lambda r_1y = 0, \quad -(p_2y')' + q_2y - \lambda^* r_2y = 0,$$
$$y(a) = 0, \quad y(b) = 0$$

the relations

$$0 < p_1(x) \leqq p_2(x), \quad r_1(x) \geqq r_2(x) > 0, \quad 0 \leqq q_1(x) \leqq q_2(x)$$

are satisfied in $[a, b]$, then

$$\lambda_k \leqq \lambda_k^*.$$

REMARK 15. While the Rayleigh quotient provides a simple possibility how to obtain an upper estimate for the first eigenvalue λ_1 (Remark 13), in applications (in stability problems, etc.) it is usually of more interest how to get a lower estimate. For this aim, Theorem 7 can be applied. However, in this way we get a rather rough estimate, as a rule. A relatively fine *two-sided estimate* (i.e., both from above and from below) can be obtained on base of the following theorem:

Theorem 8 (*Method of the Schwarz Quotients*). *Let the eigenvalue problem* (8), (7) *be symmetric and positive. Let* $F_0(x)$, $F_1(x)$ *be two comparison functions* (Definition 5) *such that*

$$M(F_1) = N(F_0).$$

Let us construct the so-called Schwarz constants

$$a_0 = \int_a^b F_0N(F_0)\,\mathrm{d}x, \quad a_1 = \int_a^b F_0N(F_1)\,\mathrm{d}x, \quad a_2 = \int_a^b F_1N(F_1)\,\mathrm{d}x$$

and Schwarz quotients

$$\kappa_1 = \frac{a_0}{a_1}, \quad \kappa_2 = \frac{a_1}{a_2}.$$

Let the first eigenvalue λ_1 *be simple and let* l_2 *be such a lower estimate of the second eigenvalue* λ_2 *that* $l_2 > \kappa_2$. *Then we have*

$$\kappa_2 - \frac{\kappa_1 - \kappa_2}{\frac{l_2}{\kappa_2} - 1} \leqq \lambda_1 \leqq \kappa_2.$$

REMARK 16. A lower estimate l_2 of the second eigenvalue λ_2 (does not matter when rather rough) can be found using Theorem 7 (see the following example). Let us note that the requirement on the functions F_0, F_1 to be comparison functions (thus to fulfill all the boundary conditions) can be essentially weakened. An exact formulation of assumptions (including the case of partial differential equations) can be found, e.g. in [389]. In particular, if the operator N is of order zero, then F_0 need satisfy no boundary conditions, what is very convenient from the numerical point of view.

Example 7. We have to find a twosided estimate of the first eigenvalue of the following (obviously symmetric and positive) problem:

$$-y'' - \lambda \frac{y}{E(1 - 0{\cdot}4|x|)^2} = 0, \quad y(-1) = 0, \quad y(1) = 0.$$

(This problem arises when considering, e.g., buckling stress of a bar with a variable cross-section; E is then the modulus of elasticity.) Here

$$M(y) = -y'', \quad N(y) = \frac{y}{E(1 - 0{\cdot}4|x|)^2}.$$

The function F_1, being a comparison function, has to satisfy

$$F_1(-1) = 0, \quad F_1(1) = 0,$$

while F_0 need not satisfy any conditions, because the operator N is of order zero (see the preceding remark). It is thus not necessary to solve the boundary value problem

$$M(F_1) = N(F_0)$$

with the comparison function F_0 chosen before, but it is sufficient to choose, for the function F_1, merely the function

$$F_1(x) = 1 - x^2$$

(which satisfies the given boundary conditions) and to determine F_0 in order to satisfy

$$-F_1'' = \frac{F_0}{E(1-0{\cdot}4|x|)^2},$$

i.e.

$$F_0(x) = 2E(1-0{\cdot}4|x|)^2.$$

According to the preceding theorem we then compute the Schwarz constants

$$a_0 = \int_{-1}^{1} 2E(1-0{\cdot}4|x|)^2 \cdot \frac{2E(1-0{\cdot}4|x|)^2}{E(1-0{\cdot}4|x|)^2}\,\mathrm{d}x = 5{\cdot}227E,$$

$$a_1 = \int_{-1}^{1} 2E(1-0{\cdot}4|x|)^2 \cdot \frac{1-x^2}{E(1-0{\cdot}4|x|)^2}\,\mathrm{d}x = 2{\cdot}667,$$

$$a_2 = \int_{-1}^{1} (1-x^2) \cdot \frac{1-x^2}{E(1-0{\cdot}4|x|)^2}\,\mathrm{d}x = \frac{1{\cdot}437}{E}$$

and the Schwarz quotients

$$\kappa_1 = 1{\cdot}960E, \quad \kappa_2 = 1{\cdot}856E.$$

To obtain l_2, let us compare the given problem with the problem

$$-y'' - \frac{\lambda y}{0{\cdot}6^2}, \quad y(-1) = 0, \quad y(1) = 0$$

that has — according to Theorem 7 — smaller eigenvalues and where the equation considered has constant coefficients, so that we easily find its second eigenvalue to be $\pi^2 E \,.\, 0{\cdot}6^2 = 3{\cdot}560E$. We thus can choose

$$l_2 = 3{\cdot}560E,$$

because obviously we have $l_2 > \kappa_2$ at the same time, as required in Theorem 8. So we come to the two-sided estimate for λ_1,

$$1{\cdot}856E - \frac{1{\cdot}960E - 1{\cdot}856E}{\frac{3{\cdot}560E}{1{\cdot}856E} - 1} \leqq \lambda_1 \leqq 1{\cdot}856E,$$

i.e.

$$1{\cdot}743E \leqq \lambda_1 \leqq 1{\cdot}856E.$$

This is a quite satisfactory estimate for practical use.

Let us note, finally, that the method of the Schwarz quotients has been excellently worked out and adapted for applications by L. Collatz in [88], and then extensively developed by his school (J. Albrecht, F. Goerisch and others).

Theorem 9. *Let a symmetric positive eigenvalue problem* (8), (7) *be given. Let the system of orthogonal eigenfunctions* $\varphi_i(x)$ (Remark 12) *be normalized in the generalized sense, i.e. let*

$$\int_a^b \varphi_i N(\varphi_k)\,\mathrm{d}x = \begin{cases} 0 & \text{for } i \neq k, \\ 1 & \text{for } i = k. \end{cases} \tag{29}$$

Let $u(x)$ *be an arbitrary comparison function and let the numbers*

$$a_k = \int_a^b uN(\varphi_k)\,\mathrm{d}x \tag{30}$$

be its so-called generalized Fourier coefficients. Then the series

$$\sum_{n=1}^{\infty} a_n \varphi_n(x) \tag{31}$$

as well as the series

$$\sum_{n=1}^{\infty} a_n \varphi_n^{(i)}(x), \quad i = 1, 2, \ldots, m-1, \tag{32}$$

arising, subsequently, by differentiating (31) *term-by-term, converge absolutely and uniformly in* $[a, b]$. (On the number m see in (6).)

REMARK 17. The sum of the series (31) need not be equal to the function u, in general. However, the equality holds, when the problem is a so-called *closed problem.* For details see e.g. the Kamke's paper in Math. Zeitschrift 1940, pp. 275-280. In particular, if we have

$$N(y) = (-1)^n \left[g_n(x)\, y^{(n)}\right]^{(n)} \quad (g_n(x) > 0 \quad \text{in } [a, b])$$

and if (for $n \geqq 1$) among the given boundary conditions the conditions

$$y(a) = y'(a) = \ldots = y^{(n-1)}(a) = 0, \quad y(b) = y'(b) = \ldots = y^{(n-1)}(b) = 0$$

occur (to which, in general, further conditions are to be added according to the degree of the operator M), then the sums of the series (31) and (32) are equal to the function u and to its derivatives $u^{(i)}$, $i = 1, 2, \ldots, m-1$, respectively.

The conditions of Theorem 8 can often be weakened. For example, when investigating the problem

$$\begin{gathered} -(py')' + qy - \lambda ry = 0, \quad p(x) > 0, \quad r(x) > 0, \quad q(x) \geqq 0, \\ y(a) = 0, \quad y(b) = 0, \end{gathered} \tag{33}$$

it is sufficient for the function $u(x)$, which is to be developed into the Fourier series (31), to have only the first derivative continuous in $[a, b]$ (and to fulfil the prescribed boundary conditions). Sometimes, it is possible to omit even the condition of satisfying the boundary conditions (what may be advantageous, because we often have to deal with the case where $u(x)$ has a sufficient number of derivatives but does not satisfy the boundary conditions). If, for example, the condition $y(a) = 0$ is replaced by the condition $y'(a) = 0$ in (33) and if $u(x)$ does not satisfy this condition, then the corresponding series (31) can be shown to converge uniformly to $u(x)$ in every interval $[c, b]$ with $a < c < b$.

On expansion theorems in suitable functional spaces see e.g. [389], Chap. 39. See also § 22.6 of this book.

On solvability of the *non-homogeneous* problem (5), (7), thus of the equation

$$M(y) - \lambda N(y) = f(x) \tag{5}$$

with boundary conditions (7), the following theorem is valid (for the operators M, N see (6), $f(x)$ is assumed to be continuous in $[a, b]$):

Theorem 10. *Let λ in* (5) *be given. Then:*

(i) *If that λ is no eigenvalue of the corresponding homogeneous problem* (8), (7), *then the given problem* (5), (7) *has exactly one solution for every right-hand side $f(x)$.*

(ii) *If λ is an eigenvalue of the corresponding problem* (8), (7), *then the problem* (5), (7) *is not solvable, in general. It is solvable (however not uniquely, in this case), if and only if the function f is orthogonal to every eigenfunction φ corresponding to that λ, i.e. exactly if*

$$(f, \varphi) = \int_a^b f(x)\varphi(x)\,\mathrm{d}x = 0 \tag{34}$$

holds for every such eigenfunction.

Example 8. The problem

$$-y'' - y = \sin x, \quad y(0) = 0, \quad y(\pi) = 0$$

is *not* solvable. In accordance with the notation of Example 3 we have $\lambda = 1$ here, so that (following the same example) λ *is* an eigenvalue of the corresponding homogeneous problem (12). Eigenfunctions, corresponding to that $\lambda = 1$, are of the form

$$\varphi = c \sin x, \quad c \neq 0.$$

Now

$$(f, \varphi) = \int_0^\pi \sin x \,.\, c \sin x\,\mathrm{d}x = c \int_0^\pi \sin^2 x\,\mathrm{d}x \neq 0,$$

so that the condition (34) is not fulfilled.

In this simple case (the given equation is an equation with constant coefficients), we can decide on solvability of the considered problem immediately, without applying Theorem 10: The general integral of the given equation is (cf. Example 17.14.3)

$$y = \tfrac{1}{2}x\cos x + C_1\cos x + C_2\sin x.$$

The condition $y(0) = 0$ yields $C_1 = 0$, so that the condition $y(\pi) = 0$ becomes

$$\frac{1}{2}\pi\cos\pi + C_2\sin\pi = 0.$$

However, $\cos\pi \neq 0$, $\sin\pi = 0$, thus this condition cannot be fulfilled for any C_2. Consequently, the given problem is not solvable.

On the other side, the problem

$$-y'' - y = \cos x, \quad y(0) = 0, \quad y(\pi) = 0$$

is solvable, because the condition (34) is fulfilled here:

$$(f, \varphi) = \int_0^\pi \cos x \,.\, c\sin x\,\mathrm{d}x = 0.$$

Also this result can be obtained without applying Theorem 10: In the same way as above we come to the solution

$$y = -\frac{1}{2}x\sin x + C_2\sin x, \quad \text{with } C_2 \text{ arbitrary.}$$

Thus there are infinitely many solutions in this case.

The problem

$$-y'' + 4x^4y = \mathrm{e}^x, \quad y(1) = 0, \quad y(2) = 0$$

is uniquely solvable. In the notation of Example 4 we have

$$a = 1, \quad b = 2, \quad M(y) = -y'', \quad N(y) = x^4y, \quad \lambda = -4.$$

However, the corresponding homogeneous problem

$$-y'' - \lambda x^4y = 0, \quad y(1) = 0, \quad y(2) = 0$$

is symmetric and positive, so that it has only positive eigenvalues. Consequently, $\lambda = -4$ cannot be its eigenvalue. By Theorem 10, the given non-homogeneous problem is uniquely solvable for every right-hand side $f(x)$, thus also for the right--hand side $f(x) = \mathrm{e}^x$.

Here the application of Theorem 10 to answering the question on solvability has been very useful, because the given equation has not constant coefficients, and the construction of the general integral is not easy.

REMARK 18. Results, presented in this section for symmetric and positive problems, can be derived with the help of the so-called Green function. This function can as well be applied to the solution of problems which need be neither symmetric nor positive:

Let us consider an equation of the form

$$L(y) = f(x), \tag{35}$$

where

$$L(y) = \sum_{i=0}^{k} p_i(x) y^{(i)},$$

with $p_i(x)$ continuous, $p_k(x) \neq 0$ in $[a, b]$, with the boundary conditions

$$\sum_{i=0}^{k-1} \left[{}^{l}\alpha_i y^{(i)}(a) + {}^{l}\beta_i y^{(i)}(b) \right] = 0 \quad (l = 0, 1, \ldots, k-1) \tag{36}$$

(cf. (7), where $k = 2m$; in our case k need not be an even number).

Definition 8. The *Green function* $G(x, \xi)$ of the problem (35), (36) is defined as follows:

1. The function $G(x, \xi)$ is defined in a square $a \leqq x \leqq b$, $a \leqq \xi \leqq b$. With the

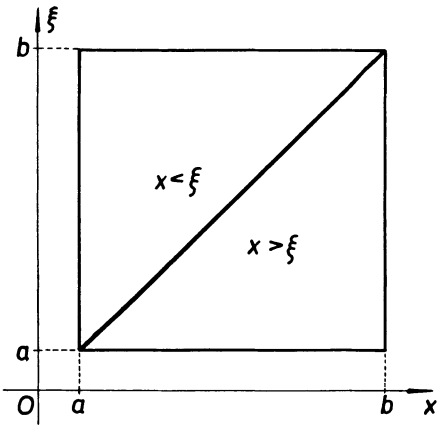


Fig. 17.5.

exception of points lying on the diagonal $x = \xi$ (Fig. 17.5), the partial derivatives

$$\frac{\partial^i G}{\partial x^i} \quad (i = 0, 1, 2, \ldots, k)$$

are continuous functions of both variables (if $i = 0$ the continuity of the function G itself is to be understood), continuously extensible to the boundaries of both triangles into which the given square is divided by the diagonal $x = \xi$.

2. For any fixed $\xi \in (a, b)$, $G(x, \xi)$ satisfies, as a function of the variable x, the equation $L(y) = 0$ everywhere in $[a, b]$ *with the exception of the point* $x = \xi$.

3. $G(x, \xi)$ satisfies, as a function of the variable x, the boundary conditions (36) (for any $\xi \in (a, b)$).

4. $G(x, \xi)$, together with its derivatives with respect to x up to the $(k-2)$-th order, is a continuous function of the variable x (ξ being fixed). The derivative (with respect to x) of the $(k-1)$-th order has a jump $1/p_k(\xi)$ at the point $x = \xi$, i.e.

$$\lim_{x \to \xi+} \frac{\partial^{k-1} G}{\partial x^{k-1}}(x, \xi) - \lim_{x \to \xi-} \frac{\partial^{k-1} G}{\partial x^{k-1}}(x, \xi) = \frac{1}{p_k(\xi)}$$

($p_k(x)$ being the coefficient of the highest derivative $y^{(k)}$ in the operator $L(y)$, Remark 18).

REMARK 19. *Construction of the Green function:* Let

$$y_1(x), y_2(x), \ldots, y_k(x)$$

be the fundamental system (Definition 17.11.2) of the equation $L(y) = 0$, i.e. of the equation

$$p_k(x)y^{(k)} + p_{k-1}(x)y^{(k-1)} + \ldots + p_0(x)y = 0. \tag{37}$$

Let the Green function be assumed to have the form

$$G(x, \xi) = \sum_{i=1}^{k} (a_i + b_i) y_i(x) \quad \text{for} \quad x \leqq \xi, \tag{38}$$

$$G(x, \xi) = \sum_{i=1}^{k} (a_i - b_i) y_i(x) \quad \text{for} \quad x \geqq \xi. \tag{39}$$

For the unknown coefficients b_i we get a system of equations

$$\begin{aligned} &\sum_{i=1}^{k} b_i y_i^{(l)}(\xi) = 0 \quad \text{for} \quad l = 0, 1, \ldots, k-2, \\ &\sum_{i=1}^{k} b_i y_i^{(k-1)}(\xi) = -\frac{1}{2p_k(\xi)} \end{aligned} \tag{40}$$

with a non-zero determinant; this system therefore has always a unique solution, $b_1 = b_1(\xi)$, $b_2 = b_2(\xi)$, ..., $b_k = b_k(\xi)$. The unknown coefficients a_i can then be found with the help of the boundary conditions (36), where we substitute the values of the function G and of its derivatives at the point a by means of (38) (because the form (38) holds true "on the left", i.e. for $x \leqq \xi$) and at the point b according to (39). If the conditions (36) are written briefly in the form

$$U_l(y) = 0,$$

then it is possible to show that the determinant of the system of equations for the evaluation of the coefficients a_i is

$$D = \begin{vmatrix} U_1(y_1), U_1(y_2), \ldots, U_1(y_k) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ U_k(y_1), U_k(y_2), \ldots, U_k(y_k) \end{vmatrix}. \tag{41}$$

Example 9. (See also Examples 10 and 11.) Let us consider a non-homogeneous problem

$$-y'' = x^2, \tag{42}$$

$$y(0) = 0, \quad y(1) = 0. \tag{43}$$

Then $L(y) = -y''$. As the fundamental system of equation $L(y) = 0$ (or $-y'' = 0$) we may take the functions

$$y_1 = x, \quad y_2 = 1.$$

Since $p_2(x) \equiv -1$, equations (40) are of the form

$$b_1\xi + b_2 = 0,$$
$$b_1 = \tfrac{1}{2},$$

whence

$$b_1 = \tfrac{1}{2}, \quad b_2 = -\tfrac{1}{2}\xi.$$

By virtue of (38), (39) we have

$$G(x, \xi) = \left(a_1 + \tfrac{1}{2}\right)x + \left(a_2 - \tfrac{1}{2}\xi\right) \quad \text{for} \quad x \leqq \xi, \tag{44}$$

$$G(x, \xi) = \left(a_1 - \tfrac{1}{2}\right)x + \left(a_2 + \tfrac{1}{2}\xi\right) \quad \text{for} \quad x \geqq \xi, \tag{45}$$

Substituting (44) into the first of equations (43) (we consider (44), because in the first of equation (43) only the "left" point $x = 0$ occurs), we get

$$a_2 = \tfrac{1}{2}\xi, \tag{46}$$

then substituting (45) into the second of equations (43) we get (now using (46))

$$a_1 = -\xi + \tfrac{1}{2}.$$

Thus

$$G(x, \xi) = -\xi x + x \quad \text{for} \quad x \leqq \xi, \quad G(x, \xi) = -\xi x + \xi \quad \text{for} \quad x \geqq \xi. \tag{47}$$

Obviously

$$D = \begin{vmatrix} 0, & 1 \\ 1, & 1 \end{vmatrix} \neq 0. \tag{48}$$

Theorem 11. *If for the problem* (35), (36) *the determinant* (41) *is different from zero, then there exists exactly one Green function, corresponding to the problem* (35), (36). *The (unique) solution of the problem is then given, for an arbitrary continuous right-hand side* $f(x)$ *(it is sufficient if* $f(x)$ *is piecewise continuous in* $[a, b]$*) by the equation*

$$y(x) = \int_a^b G(x, \xi) f(\xi)\, \mathrm{d}\xi. \tag{49}$$

The vanishing of the determinant (41) *is a necessary and sufficient condition for the corresponding homogeneous problem (i.e. for the problem* (35), (36) *with* $f(x) \equiv 0$*) to have a non-zero solution.*

(In this second case the solution of the problem (35), (36) (if it exists) is not unique. For example, a solution of the problem

$$y'' + \pi^2 y = \pi^2(2x - 1), \quad y(0) = 0, \quad y(1) = 0$$

is the function

$$y = \cos \pi x + 2x - 1,$$

but the function

$$y = \cos \pi x + 2x - 1 + 5 \sin \pi x$$

is as well a solution.)

REMARK 20. The Green function is often called the *influence function.* The reason lies in its technical meaning. Let us consider, for example, a bar supported at both ends a, b and loaded at the point $x = \xi$ by a concentrated load of unit value. Then the Green function corresponding to this problem is equal to the deflection $y(x)$, caused by this concentrated load. If, at this point, instead of a unit load a load $f(\xi)\Delta\xi$ is acting, the deflection will be equal to $G(x, \xi) f(\xi)\Delta\xi$. If a continuous load is considered, the deflection will be

$$y(x) = \int_a^b G(x, \xi) f(\xi)\, \mathrm{d}\xi.$$

This is a technical interpretation of equation (49) (in the general case the meaning is similar).

Example 10. In the case of the problem (42), (43), $D \neq 0$ (see (48)). According to Theorem 11, the solution $y(x)$ is then given by equation (49), and is unique:

$$y(x) = \int_0^1 G(x, \xi)\xi^2 \, d\xi = \int_0^x (-\xi x + \xi)\xi^2 \, d\xi + \int_x^1 (-\xi x + x)\xi^2 \, d\xi =$$
$$= -\frac{x^5}{4} + \frac{x^4}{4} - \frac{x}{4} + \frac{x}{3} + \frac{x^5}{4} - \frac{x^4}{3} = -\frac{x^4}{12} + \frac{x}{12}.$$

REMARK 21. Examples 9 and 10 are only illustrative. It is obvious that the problem (42), (43) could easily be solved without the construction of the Green function.

REMARK 22. Equation (35) often contains a parameter λ (cf. (5)). Then the Green function is a function of λ as well. (Here, the name *Green's resolvent* $G(x, \xi, \lambda)$ or simply *resolvent* is often used.)

Example 11. Let us consider the equation

$$L(y) = -y'' - \lambda y = 0 \quad (\lambda > 0), \tag{50}$$

with the boundary conditions

$$y(0) = 0, \quad y(1) = 0. \tag{51}$$

If we put $\lambda = k^2$, $k > 0$, then the functions $y_1 = \sin kx$, $y_2 = \cos kx$ form a fundamental system of the equation $L(y) = 0$, and using a procedure similar to that used in Example 9 we get

$$\begin{aligned} G(x, \xi) &= \left(\frac{\cos k\xi}{k} - \frac{\sin k\xi}{k} \cot k\right) \sin kx \quad \text{if} \quad x \leqq \xi, \\ G(x, \xi) &= \left(\frac{\cos kx}{k} - \frac{\sin kx}{k} \cot k\right) \sin k\xi \quad \text{if} \quad x \geqq \xi. \end{aligned} \tag{52}$$

The determinant (41) is equal to $\sin k$ in this case and is a function of λ. For $\lambda = \pi^2, 4\pi^2, 9\pi^2, \ldots$ we have $D = 0$ and the problem

$$-y'' - \lambda y = 0, \quad y(0) = 0, \quad y(1) = 0$$

has a non-zero solution (see Theorem 11).

Theorem 12. *If the operator L is self-adjoint* (see Definition 2) *and the boundary conditions are such that*

$$\int_a^b [uL(v) - vL(u)]\,dx = 0$$

for any two comparison functions $u(x)$, $v(x)$ (i.e. we have a symmetric problem), then $G(x, \xi)$ is a symmetric function of the variables x and ξ, i.e.

$$G(x, \xi) = G(\xi, x). \tag{53}$$

REMARK 23. This implies: If the function $G(x, \xi)$ for $x \leqq \xi$ is known, we obtain $G(x, \xi)$ for $x \geqq \xi$ by writing ξ instead of x and x instead ξ.

The symmetry of the problems treated in Examples 9 and 11 may be easily verified. The corresponding Green functions, (47) and (52) respectively, are, in fact, symmetric.

REMARK 24. Applying Green function, solution of problems discussed in this paragraph can be transformed into solution of integral equations:

Let us consider a symmetric eigenvalue problem (see Definition 4)

$$M(y) = \lambda N(y)$$

with boundary conditions (7). First of all, let

$$N(y) = g_0(x)y, \quad g_0(x) > 0 \quad \text{in} \quad [a, b].$$

For the given operator M and the given boundary conditions let us construct the Green function $G(x, \xi)$. By virtue of Theorem 12, $G(x, \xi) = G(\xi, x)$. Applying (49) (where we write $\lambda g_0(x)y$ for $f(x)$), we get, for the required function $y(x)$, an integral equation

$$y(x) = \lambda \int_a^b G(x, \xi)g_0(\xi)y(\xi)\,d\xi. \tag{54}$$

If we write now

$$\varphi(x) = \sqrt{[\,g_0(x)]}y(x), \quad K(x, \xi) = \sqrt{[\,g_0(x)g_0(\xi)]}G(x, \xi),$$

we get from (54) an integral equation with a *symmetric kernel*

$$\varphi(x) = \lambda \int_a^b K(x, \xi)\varphi(\xi)\,d\xi. \tag{55}$$

If the problem is not homogeneous, but the equation is of the form

$$M(y) = \lambda g_0(x)y + g(x),$$

we transform it into the corresponding non-homogeneous equation in the same way by writing $\lambda g_0(x)y + g(x)$ instead of $f(x)$ in (49). By this procedure, it is possible, in this simple case, to transform the investigation of a boundary value problem into the study of an integral equation.

The advantage of this procedure lies in the fact that it is often easy to construct the Green function corresponding to the operator M itself (with the corresponding boundary conditions). The analysis of the corresponding integral equation is then, as a rule, relatively easy because the theory of integral equations has been extensively developed.

In the case where the problem is regular (Remark 10), it can similarly be transformed into an equation of the form (55), where

$$\varphi(x) = \sqrt{[\, g_n(x)]}y^{(n)}(x), \quad K(x, \xi) = \frac{\partial^{2n}G(x, \xi)}{\partial x^n \partial \xi^n}\sqrt{[\, g_n(x)g_n(\xi)]}.$$

17.18. Systems of Ordinary Differential Equations

Let us consider a system of differential equations

$$F_i(x, y_1, y_1', \ldots, y_1^{(m_1)}, y_2, y_2', \ldots, y_2^{(m_2)}, \ldots, y_k, y_k', \ldots, y_k^{(m_k)}) = 0, \qquad (1)$$
$$i = 1, 2, \ldots, k,$$

for k unknown functions $y_1(x), y_2(x), \ldots, y_k(x)$. Equations (1) contain the derivatives of the required function $y_1(x)$ up to the order m_1 (although $y_1^{(m_1)}$ does not necessarily occur in every equation), the derivatives of the function $y_2(x)$ up to the order m_2, etc. The greatest of the numbers m_i is called the *order* of the given system. (In applications we frequently meet the case mentioned above, where the number of equations is equal to the number of unknown functions. In the general case, however, the number of equations need not be equal to the number of unknown functions.)

If the conditions of the theorem on implicit functions are satisfied, the system (1) can be solved with respect to the highest derivatives of the unknown functions and written in the so-called *canonical form*

$$\begin{aligned} y_1^{(m_1)} &= g_1\big(x, y_1 \ldots, y_1^{(m_1-1)}, y_2, \ldots, y_2^{(m_2-1)}, \ldots, y_k \ldots, y_k^{(m_k-1)}\big), \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_k^{(m_k)} &= g_k\big(x, y_1, \ldots, y_1^{(m_1-1)}, y_2, \ldots, y_2^{(m_2-1)} \ldots, y_k, \ldots, y_k^{(m_k-1)}\big). \end{aligned} \qquad (2)$$

Now, new unknown functions can be introduced by the relations

$$y_{11} = y_1', y_{12} = y_{11}' = y_1'', \ldots, y_{1,m_1-1} = y_1^{(m_1-1)}$$

and similarly

$$y_{21} = y_2', y_{22} = y_{21}' = y_2'', \ldots, y_{2,m_2-1} = y_2^{(m_2-1)}, \ldots,$$

so that we obtain $m_1 + m_2 + \ldots + m_k$ equations of the form

$$\begin{aligned}
y_1' &= y_{11}, \\
y_{11}' &= y_{12}, \\
&\ldots\ldots\ldots\ldots \\
y_{1,m_1-1}' &= g_1(x, y_1, y_{11}, \ldots, y_{1,m_1-1}, y_2, \ldots, y_{2,m_2-1}, \ldots, y_k \ldots, y_{k,m_k-1}), \\
y_2' &= y_{21}, \\
&\ldots\ldots\ldots\ldots
\end{aligned} \tag{3}$$

In this way, the system (2) has been transformed into the system (3) of the first order. These two systems are equivalent, i.e. every solution of the system (3) is a solution of the system (2) (with $y_{11} = y_1'$, etc.), and vice versa.

It is thus sufficient to investigate systems of the first order only, i.e. systems of the form

$$\begin{aligned}
y_1' &= f_1(x, y_1, y_2, \ldots, y_n), \\
&\ldots\ldots\ldots\ldots\ldots\ldots \\
y_n' &= f_n(x, y_1, y_2, \ldots, y_n).
\end{aligned} \tag{4}$$

Such systems of differential equations are often called *normal*. For the concept of a solution (integral) of the system (4) see Definition 17.2.4. On existence of a solution which fulfils prescribed initial conditions

$$y_1(a) = b_1, y_2(a) = b_2, \ldots, y_n(a) = b_n \tag{5}$$

see Theorem 17.2.1. In the sense of Remark 17.2.5 we often speak of an *integral curve of the system* (4) *passing through the* $(n + 1)$*-dimensional point* $P(a, b_1, b_2, \ldots, b_n)$.

Similarly as in the case of one differential equation (cf Definition 17.2.5), one often speaks (especially in technical literature) of a general integral of the system (4): Let Q be an $(n + 1)$-dimensional region constituted of such points $P(a, b_1, b_2, \ldots, b_n)$ for which (i.e. for the initial conditions (5)) the system (4) has exactly one solution in the sense of Theorem 17.2.1.

Definition 1. Under a *general integral* (*general solution, general form of solution*) of the system (4), with respect to the region Q, we understand such a system of functions

$$y_1 = g_1(x, C_1, C_2, \ldots, C_n),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$y_n = g_n(x, C_1, C_2, \ldots, C_n)$$

which contain, besides x, n parametres C_1, C_2, ..., C_n and which constitute, as functions of x, a solution of the system (4) for arbitrarily chosen values of these parametres. At the same time, the parameters C_1, C_2, ..., C_n are independent in the following sense: If an arbitrary point $P \in Q$ has been chosen, these parameters can uniquely be given such values that this solution fulfils conditions (5) given by the point P.

This concept deserves a remark similar to Remark 17.2.14: In the case of non-linear systems, this concept is not very fitting, because, generally speaking, not all solutions of the system (4) are contained in the general integral (cf. the quoted remark 17.2.14; this is one of the reasons why this concept is often not introduced in the mathematical literature at all). Very natural is this concept in the case of linear systems (because of Theorem 3 below):

Definition 2. The system

$$\begin{aligned} y_1' &= a_{11}(x)y_1 + a_{12}(x)y_2 + \ldots + a_{1n}(x)y_n + h_1(x), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n' &= a_{n1}(x)y_1 + a_{n2}(x)y_2 + \ldots + a_{nn}(x)y_n + h_n(x) \end{aligned} \tag{6}$$

is called *linear*. If all functions $h_i(x)$ are identically zero, the system is called *homogeneous*; if they are not, then it is *non-homogeneous*.

REMARK 1. In modern texts, the system (6) is written briefly in a *vector* (*matrix*) *form* in the following way (cf. § 17.2):

$$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{h},$$

where $\boldsymbol{A}$ is the square matrix formed by the coefficients $a_{ik}(x)$ of the system (6), $\boldsymbol{y}$, $\boldsymbol{y}'$, $\boldsymbol{h}$ are vectors (one-column matrices) formed by the functions $y_i(x)$, $y_i'(x)$, $h_i(x)$:

$$\boldsymbol{A} = \begin{bmatrix} a_{11}(x), & a_{12}(x), & \ldots, & a_{1n}(x) \\ a_{21}(x), & a_{22}(x), & \ldots, & a_{2n}(x) \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_{n1}(x), & a_{n2}(x), & \ldots, & a_{nn}(x) \end{bmatrix},$$

$$\boldsymbol{y} = \begin{bmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_n(x) \end{bmatrix}, \quad \boldsymbol{y}' = \begin{bmatrix} y_1'(x) \\ y_2'(x) \\ \vdots \\ y_n'(x) \end{bmatrix}, \quad \boldsymbol{h} = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_n(x) \end{bmatrix}.$$

The product $\boldsymbol{Ay}$ is the matrix product according to Definition 1.25.3. The system (4) can also be written in the vector form

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}),$$

where

$$\boldsymbol{f}(x, \boldsymbol{y}) = \begin{bmatrix} f_1(x, y_1, y_2, \ldots, y_n) \\ f_2(x, y_1, y_2, \ldots, y_n) \\ \ldots\ldots\ldots\ldots\ldots \\ f_n(x, y_1, y_2, \ldots, y_n) \end{bmatrix}.$$

If x is not explicitly contained in the functions $f_1, \ldots, f_n$ (thus if (4) can be written in the form

$$\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y})),$$

then the system (4) is called *autonomous.* An example of such a system is the system (6), provided all the coefficients a_{ik} and the members h_i are constants.

Theorem 1. *If all functions $a_{ik}(x)$, $h_i(x)$ are continuous in an interval I, then corresponding to any $a \in I$ and to n arbitrarily chosen numbers $b_1, b_2, \ldots, b_n$ there exists precisely one solution of the system* (6)*, defined in the whole interval I and satisfying the given initial conditions, i.e. there exists exactly one system of functions $y_1(x), y_2(x), \ldots, y_n(x)$ defined in the whole interval I, satisfying the system* (6) *in I and the conditions*

$$y_1(a) = b_1, y_2(a) = b_2, \ldots, y_n(a) = b_n.$$

Definition 3. Let a *homogeneous* system (6), i.e. the system

$$\begin{aligned} y_1' &= a_{11}(x)y_1 + a_{12}(x)y_2 + \ldots + a_{1n}(x)y_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n' &= a_{n1}(x)y_1 + a_{n2}(x)y_2 + \ldots + a_{nn}(x)y_n \end{aligned} \tag{6'}$$

be given. Let us have n solutions of this system, i.e. n systems of functions

$$\begin{aligned} &{}^1y_1, {}^1y_2, \ldots, {}^1y_n, \\ &\ldots\ldots\ldots\ldots\ldots \\ &{}^ny_1, {}^ny_2, \ldots, {}^ny_n. \end{aligned} \tag{7}$$

(The index i before the function y_k denotes that the function corresponds to a certain i-th solution of the given system). The system of solutions (7) is called a *fundamental system (of solutions)* of the homogeneous system (6') in the interval I, if the determinant

$$D = \begin{vmatrix} {}^1y_1, {}^1y_2, \ldots, {}^1y_n \\ \ldots\ldots\ldots\ldots\ldots \\ {}^ny_1, {}^ny_2, \ldots, {}^ny_n \end{vmatrix} \tag{8}$$

is non-zero in I, i.e. if it does not vanish at any point of I.

Theorem 2. *Let the functions $a_{ik}(x)$ be continuous in the interval I. Then $D(x)$ is either non-zero or identically equal to zero in the whole interval I.*

REMARK 2. It is thus sufficient to evaluate the determinant (8) at a single point $x_0 \in I$.

Theorem 3. *Let (7) be a fundamental system of solutions of the homogeneous system (6′) in the interval I. Then every solution of this system can be expressed, in I, in the form*

$$\begin{aligned} y_1 &= C_1{}^1y_1 + C_2{}^2y_1 + \ldots + C_n{}^ny_1, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= C_1{}^1y_n + C_2{}^2y_n + \ldots + C_n{}^ny_n, \end{aligned} \tag{9}$$

where $C_1, C_2, \ldots, C_n$ are suitable numbers.

REMARK 3. Thus *all solutions of the linear homogeneous system (6′) form an n-dimensional linear vector space.*

Because, under assumptions of Theorem 3, the determinant (8) is different from zero everywhere in I, it is possible to fulfil, in I, arbitrary initial conditions by a suitable (and unique) choice of the numbers $C_1, C_2, \ldots, C_n$. In the sense of Definition 1 we call (9) a *general integral of the system (6′)*.

If we denote by

$$\boldsymbol{c} = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_n \end{bmatrix}$$

a vector of constants, then the general integral (9) can be written in the vector form

$$\boldsymbol{y} = \boldsymbol{M}\boldsymbol{c},$$

where

$$\boldsymbol{M} = \begin{bmatrix} {}^1y_1, {}^2y_1, \ldots, {}^ny_1 \\ {}^1y_2, {}^2y_2, \ldots, {}^ny_2 \\ \ldots\ldots\ldots\ldots\ldots \\ {}^1y_n, {}^2y_n, \ldots, {}^ny_n \end{bmatrix}$$

is the so-called *fundamental matrix of the homogeneous system (6′)* (its columns are formed by rows of the scheme (7)).

REMARK 4 (*solution*). Let us consider a homogeneous system

$$\begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 + \ldots + a_{1n}y_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n' &= a_{n1}y_1 + a_{n2}y_2 + \ldots + a_{nn}y_n, \end{aligned} \tag{10}$$

where the coefficients a_{ik} are constants (complex, in general). The assumption

$$y_1 = k_1 e^{\lambda x}, y_2 = k_2 e^{\lambda x}, \ldots, y_n = k_n e^{\lambda x} \tag{11}$$

leads, after substituting (11) into (10), to the system

$$\begin{aligned} &(a_{11} - \lambda)k_1 + a_{12}k_2 + \ldots + a_{1n}k_n = 0, \\ &a_{21}k_1 + (a_{22} - \lambda)k_2 + \ldots + a_{2n}k_n = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &a_{n1}k_1 + a_{n2}k_2 + \ldots + (a_{nn} - \lambda)k_n = 0, \end{aligned} \tag{12}$$

which has a non-zero solution if and only if (cf. § 1.18)

$$\begin{vmatrix} a_{11} - \lambda, & a_{12}, & \ldots, & a_{1n} \\ a_{21}, & a_{22} - \lambda, & \ldots, & a_{2n} \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots \\ a_{n1}, & a_{n2}, & \ldots, & a_{nn} - \lambda \end{vmatrix} = 0. \tag{13}$$

I. Let the so-called *characteristic equation* (13) have n *different roots* $\lambda_1, \lambda_2, \ldots, \lambda_n$. Substituting λ_1 into (12), we get, on solving the system (12), a system of numbers ${}^1k_1, {}^1k_2, \ldots, {}^1k_n$, determined uniquely except for an *arbitrary factor*. (If ${}^1k_1 \neq 0$, then ${}^1k_1, \ldots, {}^1k_n$ are uniquely determined if we choose, e.g., ${}^1k_1 = 1$.) The first system of functions of the fundamental system is then

$${}^1y_1 = {}^1k_1 e^{\lambda_1 x}, {}^1y_2 = {}^1k_2 e^{\lambda_1 x}, \ldots, {}^1y_n = {}^1k_n e^{\lambda_1 x}. \tag{14}$$

Similarly, on substituting the values $\lambda_2, \ldots, \lambda_n$ into (12), we get further systems of functions of the fundamental system required,

$$\begin{aligned} &{}^2y_1 = {}^2k_1 e^{\lambda_2 x}, {}^2y_2 = {}^2k_2 e^{\lambda_2 x}, \ldots, {}^2y_n = {}^2k_n e^{\lambda_2 x}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &{}^ny_1 = {}^nk_1 e^{\lambda_n x}, {}^ny_2 = {}^nk_2 e^{\lambda_n x}, \ldots, {}^ny_n = {}^nk_n e^{\lambda_n x}. \end{aligned} \tag{15}$$

The general integral is then, by virtue of Theorem 3 (and Remark 3),

$$\begin{aligned} &y_1 = C_1 {}^1k_1 e^{\lambda_1 x} + C_2 {}^2k_1 e^{\lambda_2 x} + \ldots + C_n {}^nk_1 e^{\lambda_n x}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &y_n = C_1 {}^1k_n e^{\lambda_1 x} + C_2 {}^2k_n e^{\lambda_2 x} + \ldots + C_n {}^nk_n e^{\lambda_n x}. \end{aligned} \tag{16}$$

Example 1.

$$\begin{aligned} y_1' &= 4y_1 - 2y_2, \\ y_2' &= y_1 + y_2. \end{aligned} \tag{17}$$

Equation (13),

$$\begin{vmatrix} 4-\lambda, & -2 \\ 1, & 1-\lambda \end{vmatrix} = 0$$

has the roots $\lambda_1 = 2$, $\lambda_2 = 3$. The solution of (12) with $\lambda = \lambda_1 = 2$, i.e. of

$$\begin{aligned} 2k_1 - 2k_2 &= 0, \\ k_1 - k_2 &= 0, \end{aligned}$$

is $k_1 = 1$, $k_2 = 1$, so that

$$^1y_1 = \mathrm{e}^{2x}, \quad ^1y_2 = \mathrm{e}^{2x}. \tag{18}$$

(It is advisable to verify that the functions (18) satisfy equations (17).) Substituting $\lambda = \lambda_2 = 3$ into (12), we get

$$\begin{aligned} k_1 - 2k_2 &= 0, \\ k_1 - 2k_2 &= 0, \end{aligned}$$

whence $k_1 = 2$, $k_2 = 1$ and

$$^2y_1 = 2\,\mathrm{e}^{3x}, \quad ^2y_2 = \mathrm{e}^{3x}. \tag{19}$$

The general integral is then, according to (16),

$$\begin{aligned} y_1 &= C_1\,\mathrm{e}^{2x} + 2C_2\,\mathrm{e}^{3x}, \\ y_2 &= C_1\,\mathrm{e}^{2x} + C_2\,\mathrm{e}^{3x}. \end{aligned} \tag{20}$$

If the initial conditions are, for example, $y_1(0) = 3$, $y_2(0) = -1$, we get from (20)

$$C_1 + 2C_2 = 3, \quad C_1 + C_2 = -1$$

so that

$$C_1 = -5, \quad C_2 = 4$$

and the required solution is

$$y_1 = -5\,\mathrm{e}^{2x} + 8\,\mathrm{e}^{3x}, \quad y_2 = -5\,\mathrm{e}^{2x} + 4\,\mathrm{e}^{3x}.$$

REMARK 5. It may happen that the characteristic equation has complex roots. If the coefficients in (10) are *real*, the solution may nevertheless be found in real form. Let us consider first the case of two simple complex conjugate roots

$$\lambda_1 = \alpha + \mathrm{i}\beta, \quad \lambda_2 = \alpha - \mathrm{i}\beta, \quad \alpha, \beta \text{ real}.$$

To these roots, two systems of real functions correspond in the fundamental system,

$$^1y_j = \mathrm{e}^{\alpha x}\left(^1l_j \cos\beta x - {}^2l_j \sin\beta x\right) \quad (j = 1, 2, \ldots, n), \tag{21}$$

$$^2y_j = \mathrm{e}^{\alpha x}\left(^1l_j \sin\beta x + {}^2l_j \cos\beta x\right) \quad (j = 1, 2, \ldots, n), \tag{22}$$

where ${}^1l_j$, ${}^2l_j$ are numbers given by the equations

$$ {}^1k_j = {}^1l_j + \mathrm{i}\,{}^2l_j \quad \text{and} \quad {}^2k_j = {}^1l_j - \mathrm{i}\,{}^2l_j $$

and ${}^1k_j$, ${}^2k_j$ are the solutions of equations (12) for $\lambda_1 = \alpha + \mathrm{i}\beta$, $\lambda_2 = \alpha - \mathrm{i}\beta$, respectively.

Similarly, to another pair of simple conjugate values of λ, another two systems of real functions correspond in the fundamental system. For real λ, the corresponding systems of functions can then be obtained according to Remark 4 and Example 1.

Example 2. Let us consider the system

$$ \begin{aligned} y_1' &= -7y_1 + y_2, \\ y_2' &= -2y_1 - 5y_2. \end{aligned} \tag{23} $$

Equation (13) is

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

or $\lambda^2 + 12\lambda + 37 = 0$, and has roots $\lambda_1 = -6 + \mathrm{i}$, $\lambda_2 = -6 - \mathrm{i}$. Thus $\alpha = -6$, $\beta = 1$. Substituting λ_1 and λ_2 into (12), we get (apart from a constant factor)

$$ {}^1k_1 = 1, \quad {}^1k_2 = 1 + \mathrm{i}, \quad \text{and} \quad {}^2k_1 = 1, \quad {}^2k_2 = 1 - \mathrm{i}, $$

thus

$$ {}^1l_1 = 1, \quad {}^2l_1 = 0, \quad {}^1l_2 = 1, \quad {}^2l_2 = 1. $$

Then the systems of functions that form the fundamental system are, according to (21), (22),

$$ \begin{aligned} {}^1y_1 &= \mathrm{e}^{-6x}\cos x, & {}^1y_2 &= \mathrm{e}^{-6x}(\cos x - \sin x), \\ {}^2y_1 &= \mathrm{e}^{-6x}\sin x, & {}^2y_2 &= \mathrm{e}^{-6x}(\sin x + \cos x), \end{aligned} $$

and the general integral is

$$ \begin{aligned} y_1 &= \mathrm{e}^{-6x}(C_1\cos x + C_2\sin x), \\ y_2 &= \mathrm{e}^{-6x}\left[(C_1 + C_2)\cos x + (C_2 - C_1)\sin x\right]. \end{aligned} $$

II. If λ_i is a root of the characteristic equation of multiplicity r, then it can be shown that solutions which correspond to this root in the fundamental system are of the form

$$ y_j = P_{ij}\,\mathrm{e}^{\lambda_i x} \tag{24} $$

where $P_{ij}(x)$ are polynomials of degree $r-1$ at most. The practical procedure will be shown in an example:

Example 3. Let us consider the system

$$\begin{aligned} y_1' &= -y_1 + y_2, \\ y_2' &= \quad - y_2 + 4y_3, \\ y_3' &= \; y_1 \quad\quad - 4y_3. \end{aligned} \tag{25}$$

Equation (13),

$$\begin{vmatrix} -1-\lambda, & 1, & 0 \\ 0, & -1-\lambda, & 4 \\ 1, & 0, & -4-\lambda \end{vmatrix} = -\lambda^3 - 6\lambda^2 - 9\lambda = 0$$

has a double root $\lambda_1 = -3$ and a single root $\lambda_2 = 0$. According to (24), the system of solutions corresponding to the root $\lambda_1 = -3$ will be of the form

$${}^1y_1 = \mathrm{e}^{-3x}(a_1 + a_2x), \quad {}^1y_2 = \mathrm{e}^{-3x}(b_1 + b_2x), \quad {}^1y_3 = \mathrm{e}^{-3x}(c_1 + c_2x). \tag{26}$$

Substituting (26) into (25), we obtain for the unknown constants $a_1, \ldots, c_2$ the following equations

$$\begin{aligned} 2a_1 - a_2 + b_1 &= 0, & 2a_2 + b_2 &= 0, \\ 2b_1 - b_2 + 4c_1 &= 0, & 2b_2 + 4c_2 &= 0, \\ -c_1 - c_2 + a_1 &= 0, & -c_2 + a_2 &= 0. \end{aligned} \tag{27}$$

Choosing $a_2 = C_1$ (C_1 being an arbitrary non-zero constant), the second column of equations (27) gives $b_2 = -2C_1$, $c_2 = C_1$. Further, choosing $a_1 = C_2$, the first column of equations (27) gives $b_1 = C_1 - 2C_2$, $c_1 = C_2 - C_1$. Thus, to the root $\lambda_1 = -3$ there correspond the solutions

$$\begin{gathered} {}^1y_1 = \mathrm{e}^{-3x}(C_2 + C_1x), \quad {}^1y_2 = \mathrm{e}^{-3x}(C_1 - 2C_2 - 2C_1x), \\ {}^1y_3 = \mathrm{e}^{-3x}(C_2 - C_1 + C_1x). \end{gathered} \tag{28}$$

Solutions corresponding to the root $\lambda_2 = 0$ can be obtained either by virtue of Remark 4 and Example 1 or can be supposed, according to (24), to have the form ($r = 1$)

$${}^2y_1 = a, \quad {}^2y_2 = b, \quad {}^2y_3 = c.$$

Substituting into (25), we get

$${}^2y_1 = 4C_3, \quad {}^2y_2 = 4C_3, \quad {}^2y_3 = C_3. \tag{29}$$

The general integral of the system (25) is then, according to (28) and (29),

$$\begin{aligned} y_1 &= e^{-3x}(C_2 + C_1 x) + 4C_3, \\ y_2 &= e^{-3x}(C_1 - 2C_2 - 2C_1 x) + 4C_3, \\ y_3 &= e^{-3x}(C_2 - C_1 + C_1 x) + C_3. \end{aligned}$$

REMARK 6. If the coefficients of the system (10) are real and if the r-fold root λ is complex, i.e. $\lambda = \alpha + i\beta$, then equation (13) has also the r-fold root $\lambda = \alpha - i\beta$ and the solution can be supposed to have the form

$$y_j = e^{\alpha x}(P_{ij} \cos \beta x + Q_{ij} \sin \beta x),$$

where P, Q are polynomials of degree $r - 1$ at most (cf. also Remarks 5 and 7).

Let us proceed to non-homogeneous linear systems.

Theorem 4. *Let a non-homogeneous system of linear equations*

$$\begin{aligned} y_1' &= a_{11}(x)y_1 + a_{12}(x)y_2 + \ldots + a_{1n}(x)y_n + f_1(x), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n' &= a_{n1}(x)y_1 + a_{n2}(x)y_2 + \ldots + a_{nn}(x)y_n + f_n(x) \end{aligned} \tag{30}$$

be given. Let the functions $a_{ik}(x)$, $f_i(x)$ be continuous in the interval I and let

$$\begin{aligned} &{}^1y_1, {}^1y_2, \ldots, {}^1y_n, \\ &\ldots\ldots\ldots\ldots \\ &{}^ny_1, {}^ny_2, \ldots, {}^ny_n \end{aligned} \tag{31}$$

be the fundamental system (Definition 3) *of the homogeneous system* (6′) *corresponding to the system* (30) (*i.e. for $f_i(x) \equiv 0$*). *Then*

I. *the general integral of the system* (30) *is of the form*

$$\begin{aligned} y_1 &= y_{1p} + C_1 {}^1y_1 + C_2 {}^2y_1 + \ldots + C_n {}^ny_1, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= y_{np} + C_1 {}^1y_n + C_2 {}^2y_n + \ldots + C_n {}^ny_n, \end{aligned} \tag{32}$$

where $y_{1p}, y_{2p}, \ldots, y_{np}$ is a particular solution of the system (30) ($C_1, C_2, \ldots, C_n$ *are arbitrary parameters*);

II. *the fundamental system* (31) *being known, the particular solution $y_{1p}, y_{2p}, \ldots, y_{np}$ can be obtained by the method of variation of parameters*:

$$\begin{aligned} y_{1p} &= C_1(x) {}^1y_1 + C_2(x) {}^2y_1 + \ldots + C_n(x) {}^ny_1, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{np} &= C_1(x) {}^1y_n + C_2(x) {}^2y_n + \ldots + C_n(x) {}^ny_n. \end{aligned} \tag{33}$$

It suffices to choose the functions $C_1(x)$, $C_2(x)$, ..., $C_n(x)$ *so as to satisfy the system of equations*

$$\begin{aligned} &C_1'^{1}y_1 + C_2'^{2}y_1 + \ldots + C_n'^{n}y_1 = f_1(x), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &C_1'^{1}y_n + C_2'^{2}y_n + \ldots + C_n'^{n}y_n = f_n(x) \end{aligned} \tag{34}$$

(*which is uniquely solvable*).

Example 4. Let us consider the system

$$\begin{aligned} y_1' &= \quad y_2 + \cos x, \\ y_2' &= -y_1 + 1. \end{aligned}$$

The fundamental system (obtained in a similar way as in Example 2) is

$$\begin{aligned} {}^1y_1 &= \cos x, \ {}^1y_2 = -\sin x, \\ {}^2y_1 &= \sin x, \ {}^2y_2 = \quad \cos x. \end{aligned}$$

System (34):

$$\begin{aligned} C_1' \cos x + C_2' \sin x &= \cos x, \\ -C_1' \sin x + C_2' \cos x &= 1, \end{aligned}$$

whence

$$\begin{aligned} C_1'(x) &= \cos^2 x - \sin x, & C_1(x) &= \tfrac{1}{2}x + \tfrac{1}{2}\sin x \cos x + \cos x, \\ C_2'(x) &= \sin x \cos x + \cos x, & C_2(x) &= -\tfrac{1}{2}\cos^2 x + \sin x. \end{aligned}$$

(Constants of integration can be omitted, because we are trying to find only a special particular solution of (30).) Thus, by virtue of (33), we get

$$\begin{aligned} y_{1p} &= \left(\tfrac{1}{2}x + \tfrac{1}{2}\sin x \cos x + \cos x\right)\cos x + \left(-\tfrac{1}{2}\cos^2 x + \sin x\right)\sin x = \\ &= \tfrac{1}{2}x \cos x + 1, \\ y_{2p} &= -\left(\tfrac{1}{2}x + \tfrac{1}{2}\sin x \cos x + \cos x\right)\sin x + \left(-\tfrac{1}{2}\cos^2 x + \sin x\right)\cos x = \\ &= -\tfrac{1}{2}x \sin x - \tfrac{1}{2}\cos x. \end{aligned}$$

The general integral is

$$\begin{aligned} y_1 &= \tfrac{1}{2}x \cos x + 1 + C_1 \cos x + C_2 \sin x, \\ y_2 &= -\tfrac{1}{2}x \sin x - \tfrac{1}{2}\cos x - C_1 \sin x + C_2 \cos x \end{aligned}$$

where C_1, C_2 are arbitrary parameters.

REMARK 7. Linear systems of equations can often be solved by transforming them into one linear equation of the n-th order for one unknown function:

Example 5. Let us consider the system (17) (Example 1). From the second equation it follows that

$$y_1 = y_2' - y_2, \quad \text{thus} \quad y_1' = y_2'' - y_2'. \tag{35}$$

Substituting into the first equation, we get

$$y_2'' - 5y_2' + 6y_2 = 0. \tag{36}$$

The characteristic equation $\lambda^2 - 5\lambda + 6 = 0$ is the same as in Example 1. The general solution of equation (36) is

$$y_2 = C_1 \mathrm{e}^{2x} + C_2 \mathrm{e}^{3x}.$$

Substituting into the first equation (35), we then get

$$y_1 = 2C_1 \mathrm{e}^{2x} + 3C_2 \mathrm{e}^{3x} - C_1 \mathrm{e}^{2x} - C_2 \mathrm{e}^{3x} = C_1 \mathrm{e}^{2x} + 2C_2 \mathrm{e}^{3x},$$

which agrees with (20).

This procedure can be used even in the case of non-homogeneous linear equations.

REMARK 8 (*Solution of System* (10) *in Exponential Form, Notes on Matrix Analysis and on Sequences and Series of Matrices*). Similarly as in the case of one homogeneous linear equation with constants coefficients, the solution of the system (10) can be written in an exponential form. First, we introduce two short notes:

(i) (*A Note on Matrix Analysis.*) Let $\boldsymbol{A} = (a_{ij})$, $i, j = 1, 2, \ldots, n$, be a (real or complex) square matrix. Let its elements a_{ij} be functions of a (not necessarily real) variable x, $a_{ij} = a_{ij}(x)$, with the same domain of definition D. Then the matrix $\boldsymbol{A}$ is also a function of x, we write $\boldsymbol{A} = \boldsymbol{A}(x)$. Let each of the functions $a_{ij}(x)$ have a limit l_{ij} for $x \to x_0$. Then we say that the matrix $\boldsymbol{A}(x)$ has a limit at the point x_0 and write

$$\lim_{x \to x_0} \boldsymbol{A}(x) = (l_{ij}). \tag{37}$$

Similarly continuity is defined, as well as the derivative

$$\frac{\mathrm{d}\boldsymbol{A}}{\mathrm{d}x}(x_0) = \boldsymbol{A}'(x_0) = (a_{ij}'(x_0)). \tag{38}$$

It easily follows that well-known rules from analysis are preserved, thus that we have, for example,

$$\lim_{x \to x_0} (\boldsymbol{A} + \boldsymbol{B}) = \lim_{x \to x_0} \boldsymbol{A} + \lim_{x \to x_0} \boldsymbol{B},$$

$$(\boldsymbol{A} + \boldsymbol{B})' = \boldsymbol{A}' + \boldsymbol{B}', \quad (\boldsymbol{A}\boldsymbol{B})' = \boldsymbol{A}'\boldsymbol{B} + \boldsymbol{A}\boldsymbol{B}', \quad \text{etc.}$$

(ii) (*Sequences and Series of Matrices.*) Let a sequence of square matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2, \ldots, \boldsymbol{A}_k, \ldots$ of order n be given, with $\boldsymbol{A}_k = (a_{ij}^{(k)})$, $a_{ij}^{(k)}$ being numbers, first. If

$$\lim_{k\to\infty} a_{ij}^{(k)} = a_{ij} \tag{39}$$

holds for every couple i, j $(i, j = 1, 2, \ldots, n)$, then we say that this sequence of matrices is *convergent* and write

$$\lim_{k\to\infty} \boldsymbol{A}_k = (a_{ij}). \tag{40}$$

Let a series be given by

$$a_0\boldsymbol{I} + a_1\boldsymbol{A} + a_2\boldsymbol{A}^2 + \ldots + a_k\boldsymbol{A}^k + \ldots, \tag{41}$$

where $a_0, a_1, \ldots, a_k, \ldots$ are numbers, $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{A}^2 = \boldsymbol{A}\boldsymbol{A}$ is the product of the (square) matrices $\boldsymbol{A}$, $\boldsymbol{A}$ (§ 1.25), $\boldsymbol{A}^3 = \boldsymbol{A}^2\boldsymbol{A}$, etc. Denote

$$\boldsymbol{S}_k = a_0\boldsymbol{I} + a_1\boldsymbol{A} + a_2\boldsymbol{A}^2 + \ldots + a_k\boldsymbol{A}^k. \tag{42}$$

If the sequence of the partial sums (42) converges to a matrix $\boldsymbol{S}$, then we say that the *series* (41) *is convergent and has the sum* $\boldsymbol{S}$.

These definitions may easily be extended to matrices with non-constant terms.

In particular, let $\boldsymbol{A}x = x\boldsymbol{A}$ be the matrix $\boldsymbol{A} = (a_{ij})$ multiplied by the number x, thus $\boldsymbol{A}x = (xa_{ij})$. Then the series

$$\boldsymbol{I} + \frac{\boldsymbol{A}x}{1!} + \frac{(\boldsymbol{A}x)^2}{2!} + \ldots + \frac{(\boldsymbol{A}x)^k}{k!} + \ldots \tag{43}$$

can be shown to be convergent for every x (real, or complex). It is usual to denote its sum (which is thus a matrix dependent on x) by $\mathrm{e}^{\boldsymbol{A}x}$. It can further be shown that this matrix has a derivative with respect to x and that we have

$$(\mathrm{e}^{\boldsymbol{A}x})' = \boldsymbol{A}\,\mathrm{e}^{\boldsymbol{A}x}. \tag{44}$$

(iii) If we write the system (10) with constant coefficients in the form

$$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y} \tag{45}$$

(Remark 1), then using (44), its general integral can be written in the form

$$\boldsymbol{y} = \mathrm{e}^{\boldsymbol{A}x}\,\boldsymbol{C}, \tag{46}$$

where $\boldsymbol{C}$ is the vector from Remark 3 with components $C_1, C_2, \ldots, C_n$. Thus the solution of the system (10) can be written in a very compact form.

17.19. Dependence of Solutions of Systems of Differential Equations on Initial Conditions and on Parameters of the System. Stability of Solutions

Let us consider the system of equations

$$\begin{aligned} y_1' &= f_1(x, y_1, y_2, \ldots, y_n), \\ y_2' &= f_2(x, y_1, y_2, \ldots, y_n), \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n' &= f_n(x, y_1, y_2, \ldots, y_n) \end{aligned} \tag{1}$$

with initial conditions

$$y_1(a) = b_1, y_2(a) = b_2, \ldots, y_n(a) = b_n. \tag{2}$$

Denote

$$\begin{aligned} y_1 &= \varphi_1(x, a, b_1, b_2, \ldots, b_n), \\ y_2 &= \varphi_2(x, a, b_1, b_2, \ldots, b_n), \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= \varphi_n(x, a, b_1, b_2, \ldots, b_n) \end{aligned} \tag{3}$$

the solution of system (1) satisfying conditions (3), thus (in the terminology of §17.2) the integral curve of system (1) passing through the point $P(a, b_1, b_2, \ldots, b_n)$. It can be shown that, the functions $f_1, f_2, \ldots, f_n$ being "reasonable", the solution (3) depends continuously on initial conditions (2), i.e. if the numbers $a, b_1, b_2, \ldots, b_n$ in (2) have been "slightly" changed, then also the functions (3) change only "slightly". In details:

Theorem 1. *Let Ω be an $(n+1)$-dimensional region in which existence and uniqueness of solutions are guaranteed, i.e. let exactly one integral curve of the system (1) pass (locally) through every point $P(a, b_1, b_2, \ldots, b_n) \in \Omega$. If the functions f_i $(i = 1, 2, \ldots, n)$ are continuous in Ω (as functions of their $n+1$ variables), then the functions φ_i are as well continuous (as functions of their $n+2$ variables) for $(a, b_1, b_2, \ldots, b_n) \in \Omega$ and for x from a certain neighbourhood of the point a. Moreover, if the functions f_i have continuous partial derivatives with respect to their $n+1$ variables up to the order k, then the functions φ_i have as well continuous partial derivatives with respect to their $n+2$ variables up to the order k.*

REMARK 1. If, in addition, the functions f_i contain parameters $p_1, p_2, \ldots, p_s$, then also the functions φ_i are functions of these parameters. If the functions f_i have continuous partial derivatives, with respect to p_j $(j = 1, 2, \ldots, s)$, up to the order k, then the same holds for the functions φ_i.

Let us change the initial conditions (2) into the conditions

$$y_1(a) = c_1, y_2(a) = c_2, \ldots, y_n(a) = c_n, \tag{4}$$

so that the corresponding solution becomes

$$\begin{aligned} y_1 &= \varphi_1(x, a, c_1, c_2, \ldots, c_n), \\ y_2 &= \varphi_2(x, a, c_1, c_2, \ldots, c_n), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= \varphi_n(x, a, c_1, c_2, \ldots, c_n). \end{aligned} \tag{5}$$

If the solutions (3) and (5) are defined for x not only from a certain neighbourhood of the point a, but in a whole interval $[a_0, +\infty)$ and if for every $a \geqq a_0$ the difference between the solutions (3) and (5) is "small" in the interval $[a, +\infty)$ whenever the difference between the numbers b_i and c_i is "small", we speak about *stability* of the solution (3). In details:

Definition 1. We say that the solution (3) is *stable in the sense of Liapunov* (briefly *Liapunov stable*, or only *stable*), if it is defined for all $x \geqq a_0$ and if to every $\varepsilon > 0$ and $a \geqq a_0$ such a $\delta(\varepsilon, a) > 0$ exists that we have, for all $x \geqq a$,

$$|\varphi_i(x, a, c_1, c_2, \ldots, c_n) - \varphi_i(x, a, b_1, b_2, \ldots, b_n)| < \varepsilon, \quad i = 1, 2, \ldots, n, \tag{6}$$

whenever

$$|c_i - b_i| < \delta, \quad i = 1, 2, \ldots, n.$$

If, moreover, there is such an $\eta > 0$ that we have even

$$\lim_{x \to +\infty} |\varphi_i(x, a, c_1, c_2, \ldots, c_n) - \varphi_i(x, a, b_1, b_2, \ldots, b_n)| = 0, \quad i = 1, 2, \ldots, n, \tag{6'}$$

when

$$|c_i - b_i| < \eta, \quad i = 1, 2, \ldots, n,$$

then we say that the solution (3) is *asymptotically stable in the Liapunov sense* (*Liapunov asymptotically stable*, in brief).

REMARK 2. If, in Definition 1, δ can be chosen dependent only on ε (thus independent of a), we speak about a *uniform Liapunov stability.* It can be shown that if the system (1) is *autonomous* (i.e. if the functions f_i, $i = 1, 2, \ldots, n$, are independent of x, see Remark 17.18.1), then the Liapunov stability implies uniform Liapunov stability.

In applications, we are most often interested in the stability of the so-called *zero-solution*

$$y_1(x) \equiv 0, y_2(x) \equiv 0, \ldots, y_n(x) \equiv 0 \tag{7}$$

of the system (1), as far as this system does have such a solution at all, i.e. as far as

$$f_i(x, 0, 0, \ldots, 0) = 0$$

holds for all $i = 1, 2, \ldots, n$. (It can be shown that if it is not the case, the problem of stability of a non-zero solution can be easily converted into that of a zero solution by a simple transformation of system (1) into another system which already has the zero solution.) A typical example of a problem of stability of a zero solution is that for a system of homogeneous linear equations of the form

$$\begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 + \ldots + a_{1n}y_n, \\ y_2' &= a_{21}y_1 + a_{22}y_2 + \ldots + a_{2n}y_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n' &= a_{n1}y_1 + a_{n2}y_2 + \ldots + a_{nn}y_n, \end{aligned} \tag{8}$$

a_{ik} being constants (complex, in general).

Theorem 2. *If all roots of the equation*

$$\begin{vmatrix} a_{11} - \lambda, & a_{12}, & \ldots, & a_{1n} \\ a_{21}, & a_{22} - \lambda, & \ldots, & a_{2n} \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots \\ a_{n1}, & a_{n2}, & \ldots, & a_{nn} - \lambda \end{vmatrix} = 0 \tag{9}$$

have negative real parts, then the zero solution (7) *of the system* (8) *is Liapunov asymptotically* (*and, consequently, Liapunov*) *stable.*

A simple criterion for equation (9) to have all roots with negative real parts, is given in Theorem 3 below: Let us consider a polynomial

$$f(x) = a_0 + a_1x + \ldots + a_{n-1}x^{n-1} + a_nx^n, \quad n \geqq 1, \quad a_0 > 0, \quad a_n \neq 0. \tag{10}$$

We call this polynomial a *Hurwitz polynomial* if all its roots have negative real parts. Under its *Hurwitz matrix* we understand the matrix

$$\begin{bmatrix} a_1, & a_0, & 0, & 0, & \ldots, & 0 \\ a_3, & a_2, & a_1, & a_0, & \ldots, & 0 \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots \\ a_{2n-1}, & a_{2n-2}, & a_{2n-3}, & a_{2n-4}, & \ldots, & a_n \end{bmatrix}, \tag{11}$$

where $a_s = 0$ if $s < 0$ or $s > n$. Denote the determinant of the matrix (11) by D.

Theorem 3 (*Hurwitz Test*). *The polynomial* (10) *is a Hurwitz polynomial* (thus having all roots with negative real parts) *if and only if all principal minors of the matrix* (11) *are positive, i.e. if and only if*

$$a_1 > 0, \begin{vmatrix} a_1, & a_0 \\ a_3, & a_2 \end{vmatrix} > 0, \ldots, D > 0. \tag{12}$$

Example 1. Consider the system

$$\begin{aligned} y_1' &= -7y_1 + y_2, \\ y_2' &= -2y_1 - 5y_2, \end{aligned} \tag{13}$$

(see Example 17.18.2). Equation (9) is of the form

$$\begin{vmatrix} -7-\lambda, & 1 \\ -2, & -5-\lambda \end{vmatrix} = 0, \tag{14}$$

i.e.

$$\lambda^2 + 12\lambda + 37 = 0. \tag{15}$$

For the polynomial $37 + 12\lambda + \lambda^2$ we have (see (10)) $n = 2$, $a_0 = 37$, $a_1 = 12$, $a_2 = 1$, $a_3 = 0, \ldots$, so that its Hurwitz matrix is

$$\begin{bmatrix} 12, & 37 \\ 0, & 1 \end{bmatrix}. \tag{16}$$

Evidently

$$a_1 = 12 > 0, \quad D = \begin{vmatrix} 12, & 37 \\ 0, & 1 \end{vmatrix} = 12 > 0. \tag{17}$$

Thus the polynomial (15) is a Hurwitz polynomial. According to Theorem 2, the zero solution

$$y_1(x) \equiv 0, \quad y_2(x) \equiv 0$$

of the system (13) is Liapunov asymptotically stable (and, consequently, also Liapunov stable).

Theorem 2 can be extended to systems with "small perturbations":

Theorem 4. *Let us consider the system of equations*

$$\begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 + \ldots + a_{1n}y_n + \psi_1(x, y_1, y_2, \ldots, y_n), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n' &= a_{n1}y_1 + a_{n2}y_2 + \ldots + a_{nn}y_n + \psi_n(x, y_1, y_2, \ldots, y_n), \end{aligned} \tag{18}$$

where the a_{ik} are constants (in general, complex). Let the functions ψ_i and the constants a_{ik} be such that:

I. *For $x \geqq a_0$, $|y_i| < K$ ($i = 1, 2, \ldots, n$, $K = const.$), the functions ψ_i are continuous and*

$$|\psi(x, y_1, y_2, \ldots, y_n)| \leqq L\big(|y_1| + |y_2| + \ldots + |y_n|\big)$$

($i = 1, 2, \ldots, n$; L *being a constant*). (*In particular* $\psi_i(x, 0, 0, \ldots, 0) = 0$, *and* $y_1 \equiv 0, y_2 \equiv 0, \ldots, y_n \equiv 0$ *is a solution of the system* (18).)

II. *Corresponding to an arbitrary* $\varepsilon > 0$ *there exist numbers* δ_ε *and* T_ε *such that for* $|y_1| < \delta_\varepsilon, |y_2| < \delta_\varepsilon, \ldots, |y_n| < \delta_\varepsilon, x \geqq T_\varepsilon$ *the inequality* $|\psi(x, y_1, y_2, \ldots, y_n)| \leqq \leqq \varepsilon(|y_1| + |y_2| + \ldots + |y_n|)$ *holds.*

III. *All roots of the equation*

$$\begin{vmatrix} a_{11} - \lambda, & a_{12}, & \ldots, & a_{1n} \\ a_{21}, & a_{22} - \lambda, & \ldots, & a_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}, & a_{n2}, & \ldots, & a_{nn} - \lambda \end{vmatrix} = 0 \tag{19}$$

have negative real parts.

Then the solution $y_1 \equiv 0, y_2 \equiv 0, \ldots, y_n \equiv 0$ *is Liapunov asymptotically* (*and thus Liapunov*) *stable.*

17.20. First Integrals of a System of Differential Equations

Definition 1. Let us consider the system

$$\begin{aligned} y_1' &= f_1(x, y_1, y_2, \ldots, y_n), \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n' &= f_n(x, y_1, y_2, \ldots, y_n). \end{aligned} \tag{1}$$

Let $\psi(x, y_1, y_2, \ldots, y_n)$ be a differentiable function of its $n+1$ variables in a region Ω in which the system (1) is considered. We say that the equation

$$\psi(x, y_1, y_2, \ldots, y_n) = C$$

gives (is, represents) a *first integral of the system* (1) (in the region considered) if the function ψ is not identically constant in $x, y_1, y_2, \ldots, y_n$ and assumes a constant value if we substitute an arbitrary solution of the system (1) for $y_1, y_2, \ldots, y_n$.

REMARK 1. The first integrals of the system (1) are often defined in the following way: Let

$$\begin{aligned} y_1 &= \varphi_1(x, C_1, C_2, \ldots, C_n), \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= \varphi_n(x, C_1, C_2, \ldots, C_n) \end{aligned} \tag{2}$$

be the general integral of the system (1). Let us determine C_i from (2):

$$\begin{aligned} &\psi_1(x, y_1, y_2, \ldots, y_n) = C_1, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\psi_n(x, y_1, y_2, \ldots, y_n) = C_n. \end{aligned} \tag{3}$$

Then each of the relations (3) is called a *first integral of the system* (1).

These two definitions are not equivalent, in general. In simple cases, however, it can be verified that both of them give the same result.

Example 1. If we calculate C_1 and C_2 from (17.18.20) in Example 17.18.1, we get

$$\begin{aligned} (2y_2 - y_1)\,\mathrm{e}^{-2x} &= C_1, \\ (y_1 - y_2)\,\mathrm{e}^{-3x} &= C_2. \end{aligned} \tag{4}$$

Each of the relations (4) gives a first integral of the system (17.18.17) of the above-mentioned example. The left-hand sides of equations (4) are constants for each solution of (17.18.17). For example, for the solution

$$y_1 = -5\,\mathrm{e}^{2x} + 8\,\mathrm{e}^{3x}, \quad y_2 = -5\,\mathrm{e}^{2x} + 4\,\mathrm{e}^{3x}$$

of the example under consideration, we have

$$(2y_2 - y_1)\,\mathrm{e}^{-2x} = \mathrm{e}^{-2x}(-10\,\mathrm{e}^{2x} + 8\,\mathrm{e}^{3x} + 5\,\mathrm{e}^{2x} - 8\,\mathrm{e}^{3x}) = -5.$$

Theorem 1. *The necessary and sufficient condition for the equation*

$$\psi(x, y_1, y_2, \ldots, y_n) = C \tag{5}$$

to represent a first integral of the system (1) *(in the domain Ω of the variables x, $y_1, y_2, \ldots, y_n$) is that*

$$\frac{\partial \psi}{\partial x} + f_1(x, y_1, \ldots, y_n)\frac{\partial \psi}{\partial y_1} + f_2(x, y_1, \ldots, y_n)\frac{\partial \psi}{\partial y_2} + \ldots + f_n(x, y_1, \ldots, y_n)\frac{\partial \psi}{\partial y_n} = 0 \tag{6}$$

be satisfied identically in Ω. (It is assumed, naturally, that ψ and f_k are differentiable functions of their respective variables.)

Example 2. Following Theorem 1, we can easily verify that each of equations (4) gives a first integral of the system (17.18.17) in Example 17.18.1. For example, writing down the condition (6) for the second of equations (4), we get

$$-3(y_1 - y_2)\,\mathrm{e}^{-3x} + (4y_1 - 2y_2)\,\mathrm{e}^{-3x} - (y_1 + y_2)\,\mathrm{e}^{-3x} \equiv 0.$$

REMARK 2. If one first integral of the system (1) is known, it is generally possible to determine from its equation one unknown function $y_k(x)$ as a function of x and of the other unknown functions $y_1, \ldots, y_{k-1}, y_{k+1}, \ldots, y_n$. Substituting this into the given system, the number of equations of the system is then reduced by one. Similarly, if j (independent) first integrals are known, the number of equations of the system can be reduced by j. If n (independent) first integrals are known, we can obtain the general integral without any integration; it is sufficient to determine $y_1, \ldots, y_n$ from (3), as functions of $x, C_1, \ldots, C_n$.

It can be shown that corresponding to any system (1) satisfying the conditions of Theorem 17.2.1, there exists a system of n independent first integrals. Any system of $n+1$ first integrals is dependent.

REMARK 3. From Definition 1 it follows that: If (3) are the first integrals of the system (1), then

$$F(\psi_1, \psi_2, \ldots, \psi_n) = C,$$

where F is an arbitrary (not identically constant) differentiable function, also constitutes a first integral of the system (1). Since all ψ_k assume constant values for any solution, then F also assumes a constant value for any solution. The same is valid if F is a function of only k of the integrals (3).

REMARK 4 (*construction if first integrals*). The first integrals of the system (1) can often be obtained easily by writing the system (1) in the differential form

$$\frac{\mathrm{d}y_1}{f_1(x, y_1, \ldots, y_n)} = \frac{\mathrm{d}y_2}{f_2(x, y_1, \ldots, y_n)} = \ldots = \frac{\mathrm{d}y_n}{f_n(x, y_1, \ldots, y_n)} = \mathrm{d}x \qquad (7)$$

and by combining equations (7) in such a way that we obtain complete differentials. For the determination of further first integrals it is possible to use the first integrals already found. In particular, the problem becomes simpler if the functions f_i do not depend explicitly on x. Then, in (7), the last term $\mathrm{d}x$ can be omitted and the system can be written in the "symmetric form"

$$\frac{\mathrm{d}y_1}{f_1(y_1, \ldots, y_n)} = \frac{\mathrm{d}y_2}{f_2(y_1, \ldots, y_n)} = \ldots = \frac{\mathrm{d}y_n}{f_n(y_1, \ldots, y_n)}. \qquad (8)$$

The system (8) has $n-1$ independent first integrals, while, obviously, x does not occur in them. (This fact is important in mechanics, where systems of this type are often met; x then denotes time. The given system has thus at least $n-1$ first integrals independent of time.)

In the case where a system is written in the form (8) we need not assume that all the functions $f_1, f_2, \ldots, f_n$ are different from zero in the region considered, that is

to say, (8) is only a brief form of transcription and if, for example, $f_n \neq 0$, it really represents the system

$$\frac{dy_1}{dy_n} = \frac{f_1}{f_n}, \frac{dy_2}{dy_n} = \frac{f_2}{f_n}, \ldots, \frac{dy_{n-1}}{dy_n} = \frac{f_{n-1}}{f_n}.$$

The points at which all the functions $f_1, f_2, \ldots, f_n$ are simultaneously zero (the so-called *singular points* of the given system) are excluded from our considerations.

Example 3. Let us consider the system

$$y_1' = y_2, \quad y_2' = -y_1.$$

The system (8) is:

$$\frac{dy_1}{y_2} = -\frac{dy_2}{y_1},$$

whence $y_1\,dy_1 = -y_2\,dy_2$ and the first integral is

$$y_1^2 + y_2^2 = C.$$

(A singular point is the point $y_1 = 0$, $y_2 = 0$.)

Example 4. Let us consider the system

$$\frac{dy_1}{dx} = 1 - \frac{1}{y_2}, \quad \frac{dy_2}{dx} = \frac{1}{y_1 - x} \quad (y_1 \neq x,\ y_2 \neq 0). \tag{9}$$

The system can be written in the differential form (7) (after multiplying by dx):

$$dy_1 - dx = -\frac{dx}{y_2}, \quad \frac{dx}{y_1 - x} = dy_2. \tag{10}$$

Multiplying together both sides of these equations, we get an integrable combination

$$\frac{dy_1 - dx}{y_1 - x} = -\frac{dy_2}{y_2}$$

whence we obtain the first integral

$$(y_1 - x)y_2 = C_1 \quad (C_1 \neq 0). \tag{11}$$

We can use (11) for calculating the second first integral. From (11), it follows that

$$y_2 = \frac{C_1}{y_1 - x}, \tag{12}$$

which, when substituted into the first equation (10), gives

$$\mathrm{d}y_1 - \mathrm{d}x = -\frac{(y_1 - x)\,\mathrm{d}x}{C_1}$$

and

$$(y_1 - x)\,\mathrm{e}^{x/C_1} = C_2 \quad (C_2 \neq 0). \tag{13}$$

Now (13) is not a first integral of the system (9) (it contains two constants). We therefore substitute for C_1 from (11) into (13). In this way we get the second first integral of (9),

$$(y_1 - x)\,\mathrm{e}^{x/[y_2(y_1-x)]} = C_2. \tag{14}$$

By virtue of Remark 2, the general integral of the system (9) can be determined from (12) and (14). Here, however, it is more advantageous to use (13) and (12). From (13) it follows that

$$y_1 = x + C_2\,\mathrm{e}^{-x/C_1} \tag{15}$$

and on substituting for $y_1 - x$ into (12), we have

$$y_2 = \frac{C_1}{C_2}\,\mathrm{e}^{x/C_1}. \tag{16}$$

Equations (15) and (16) constitute the general integral of the system (9). (The above procedure can obviously be applied, because (12) and (14) follow from (12) and (13), and conversely.)

REMARK 5. First integrals of ordinary differential equations are frequently used in theoretical mechanics and in the theory of partial differential equations of the first order.

17.21. Table of Solved Differential Equations

See also [250]; in this paragraph m, n denote integers. Many differential equations which the reader will meet in applications can easily be transformed into equations given in the following table. For example, if we encounter the equation

$$y' + y^2 = 16,$$

we can make the substitutions

$$y = 4u, \quad 4x = t, \tag{1}$$

which give

$$y' = \frac{\mathrm{d}y}{\mathrm{d}x} = 4\frac{\mathrm{d}u}{\mathrm{d}x} = 4\frac{\mathrm{d}u}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}x} = 16\frac{\mathrm{d}u}{\mathrm{d}t} = 16\dot{u}$$

so that the given equation becomes $\dot{u} + u^2 = 1$, which is equation 7 of the table. (From the first of the solutions of 7, $u = \tanh(t + C)$, we then get, by substitution (1), $y/4 = \tanh(4x + C)$, etc.)

(a) Equations of the First Order

1. $y' + ay = c\,e^{bx}$

$$y = \begin{cases} \dfrac{c}{a+b}\,e^{bx} + C\,e^{-ax} & \text{if } a + b \neq 0, \\ cx\,e^{bx} + C\,e^{-ax} & \text{if } a + b = 0. \end{cases}$$

2. $\dfrac{dy}{dt} + \dfrac{R}{L}y = \dfrac{E}{L}\sin \omega t.$

(*Equation of an electric circuit*; $y(t)$ denotes the current in the circuit, R the resistance, L the self-inductance, $E \sin \omega t$ the alternating voltage.) If $y = y_0$ for $t = 0$, the solution is

$$y = \left(y_0 + \frac{\omega LE}{R^2 + \omega^2 L^2}\right) e^{-(R/L)t} + \frac{E}{\sqrt{(R^2 + \omega^2 L^2)}} \sin(\omega t - \gamma), \quad \text{where}$$

$\tan \gamma = \dfrac{\omega L}{R}$ and $0 < \gamma < \dfrac{\pi}{2}$. For large t, $y \approx \dfrac{E}{\sqrt{(R^2 + \omega^2 L^2)}} \sin(\omega t - \gamma)$.

3. $y' + 2xy = x\,e^{-x^2}$;

$y = e^{-x^2}\left(\frac{1}{2}x^2 + C\right)$.

4. $y' + y \cos x = e^{-\sin x}$;

$y = (x + C)\,e^{-\sin x}$.

5. $y' + f'(x)y = f(x)f'(x)$;

$y = f(x) - 1 + C\,e^{-f(x)}$.

6. $y' + f(x)y = g(x)$;

$$y = e^{-F(x)} \int g(x)\,e^{F(x)}\,dx \quad \text{where} \quad F(x) = \int f(x)\,dx;$$

the integral curve passing through the point (x_0, y_0) is

$$y = e^{-F_0(x)} \left(y_0 + \int_{x_0}^{x} g(t)\,e^{F_0(t)}\,dt\right), \quad \text{where} \quad F_0(x) = \int_{x_0}^{x} f(t)\,dt.$$

7. $y' + y^2 = 1$;

$y = \tanh(x + C)$; $y = \coth(x + C)$; $y = \pm 1$.

8. $y' + y^2 - 2x^2 y + x^4 - 2x - 1 = 0$;

substituting $u = y - x^2$, we obtain $u' + u^2 = 1$, see 7.

9. $y' = (y + x)^2$:

$y = -x + \tan(x + C)$.

10. $y' + ay^2 = b$;

the integral curve passing through the point (x_0, y_0) is

$$y = \begin{cases} y = y_0 + b(x - x_0) & \text{if } a = 0, \\ \dfrac{y_0}{1 + ay_0(x - x_0)} & \text{if } b = 0, \\ \dfrac{y_0\sqrt{(ab)} + b\tanh\left[\sqrt{(ab)}\,(x - x_0)\right]}{\sqrt{(ab)} + ay_0\tanh\left[\sqrt{(ab)}\,(x - x_0)\right]} & \text{if } ab > 0, \\ \dfrac{y_0\sqrt{(-ab)} + b\tan\left[\sqrt{(-ab)}\,(x - x_0)\right]}{\sqrt{(-ab)} + ay_0\tan\left[\sqrt{(-ab)}\,(x - x_0)\right]} & \text{if } ab < 0. \end{cases}$$

11. $y' - ax^r(y^2 + 1) = 0$;

$$y = \begin{cases} \tan\left(\dfrac{a}{r+1}x^{r+1} + C\right) & \text{if } r \neq -1, \\ \tan(a \ln Cx) & \text{if } r = -1. \end{cases}$$

12. $y' + f(x)y^2 + g(x)y = 0$;

$$\frac{1}{y} = E(x)\int \frac{f(x)}{E(x)}\,dx, \quad \text{where} \quad E(x) = e^{\int g(x)\,dx}; \quad y = 0.$$

13. $y' = a\cos y + b$, $b > |a| > 0$;

$$\arctan\left[\sqrt{\left(\frac{b+a}{b-a}\right)}\cot\frac{y}{2}\right] + \frac{x}{2}\sqrt{(b^2 - a^2)} = C; \quad \cos y = -\frac{b}{a}.$$

14. $y' = \cos(ay + bx)$, $a \neq 0$;

substituting $u = ay + bx$, we get $u' = a\cos u + b$, see 13.

15. $xy' + y = x \sin x;$

$$y = \frac{\sin x}{x} - \cos x + \frac{C}{x}.$$

16. $xy' - y = x^2 \sin x;$

$$y = x(C - \cos x).$$

17. $xy' + ay + bx^r = 0, \quad x > 0;$

$$y = \begin{cases} Cx^{-a} - \dfrac{b}{r+a}x^r & \text{if } a \neq -r, \\ Cx^{-a} - bx^{-a} \ln x & \text{if } a = -r. \end{cases}$$

18. $xy' - y^2 + 1 = 0;$

$$y = \frac{1 - Cx^2}{1 + Cx^2}; \quad y = \pm 1.$$

19. $xy' + xy^2 - y = 0;$

$$y = \frac{2x}{x^2 + C}; \quad y = 0.$$

20. $xy' - y^2 \ln x + y = 0;$

$$\frac{1}{y} = 1 + \ln x + Cx; \quad y = 0.$$

21. $xy' - y(2y \ln x - 1) = 0;$

$$\frac{1}{y} - 2(1 + \ln x) = Cx; \quad y = 0.$$

22. $xy' = x \sin \frac{y}{x} + y;$

$$y = 2x \arctan Cx.$$

23. $xy' + x \cos \frac{y}{x} - y + x = 0;$

$$\cos \frac{y}{x} - (C + \ln x) \sin \frac{y}{x} = 1.$$

24. $xy' + x\tan\frac{y}{x} - y = 0;$

$$x\sin\frac{y}{x} = C; \quad y = 0.$$

25. $x^2y' + y - x = 0;$

$$y = e^{1/x}\left(C + \int \frac{1}{x}\, e^{-1/x}\, dx\right).$$

26. $x^2y' - y^2 - xy = 0;$

$$y = -\frac{x}{\ln Cx}; \quad y = 0.$$

27. $(x^2 + 1)y' + xy - 1 = 0;$

$$y = \frac{C + \ln\left[x + \sqrt{(x^2+1)}\right]}{\sqrt{(x^2+1)}}.$$

28. $(x^2 - 1)y' + 2xy - \cos x = 0;$

$$y = \frac{\sin x + C}{x^2 - 1}.$$

29. $(x^2 - 1)y' - y(y - x) = 0;$

$$y = \frac{1}{x + C\sqrt{|x^2 - 1|}}; \quad y = 0.$$

30. $(x^2 - 1)y' + axy^2 + xy = 0;$

$$y = \frac{1}{C\sqrt{|x^2 - 1|} - a}; \quad y = 0.$$

31. $(x^2 - 1)y' = 2xy\ln y;$

$$y = e^{C(x^2-1)}.$$

32. $yy' + xy^2 - 4x = 0;$

$$y^2 = 4 + Ce^{-x^2}.$$

33. $(y - x^2)y' = x;$

$$x^2 = y - \tfrac{1}{2} + Ce^{-2y}.$$

34. $ayy' + by^2 + f(x) = 0;$

substituting $u = y^2$, we obtain the linear equation

$au' + 2bu + 2f(x) = 0.$

35. $y'^2 + y^2 = a^2;$

$$y = a\frac{1 - C^2}{1 + C^2}\sin x + a\frac{2C}{1 + C^2}\cos x; \quad y = \pm a.$$

36. $y'^2 = y^3 - y^2;$

$$y = \left(\cos\frac{x + C}{2}\right)^{-2}; \quad y = 0, \quad y = 1.$$

37. $y'^2 - 4y^3 + ay + b = 0;$

$$x = C \pm \int \frac{dy}{\sqrt{(4y^3 - ay - b)}};$$ the integral is an elliptic integral.

38. $y'^2 - 2y' - y^2 = 0;$

$$1 \mp \sqrt{(y^2 + 1)} + y\ln\left[\sqrt{(y^2 + 1)} \pm y\right] = y(x + C); \quad y = 0.$$

39. $y'^2 + ay' + bx = 0, \quad b \neq 0;$

the solution (in parametric form) is

$bx = -t^2 - at, \quad by = C - \frac{2}{3}t^3 - \frac{1}{2}at^2.$

40. $y'^2 + ay' + by = 0, \quad b \neq 0;$

the solution (in parametric form) is

$bx = -2t - a\ln t + C, \quad by = -t^2 - at.$

A further solution is $y = 0$.

41. $y'^2 + (x - 2)y' - y + 1 = 0;$

$$y = C(x - 2) + C^2 + 1; \quad y = x - \frac{x^2}{4}.$$

42. $y'^2 + (x + a)y' - y = 0;$

$y = C(x + a) + C^2, \quad 4y = -(x + a)^2.$

43. $y'^2 - (x+1)y' + y = 0;$

$y = Cx + C(1-C); \quad y = \frac{1}{4}(x+1)^2.$

44. $y'^2 - 2xy' + y = 0;$

the solution (in parametric form) is

$$x = \tfrac{2}{3}t + \frac{C}{t^2}, \quad y = 2xt - t^2.$$

Further solutions are $y = 0$, $y = \frac{3}{4}x^2$.

45. $y'^2 + 2xy' - y = 0;$

substituting $u = -y$, the equation is transformed into 44.

46. $y'^2 + axy' = bx^2 + c, \quad a^2 + 4b > 0;$

$$y = C - \tfrac{1}{4}ax^2 + \tfrac{1}{4}x\sqrt{\left[(a^2+4b)x^2 + 4c\right]} + \\ + \frac{c}{\sqrt{(a^2+4b)}} \ln\left[x + \sqrt{\left(x^2 + \frac{4c}{a^2+4b}\right)}\right].$$

47. $y'^2 + (ax+b)y' - ay + c = 0, \quad a \neq 0;$

$$y = (ax+b)C + aC^2 + \frac{c}{a}; \quad 4ay = 4c - (ax+b)^2.$$

48. $y'^2 - 2yy' - 2x = 0;$

the solution (in parametric form) is

$$x = \frac{t^2}{2} - yt, \quad y = \frac{t}{2} + \frac{1}{\sqrt{(t^2+1)}}\left(C - \frac{1}{2}\operatorname{arsinh} t\right).$$

49. $y'^2 - (4y+1)y' + (4y+1)y = 0;$

$y = C^2 e^{2x} + C e^x; \quad y = -\tfrac{1}{4}.$

50. $y'^2 - xyy' + y^2 \ln ay = 0;$

$$y = \frac{e^{Cx - C^2}}{a}; \quad y = \frac{e^{x^2/4}}{a}.$$

51. $y'^2 + 2yy' \cot x - y^2 = 0;$

$y(1 \pm \cos x) = C.$

52. $ay'^2 + by' - y = 0;$

the solution (in parametric form) is

$x = 2at + b \ln |t| + C, \quad y = at^2 + bt.$

A further solution is $y = 0.$

53. $ay'^2 - yy' - x = 0;$

the solution (in parametric form) is

$$x = \frac{t}{\sqrt{(t^2+1)}} \{C + a \ln [t + \sqrt{(t^2+1)}]\} = \frac{t}{\sqrt{(t^2+1)}} (C + a \operatorname{arsinh} t),$$

$$y = at - \frac{x}{t}.$$

54. $xy'^2 = y;$

$(y - x)^2 - 2C(y + x) + C^2 = 0; \quad y = 0.$

55. $xy'^2 - 2y + x = 0;$

the solution (in parametric form) is

$$x = \frac{C}{(t-1)^2} e^{2/(t-1)}, \quad y = \frac{x}{2}(t^2+1).$$

A further solution is $y = x.$

56. $xy'^2 - 2y' - y = 0;$

the solution (in parametric form) is

$$x = \frac{2t - 2 \ln |t| + C}{(t-1)^2}, \quad y = xt^2 - 2t.$$

Further solutions are $y = 0$, $y = x - 2.$

57. $xy'^2 + 4y' - 2y = 0;$

the solution (in parametric form) is

$$x = \frac{2y - 4t}{t^2}, \quad y = \left(\frac{t}{t-2}\right)^2 \left(C + 4\ln|t| + \frac{8}{t}\right).$$

Further solutions are $y = 0$, $y = 2x + 4$.

58. $xy'^2 + xy' - y = 0$;

the solution (in parametric form) is

$$x = Ct^2\,\mathrm{e}^t, \quad y = C(t+1)\,\mathrm{e}^t.$$

A further solution is $y = 0$.

59. $xy'^2 + yy' + a = 0$;

the solution (in parametric form) is

$$x = \frac{C}{\sqrt{t}} - \frac{a}{3t^2}, \quad y = -C\sqrt{(t)} - \frac{2a}{3t}.$$

60. $xy'^2 + yy' - y^4 = 0$;

$$y(x - C^2) = C.$$

61. $xy'^2 - yy' + a = 0$;

$$y = Cx + \frac{a}{C}; \quad y = \pm 2\sqrt{ax}\,.$$

62. $xy'^2 - yy' + ay = 0$;

the solution (in parametric form) is

$$x = C(t-a)\,\mathrm{e}^{-t/a}, \quad y = Ct^2\,\mathrm{e}^{-t/a}.$$

63. $xy'^2 - 2yy' + a = 0, \quad a \neq 0$;

$$16ax^3 - 12x^2y^2 - 12Caxy + 8Cy^3 + C^2a^2 = 0$$

$$\left(\text{in parametric form } x = Ct + \frac{a}{3t^2}, \quad y = \frac{xt}{2} + \frac{a}{2t}\right);$$

$$y = \pm\frac{2}{\sqrt{3}}\sqrt{(ax)}\,.$$

64. $xy'^2 - 2yy' - x = 0$;

$$y = \frac{C}{2}x^2 - \frac{1}{2C}.$$

65. $x^2y'^2 - y^4 + y^2 = 0;$

$$y = \frac{1}{\sin(\ln Cx)}, \quad Cx > 0, \quad \sin(\ln Cx) \neq 0; \quad y = 0.$$

66. $(x^2 - 1)y'^2 = 1;$

$$y = \pm \operatorname{arcosh} x + C.$$

67. $(x^2 - 1)y'^2 = y^2 - 1;$

$$x^2 + y^2 - 2Cxy + C^2 = 1; \quad y = \pm 1.$$

68. $\mathrm{e}^{-2x}\, y'^2 - (y' - 1)^2 + \mathrm{e}^{-2y} = 0;$

$$\mathrm{e}^y = C\,\mathrm{e}^x \pm \sqrt{(1 + C^2)}\,; \quad \mathrm{e}^{2y} + \mathrm{e}^{2x} = 1.$$

69. $yy'^2 = 1;$

$$4y^3 = 9(x + C)^2.$$

70. $yy'^2 = \mathrm{e}^{2x};$

$$4y^3 = 9(\mathrm{e}^x + C)^2.$$

71. $yy'^2 + 2xy' - y = 0;$

$$y^2 = 2Cx + C^2.$$

72. $yy'^2 + 2xy' - 9y = 0;$

the solution (in parametric form) is

$$x = \frac{t}{14} + Ct^{1/8}, \quad y^2 = \frac{t}{9}x + \frac{t^2}{36}.$$

A further solution is $y = 0$.

73. $yy'^2 - 2xy' + y = 0;$

$$y^2 = 2Cx - C^2; \quad y = \pm x.$$

74. $yy'^2 - 4xy' + y = 0;$

$$y^6 - 3x^2y^4 + 2Cx(3y^2 - 8x^2) + C^2 = 0$$

$$\left(\text{in parametric form, } x = y\frac{t^2 + 1}{4t},\ y^3 = \frac{C}{t(t^2 - 3)}\right).$$

75. $yy'^2 + x^3y' - x^2y = 0;$

$y^2 + Cx^2 = C^2.$

76. $ayy'^2 + (2x - b)y' - y = 0;$

$y^2 = C(2x - b + aC).$

Further solutions, for $a < 0$, are $\pm 2\sqrt{(-a)}\, y = 2x - b.$

77. $y^2y'^2 + y^2 - a^2 = 0;$

$(x - C)^2 + y^2 = a^2; \quad y = \pm a.$

78. $(y^2 - a^2)y'^2 + y^2 = 0;$

$$a \ln \left| \frac{a \pm \sqrt{(a^2 - y^2)}}{y} \right| \mp \sqrt{(a^2 - y^2)} = x + C; \quad y = 0.$$

79. $y'^3 + y' - y = 0;$

the solution (in parametric form) is

$x = C + \frac{3}{2}t^2 + \ln|t|, \quad y = t^3 + t.$

A further solution is $y = 0.$

80. $y'^3 + xy' - y = 0;$

$y = Cx + C^3.$

Further solutions, for $x < 0$, are $y = \pm 2(-x/3)^{3/2}.$

81. $y'^3 - (x + 5)y' + y = 0;$

$y = Cx + C(5 - C^2); \quad 27y^2 = 4(x + 5)^3.$

82. $y'^3 - axy' + x^3 = 0, \quad a \neq 0;$

the solution (in parametric form) is

$$x = \frac{at}{t^3 + 1}, \quad y = C + \frac{a^2}{6} \frac{4t^3 + 1}{(t^3 + 1)^2}.$$

83. $y'^3 - 2yy' + y^2 = 0;$

the solution (in parametric form) is

$$x = C \pm 3\sqrt{(1-t)} + 2\ln\left[1 \mp \sqrt{(1-t)}\right], \quad y = t\left[1 \pm \sqrt{(1-t)}\right].$$

A further solution is $y = 0$.

84. $y'^3 - axyy' + 2ay^2 = 0;$

$$y = \frac{a}{4}C(x-C)^2; \quad y = \frac{a}{27}x^3.$$

85. $y'^3 - yy'^2 + y^2 = 0;$

the solution (in parametric form) is

$$x = t \pm \sqrt[4]{(t^2-4t)} \mp \ln\left|\sqrt[4]{(t^2-4t)} + t - 2\right| + C,$$

$$y = \tfrac{1}{2}t^2 \pm \tfrac{1}{2}t\sqrt[4]{(t^2-4t)}.$$

A further solution is $y = 0$.

86. $xy'^3 - yy'^2 + a = 0;$

$$y = Cx + \frac{a}{C^2}; \quad 4y^3 = 27ax^2.$$

87. $ay'^r + by'^s = y \quad (r \neq 1, s \neq 1);$

the solution (in parametric form) is

$$x = C + \frac{ar}{r-1}t^{r-1} + \frac{bs}{s-1}t^{s-1}, \quad y = at^r + bt^s.$$

88. $\sqrt{(y'^2+1)} - xy'^2 + y = 0;$

the solution (in parametric form) is

$$x = \frac{\sqrt{(t^2+1)} - \ln\left[t + \sqrt{(t^2+1)}\right] + C}{(t-1)^2}, \quad y = xt^2 - \sqrt{(t^2+1)}\,.$$

89. $\ln y' + a(xy' - y) = 0;$

$$y = Cx + \frac{1}{a}\ln C \quad (C > 0), \qquad ay + 1 + \ln(-ax) = 0 \quad (ax < 0).$$

(b) Linear Equations of the Second Order.

90. $y'' + a^2 y = 0 \quad (a > 0);$

$$y = C_1 \cos ax + C_2 \sin ax \quad \text{(harmonic vibrations)}.$$

91. $y'' + a^2 y = b \sin cx;$

$$y = \begin{cases} \frac{b}{a^2 - c^2} \sin cx + C_1 \cos ax + C_2 \sin ax & \text{if } a^2 \neq c^2, \\ -\frac{b}{2c} x \cos cx + C_1 \cos cx + C_2 \sin cx & \text{if } a^2 = c^2 \end{cases}$$

(undamped forced vibrations).

92. $y'' + a^2 y = b \cos cx;$

$$y = \begin{cases} \frac{b}{a^2 - c^2} \cos cx + C_1 \cos ax + C_2 \sin ax & \text{if } a^2 \neq c^2, \\ \frac{b}{2c} x \sin cx + C_1 \cos cx + C_2 \sin cx & \text{if } a^2 = c^2 \end{cases}$$

(undamped forced vibrations).

93. $y'' + y = \sin ax \sin bx$

$$y = \frac{\cos(a-b)x}{2 - 2(a-b)^2} - \frac{\cos(a+b)x}{2 - 2(a+b)^2} + C_1 \cos x + C_2 \sin x$$

$(|a+b| \neq 1, |a-b| \neq 1).$

94. $y'' - a^2 y = 0;$

$$y = C_1 \, e^{ax} + C_2 \, e^{-ax} = k_1 \cosh ax + k_2 \sinh ax.$$

95. $y'' + \lambda y = 0;$

$$y = \begin{cases} C_1 \, e^{\sqrt{(-\lambda)}x} + C_2 \, e^{-\sqrt{(-\lambda)}x} = k_1 \cosh \sqrt{(-\lambda)}\, x + k_2 \sinh \sqrt{(-\lambda)}\, x & \text{if } \lambda < 0, \\ C_1 + C_2 x & \text{if } \lambda = 0, \\ y = C_1 \sin \sqrt{(\lambda)}\, x + C_2 \cos \sqrt{(\lambda)}\, x & \text{if } \lambda > 0. \end{cases}$$

96. $y'' - (a^2 x^2 + a) y = 0;$

$$y = e^{ax^2/2} \left(C_1 + C_2 \int e^{-ax^2} \, dx \right).$$

97. $y'' + 2ay' + b^2 y = 0;$

$$y = \begin{cases} e^{-ax}(C_1 \cos \alpha x + C_2 \sin \alpha x) & \text{if } \alpha^2 = b^2 - a^2 > 0 \\ \qquad \text{(damped vibrations, see also § 4.13),} & \\ e^{-ax}(C_1 x + C_2) & \text{if } b^2 - a^2 = 0, \\ e^{-ax}\left(C_1 \, e^{\beta x} + C_2 \, e^{-\beta x}\right) & \text{if } \beta^2 = a^2 - b^2 > 0. \end{cases}$$

98. $y'' + 2ay' + b^2y = A\sin\omega x;$

$$y = K\sin(\omega x + \varphi) + y_h,$$

$$\text{where } K = \frac{A}{\sqrt{[(b^2-\omega^2)^2 + 4a^2\omega^2]}},$$

$$\sin\varphi = \frac{-2a\omega}{\sqrt{[(b^2-\omega^2)^2 + 4a^2\omega^2]}}, \quad \cos\varphi = \frac{b^2-\omega^2}{\sqrt{[(b^2-\omega^2)^2 + 4a^2\omega^2]}}$$

and y_h is a solution of the corresponding homogeneous equation (see 97). (Forced damped vibrations; see also § 4.13).

99. $y'' + 2ay' + b^2y = A\cos\omega x;$

$$y = K\sin(\omega x + \varphi) + y_h,$$

$$\text{where } K = \frac{A}{\sqrt{[(b^2-\omega^2)^2 + 4a^2\omega^2]}},$$

$$\cos\varphi = \frac{2a\omega}{\sqrt{[(b^2-\omega^2)^2 + 4a^2\omega^2]}}, \quad \sin\varphi\frac{b^2-\omega^2}{\sqrt{[(b^2-\omega^2)^2 + 4a^2\omega^2]}}$$

and y_h is a solution of the corresponding homogeneous equation (see 97). (Forced damped vibrations.)

100. $y'' + xy' + (n+1)y = 0;$

$$y = \frac{d^n}{dx^n}\left[e^{-x^2/2}\left(C_1 + C_2\int e^{x^2/2}\,dx\right)\right].$$

101. $y'' - xy' + 2y = 0;$

$$y = (x^2-1)\left(C_1 + C_2\int \frac{1}{(x^2-1)^2}\,e^{x^2/2}\,dx\right).$$

102. $y'' - xy' + (x-1)y = 0;$

$$y = C_1\,e^x + C_2\,e^x\int e^{(x^2/2)-2x}\,dx.$$

103. $y'' + 4xy' + (4x^2+2)y = 0;$

$$y = (C_1 + C_2x)\,e^{-x^2}.$$

104. $y'' - 4xy' + (4x^2 - 1)y = e^{x^2}$;

$y = e^{x^2}(1 + C_1 \cos x + C_2 \sin x)$.

105. $y'' - x^2y' + xy = 0$;

$y = C_1 x + C_2\left(e^{x^3/3} - x \int x\, e^{x^3/3}\, dx\right)$.

106. $y'' + \left(\frac{2 - x^2}{4} + n\right) y = 0$;

$y = C\, e^{-x^2/4}\, H_n\left(\frac{x}{\sqrt{2}}\right)$,

where $H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}$ is the Hermite polynomial of degree n.

107. $x(y'' + y) = \cos x$;

$y = \sin x \int \frac{\cos^2 x}{x}\, dx - \cos x \int \frac{\sin 2x}{2x}\, dx + C_1 \sin x + C_2 \cos x$.

108. $xy'' + y' = 0$;

$y = C_1 + C_2 \ln |x|$.

109. $xy'' + y' + \lambda y = 0$

with conditions $y(1) = 0$, y bounded for $x \to 0$; the eigenfunctions are $y = C_0\, J_0(2\sqrt{(\lambda x)})$, where J_0 is the Bessel function of the first kind (§ 16.4) and the eigenvalues λ are given by the equation $J_0(2\sqrt{\lambda}) = 0$.

110. $xy'' + 2y' - xy = e^x$;

$y = \frac{1}{2} e^x + \frac{1}{x}(C_1 e^x + C_2 e^{-x})$.

111. $xy'' + 2y' + axy = 0$;

substituting $u = xy$, the equation is transformed into the equation $u'' + au = 0$.

112. $xy'' + (1 - a)y' + a^2x^{2a-1}y = 0$;

$y = C_1 \cos(x^a + C_2)$.

113. $2xy'' + y' + ay = 0, \quad a \neq 0;$

$$y = \begin{cases} C_1 \cos \sqrt{(2ax)} + C_2 \sin \sqrt{(2ax)} & \text{if } ax > 0, \\ C_1 \cosh \sqrt{(-2ax)} + C_2 \sinh \sqrt{(-2ax)} & \text{if } ax < 0. \end{cases}$$

114. $x^2y'' - 6y = 0;$

$$y = C_1x^3 + C_2x^{-2}.$$

115. $x^2y'' - 12y = 0;$

$$y = C_1x^4 + C_2x^{-3}.$$

116. $x^2y'' + ay = 0;$

$$y = \begin{cases} \sqrt{(x)}\,[C_1 \cos(b \ln|x|) + C_2 \sin(b \ln|x|)] & \text{if } b^2 = a - \frac{1}{4} > 0, \\ \sqrt{(x)}\,(C_1x^c + C_2x^{-c}) & \text{if } c^2 = \frac{1}{4} - a > 0, \\ \sqrt{(x)}\,(C_1 + C_2 \ln|x|) & \text{if } c^2 = \frac{1}{4} - a = 0. \end{cases}$$

117. $x^2y'' + xy' + (x^2 - r^2)y = 0$ (*Bessel's equation*; see § 17.15);

$$y = \begin{cases} \mathrm{J}_r(x) & \text{(Bessel's function of the first kind, § 16.4),} \\ \mathrm{Y}_r(x) & \text{(Bessel's function of the second kind, Remark 16.4.12, § 17.15).} \end{cases}$$

The general integral of Bessel's equation is

$$\mathrm{Z}_r(x) = C_1\,\mathrm{J}_r(x) + C_2\mathrm{Y}_r(x).$$

If r is not an integer, the general integral is

$$\mathrm{Z}_r(x) = C_1\,\mathrm{J}_r(x) + C_2\,\mathrm{J}_{-r}(x) \quad (\S\,17.15).$$

By proper transformations, the following equations can be transformed into Bessel's equation:

118. $x^2y'' + xy' - (x^2 + r^2)y = 0$ (*modified Bessel's equation*);

$y = \mathrm{Z}_r(\mathrm{i}x)$ (see 117).

119. $x^2y'' + xy' - (x^2 + r^2)y = 0;$

$y = \mathrm{Z}_{2r}(2\mathrm{i}\sqrt{x})$ (see 117).

120. $x^2y'' + xy' + 4(x^4 - r^2)y = 0;$

$y = \mathrm{Z}_r(x^2)$ (see 117).

121. $x^2y'' + (1 - 2r)xy' + r^2(x^{2r} + 1 - r^2)y = 0;$

$y = x^r \mathrm{Z}_r(x^r)$ (see 117).

122. $x^2y'' - [cx^2 + p(p-1)]y = 0;$

$y = \sqrt{(x)}\, \mathrm{Z}_{p-\frac{1}{2}}(\mathrm{i}\sqrt{(c)}\, x)$ (see 117).

123. $xy'' + (1 - 2r)y' + xy = 0;$

$y = x^r \mathrm{Z}_r(x)$ (see 117).

124. $xy'' - 2py' - cxy = 0;$

$y = x^{p+\frac{1}{2}} \mathrm{Z}_{p+\frac{1}{2}}(\mathrm{i}\sqrt{(c)}\, x)$ (see 117).

125. $y'' - cx^{2p-2}y = 0;$

$$y = \sqrt{(x)}\, \mathrm{Z}_{p/2}\left(\mathrm{i}\sqrt{(c)}\, \frac{x^p}{p}\right) \quad \text{(see 117).}$$

126. $y'' + xy = 0;$

$y = \sqrt{(x)}\, \mathrm{Z}_{1/3}\left(\frac{2}{3}x^{3/2}\right)$ (see 117).

127. $y'' - xy = 0;$

$y = \sqrt{(x)}\, \mathrm{Z}_{1/3}\left(\frac{2}{3}\mathrm{i}x^{3/2}\right)$ (see 117).

128. $(x^2 - 1)y'' + xy' + ay = 0;$

$$y = \begin{cases} C_1 \cos(\alpha \operatorname{arcosh} |x|) + C_2 \sin(\alpha \operatorname{arcosh} |x|) & \text{if } |x| > 1, \\ C_1\, \mathrm{e}^{\alpha \arccos x} + C_2\, \mathrm{e}^{-\alpha \arccos x} & \text{if } |x| < 1, \end{cases} \Bigg\} a = \alpha^2 > 0,$$

$$\begin{cases} C_1\, \mathrm{e}^{\beta \operatorname{arcosh} |x|} + C_2\, \mathrm{e}^{-\beta \operatorname{arcosh} |x|} & \text{if } |x| > 1, \\ C_1 \cos(\beta \arccos x) + C_2 \cos(\beta \arcsin x) & \text{if } |x| < 1, \end{cases} \Bigg\} a = -\beta^2 < 0.$$

For $a = -n^2$ (n an integer), Chebyshev polynomials

$y = \mathrm{T}_n(x) = 2^{-n+1} \cos(n \arccos x)$

constitute solutions of 128.

129. $(x^2-1)y''+2xy'-r(r+1)y=0$; *Legendre's equation*, cf. § 17.15;

by the transformation $x=\cos\xi$, $\eta(\xi)=y(x)$, the equation

130. $\eta''\sin\xi+\eta'\cos\xi+r(r+1)\eta\sin\xi=0$,

also often called *Legendre's equation*, is transformed into 129.

A. If $|x|<1$ the general integral of equation 129 is

$$y=C_1F\left(-\frac{r}{2},\frac{1+r}{2},\frac{1}{2},x^2\right)+C_2xF\left(\frac{1-r}{2},1+\frac{r}{2},\frac{3}{2},x^2\right),$$

where $F(\alpha,\beta,\gamma,x)$ is the hypergeometric series (§ 16.6).

B. If $|x|>1$ we define (for r see below)

$$y_r(x)=x^r+\sum_{k=1}^{\infty}(-1)^k\frac{\binom{r}{2k}\binom{r}{k}}{\binom{2r}{2k}}x^{r-2k},$$

$$y_r^*(x)=\lim_{s\to r}\frac{(s-r)y_s-\frac{1}{2}\lambda_s(r)y_{-s-1}}{s-r},$$

where $\lambda_s(r)=(-1)^{r+\frac{1}{2}}\dfrac{s(s-1)\dots(s-2r)}{2^{r+\frac{1}{2}}\left(r+\frac{1}{2}\right)!(2s-1)(2s-3)\dots(2s-2r+2)}$.

The general integral of equation 129 for $|x|>1$ is then

$$y=\begin{cases}C_1y_r+C_2y_{-r-1} & \text{if } 2r \text{ is not an odd number,}\\ C_1y_r^*+C_2y_{-r-1} & \text{if } 2r=2p+1,\ p\geqq 0 \text{ an integer,}\\ C_1y_r+C_2y_{-r-1}^* & \text{if } 2r=-(2n+1),\ n \text{ is a positive integer,}\\ C_1y_{-1/2}+C_2y_{-1/2}^* & \text{if } r=-\frac{1}{2}.\end{cases}$$

If $r=n$ (n integral), Legendre polynomials (§ 16.5)

$$\mathrm{P}_n(x)=\frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n$$

are solutions of equation 129.

The following equations may easily be transformed into equation 129. In equations 131–139, $L(x)$ means the solution of that equation.

131. $(x^2+1)y''+2xy'-r(r+1)y=0$;

$y=L(\mathrm{i}x)$.

132. $(x^2-1)y''+2(n+1)xy'-r(r+n+1)(r-n)y=0;$

$y=L^{(n)}(x).$

133. $x(x^2+1)y''+(2x^2+1)y'-r(r+1)xy=0;$

$y=L(\sqrt{(x^2+1)}).$

134. $x(x^2+1)y''+[2(n+1)x^2+2n+1]y'-(r-n)(r+n+1)xy=0;$

$y=L^{(n)}(\sqrt{(x^2+1)}).$

135. $x^2(x^2+1)y''+x(2x^2+1)y'-[r(r+1)x^2+n^2]y=0;$

$y=x^nL^{(n)}(\sqrt{(x^2+1)}).$

136. $x^2(x^2-1)y''+2x^3y'+r(r+1)y=0;$

$y=L\left(\frac{1}{x}\right).$

137. $x^2(x^2-1)y''+2x^3y'-r(r+1)(x^2-1)y=0;$

$y=L\left(\frac{x^2+1}{2x}\right).$

138. $x^2(x^2-1)y''+2x\big[(1-a)x^2+a\big]y'+$

$+\{[a(a-1)-r(r+1)]x^2-a(a+1)\}y=0;$

$y=x^aL(x).$

139. $(x^2-1)^2y''+2x(x^2-1)y'-[r(r+1)(x^2-1)+n^2]y=0$
($n\geqq 0$ is an integer);

$y=|x^2-1|^{n/2}L^{(n)}(x).$

140. $x(x-1)y''+\big[(\alpha+\beta+1)x-\gamma\big]y'+\alpha\beta y=0;$

the hypergeometric (Gauss's) equation, see § 17.15; brief notation:

$H(\alpha,\beta,\gamma,y,x)=0.$

For $|x|<1$, the solution is given by the hypergeometric series

$y=F(\alpha,\beta,\gamma,x)$ (see § 16.6).

If $0 < x < 1$, the general integral is

$$y = \begin{cases} y_1 = C_1 F(\alpha, \beta, \gamma, x) + C_2 x^{1-\gamma} F(\alpha - \gamma + 1, \beta - \gamma + 1, 2 - \gamma, x) \\ \qquad \text{if } \gamma \text{ is not an integer.} \\ C_1 y_1 + C_2 y_2, \quad \text{if } \gamma = -c,\ c \geqq -1 \text{ is an integer,} \end{cases}$$

where

$$y_2 = \begin{cases} \lim\limits_{\gamma \to -c} \left[F(\alpha, \beta, \gamma, x) - \dfrac{\lambda_\gamma}{\gamma + c} x^{1-\gamma} F(\alpha - \gamma + 1, \beta - \gamma + 1, 2 - \gamma, x) \right] \\ \qquad \text{if } c \geqq 0, \\ \lim\limits_{\gamma \to 1} \dfrac{1}{\gamma - 1} \left[F(\alpha, \beta, \gamma, x) - x^{1-\gamma} F(\alpha - \gamma + 1, \beta - \gamma + 1, 2 - \gamma, x) \right] \\ \qquad \text{if } c = -1, \end{cases}$$

where

$$\lambda_\gamma = \begin{cases} \dbinom{\alpha + c}{c + 1} \dfrac{\beta(\beta + 1) \dots (\beta + c)}{\gamma(\gamma + 1) \dots (\gamma + c - 1)} & \text{for } c \geqq 1, \\ \alpha\beta & \text{for } c = 0, \\ 1 & \text{for } c = -1. \end{cases}$$

The case $\gamma = c$, where $c \geqq 2$ is an integer, can be transformed by the substitution $y(x) = x^{1-\gamma}\eta(x)$ into the above-mentioned case.

In some special cases the solution can be written in a closed form (for notation see 140):

141. $H(\alpha, \alpha + \frac{1}{2}, 2\alpha + 1, y, x) = 0, \quad \alpha \neq 0;$

$$y = C_1 (1 + \sqrt{(1 - x)})^{-2\alpha} + C_2 x^{-2\alpha} (1 + \sqrt{(1 - x)})^{2\alpha} \quad \text{(see 140).}$$

142. $H(\alpha, \alpha - \frac{1}{2}, \frac{1}{2}, y, x) = 0;$

$$y = C_1 (1 + \sqrt{x})^{1-2\alpha} + C_2 (1 - \sqrt{x})^{1-2\alpha} \quad \text{(see 140).}$$

143. $H(\alpha, \alpha + \frac{1}{2}, \frac{3}{2}, y, x) = 0;$

$$y = C_1 \frac{1}{\sqrt{x}} (1 + \sqrt{x})^{1-2\alpha} + C_2 \frac{1}{\sqrt{x}} (1 - \sqrt{x})^{1-2\alpha} \quad \text{(see 140).}$$

144. $H(1, \beta, \gamma, y, x) = 0;$

$$y = x^{1-\gamma} (1 - x)^{\gamma - \beta - 1} \left[C_1 + C_2 \int x^{\gamma - 2} (1 - x)^{\beta - \gamma} \, dx \right] \quad \text{(see 140).}$$

145. $H(\alpha, \beta, \alpha, y, x) = 0$;

$$y = (1-x)^{-\beta}\left[C_1 + C_2 \int x^{-\alpha}(1-x)^{\beta-1}\,dx\right] \quad \text{(see 140)}.$$

146. $H(\alpha, \beta, \alpha+1, y, x) = 0$;

$$y = x^{-\alpha}\left[C_1 + C_2 \int x^{\alpha-1}(1-x)^{-\beta}\,dx\right] \quad \text{(see 140)}.$$

Related equations (notation: $y(\alpha, \beta, \gamma, x)$ is the solution of equation 140):

147. $x(x-1)y'' + (2x-1)y' - r(r+1)y = 0$;

$y = y(r+1, -r, 1, x)$ (see 140).

148. $x(x-1)y'' + \left[(a+b+1)x + (\alpha+\beta-1)\right]xy' + (abx - \alpha\beta)y = 0$;

$y = x^{\alpha}y(a+\alpha, b+\alpha, \alpha-\beta+1, x)$ (see 140).

149. $x(x^2-1)y'' + (ax^2+b)y' + cxy = 0$;

$$y = y\left(\frac{a-1}{2} + \sqrt{\left[\tfrac{1}{4}(a-1)^2 - c\right]}\,,\ \frac{a-1}{2} - \sqrt{\left[\tfrac{1}{4}(a-1)^2 - c\right]}\,,\ \frac{1-b}{2},\ x^2\right)$$

(see 140).

150. $16(x^3-1)^2y'' + 27xy = 0$;

$y = (x^3-1)^{1/4}y\left(\frac{1}{12}, -\frac{1}{4}, -\frac{1}{3}, x^3\right)$ (see 140).

151. $x(x-1)y'' + \left[(\alpha+\beta+2n+1)x - (\gamma+n)\right]y' + (\alpha+n)(\beta+n)y = 0$;

$y = y^{(n)}(\alpha, \beta, \gamma, x) = y(\alpha+n, \beta+n, \gamma+n, x)$ (see 140).

152. $(x^2 \pm a^2)^2y'' + b^2y = 0$ (the bending flexion of a bar of parabolic cross-section):

$$y = \begin{cases} \sqrt{(x^2+a^2)}\,(C_1\cos u + C_2 \sin u), \\ \qquad \text{where} \quad u = \dfrac{\sqrt{(a^2+b^2)}}{a} \arctan\dfrac{x}{a} \quad \text{for the sign } +, \\ \sqrt{(a^2-x^2)}\,(C_1\cos v + C_2 \sin v), \\ \qquad \text{where} \quad v = \dfrac{\sqrt{(b^2-a^2)}}{2a} \ln\dfrac{a+x}{a-x} \quad (|x| < a) \quad \text{for the sign } -. \end{cases}$$

153. $(e^x+1)y''=y$;

$$y=C_1(1+e^{-x})+C_2[-1+(1+e^{-x})\ln(1+e^x)].$$

154. $xy''\ln x-y'-xy\ln^3 x=0$;

$$y=C_1\left(\frac{x}{e}\right)^x+C_2\left(\frac{e}{x}\right)^x.$$

155. $y''\sin x-2y=0$;

$$y=C_1\cot x+C_2(1-x\cot x).$$

(c) Linear Equations of Higher Orders. Nonlinear Equations. Systems.

156. $y'''+\lambda y=0$;

$$y=\begin{cases}C_1+C_2x+C_3x^2 & \text{for } \lambda=0,\\ C_1e^{-\sqrt[3]{(\lambda)}x}+e^{\frac{1}{2}\sqrt[3]{(\lambda)}x}\left(C_2\cos\left[\frac{\sqrt{3}}{2}\sqrt[3]{(\lambda)}x\right]+C_3\sin\left[\frac{\sqrt{3}}{2}\sqrt[3]{(\lambda)}x\right]\right) & \text{for } \lambda\neq 0.\end{cases}$$

157. $y'''+3y'-4y=0$;

$$y=C_1e^x+\left(C_2\cos\frac{\sqrt{15}}{2}x+C_3\sin\frac{\sqrt{15}}{2}x\right)e^{-x/2}.$$

158. $y'''-a^2y'=e^{2ax}\sin^2 x$;

$$y=C_1+C_2e^{ax}+C_3e^{-ax}+$$
$$+\left(\frac{1}{12a^3}+\frac{(4-11a^2)\sin 2x+3a(4-a^2)\cos 2x}{4(a^2+1)(a^2+4)(9a^2+4)}\right)e^{2ax}.$$

159. $y'''-2y''-a^2y'+2a^2y=\sinh x$;

$$y=\begin{cases}C_1e^{2x}+C_2e^{ax}+C_3e^{-ax}+\dfrac{2\sinh x+\cosh x}{3(a^2-1)} & \text{for } a^2\neq 1,\ a^2\neq 4,\\ e^{2x}(C_1x+C_2)+C_3e^{-2x}+\dfrac{2\sinh x+\cosh x}{9} & \text{for } a^2=4,\\ C_1e^{2x}+C_2e^x+C_3e^{-x}-\dfrac{x+1}{4}e^x-\dfrac{3x+1}{36}e^{-x} & \text{for } a^2=1.\end{cases}$$

160. $y^{(4)}=0$;

$$y=C_0+C_1x+C_2x^2+C_3x^3.$$

161. $y^{(4)} + 4y = f(x)$;

$y = C_1 \cos x \cosh x + C_2 \cos x \sinh x + C_3 \sin x \cosh x + C_4 \sin x \sinh x + y_p$,

where y_p is a particular integral of equation 161.

162. $y^{(4)} - k^4 y = 0$ (transverse vibrations of a rod); solutions are given in (1)–(6) below for various sets of boundary conditions:

(1) $y(a) = y'(a) = y(b) = y'(b) = 0$;

the eigenvalues are to be calculated from the equation

$\cos k(b-a) \cosh k(b-a) = 1, \quad k \neq 0$;

the eigenfunctions are:

$y = [\cosh k(b-a) - \cos k(b-a)][\sinh k(x-a) - \sin k(x-a)] -$

$- [\sinh k(b-a) - \sin k(b-a)][\cosh k(x-a) - \cos k(x-a)]$.

(2) $y(a) = y'(a) = y(b) = y''(b) = 0$;

the eigenvalues are to be calculated from the equation

$\cos k(b-a) \sinh k(b-a) = \sin k(b-a) \cosh k(b-a)$;

for the eigenfunctions see (1) (the preceding case).

(3) $y(a) = y'(a) = y''(b) = y'''(b) = 0$;

the eigenvalues are to be calculated from the equation

$\cos k(b-a) \cosh k(b-a) = -1$;

the eigenfunctions are:

$y = [\cosh k(b-a) + \cos k(b-a)][\sinh k(x-a) - \sin k(x-a)] -$

$- [\sinh k(b-a) + \sin k(b-a)][\cosh k(x-a) - \cos k(x-a)]$.

(4) $y(a) = y''(a) = y(b) = y''(b) = 0$;

eigenvalues: $k = \dfrac{n\pi}{b-a}, \quad n = 1, 2, 3, \ldots$;

eigenfunctions: $y = \sin k(x-a)$.

(5) $y(a) = y''(a) = y''(b) = y'''(b) = 0$;

the eigenvalues are to be calculated from the equation

$\cos k(b-a)\sinh k(b-a) = \sin k(b-a)\cosh k(b-a)$;

the eigenfunctions are:

$y = \sin k(b-a)\sinh k(x-a) + \sinh k(b-a)\sin k(x-a)$.

(6) $y''(a) = y'''(a) = y''(b) = y'''(b) = 0$;

the eigenvalues are to be calculated from the equation

$\cos k(b-a)\cosh k(b-a) = 1$;

the eigenfunctions are:

$y = C_1 + C_2 x$ if $k = 0$; for $k \neq 0$, they are:

$y = [\sinh k(b-a) - \sin k(b-a)][\cosh k(x-a) + \cos k(x-a)] -$

$- [\cosh k(b-a) - \cos k(b-a)][\sinh k(x-a) + \sin k(x-a)]$.

163. $y^{(5)} + 2y''' + y' = ax + b\sin x + c\cos x$;

$$y = \frac{a}{2}x^2 + \frac{b}{8}x^2\cos x - \frac{c}{8}x^2\sin x +$$

$$+ C_1 + C_2\sin x + C_3\cos x + C_4 x\sin x + C_5 x\cos x.$$

164. $y^{(6)} + y = \sin\frac{3}{2}x\sin\frac{1}{2}x$;

$$y = \frac{x}{12}\sin x + \frac{1}{126}\cos 2x + C_1\cos(x + C_2) + C_3\,e^{x\sqrt{(3)}/2}\cos\left(\frac{x}{2} + C_4\right) +$$

$$+ C_5\,e^{-x\sqrt{(3)}/2}\cos\left(\frac{x}{2} + C_6\right).$$

165. $x^{2n}y^{(n)} = ay, \quad a \neq 0$;

$y = x^{n-1}\,e^{-r/x}$, where $r^n = a$; the n different roots of this equation yield n linearly independent solutions of equation 165.

166. $y'' = a(y'^2 + 1)^{3/2}$;

$$y = C_1 - \sqrt{\left[\frac{1}{a^2} - (x - C_2)^2\right]}.$$

167. $y'' = 2ax(y'^2 + 1)^{3/2}$;

$$y = C_1 + \int \frac{ax^2 + C_2}{\sqrt{[1 - (ax^2 + C_2)^2]}}\,\mathrm{d}x.$$

168. $8y'' + 9y'^4 = 0$;

$$(y + C_1)^3 = (x + C_2)^2.$$

169. $2y'y''' - 3y''^2 = 0$;

$$y = \frac{C_1 x + C_2}{C_3 x + C_4}.$$

170. $x'(t) = ay(t), \quad y'(t) = bx(t), \quad a \neq 0, \quad b \neq 0$;

$$x = C_1 a\,\mathrm{e}^{\sqrt{(ab)}t} + C_2 a\,\mathrm{e}^{-\sqrt{(ab)}t},$$

$$y = C_1\sqrt{(ab)}\,\mathrm{e}^{\sqrt{(ab)}t} - C_2\sqrt{(ab)}\,\mathrm{e}^{-\sqrt{(ab)}t} \qquad \text{if } ab > 0,$$

$$x = C_1 a\cos\sqrt{(-ab)}t + C_2 a\sin\sqrt{(-ab)}t\,,$$

$$y = C_2\sqrt{(-ab)}\cos\sqrt{(-ab)}t - C_1\sqrt{(-ab)}\sin\sqrt{(-ab)}t \qquad \text{if } ab < 0.$$

171. $x'(t) = ax(t) - y(t), \quad y'(t) = x(t) + ay(t)$;

$$x = \mathrm{e}^{at}(C_1\sin t + C_2\cos t), \quad y = \mathrm{e}^{at}(C_2\sin t - C_1\cos t)$$

(see § 17.18).

172. $ax'(t) + by'(t) = \alpha x(t) + \beta y(t), \quad bx'(t) - ay'(t) = \beta x(t) - \alpha y(t)$;

$$x = \mathrm{e}^{At}(C_1\cos Bt + C_2\sin Bt), \quad y = \mathrm{e}^{At}(C_2\cos Bt - C_1\sin Bt)$$

if $a\beta - b\alpha \neq 0$, with $A = \dfrac{a\alpha + b\beta}{a^2 + b^2}, \quad B = \dfrac{a\beta - b\alpha}{a^2 + b^2}$;

if $a\beta - b\alpha = 0$, $a^2 + b^2 > 0$, we get

$x = C_1\,\mathrm{e}^{\lambda t}$, $y = C_2\,\mathrm{e}^{\lambda t}$, where $\alpha = \lambda a$, $\beta = \lambda b$.

173. $x'(t) = -y(t), \quad y'(t) = 2x(t) + 2y(t)$;

$$x = \mathrm{e}^t(C_1\sin t + C_2\cos t), \quad y = \mathrm{e}^t\big[(C_2 - C_1)\sin t - (C_2 + C_1)\cos t\big].$$

174. $x'(t) + 3x(t) + 4y(t) = 0, \quad y'(t) + 2x(t) + 5y(t) = 0$;

$$x = 2C_1\,\mathrm{e}^{-t} + C_2\,\mathrm{e}^{-7t}, \quad y = -C_1\,\mathrm{e}^{-t} + C_2\,\mathrm{e}^{-7t}.$$

175. $x'(t) = -5x(t) - 2y(t), \quad y'(t) = x(t) - 7y(t);$

$x = (2C_1 \cos t + 2C_2 \sin t)\,\mathrm{e}^{-6t}, \quad y = \left[(C_1 - C_2)\cos t + (C_1 + C_2)\sin t\right]\mathrm{e}^{-6t}.$

176. $x'(t) + 2y(t) = 3t, \quad y'(t) - 2x(t) = 4;$

$x = -\frac{5}{4} + C_1 \cos 2t - C_2 \sin 2t, \qquad y = \frac{3}{2}t + C_1 \sin 2t + C_2 \cos 2t.$

177. $x'(t) + y(t) = t^2 + 6t + 1, \quad y'(t) - x(t) = -3t^2 + 3t + 1;$

$x = 3t^2 - t - 1 + C_1 \cos t + C_2 \sin t, \qquad y = t^2 + 2 + C_1 \sin t - C_2 \cos t.$

178. $x'(t) + y'(t) - y(t) = \mathrm{e}^t, \quad 2x'(t) + y'(t) + 2y(t) = \cos t;$

$x = \mathrm{e}^t + \frac{5}{17}\sin t - \frac{3}{17}\cos t + C_1 + 3C_2\,\mathrm{e}^{4t},$

$y = -\frac{2}{3}\,\mathrm{e}^t - \frac{1}{17}\sin t + \frac{4}{17}\cos t - 4C_2\,\mathrm{e}^{4t}.$

179. $x'(t) = x(t)f(t) + y(t)g(t), \quad y'(t) = -x(t)g(t) + y(t)f(t);$

$x = (C_1 \cos G + C_2 \sin G)F, \quad y = (-C_1 \sin G + C_2 \cos G)F,$

where $F = \mathrm{e}^{\int f(t)\,\mathrm{d}t}, \quad G = \int g(t)\,\mathrm{d}t.$

180. $tx'(t) + y(t) = 0, \quad ty'(t) + x(t) = 0;$

$x = C_1 t + \frac{C_2}{t}, \quad y = -C_1 t + \frac{C_2}{t}.$

181. $tx'(t) + 2x(t) = t, \quad ty'(t) - (t+2)x(t) - ty(t) = -t;$

$x = \frac{t}{3} + \frac{C_1}{t^2}, \quad y = -x + C_2\,\mathrm{e}^t.$

182. $tx'(t) + 2\big(x(t) - y(t)\big) = t, \quad ty'(t) + x(t) + 5y(t) = t^2;$

$x = \frac{3t}{10} + \frac{t^2}{15} + 2\frac{C_1}{t^3} + \frac{C_2}{t^4}, \quad y = -\frac{t}{20} + \frac{2t^2}{15} - \frac{C_1}{t^3} - \frac{C_2}{t^4}.$

183. $x''(t) + a^2 y(t) = 0, \quad y''(t) - a^2 x(t) = 0, \quad a \neq 0;$

$x = (C_1 \cos \alpha t + C_2 \sin \alpha t)\,\mathrm{e}^{\alpha t} + (C_3 \cos \alpha t + C_4 \sin \alpha t)\,\mathrm{e}^{-\alpha t},$

$y = (C_1 \sin \alpha t - C_2 \cos \alpha t)\,\mathrm{e}^{\alpha t} + (-C_3 \sin \alpha t + C_4 \cos \alpha t)\,\mathrm{e}^{-\alpha t},$

where $2\alpha^2 = a^2.$

184. $x''(t) - ay'(t) + bx(t) = 0, \quad y''(t) + ax'(t) + by(t) = 0, \quad a^2 + 4b > 0;$

$x = C_1 \cos \alpha t + C_2 \sin \alpha t + C_3 \cos \beta t + C_4 \sin \beta t,$

$y = -C_1 \sin \alpha t + C_2 \cos \alpha t - C_3 \sin \beta t + C_4 \cos \beta t,$

where $2\alpha = a + \sqrt{(a^2 + 4b)}\,, \quad 2\beta = a - \sqrt{(a^2 + 4b)}\,.$

185. $x''(t) - x'(t) + y'(t) = 0, \quad x''(t) + y''(t) - x(t) = 0;$

$x = C_1 \, \mathrm{e}^t + C_2 \alpha \, \mathrm{e}^{\alpha t} + C_3 \beta \, \mathrm{e}^{\beta t}, \quad y = C_4 - C_2 \, \mathrm{e}^{\alpha t} - C_3 \, \mathrm{e}^{\beta t},$

where $2\alpha = 1 + \sqrt{5}\,, \quad 2\beta = 1 - \sqrt{5}\,.$

186. $x''(t) = kx(t)/r^3(t), \quad y''(t) = ky(t)/r^3(t) \quad (r^2(t) = x^2(t) + y^2(t)$ motion of a particle in a central gravitational field);

on substituting $x = r \cos \varphi$, $y = \sin \varphi$, we get

$$r^2 \varphi' = C_1, \quad r'^2 + r^2 \varphi'^2 = -\frac{2k}{r} + C_2 \quad \left(\varphi' = \frac{\mathrm{d}\varphi}{\mathrm{d}t}, \quad r' = \frac{\mathrm{d}r}{\mathrm{d}t}\right);$$

$$r\left[C \cos(\varphi - \varphi_0) - k\right] = C_1^2 \quad (C^2 = C_2 C_1^2 + k^2),$$

which is the equation of a conic.

18. PARTIAL DIFFERENTIAL EQUATIONS

By KAREL REKTORYS

References: [14], [22], [26], [27], [29], [37], [41], [45], [60], [61], [65], [67], [76], [77], [79], [84], [85], [98], [99], [109], [112], [113], [115], [125], [132], [136], [144], [151], [155], [156], [157], [159], [160], [162], [172], [174], [181], [184], [185], [190], [211], [216], [220], [222], [223], [239], [246], [248], [251], [252], [259], [261], [264], [270], [272], [274], [277], [281], [282], [283], [284], [285], [287], [288], [290], [299], [303], [320], [324], [337], [339], [343], [346], [348], [349], [350], [353], [365], [369], [379], [384], [385], [389], [390], [393], [400], [413], [418], [426], [429], [430], [431], [432], [436], [437], [438], [444], [454], [458], [462], [465], [470], [471], [472], [486], [501], [502], [510], [515].

INFORMATIVE REMARK. Equations of the *first* order are treated in § 18.2, by standard classical methods. These equations are encountered mostly in problems of geometry.

As concerns equations of the second and higher orders, motivated by problems of physics, engineering, etc. (the so-called *equations of mathematical physics*), the whole problematics is very broad, from the point of view of applications as well as of the mathematical theory itself (classical approach, modern functional-analytical methods). Therefore, it was not easy to decide how to treat this rich and often very diverse problematics. Moreover, in contrast to problems in ordinary differential equations, solutions of problems in partial differential equations can be found in a "closed" form (i.e., expressed by an explicit formula, in the form of an infinite series, etc.) in simplest cases only. As usual it is necessary to solve them approximately. Thus we have chosen the following way of explanation: In §§ 18.4 – 18.6 the rather elementary classical theory is given for typical equations of the second order (the Laplace, Poisson, wave and heat equations). At the end of each of these paragraphs, remarks are added concerning possibility of some generalizations and choice of approximate methods of solution. In § 18.7, systems of equations are briefly treated. In § 18.8, 18.9 we present fundamental ideas of functional-analytical approach to the solution of sufficiently general elliptic equations. We deal

(i) with the theory based on the theorem on minimum of functional of energy, leading to the concept of a generalized solution,

(ii) with the Lax-Milgram theory, leading to the concept of a weak solution, and

(iii) with application of the Gâteaux differential to the solution of nonlinear problems.

This all gives a theoretical basis for approximate and numerical methods of solution, given in Chaps. 24 – 27. In § 18.10, the so-called method of discretization in time is briefly treated, giving a powerful theoretical and numerical tool for solving sufficiently general parabolic and hyperbolic problems.

Unless the contrary is stated, the equations as well as their solutions, dealt with in this chapter, are assumed to be *real.*

18.1. Partial Differential Equations in General. Basic Concepts. Questions Concerning the Concept of General Solution. Cauchy's Problem, Boundary Value Problems, Mixed Problems. Kovalewski's Theorem. Characteristics. Well-posed Problems.

Definition 1. We shall refer to an equation of the form

$$F\left(x_1, x_2, \ldots, x_n, z, \frac{\partial z}{\partial x_1}, \ldots, \frac{\partial z}{\partial x_n}, \frac{\partial^2 z}{\partial x_1^2}, \ldots, \frac{\partial^k z}{\partial x_1^k}, \ldots\right) = 0 \tag{1}$$

which relates the unknown function $z(x_1, x_2, \ldots, x_n)$ $(n \geqq 2)$ and its derivatives as a *partial differential equation.*

Definition 2. The highest order of the derivatives which appears in the equation is called the *order* of the equation.

REMARK 1. More generally, a *system of partial differential equations* for unknown functions $z_1, z_2, \ldots, z_r$ may be considered. If the number of equations is not equal to the number of functions to be found, then, generally speaking, there is no system of functions satisfying these equations. See e.g. Remark 18.7.1.

Example 1. The equation

$$\frac{\partial^2 z}{\partial t^2} = \frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2}$$

is a partial differential equation of the second order for the unknown function $z(x, y, t)$.

Example 2. The system

$$\frac{\partial z_1}{\partial t} = \frac{\partial z_1}{\partial x} - \frac{\partial^2 z_2}{\partial x^2}, \qquad \frac{\partial z_2}{\partial t} = \frac{\partial z_2}{\partial x} + \frac{\partial^2 z_1}{\partial x^2}$$

is a system of partial differential equations (of the second order) for two unknown functions $z_1(x, t)$, $z_2(x, t)$.

Definition 3. A function $z(x_1, x_2, \ldots, x_n)$ is called a (*classical*) *solution* (or *integral*) of equation (1) (in a given domain) if, on substituting in (1) for z and for its derivatives of the required orders, equation (1) is satisfied identically (i.e. for all points $(x_1, x_2, \ldots, x_n)$ of the domain in question).

Definition 4. By a *solution of a system of equations for r unknown functions* we understand any system of r functions having derivatives of the required orders and satisfying identically all equations of the given system.

As in the theory of ordinary differential equations, we may speak about a solution in an implicit form (cf. Remark 17.2.3).

REMARK 2. Similarly as in the case of ordinary differential equations, to solve a given partial equation means to find *all* its solutions. In contrast to the case of ordinary differential equations, where it is possible to find the general solution from which any other solution may be obtained (at least in a certain domain) by a suitable choice of some parameters, in the case of partial differential equations the situation is rather different. For some simple types of equations, a general form of the solution can be found. It then can be shown that this "general solution" depends on one or more arbitrary functions. However, in the general case it is not possible to find a general solution of an equation from which any solution of this equation (in the domain considered) might be found by specifying one or more "arbitrary" functions.

Example 3. On integrating the equation

$$\frac{\partial^2 z}{\partial x \partial y} = 0, \tag{2}$$

with respect to y, we obtain

$$\frac{\partial z}{\partial x} = f(x)$$

and by a further integration, with respect to x,

$$z = \int f(x)\,\mathrm{d}x + g(y) = F(x) + g(y). \tag{3}$$

This is the *general* solution of equation (1) which depends on two "arbitrary" functions F and g. Of course, one cannot conclude from this example that all solutions of a partial differential equation of the second order may be obtained from a certain "general form of the solution" involving two arbitrary functions by a suitable choice of these functions. It can be shown that under certain assumptions equations of the *first order* do possess a general solution depending on an arbitrary function (Remark 18.2.8).

REMARK 3. The analogue of the Cauchy problem (= of the problem with initial conditions) for ordinary differential equations, or of the boundary value problem, is here again the *Cauchy problem* and *boundary value problem*, respectively. One often meets with so-called *mixed problems* where initial conditions and boundary conditions are prescribed simultaneously.

Definition 5 (*The Special Cauchy Problem*; for important particular cases see Examples 4 and 5). Suppose we are given an equation of the m-th order, written in the form

$$\frac{\partial^m z}{\partial x_1^m} = f\left(x_1, x_2, \ldots, x_n, z, \ldots, \frac{\partial^k z}{\partial x_1^{k_1} \partial x_2^{k_2} \ldots \partial x_n^{k_n}}, \ldots\right), \tag{4}$$

$$k_1 + k_2 + \ldots + k_n = k \leqq m, \quad k_1 < m. \tag{5}$$

The *special Cauchy problem* is to find a solution of equation (4) (in a certain region) that satisfies the initial conditions

$$\begin{aligned} z(x_1^0, x_2, \ldots, x_n) &= f_0(x_2, \ldots, x_n), \\ \frac{\partial z}{\partial x_1}(x_1^0, x_2, \ldots, x_n) &= f_1(x_2, \ldots, x_n), \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ \frac{\partial^{m-1} z}{\partial x_1^{m-1}}(x_1^0, x_2, \ldots, x_n) &= f_{m-1}(x_2, \ldots, x_n). \end{aligned} \tag{6}$$

REMARK 4. The variable x_1 (which in applications often denotes time and is then usually denoted by t) plays a special role in equation (4) and initial conditions (6); particularly, the equation is assumed to be explicitly solved with respect to $\partial z^m / \partial x_1^m$ which is the highest derivative with respect to x_1 occurring in the equation.

Example 4. An example of the Cauchy problem is to find the solution of the equation

$$\frac{\partial^2 z}{\partial x^2} = f\left(x, y, z, \frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}, \frac{\partial^2 z}{\partial x \partial y}, \frac{\partial^2 z}{\partial y^2}\right) \tag{7}$$

with the conditions

$$z(x_0, y) = f_0(y), \qquad \frac{\partial z}{\partial x}(x_0, y) = f_1(y). \tag{8}$$

Example 5. An other example is the equation of the first order

$$\frac{\partial z}{\partial x} = f\left(x, y, z, \frac{\partial z}{\partial y}\right), \tag{9}$$

with the initial condition

$$z(x_0, y) = f_0(y). \tag{10}$$

REMARK 5. In the case of the problem (9), (10) for equations of the first order we often say that we are looking for a surface $z = h(x, y)$ that passes through the curve

$$x = x_0, \qquad z = f_0(y). \tag{11}$$

For an equation of the second order (Example 4) a further condition is attached, namely that the first derivative of the function z with respect to x is a prescribed function of y.

REMARK 6. If we choose a point $(x_1^0, x_2^0, \ldots, x_n^0)$, then all the derivatives occurring on the right-hand side of equation (4) and containing a derivative with respect to x_1, are determined, at this point, by conditions (6). For example (see Example 4; the interchangeability of mixed derivatives is assumed)

$$\frac{\partial^2 z}{\partial y \partial x}(x_0, y_0) = \left[\frac{\partial}{\partial y}\frac{\partial z}{\partial x}\right]_{\substack{x=x_0 \\ y=y_0}} = \left[\frac{\mathrm{d}}{\mathrm{d}y}\frac{\partial z}{\partial x}(x_0, y)\right]_{y=y_0} = f_1'(y_0). \tag{12}$$

Let us write

$$\left[\frac{\partial^{k-k_1} f_{k_1}}{\partial x_2^{k_2} \partial x_3^{k_3} \ldots \partial x_n^{k_n}}\right]_{x_i = x_i^0} = f^0_{k_1, k_2, \ldots, k_n} \qquad (i = 2, \ldots, n;\ k_1 + k_2 + \ldots + k_n = k). \tag{13}$$

For instance, in (12) we have $f_1'(y_0) = f^0_{1,1}$ since $k = 2$, $k_1 = 1$, $k_2 = 1$, $f_{k_1} = f_1(y)$.

Theorem 1 (*The Cauchy-Kovalewski Theorem*). *Let the function f appearing in the equation* (4) *be analytic with respect to all variables in the neighbourhood of the point*

$$(x_1^0, x_2^0, \ldots, x_n^0, \ldots, f^0_{k_1, k_2, \ldots, k_n, \ldots}) \tag{14}$$

(i.e. the function f can be developed into a power series in this neigbourhood) and let the functions $f_0, f_1, \ldots, f_{m-1}$ (see Definition 5) *be analytic in the neighbourhood of the point $(x_2^0, \ldots, x_n^0)$. Then there exists an analytic solution of the Cauchy problem in a neighbourhood of the point $(x_1^0, x_2^0, \ldots, x_n^0)$ and this solution is unique in the class of analytic functions.*

REMARK 7. Thus, in Example 4, we investigate the function f at the point (see (12) and (13))

$$(x_0, y_0, f^0_{0,0}, f^0_{1,0}, f^0_{0,1}, f^0_{1,1}, f^0_{0,2}), \tag{15}$$

where, by (13), we have

$$f^0_{0,0} = f_0(y_0), \quad f^0_{1,0} = f_1(y_0), \quad f^0_{0,1} = f_0'(y_0), \quad f^0_{1,1} = f_1'(y_0), \quad f^0_{0,2} = f_0''(y_0).$$

REMARK 8. A typical example of a Cauchy problem is that of solving the equation of a vibrating string

$$\frac{\partial^2 z}{\partial x^2} = a^2 \frac{\partial^2 z}{\partial y^2} \tag{16}$$

(here x denotes time, and y denotes the space variable), subject to the initial conditions

$$z(0, y) = f_0(y) \qquad \text{(the initial position of the string)},$$
$$\frac{\partial z}{\partial x}(0, y) = f_1(y) \qquad \text{(the initial velocity of the string)}$$

(cf. § 26.1). The right-hand side of equation (4) is of a very special form; consequently the investigation of the function f at the point (14) can be omitted and the existence of an analytic solution is ensured in the neighbourhood of the point $(0, y_0)$, provided only that the functions f_0 and f_1 are analytic in the neighbourhood of the point y_0.

REMARK 9. For generalizations of Theorem 1 to systems of equations see, e.g., [369]. The reader will also find there a statement of conditions (which are almost always satisfied in practice) which ensure uniqueness, but not existence, of the solution in a more general class of functions than is the class of analytic functions.

REMARK 10 (*The Generalized Cauchy Problem*). In this problem, the initial conditions are not prescribed on an $(n-1)$-dimensional hyperplane $x_1 = x_1^0$ (e.g., for $n = 2$, on a straight line) but on an $(n-1)$-dimensional surface (for $n = 2$, on a curve). (For details see [369].) This problem can, in general, be reduced to the previous one by a suitable transformation of coordinates. This reduction, however, is not practicable for some surfaces (curves) typical for the given differential equation; these surfaces (curves) are called *characteristics.* (For a strict definition see [369].) The direction of the normal to such a surface (curve) at a given point is called a *characteristic direction.* In the case of *linear* equations, the direction cosines $\alpha_1, \alpha_2, \ldots, \alpha_n$ of a characteristic direction are determined as follows: Let the equation be of the form

$$\sum_{k_1+k_2+\ldots+k_n=m} A^{(m)}_{k_1,k_2,\ldots,k_n}(x_1, x_2, \ldots, x_n) \frac{\partial^m z}{\partial x_1^{k_1} \partial x_2^{k_2} \ldots \partial x_n^{k_n}} + \ldots = 0 \tag{17}$$

(writing only the highest order terms). Then the direction cosines are given by the equations

$$\sum_{k_1+k_2+\ldots+k_n=m} A^{(m)}_{k_1,k_2,\ldots,k_n}(x_1, x_2, \ldots, x_n) \alpha_1^{k_1} \alpha_2^{k_2} \ldots \alpha_n^{k_n} = 0, \tag{18}$$

$$\alpha_1^2 + \alpha_2^2 + \ldots + \alpha_n^2 = 1. \tag{19}$$

Characteristics of some important equations with constant coefficients:

Example 6.

$$\frac{\partial^2 z}{\partial x^2} - \frac{\partial^2 z}{\partial y^2} = 0. \tag{20}$$

Equations (18) and (19) now become ($m = 2$, $A^{(2)}_{2,0} = 1$, $A^{(2)}_{0,2} = -1$, $A^{(2)}_{1,1} = 0$)

$$\alpha_1^2 - \alpha_2^2 = 0, \qquad \alpha_1^2 + \alpha_2^2 = 1 \quad \Rightarrow \quad \alpha_1 = \pm\frac{\sqrt{2}}{2}, \qquad \alpha_2 = \pm\frac{\sqrt{2}}{2}.$$

The characteristics run perpendicular to these directions; hence they are parallel to the straight lines $y = \pm x$ (Fig. 18.1).

Example 7.

$$\frac{\partial z}{\partial x} - \frac{\partial^2 z}{\partial y^2} = 0. \tag{21}$$

Equations (18) and (19) become ($m = 2$, $A^{(2)}_{2,0} = 0$, $A^{(2)}_{0,2} = -1$, $A^{(2)}_{1,1} = 0$):

$$\alpha_2^2 = 0, \qquad \alpha_1^2 + \alpha_2^2 = 1 \quad \Rightarrow \quad \alpha_2 = 0, \qquad \alpha_1 = \pm 1.$$

The characteristic direction is parallel to the x-axis, so that characteristics are parallel to the y-axis (Fig. 18.2).

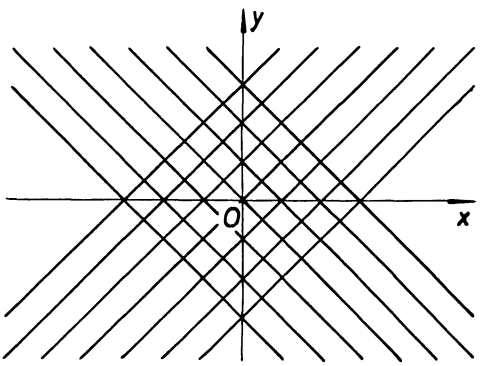


Fig. 18.1.

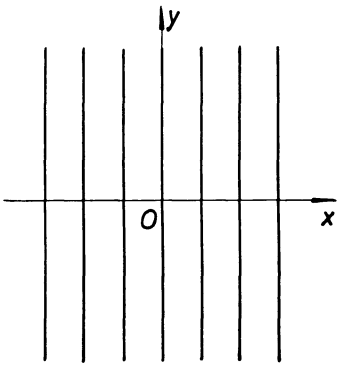


Fig. 18.2.

Example 8.

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = 0. \tag{22}$$

$\alpha_1^2 + \alpha_2^2 = 0, \qquad \alpha_1^2 + \alpha_2^2 = 1 \quad \Rightarrow$ real characteristics do not exist.

Example 9.

$$\frac{\partial z}{\partial x} = 0. \tag{23}$$

Equations (18) and (19) now become ($m = 1$, $A_{1,0}^{(1)} = 1$, $A_{0,1}^{(1)} = 0$):

$$\alpha_1 = 0, \qquad \alpha_1^2 + \alpha_2^2 = 1 \quad \Rightarrow \quad \alpha_1 = 0, \qquad \alpha_2 = \pm 1.$$

The characteristics are parallel to the x-axis.

REMARK 11. For *nonlinear* equations the problem of characteristics is made more complicated by the fact that the characteristics depend not only on the given equation but also on the given solution. For the *quasilinear equations of the first order*

$$a(x, y, z)\frac{\partial z}{\partial x} + b(x, y, z)\frac{\partial z}{\partial y} = c(x, y, z)$$

the characteristic direction at the point (x, y, z) is given by

$$a(x, y, z)\alpha_1 + b(x, y, z)\alpha_2 = 0, \qquad \alpha_1^2 + \alpha_2^2 = 1,$$

so that at the given point (x, y) it depends on the solution $z(x, y)$ in question.

Theorem 2. *If the initial values are given on a characteristic, then the Cauchy problem has either an infinite number of solutions, or it does not possess any solution.*

Example 10. A solution of the equation

$$\frac{\partial^2 z}{\partial x^2} - \frac{\partial^2 z}{\partial y^2} = 0,$$

which takes zero values on the stright line $y = x$ (which is a characteristic; see Example 6) and satisfies also

$$\frac{\partial z}{\partial x} = 0$$

on this line, is

$$z = k(x - y)^2$$

(for any value of k). Thus we have an infinite number of solutions.

Example 11. Find the solution of the equation

$$\frac{\partial z}{\partial x} = 0 \tag{24}$$

such that $z = x$ at points on the x-axis (which is a characteristic; see Example 9). It is readily verified that such a function $z(x, y)$ does not exist, because this condition requires

$$\frac{\partial z}{\partial x} = 1$$

on the x–axis, while (24) implies

$$\frac{\partial z}{\partial x} = 0.$$

REMARK 12. In applications, the Cauchy problem arises most frequently in the theory of equations of the first order and of hyperbolic and parabolic equations of the second order (§§ 18.5 and 18.6). *Boundary value problems* (thus problems with prescribed conditions on the boundary of the region where the solution is to be found) arise especially in connection with equations of mathematical physics (where also mixed problems are often encountered). An example is the problem of finding a solution of the equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \tag{25}$$

taking prescribed values on the boundary S of a region Ω in which equation (25) is to be satisfied (the so-called *Dirichlet problem*). Problems of this kind will be formulated and analysed separately for individual types of equations.

REMARK 13. It is important that the given problem be "well posed". This means – roughly speaking – the following: The problem is *well posed* if a slight change in the boundary and initial conditions produces only a slight change in the solution (we then speak of the *continuous dependence of the solution on the initial and boundary conditions*; physically: if we make only a small error in measuring the boundary conditions or the initial conditions, then the solution will remain almost unchanged). What we mean by a "well-posed" problem exactly, will be specified for individual types of equations separately. Next we state a definition (one of many possible ones) of a well-posed Cauchy problem for linear equations.

Definition 6. Let the Cauchy problem (4), (6) be given (see Definition 5). Let $\overline{\Omega}$ be a closed region in the space $x_1, x_2, \ldots, x_n$, the boundary of which contains the $(n-1)$-dimensional region G of the hyperplane $x_1 = x_1^0$. For any system of functions $f_0, f_1, \ldots, f_{m-1}$ which are sufficiently smooth in G let there be one and only one solution of the given problem. Suppose that for every $\varepsilon > 0$, there is a $\delta > 0$, such that the maximal change in values of the solution and of its derivatives up to the order k is less than ε in $\overline{\Omega}$ if the maximal change in the values of the functions f_0, $f_1, \ldots, f_{m-1}$ and of their derivatives up to the order p is less than δ in $\overline{G}$. Then the Cauchy problem for the given equation is said to be *well posed* (in details: *well posed* (p, k)).

REMARK 14. J. Hadamard proved (see e.g. [369]) that the Cauchy problem for the equation

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = 0$$

is not well posed. It will be shown later that almost all "reasonable" problems arising in physics (the so-called *problems of mathematical physics*) are well posed.

REMARK 15. It will be shown (Remark 18.5.7) that: If the boundary conditions or initial conditions are not smooth enough, then in the case of some equations the classical solution of the given problem need not exist even if the coefficients of these equations are very smooth. That is to say, it is not always possible to find a function having derivatives of the required orders and satisfying both the given equation in the region considered and the prescribed initial and boundary conditions. A typical example of such equations are hyperbolic equations. Because of this, the concept of a solution is generalized in a proper way, see, in particular, Definition 18.5.3. On the other hand, solutions of some equations are classical even if boundary or initial conditions are not smooth enough. To this type of equations there belong, first of all, the equations

$$\Delta u = 0 \tag{26}$$

and

$$\frac{\partial u}{\partial t} = a\Delta u \qquad (a > 0 \quad \text{being a constant}), \tag{27}$$

where Δ stands for the Laplace operator (Definition 18.4.1) in n-dimensional space. Solutions of equations (26) and (27) have even derivatives of all orders in the regions considered. In addition, the solutions of equation (26) are *analytic*, i.e. they can be expanded into Taylor series (in n variables) in a neighbourhood of any point of the region in question.

18.2. Partial Differential Equations of the First Order. Homogeneous and Nonhomogeneous Linear Equations, Nonlinear Equations. Complete, General and Singular Integrals. Solution of the Cauchy Problem

Definition 1. The equation of the form

$$a(x, y)\frac{\partial z}{\partial x} + b(x, y)\frac{\partial z}{\partial y} = 0 \tag{1}$$

is called a *homogeneous linear equation of the first order in two variables.*

Let us assume that $a(x, y)$ and $b(x, y)$ are continous in the region Ω in question and do not vanish simultaneously in Ω.

REMARK 1 (*Solution*). We solve the system of equations

$$\frac{\mathrm{d}x}{a(x, y)} = \frac{\mathrm{d}y}{b(x, y)} \tag{2}$$

(see Remark 17.20.4).

Theorem 1. *Let*

$$\psi(x, y) = C \tag{3}$$

be the first integral (§ 17.20) *of the system* (2). *Then the function*

$$z = F\big(\psi(x, y)\big),$$

where $F(t)$ *is an arbitrary differentiable function, is a solution of equation* (1). *In particular, if* $F(t) \equiv t$, *we have* $z = \psi(x, y)$.

REMARK 2 (*The Cauchy Problem*). The solution of equation (1) is to be found such that

$$z(x, y_0) = f(x). \tag{4}$$

Let the equation

$$\psi(x, y_0) = \varphi, \tag{5}$$

where ψ stands for the left-hand side of equation (3), be solvable with respect to x giving $x = g(\varphi)$. Then the required solution of the Cauchy problem is given by

$$z = f\big(g(\psi)\big). \tag{6}$$

Example 1. Find the solution of the equation

$$y\frac{\partial z}{\partial x} - x\frac{\partial z}{\partial y} = 0$$

satisfying $z(x, 0) = x^4$ $(1 \leqq x \leqq 4)$.

The first integral of the system

$$\frac{\mathrm{d}x}{y} = -\frac{\mathrm{d}y}{x}$$

is (see Example 17.20.3)

$$\psi(x, y) = x^2 + y^2 = C.$$

According to (5) we have

$$x^2 = \varphi \quad \text{or} \quad x = +\sqrt{\varphi}$$

and by (6)

$$z = (\sqrt{\psi})^4 = \left[\sqrt{(x^2 + y^2)}\right]^4 = (x^2 + y^2)^2.$$

REMARK 3. More generally, let the curve

$$x = \varphi(t), \qquad y = \psi(t), \qquad z = \chi(t) \tag{7}$$

be given instead of the curve (5). Let us assume that the vector $\big(\varphi'(t), \psi'(t)\big)$ is not proportional to the vector $\big(a\big(\varphi(t), \psi(t)\big), b\big(\varphi(t), \psi(t)\big)\big)$ for any t, i.e. the curve $x = \varphi(t)$, $y = \psi(t)$ is nowhere tangential to the direction field given by the functions a, b. Let

$$\Psi(x, y) = C \tag{8}$$

be a first integral of the system (2). We substitute for x and y from (7) and (8),

$$\Psi\big(\varphi(t), \psi(t)\big) = C. \tag{8'}$$

If we obtain t from (8') as a function of C, i.e. $t = h(C)$, substitute this expression into the equation $z = \chi(t)$ and write the left-hand side of equation (8) instead of C, we obtain the required solution

$$z = \chi\Big(h\big(\Psi(x, y)\big)\Big).$$

In the case of Example 1 we have: $x = t$, $y = 0$, $z = t^4$. Equation (8): $x^2 + y^2 = C$. Equation (8'): $t^2 = C$. Substituting for t, we obtain $z = C^2$ and using (8), we have

$$z = (x^2 + y^2)^2.$$

REMARK 4. In the general case of n variables, the investigation of the equation

$$a_1(x_1, x_2, \ldots, x_n)\frac{\partial z}{\partial x_1} + \ldots + a_n(x_1, x_2, \ldots, x_n)\frac{\partial z}{\partial x_n} = 0$$

reduces to that of the system

$$\frac{\mathrm{d}x_1}{a_1(x_1, x_2, \ldots, x_n)} = \ldots = \frac{\mathrm{d}x_n}{a_n(x_1, x_2, \ldots, x_n)}. \tag{9}$$

Let us assume that the functions $a_1, a_2, \ldots, a_n$ are continuous in the region Ω considered and that they do not vanish simultaneously anywhere in Ω. If the equations

$$\begin{aligned}
\psi_1(x_1, x_2, \ldots, x_n) &= C_1, \\
\psi_2(x_1, x_2, \ldots, x_n) &= C_2, \\
&\ldots\ldots\ldots\ldots\ldots\ldots \\
\psi_{n-1}(x_1, x_2, \ldots, x_n) &= C_{n-1}
\end{aligned}$$

constitute a system of independent first inregrals of the system (9), then

$$z = F\big(\psi_1(x_1, x_2, \ldots, x_n), \ldots, \psi_{n-1}(x_1, x_2, \ldots, x_n)\big),$$

(where F is an arbitrary differentiable function) is a solution of the given equation. For details the reader is referred to [444] which also shows the construction of the solution of the Cauchy problem.

Definition 2. The equation of the form

$$P(x, y, z)\frac{\partial z}{\partial x} + Q(x, y, z)\frac{\partial z}{\partial y} = R(x, y, z) \tag{10}$$

is called a *non-homogeneous linear equation of the first order in two variables* x, y. We assume that the functions P, Q, R are continuous in the region Ω in question, P, Q nowhere simultaneously vanishing, and $R \not\equiv 0$.

(In the theory of first order equations it is common to refer to equation (10) as a *linear equation* though it is, in fact, quasilinear, because the functions P, Q, R involve the unknown function z.)

REMARK 5 (*Solution*). To equation (10), there corresponds the system

$$\frac{\mathrm{d}x}{P(x, y, z)} = \frac{\mathrm{d}y}{Q(x, y, z)} = \frac{\mathrm{d}z}{R(x, y, z)}. \tag{11}$$

Theorem 2. *If the equations*

$$\varphi_1(x, y, z) = C_1, \qquad \varphi_2(x, y, z) = C_2 \tag{12}$$

constitute two independent first integrals (§ 17.20) *of the system* (11), *then the relation*

$$F\big(\varphi_1(x, y, z), \varphi_2(x, y, z)\big) = 0, \tag{13}$$

where F stands for an arbitrary differentiable function, gives a solution of equation (10).

REMARK 6 (The *Cauchy problem*). Let us find the solution of equation (10) that passes through the curve

$$x = f_1(t), \qquad y = f_2(t), \qquad z = f_3(t). \tag{14}$$

We assume that there is no value of t for which the vector $\big(f_1'(t), f_2'(t), f_3'(t)\big)$ is proportional to the vector

$$\Big(P\big(f_1(t), f_2(t), f_3(t)\big), Q\big(f_1(t), f_2(t), f_3(t)\big), R\big(f_1(t), f_2(t), f_3(t)\big)\Big)$$

(cf. Remark 3). We substitute from (14) for x, y and z into (12) and obtain C_1 and C_2 as functions of the variable t. If we obtain by eliminating t a relation between C_1 and C_2,

$$g(C_1, C_2) = 0, \tag{15}$$

then substituting (12) into (15), we get the solution:

$$g\big(\varphi_1(x, y, z), \varphi_2(x, y, z)\big) = 0.$$

For the linear non-homogeneous equation in n variables,

$$a_1(x_1, x_2, \ldots, x_n, z)\frac{\partial z}{\partial x_1} + \ldots + a_n(x_1, x_2, \ldots, x_n, z)\frac{\partial z}{\partial x_n} = b(x_1, x_2, \ldots, x_n, z),$$

the procedure is similar (a_k, b are assumed to be continuous in the region under consideration and the a_k not to vanish simultaneously anywhere). If the relations

$$\psi_0(x_1, x_2, \ldots, x_n, z) = C_1, \ \psi_1(x_1, x_2, \ldots, x_n, z) = C_2, \ \ldots$$
$$\ldots, \ \psi_{n-1}(x_1, x_2, \ldots, x_n, z) = C_{n-1}$$

constitute first integrals of the system

$$\frac{\mathrm{d}x_1}{a_1} = \frac{\mathrm{d}x_2}{a_2} = \ldots = \frac{\mathrm{d}x_n}{a_n} = \frac{\mathrm{d}z}{b},$$

then the relation

$$F\big(\psi_0(x_1, x_2, \ldots, x_n, z), \ldots, \psi_{n-1}(x_1, x_2, \ldots, x_n, z)\big) = 0$$

(F being an arbitrary differentialble function) gives the solution of the given equation.

For details see e.g. [444].

Definition 3. A *nonlinear equation of the first order* in two variables x, y is, in general, an equation of the form

$$f(x, y, z, p, q) = 0. \tag{16}$$

Here, we employ the brief notation

$$\frac{\partial z}{\partial x} = p, \qquad \frac{\partial z}{\partial y} = q. \tag{17}$$

The function f is assumed to be differentiable in the region in question.

REMARK 7. The previous cases of linear homogeneous and non-homogeneous equations in two variables are special cases of (16).

Definition 4. A function

$$z = \varphi(x, y, a, b), \quad \text{or given by} \quad V(x, y, z, a, b) = 0, \tag{18}$$

satisfying equation (16) and involving two arbitrary (independent) parameters is called the *complete integral* (*solution*) of equation (16).

REMARK 8. Each of equations (18) defines a two-parameter family of surfaces. However, (18) does *not* represent the general solution of equation (16); it can be shown that there always exist solutions which cannot be obtained from (18) by selecting appropriate values for the parameters a and b. On the other hand, all the solutions may be obtained from (18) by the so-called *method of variation of a parameter*: Let us choose an arbitrary (differentiable) function ω and set

$$b = \omega(a). \tag{19}$$

If we substitute (19) into the second of the equations (18), we obtain (keeping ω fixed) a one-parameter family of surfaces. If the envelope of this family is now obtained from the equations

$$V\big(x, y, z, a, \omega(a)\big) = 0, \qquad \frac{\partial V}{\partial a} + \frac{\partial V}{\partial b}\omega'(a) = 0 \tag{20}$$

by eliminating the parameter a, then this surface is an integral surface of equation (16). The system of all integrals obtained by specifying the arbitrary (differentiable) function ω is called the *general integral* of equation (16).

By eliminating both constants a and b from the equations

$$V = 0, \qquad \frac{\partial V}{\partial a} = 0, \qquad \frac{\partial V}{\partial b} = 0, \tag{21}$$

we obtain an envelope of the two-parameter family (18) (the so-called *singular integral* of equation (16)).

Example 2. One can readily verify by substitution, that the equation

$$(x - a)^2 + (y - b)^2 + z^2 - R^2 = 0 \tag{22}$$

constitutes a complete integral of the equation

$$z^2(1 + p^2 + q^2) = R^2;$$

(22) represents a two-parameter family of spheres having their centres at the points $(a, b, 0)$ (lying in the plane xy) and radius R. If we choose the function ω arbitrarily,

then (19) defines a curve on which the centres of the spheres of the system (22) lie. If we choose, for instance, $b = \omega(a) = a$, then the one-parameter family obtained is

$$(x-a)^2 + (y-a)^2 + z^2 - R^2 = 0. \tag{22'}$$

Eliminating a between (22') and the equation

$$-2(x-a) - 2(y-a) = 0$$

(which is obtained by differentiation of (22') with respect to a) we get

$$2(x-y)^2 + 4z^2 = 4R^2.$$

This is a cylindrical surface which envelops the family (22'). The singular integral is obtained from equations (21), namely

$$(x-a)^2 + (y-b)^2 + z^2 - R^2 = 0, \qquad -2(x-a) = 0, \qquad -2(x-b) = 0$$

that is

$$z^2 = R^2,$$

and this represents two planes $z = R$, $z = -R$, both enveloping the system (22).

REMARK 9. The complete integral of equation (16) can often be determined in a simple way, if equation (16) is of a particular form. This is done in the following Remark 10 for the cases where equation (16) takes the form (23), (26), or (28), respectively.

REMARK 10. Determination of the complete integral in some special cases.

1.

$$F(p, q) = 0. \tag{23}$$

We choose $p = a$, solve the equation $F(a, q) = 0$ for q, giving $q = g(a)$, and obtain the complete integral from the relation

$$\mathrm{d}z = p\,\mathrm{d}x + q\,\mathrm{d}y; \tag{24}$$

in other words

$$z = ax + g(a)y + b. \tag{25}$$

2.

$$\varphi(x, p) = \psi(y, q). \tag{26}$$

Choosing $\varphi(x, p) = a$, we have $\psi(y, q) = a$; from the first equation, it follows (under obvious assumptions) that $p = g_1(x, a)$ and from the second that $q = g_2(y, a)$. Integrating (24), we have

$$z = \int g_1(x, a)\,dx + \int g_2(y, a)\,dy + b. \tag{27}$$

3. *The generalized Clairaut equation,*

$$z = xp + yq + g(p, q), \tag{28}$$

has a complete integral

$$z = ax + by + g(a, b) \tag{29}$$

(see also Example 3 below).

REMARK 11 (*Solution of the Cauchy Problem using the Complete Interal*). It is required to find the integral surface of equation (16) passing through the curve

$$x = f_1(t), \qquad y = f_2(t), \qquad z = f_3(t) \tag{30}$$

when a complete integral of the equation (16),

$$V(x, y, z, a, b) = 0, \tag{31}$$

is known. Substituting (30) into (31), we obtain an equation relating a, b, t, say

$$G(a, b, t) = 0. \tag{32}$$

Then, eliminating t between (32) and the equation

$$\frac{\partial G}{\partial t}(a, b, t) = 0 \tag{33}$$

we obtain a relation between a and b, say

$$H(a, b) = 0 \qquad \text{or} \qquad b = \omega(a). \tag{34}$$

Substituting (34) into (31) gives us a one-parameter family of surfaces

$$K(x, y, z, a) = 0. \tag{35}$$

Finally, eliminating a between (35) and the equation

$$\frac{\partial K}{\partial a}(x, y, z, a) = 0, \tag{36}$$

we obtain the equation of the envelope of the system (35), which, as a rule, is the required integral surface.

It must be realized that the procedure employed above to obtain the solution of the given problem is only formal (notice that we have made use of such vague notions as the elimination of variables, etc.) so that the result obtained must be analysed carefully. In particular we must show that it really represents the solution.

The same is true for the Cauchy method (Remark 15 below).

Example 3. Find the integral surface of the equation

$$z = px + qy + pq \tag{37}$$

passing through the curve

$$x = 0, \qquad z = y^2. \tag{38}$$

By (29) a complete integral is

$$ax + by - z + ab = 0. \tag{39}$$

Now parametric equations of the curve (38) are

$$x = 0, \qquad y = t, \qquad z = t^2, \tag{40}$$

and by substituting these into (39) we get

$$bt - t^2 + ab = 0. \tag{41}$$

Equation (33) then becomes

$$b - 2t = 0. \tag{42}$$

Next, eliminating t between (41) and (42), we have

$$a = -\tfrac{1}{4}b, \tag{43}$$

i.e.

$$b = -4a. \tag{44}$$

Substituting into (39), we obtain (as in(35)):

$$ax - 4ay - z - 4a^2 = 0. \tag{45}$$

Differentiating this with respect to a,

$$x - 4y - 8a = 0, \tag{46}$$

and eliminating a between (45) and (46) (that is, substituting $8a = x - 4y$ from (46) into (45)), we obtain

$$z = \left(y - \frac{x}{4}\right)^2. \tag{47}$$

REMARK 12. For the equation

$$f(x_1, x_2, \ldots, x_n, z, p_1, p_2, \ldots, p_n) = 0, \tag{48}$$

where

$$p_1 = \frac{\partial z}{\partial x_1}, \quad p_2 = \frac{\partial z}{\partial x_2}, \quad \ldots, \quad p_n = \frac{\partial z}{\partial x_n},$$

the complete integral

$$z = \varphi(x_1, x_2, \ldots, x_n, a_1, a_2, \ldots, a_n), \tag{49}$$

or

$$V(x_1, x_2, \ldots, x_n, z, a_1, a_2, \ldots, a_n) = 0 \tag{50}$$

is defined similary as for the equation (16).

REMARK 13. Let us note that, in the literature, the complete integral of equation (48) is sometimes defined as such a solution (49) (or (50)) of this equation, that by eliminating $a_1, a_2, \ldots, a_n$ between (49) (or (50)) and the equations obtained by differentiating (49) (or (50)) with respect to $x_1, x_2, \ldots, x_n$, one arrives at the equation (48).

REMARK 14. Complete integrals for equations with several variables play an important role in analytical mechanics, e.g. when integrating the so-called *Hamiltonian equations of motion.*

REMARK 15 (*The Cauchy (Lagrange-Charpit) Method of Solution of the Cauchy Problem in Two Varibles*). If we are given the equation

$$f(x, y, z, p, q) = 0, \tag{51}$$

we write down the following system of ordinary equations

$$\frac{\mathrm{d}x}{P} = \frac{\mathrm{d}y}{Q} = \frac{\mathrm{d}z}{Pp + Qq} = -\frac{\mathrm{d}p}{X + Zp} = -\frac{\mathrm{d}q}{Y + Zq} \tag{52}$$

or, which amounts the same thing, the system

$$\frac{\mathrm{d}x}{\mathrm{d}u} = P, \quad \frac{\mathrm{d}y}{\mathrm{d}u} = Q, \quad \frac{\mathrm{d}z}{\mathrm{d}u} = Pp + Qq, \quad \frac{\mathrm{d}p}{\mathrm{d}u} = -X - Zp, \quad \frac{\mathrm{d}q}{\mathrm{d}u} = -Y - Zq. \tag{53}$$

Here we have employed the abbreviated notation

$$\frac{\partial f}{\partial x} = X, \quad \frac{\partial f}{\partial y} = Y, \quad \frac{\partial f}{\partial z} = Z, \quad \frac{\partial f}{\partial p} = P, \quad \frac{\partial f}{\partial q} = Q. \tag{54}$$

Any solution

$$x = x(u), \quad y = y(u), \quad z = z(u), \quad p = p(u), \quad q = q(u) \tag{55}$$

of the system (53) is said to be a *characteristic strip* (or a *characteristic of the first order*) corresponding to equation (51). Its geometric significance is as follows: At every point of the curve

$$x = x(u), \quad y = y(u), \quad z = z(u)$$

the values of $p(u)$ and $q(u)$ are known, thus a "tangent element" of the surface is given. If, for a certain value u_0 of the variable u, the functions (55) of the characteristic strip satisfy the equation (51) (i.e. if for the corresponding values $x_0 = x(u_0), \ldots$, the equation

$$f(x_0, y_0, z_0, p_0, q_0) = 0 \tag{56}$$

is satisfied), then this equation is satisfied at all points of the characteristic strip. In this case, the characteristic strip is called the *integral strip of equation* (51) and the tangent elements which constitute it are called *integral elements of equation* (51). An integral strip is uniquely determined by the initial integral element of the given equation. If two integral surfaces have one common integral element, then they touch each other along the whole integral strip.

Cauchy's method of solving the Cauchy problem consists in the construction of integral surfaces from integral strips. We are given equation (51) and the curve

$$x = f_1(t), \quad y = f_2(t), \quad z = f_3(t). \tag{57}$$

The functions $p(t)$ and q(t) are then determined from the equations

$$\frac{\mathrm{d}z}{\mathrm{d}t} = p(t)\frac{\mathrm{d}x}{\mathrm{d}t} + q(t)\frac{\mathrm{d}y}{\mathrm{d}t}, \qquad f\big(x(t), y(t), z(t), p(t), q(t)\big) = 0. \tag{58}$$

(If equation (51) is nonlinear, equations (58) may yield many solutions or may not yield any real solution at all). Then we find the solution (55) of (53) such that for $u = u_0$ (and for all t) the conditions

$$x(u_0) = f_1(t), \quad y(u_0) = f_2(t), \quad z(u_0) = f_3(t), \quad p(u_0) = p(t), \quad q(u_0) = q(t)$$

are satisfied (we usually choose $u_0 = 0$). In this way, x, y, z, p, q are obtained as functions of two parameters u and t. Eliminating u and t from the equations

$$x = x(u, t), \quad y = y(u, t), \quad z = z(u, t), \tag{59}$$

we obtain the required relation between x, y, z.

Example 4. Let us solve the problem of Example 3 by Cauchy's method. That is, we have to find a solution of the equation

$$px + qy + pq - z = 0 \tag{60}$$

passing through the curve

$$x = 0 \quad z = y^2 \quad \text{(or, in parametric form, } x = 0,\ y = t,\ z = t^2). \tag{61}$$

We have

$$X = p, \quad Y = q, \quad P = x + q, \quad Q = y + p, \quad Z = -1. \tag{62}$$

Equations (53) now take the form:

$$\begin{gathered} \frac{\mathrm{d}x}{\mathrm{d}u} = x + q, \quad \frac{\mathrm{d}y}{\mathrm{d}u} = y + p, \quad \frac{\mathrm{d}z}{\mathrm{d}u} = (x + q)p + (y + p)q, \\ \frac{\mathrm{d}p}{\mathrm{d}u} = -(p - p) = 0, \quad \frac{\mathrm{d}q}{\mathrm{d}u} = -(q - q) = 0. \end{gathered} \tag{63}$$

From the latter equations, it follows that

$$p = C_1 = \text{const.}, \quad q = C_2 = \text{const.}; \tag{64}$$

on substituting these expressions into first three equations of (63), we get

$$\frac{\mathrm{d}x}{\mathrm{d}u} = x + C_2, \quad \frac{\mathrm{d}y}{\mathrm{d}u} = y + C_1, \quad \frac{\mathrm{d}z}{\mathrm{d}u} = C_1 x + C_2 y + 2C_1 C_2. \tag{65}$$

The first two equations are linear so that

$$x = -C_2 + C_3 \mathrm{e}^u, \quad y = -C_1 + C_4 \mathrm{e}^u. \tag{66}$$

When we insert (66) into the third equation (65) and integrate we obtain

$$z = (C_1 C_3 + C_2 C_4)\mathrm{e}^u + C_5. \tag{67}$$

We now choose the constants $C_1, C_2, \ldots, C_5$ depending on t in such a way that the conditions (61) be satisfied. At the same time the conditions (58) should be fulfilled. It follows from (61) that

$$\frac{\mathrm{d}x}{\mathrm{d}t} = 0, \quad \frac{\mathrm{d}y}{\mathrm{d}t} = 1, \quad \frac{\mathrm{d}z}{\mathrm{d}t} = 2t$$

so that the conditions (58) take the form

$$2t = p \,.\, 0 + q \,.\, 1, \quad p \,.\, 0 + q \,.\, t + pq - t^2 = 0,$$

whence

$$q(t) = 2t, \quad p(t) = -\frac{t}{2}.$$

For $u = 0$ we then require that

$$x = 0, \quad y = t, \quad z = t^2, \quad p = -\frac{t}{2}, \quad q = 2t$$

so that by (64), (66) and (67)

$$C_1 = -\frac{t}{2}, \quad C_2 = 2t, \quad C_3 = 2t, \quad C_4 = \frac{t}{2}, \quad C_5 = t^2.$$

Hence

$$x = -2t(1 - \mathrm{e}^u), \quad y = \frac{t}{2}(1 + \mathrm{e}^u), \quad z = t^2, \quad p = -\frac{t}{2}, \quad q = 2t. \tag{68}$$

Eliminating e^u from the first two of equations (68), we have

$$-x + 4y = 4t, \quad \text{that is} \quad \left(y - \frac{x}{4}\right)^2 = t^2,$$

and by the third equation (68)

$$z = \left(y - \frac{x}{4}\right)^2$$

which agrees with the result of Example 3.

REMARK 16. In the particular case where equation (51) is linear,

$$a(x, y, z)p + b(x, y, z)q - r(x, y, z) = 0, \tag{69}$$

Cauchy's method may be considerably simplified because the first three equations (53),

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}u} &= a(x, y, z), \quad \frac{\mathrm{d}y}{\mathrm{d}u} = b(x, y, z), \\ \frac{\mathrm{d}z}{\mathrm{d}u} &= a(x, y, z)p + b(x, y, z)q = r(x, y, z), \end{aligned} \tag{70}$$

are now sufficient to determine the functions

$$x = x(u, C_1, C_2, C_3), \quad y = y(u, C_1, C_2, C_3), \quad z = z(u, C_1, C_2, C_3). \tag{71}$$

Equations (70), (71) define the *characterictics* of equation (69). (To be exact, we should, according to Remark 18.1.10, define the characteristics as the projections of the curves (71) onto the xy-plane. However, the above terminology is in common use.) The initial curve

$$x = f_1(t), \quad y = f_2(t), \quad z = f_3(t), \tag{72}$$

(the projection of which onto the xy-plane is supposed not to touch the projection of any characteristic), together with the characteristics (71), determines the integral surface.

Example 5. Find the solution of the equation

$$xp + yq - z = 0 \tag{73}$$

passing through the curve

$$x = t, \quad y = t^2, \quad z = t^3 \quad (t > 0). \tag{74}$$

The first three equations (53) read

$$\frac{\mathrm{d}x}{\mathrm{d}u} = x, \quad \frac{\mathrm{d}y}{\mathrm{d}u} = y, \quad \frac{\mathrm{d}z}{\mathrm{d}u} = xp + yq = z. \tag{75}$$

The solution satisfying the initial conditions (74) for $u = 0$ is

$$x = t\mathrm{e}^u, \quad y = t^2\mathrm{e}^u, \quad z = t^3\mathrm{e}^u; \tag{76}$$

eliminating t and u we obtain

$$z = \frac{y^2}{x}.$$

REMARK 17. The case of the linear *homogeneous* equation

$$a(x, y)p + b(x, y)q = 0 \tag{77}$$

is even simpler, for we need only solve the following two equations:

$$\frac{\mathrm{d}x}{\mathrm{d}u} = a(x, y), \quad \frac{\mathrm{d}y}{\mathrm{d}u} = b(x, y). \tag{78}$$

The third equation (53) now reads

$$\frac{\mathrm{d}z}{\mathrm{d}u} = 0, \quad \text{or} \quad z = C = \text{const.},$$

so that the characteristics (using the terminology of Remark 16) are parallel to the xy-plane.

REMARK 18. The existence and uniqueness of solutions (in the class of differentiable functions) is ensured for linear equations unless the projection of the curve (72) touches the projection of a characteristic, i.e. unless the numbers $\mathrm{d}x/\mathrm{d}t$, $\mathrm{d}y/\mathrm{d}t$ at a point (x, y, z) are proportional to the numbers $\mathrm{d}x/\mathrm{d}u$, $\mathrm{d}y/\mathrm{d}u$ given by equations (70) at the same point (x, y, z). On the other hand, if this proportionality holds at all points of (72) and at the same time the relation

$$\frac{\mathrm{d}x}{\mathrm{d}t} : \frac{\mathrm{d}y}{\mathrm{d}t} : \frac{\mathrm{d}z}{\mathrm{d}t} = \frac{\mathrm{d}x}{\mathrm{d}u} : \frac{\mathrm{d}y}{\mathrm{d}u} : \frac{\mathrm{d}z}{\mathrm{d}u} \tag{79}$$

does not hold, then the given problem has no solution. If the relation (79) is satisfied for all points of the curve (72), i.e. the curve (72) itself is a characteristic, then the problem has an infinite number of solutions.

Example 6. Find, first, the solution of Equation (73) passing through the straight line

$$x = t, \quad y = t, \quad z = 1. \tag{80}$$

The solution of equations (75) satisfying conditions (80) is

$$x = t\mathrm{e}^u, \quad y = t\mathrm{e}^u, \quad z = \mathrm{e}^u,$$

and from these equations z cannot be expressed as a function of the variables x, y. (In (80), $x = y$ for every t, so that by (75)

$$\frac{\mathrm{d}x}{\mathrm{d}t} : \frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\mathrm{d}x}{\mathrm{d}u} : \frac{\mathrm{d}y}{\mathrm{d}u}$$

for every t; however, the relation (79),

$$\frac{\mathrm{d}x}{\mathrm{d}t} : \frac{\mathrm{d}y}{\mathrm{d}t} : \frac{\mathrm{d}z}{\mathrm{d}t} = \frac{\mathrm{d}x}{\mathrm{d}u} : \frac{\mathrm{d}y}{\mathrm{d}u} : \frac{\mathrm{d}z}{\mathrm{d}u}$$

does not hold in this case.)

If the initial curve is replaced by the curve

$$x = t, \quad y = t, \quad z = t, \tag{81}$$

then the problem has an infinite number of solutions. An integral surface passing through (81) is any surface

$$z = k_1 x + k_2 y, \quad \text{where} \quad k_1 + k_2 = 1.$$

REMARK 19. In the case of several variables, the characteristic strips (characteristics of the first order) are defined by the equations

$$\frac{dx_1}{P_1} = \ldots = \frac{dx_n}{P_n} = \frac{dz}{\sum_{k=1}^{n} P_k p_k} = -\frac{dp_1}{X_1 + Zp_1} = \ldots = \frac{dp_n}{X_n + Zp_n}, \tag{82}$$

where

$$p_k = \frac{\partial z}{\partial x_k}, \quad P_k = \frac{\partial f}{\partial p_k}, \quad X_k = \frac{\partial f}{\partial x_k}, \quad Z = \frac{\partial f}{\partial z}. \tag{83}$$

The Cauchy problem: Given an $(n-1)$-dimensional surface

$$x_k = x_k(v_1, v_2, \ldots, v_{n-1}), \quad z = z(v_1, v_2, \ldots, v_{n-1}); \tag{84}$$

the problem of finding an integral surface $z = z(x_1, x_2, \ldots, x_n)$ passing through the surface (84), is called the *Cauchy problem.*

When solving this problem, we first determine

$$p_k(v_1, v_2, \ldots, v_{n-1}) \quad (k = 1, 2, \ldots, n) \tag{85}$$

from the equations

$$f(x_1, \ldots, x_n, z, p_1, \ldots, p_n) = 0, \quad \frac{\partial z}{\partial v_i} - \sum_{k=1}^{n} p_k \frac{\partial x_k}{\partial v_i} = 0 \quad (i = 1, 2, \ldots, n-1) \tag{86}$$

which correspond to equations (58) and then determine a solution

$$\begin{gathered} x_k = x_k(u, v_1, \ldots, v_{n-1}), \quad z = z(u, v_1, \ldots, v_{n-1}), \\ p_k = p_k(u, v_1, \ldots, v_{n-1}) \end{gathered} \tag{87}$$

of the system (82) such that the functions (87) assume the initial values (84) and (85) for $u = u_0$ (usually we choose $u_0 = 0$). By eliminating the parameters u, v_1, $\ldots$, v_{n-1} form the first $n+1$ equations (87), i.e. from the equations for x_k and z, we obtain the desired integral surface, i.e. the relation $g(x_1, x_2, \ldots, x_n, z) = 0$. (See, however, Remark 11, p. 163.)

For details see e.g. [444].

18.3. Linear Equations of the Second Order. Classification

Definition 1. The equation of the form

$$A_{11}(x, y)\frac{\partial^2 z}{\partial x^2} + 2A_{12}(x, y)\frac{\partial^2 z}{\partial x \partial y} + A_{22}(x, y)\frac{\partial^2 z}{\partial y^2} + B_1(x, y)\frac{\partial z}{\partial x} + \\ + B_2(x, y)\frac{\partial z}{\partial y} + C(x, y)z + D(x, y) = 0 \tag{1}$$

is called a *linear equation of the second order for the function* $z(x, y)$.

REMARK 1. The coefficients $A_{11}, \ldots, D$ are assumed to be continuous (as functions of the variables x, y) in the region Ω in question.

Definition 2. If everywhere in Ω

$$\left.\begin{array}{l} A_{11}A_{22} - A_{12}^2 > 0 \\ A_{11}A_{22} - A_{12}^2 < 0 \\ A_{11}A_{22} - A_{12}^2 = 0 \end{array}\right\}, \text{ the equation is said to be } \left\{\begin{array}{l} \textit{elliptic} \\ \textit{hyperbolic} \\ \textit{parabolic} \end{array}\right\} \text{ in } \Omega.$$

Theorem 1. *By a suitable transformation of variables, every elliptic, hyperbolic, or parabolic equation in* Ω *can be reduced in a neighbourhood of any point* $(x_0, y_0) \in \Omega$, *to the so-called canonical form* (2), (3) *or* (4), *respectively*:

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} + a_1(x, y)\frac{\partial z}{\partial x} + b_1(x, y)\frac{\partial z}{\partial y} + c_1(x, y)z + d_1(x, y) = 0, \tag{2}$$

$$\frac{\partial^2 z}{\partial x^2} - \frac{\partial^2 z}{\partial y^2} + a_2(x, y)\frac{\partial z}{\partial x} + b_2(x, y)\frac{\partial z}{\partial y} + c_2(x, y)z + d_2(x, y) = 0, \tag{3}$$

$$\frac{\partial^2 z}{\partial y^2} + a_3(x, y)\frac{\partial z}{\partial x} + b_3(x, y)\frac{\partial z}{\partial y} + c_3(x, y)z + d_3(x, y) = 0, \quad a_3(x, y) \neq 0. \tag{4}$$

REMARK 2. Linear equations of the second order for a function $z(x_1, x_2, \ldots, x_n)$ of more than two variables can be transformed by a suitable change of variables to canonical forms similar to those of (2), (3), (4). However, in the general case it is not possible to find a transformation which reduces the given equation to its canonical form in a whole neighbourhood of the given point but only such that it does so at the point itself. Consequently, an equation for several variables is called elliptic, hyperbolic or parabolic at the point $(x_1^0, x_2^0, \ldots, x_n^0)$ if, by a suitable transformation, it can be reduced at this point to the form (5), (6) or (7), respectively:

$$\sum_{k=1}^{n} \frac{\partial^2 z}{\partial x_k^2} + \ldots = 0, \tag{5}$$

$$\sum_{k=1}^{n-1} \frac{\partial^2 z}{\partial x_k^2} - \frac{\partial^2 z}{\partial x_n^2} + \ldots = 0, \tag{6}$$

$$\sum_{k=2}^{n} \frac{\partial^2 z}{\partial x_k^2} + a_1 \frac{\partial z}{\partial x_1} + \ldots = 0. \tag{7}$$

Only the terms involving second order derivatives are written in equations (5) and (6); in equation (7) the coefficient a_1 of the derivative $\partial z/\partial x_1$ is required to be different from zero. If an equation satisfies (5), (6) or (7) at every point of the region in question, it is called elliptic, hyperbolic, or parabolic in this region, respectively.

REMARK 3. If, in equation (6), the minus sign as well as the plus sign stand before more than one of the second order derivatives (while at the same time second derivatives with respect to all n variables are present), the equation is said to be *ultrahyperbolic*. If, in equation (7), more than one of the second derivatives is absent, the equation is said to be *parabolic in a wider sense*.

Example 1. The equation

$$(1+y^2)\frac{\partial^2 z}{\partial x^2} + (1+y^2)\frac{\partial^2 z}{\partial y^2} - \frac{\partial z}{\partial x} = 0$$

is an elliptic equation for the function $z(x, y)$ in the whole plane xy (by Definition 2, because $A_{11}A_{22} - A_{12}^2 = (1+y^2)^2 > 0$);

the equation

$$\frac{\partial^2 z}{\partial x_1^2} + \frac{\partial^2 z}{\partial x_2^2} - \frac{\partial^2 z}{\partial x_3^2} = 0$$

is a hyperbolic equation for the function $z(x_1, x_2, x_3)$ according to (6);

the equation

$$\frac{\partial^2 z}{\partial x_1^2} + \frac{\partial^2 z}{\partial x_2^2} - \frac{\partial^2 z}{\partial x_3^2} - \frac{\partial^2 z}{\partial x_4^2} = 0$$

is an ultrahyperbolic equation for the function $z(x_1, x_2, x_3, x_4)$ (by Remark 3);

the equation

$$\frac{\partial^2 z}{\partial x_2^2} + \frac{\partial z}{\partial x_1} + \frac{\partial z}{\partial x_3} = 0$$

is a parabolic equation in a wider sense for the function $z(x_1, x_2, x_3)$ and

the equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}$$

is a parabolic equation for the function $u(x, y, z, t)$.

18.4. Elliptic Equations. The Laplace Equation, the Poisson Equation. The Dirichlet and Neumann Problems. Properties of Harmonic Functions. The Fundamental Solution. Green's Function. Potentials

REMARK 1 (*informative*; see also the informative remark at the beginning of this chapter). In this paragraph, we deal mainly with the Laplace and Poisson equations. On some generalization of the results and on more general elliptic equations see Remarks 22 – 24 and, in particular, §§ 18.8 and 18.9.

Definition 1. The equation

$$\Delta u = 0, \tag{1}$$

where

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \ldots + \frac{\partial^2 u}{\partial x_n^2} \tag{2}$$

and $n \geqq 2$, is called the *Laplace (differential) equation;* Δ *is the so-called Laplace operator.*

(The symbol ∇^2 is often used in place of Δ.)

In particular, the Laplace equation for two and three variables reads

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{3}$$

and

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0, \tag{4}$$

respectively.

Definition 2. The equation of the form

$$\Delta u = f(x_1, x_2, \ldots, x_n) \tag{5}$$

is called the *Poisson* (differential) *equation.* In particular, the Poisson equation for two or three variables reads

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y), \tag{6}$$

or

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = f(x, y, z), \tag{7}$$

respectively.

Note that the Laplace equation is a special case of the Poisson equation.

Definition 3. Let $f(x_1, x_2, \ldots, x_n)$ be a continuous function in the region Ω. Under a *solution of equation* (5) in this region, we understand every function $u(x_1, x_2, \ldots, x_n)$ with continuous derivatives of the second order in Ω which satisfies equation (5) everywhere in Ω.

In particular, a solution of equation (1) is every function with continuous second order derivatives in Ω which satisfies (1) in Ω.

On generalization of the concept of the solution see, especially, §§ 18.8 and 18.9. The just defined solution is often called *classical*.

Definition 4. Every function which is a solution of the Laplace equation in Ω, is called *harmonic* (in Ω).

Examples of harmonic functions, in the xy-plane, are the functions $u \equiv 1$, $u = x^2 - y^2$, $u = e^y \sin x$.

Definition 5. Let us write

$$r = \sqrt{(x^2 + y^2 + z^2)}\,.$$

The function $u(x, y, z)$ (defined for all sufficiently large r) is said to *vanish at infinity* if for every $\varepsilon > 0$ there exists such $R > 0$ that $|u(x, y, z)| < \varepsilon$ whenever the point (x, y, z) is such that $r > R$.

For the case of n varibles the meaning of the statement "*u vanishes at infinity*" is similar; r is then defined by

$$r = \sqrt{(x_1^2 + x_2^2 + \ldots + x_n^2)}\,.$$

Theorem 1. *If the function $f(x, y, z)$ is continuously differentiable in the entire three-dimensional space and if, for large r, the inequality*

$$|f(x, y, z)| < \frac{A}{r^{2+\alpha}} \tag{8}$$

holds, where A and α are positive constants, then the function

$$u(x, y, z) = -\frac{1}{4\pi} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{f(\xi, \eta, \zeta)}{\sqrt{[(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2]}}\, d\xi\, d\eta\, d\zeta \tag{9}$$

satisfies equation (7) *everywhere. Moreover,* (9) *is the only solution of equation* (7) *which vanishes at infinity.*

REMARK 2. The integral (9) is called the *volume* (*Newton*) *potential*.

REMARK 3. In contrast with the three-dimensional case, the two-dimensional problem of finding a solution of equation (6) that vanishes at infinity is not solvable, in general. The integral (similar to that of (9))

$$u(x, y) = -\frac{1}{2\pi} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\xi, \eta) \ln \frac{1}{\sqrt{[(x-\xi)^2 + (y-\eta)^2]}}\, d\xi\, d\eta \tag{10}$$

(the so-called *logarithmic potential*) satisfies, under similar assumptions on the function f as in Theorem 1, equation (6) everywhere but does not, in general, vanish at infinity.

REMARK 4. The integral

$$v(x, y) = -\frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{f(\xi, \eta)}{\sqrt{[(x-\xi)^2 + (y-\eta)^2]}} \mathrm{d}\xi \, \mathrm{d}\eta \tag{11}$$

is *not* solution of equation (6).

Definition 6. The *Dirichlet problem* for the Laplace (or Poisson) equation is the problem of finding such a solution of this equation which assumes, on the boundary of the given region, prescribed values. In details: Let a region Ω with the boundary S be given. We have to find a function which is a solution of the Laplace (or Poisson) equation in Ω (Definition 3), is continuous in $\overline{\Omega} = \Omega + S$ and assumes, on S, prescribed values (given by a continuous function g).

Definition 7. The *Neumann problem* is to find a solution of the given equation such that it is continuous together with its first derivatives in $\overline{\Omega}$ and that the outward normal derivative $\partial u/\partial n$ assumes, on S, prescribed values given by a continuous function h.

REMARK 5. The Dirichlet and Neumann problems may also be defined for more general elliptic equations than for the Laplace or Poisson equation. Moreover, other problems may be solved for a given equation; for instance, a relation between the function u and its outward normal derivative may be prescribed on S (the *Newton problem*). Mixed problems are also encountered, for instance, when Dirichlet's condition is prescribed on one part of the boundary and Neumann's condition on the remaining part.

REMARK 6. We shall be mainly concerned with two and three-dimensional problems, that is, with equations (3), (4), (6), (7). If the region Ω is *bounded* (multiply connected regions are allowed), the Dirichlet and Neumann problems are referred to as *interior problems*; if Ω is the *exterior* of a simple closed curve (or of a simple closed surface which is the boundary of a simply connected region in three-dimensional space), we refer to an *exterior problem*. The *outward normal* is understood to be oriented (at a given point of the boundary) in a direction out of the given region. (For instance, if Ω is the exterior of the circle k with centre at the origin, then the outward normal points to the origin.) If a_1, a_2, or a_1, a_2, a_3 in the three-dimensional case, are direction cosines of the outward normal, then

$$\frac{\partial u}{\partial n} = a_1 \frac{\partial u}{\partial x} + a_2 \frac{\partial u}{\partial y}, \quad \text{or} \quad \frac{\partial u}{\partial n} = a_1 \frac{\partial u}{\partial x} + a_2 \frac{\partial u}{\partial y} + a_3 \frac{\partial u}{\partial z}, \tag{12}$$

respectively. It is naturally possible to define an outward normal also in the n-dimensional case.

Theorem 2 (*The Maximum Principle for Harmonic Functions*). *Let the function u be continuous in a bounded closed region $\overline{\Omega}$ and harmonic in Ω. Write M, or m for the maximum and minimum of the function u on the boundary of Ω, respectively. Then the inequalities $m \leqq u \leqq M$ hold everywhere in Ω, i.e. u attains its maximum and minimum in $\overline{\Omega}$ on the boundary of this region. If u is not constant in $\overline{\Omega}$, then even the strict inequality $m < u < M$ holds everywhere in Ω.*

The same is true for the two-dimensional case if Ω is the exterior of a closed curve provided that u is bounded in Ω. If Ω is the exterior of a closed surface in the space and if u is harmonic in Ω, vanishes at infinity (Definition 5) *and $|u| \leqq M$ holds on the boundary of Ω, then $|u| \leqq M$ holds everywhere in Ω.*

REMARK 7 (*Uniqueness of the Solution of the Dirichlet and Neumann Problems*). In this remark, when we are considering an exterior problem (either the Dirichlet or the Neumann problem) we make certain assumptions regarding the behaviour of the function u at infinity. Namely, if a plane problem is under consideration, we assume that u is bounded, and if a space problem is considered, we assume that u vanishes at infinity.

Uniqueness of the solution of both the interior and the exterior Dirichlet problem for equations (3), (4), (6), (7) is guaranteed as far as the exterior problem is subject to the above conditions.

Under the same conditions, these problems are also *well-posed problems* (that is, the solution depens continuously on the boundary conditions): If the absolute value of the change of the given boundary condition is less than ε, then the same is true for the solution of the given problem in Ω (thus, if the boundary conditions have been measured with a small error, then the error of the solution is also small).

If the function u is a solution of the Neumann problem, then $u+C$ (where C is an arbitrary constant) is also a solution of the same problem. To be able to guarantee uniqueness of the solution, we have to impose a further condition. Usually it is required that the function u takes a prescribed value at a given point of the region. In the case of an exterior problem, nothing is imposed on the function u except the condition that u vanishes at infinity. Then uniqueness of the solution of the Neumann problem (for both the interior and exterior case) for equations (3), (4), (6), (7) is ensured.

REMARK 8 (*Existence of Solution of the Dirichlet and Neumann Problems*). (As in Remark 7, when solving exterior problems we consider only solutions that are, respectively, bounded in Ω or vanishing at infinity, according to whether a plane or a space problem is in question.) The solution of the *interior* Dirichlet problem for equations (3), (4), (6), (7) always exists, if the boundary of the region Ω is

smooth enough, i.e. if it has a continuously changing tangent everywhere (or a tangent plane, if the three-dimensional problem is considered; see, however, Remark 23) and if the function f has continuous derivatives of the first order in $\overline{\Omega}$ when Poisson's problem is considered. (For a generalization or a different formulation of the considered problems see §§ 18.8 and 18.9.)

The solution of the *exterior* Dirichlet problem for the equations considered may be reduced to the solution of the interior problem as follows: Let us consider first the plane problem and let Ω be the region exterior to the curve k. Let the origin lie in

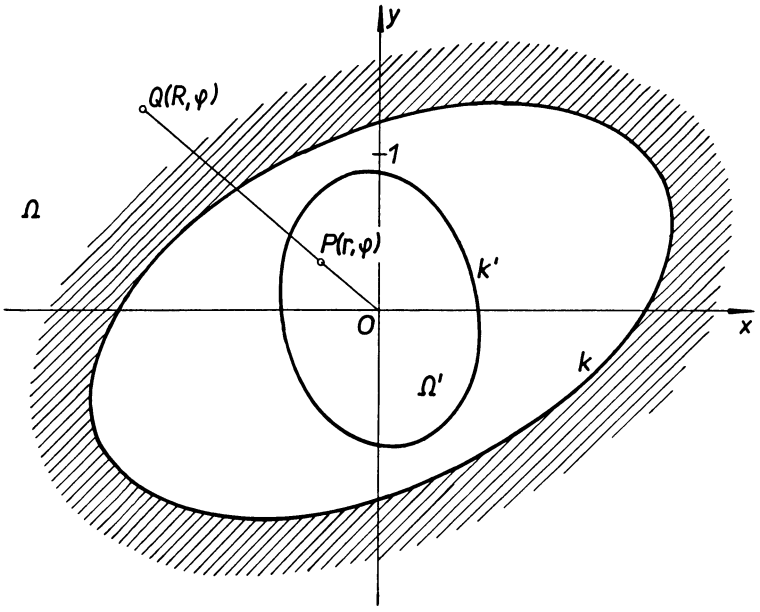


Fig. 18.3.

the interior of that curve. (Otherwise a translation of the coordinate system should be carried out.) Transformation to polar coordinates R, φ gives (see Example 12.11.3)

$$\Delta u = \frac{\partial^2 u}{\partial R^2} + \frac{1}{R}\frac{\partial u}{\partial R} + \frac{1}{R^2}\frac{\partial^2 u}{\partial \varphi^2} = \sigma(R, \varphi). \tag{13}$$

The transformation $r = 1/R$ (Fig. 18.3, where the point $P(r, \varphi)$ corresponds to the point $Q(R, \varphi)$) carries the curve k into a curve k' and its exterior Ω into the interior Ω' of the curve k'. It may be easily proved that the function

$$u(R, \varphi) = u\left(\frac{1}{r}, \varphi\right) = v(r, \varphi) \tag{14}$$

satisfies the equation

$$\Delta v \equiv \frac{\partial^2 v}{\partial r^2} + \frac{1}{r}\frac{\partial v}{\partial r} + \frac{1}{r^2}\frac{\partial^2 v}{\partial \varphi^2} = \frac{1}{r^4}\sigma\left(\frac{1}{r}, \varphi\right) \tag{15}$$

in Ω'. In particular, if $\sigma \equiv 0$ in Ω, so that u is harmonic in Ω, then v is harmonic in Ω'. It follows readily from equation (15) and from what has been said at the beginning of this remark that: If the function $\sigma(R, \varphi)$ tends to zero rapidly enough as $R \to \infty$ so that $r^{-4} \cdot \sigma(r^{-1}, \varphi)$ is continuous together with its first derivatives in Ω (more precisely: if $r^{-4} \cdot \sigma(r^{-1}, \varphi)$ has a removable singularity at $r = 0$), then existence of the solution of the exterior Dirichlet problem is ensured also for Poisson's (and, thus, also for Laplace's) equation. The boundary function g prescribed on the boundary k is transformed into a boundary function G on k' in such a way that G assumes, at every point $M \in k'$, the same value as g does at the corresponding point $N \in k$ (with the same value of φ). Hence, having determined the function $v(r, \varphi)$, i.e. the solution of the interior Dirichlet problem for equation (15) under the boundary condition $v = G$, equation (14) directly yields the desired solution $u(R, \varphi)$.

The treatment of the three-dimensional case is similar. We make use of spherical coordinates R, ϑ, φ, carry out the transformation $r = 1/R$ and define the function v by the relation

$$u(R, \vartheta, \varphi) = u\Big(\frac{1}{r}, \vartheta, \varphi\Big) = rv(r, \vartheta, \varphi). \tag{16}$$

The function v satisfies Poisson's equation

$$\Delta v \equiv \frac{\partial^2 v}{\partial r^2} + \frac{2}{r}\frac{\partial v}{\partial r} + \frac{1}{r^2 \sin^2 \vartheta}\frac{\partial^2 v}{\partial \varphi^2} + \frac{1}{r^2 \sin \vartheta}\frac{\partial}{\partial \vartheta}\left[\frac{\partial v}{\partial \vartheta}\sin \vartheta\right] = \\ = \frac{1}{r^5}\sigma\Big(\frac{1}{r}, \vartheta, \varphi\Big). \tag{17}$$

Again, we obtain the appropriate function G on the boundary of the region Ω' easily from the function g by defining $G = Rg$ at corresponding boundary points (i.e. with the same values of ϑ andφ). Having solved the interior Dirichlet problem for Poisson's equation (17), $\Delta v = r^{-5}\sigma(r^{-1}, \vartheta, \varphi)$ with boundary condition $v = G$, we obtain the desired solution by (16):

$$u(R, \vartheta, \varphi) = \frac{1}{R}v\Big(\frac{1}{R}, \vartheta, \varphi\Big).$$

The existence of the solution of the Neumann problem for the Laplace equation (3) *or* (4): Let the given region have a smooth boundary. (For the exterior problem see the supplementary conditions mentioned at the beginning of this Remark.) A necessary and sufficient condition for existence of solution is that the integral of the boundary function h over the boundary S of Ω be equal to zero, i.e.

$$\int_S h(Q)\,\mathrm{d}S = 0 \tag{18}$$

(Q being a variable point of the boundary.) For plane problems the boundary function h is usually given as a function of the arc length, so that (18) reads

$$\int_0^l h(s)\,\mathrm{d}s = 0, \tag{19}$$

where l stands for the length of the boundary; in the three-dimensional case, (18) represents a surface integral, see Definition 14.8.2. In the case of the exterior problem in three-dimensional space, condition (18) is omitted.

Even if the function f is smooth enough, the solution of the Neumann problem for *Poisson's* equation need not exist. For instance, in the case of the Neumann interior problem in three-dimensional space, a necessary and sufficient condition for existence of a solution is

$$\iiint_\Omega f(x, y, z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z - \iint_S h(Q)\,\mathrm{d}S = 0. \tag{20}$$

REMARK 9 (*Properties of Harmonic Functions*). The most important property has been formulated in Theorem 2 (the Maximum Principle). Equations (15) and (17) (for $\sigma = 0$) express another property: If $u(x, y)$ is harmonic in a two--dimensional region, then the inversion $r = 1/R$ carries this function again into a harmonic function. In the three-dimensional case the situation is rather different, for, if $u(x, y, z)$ is harmonic, then to obtain a harmonic function v by the inversion $r = 1/R$, the transformed function must be divided by r (see(16)).

Further properties of harmonic functions:

1. (*Mean Value Theorem*). If a function is harmonic in an n-dimensional sphere K and continuous in the closed sphere $\overline{K}$, then its value, at the centre of this sphere, is equal to its average value over the boundary of the sphere. In particular, for $n = 2$ (when K is a circle with centre at the point (x_0, y_0) and radius R) we have

$$u(x_0, y_0) = \frac{1}{2\pi R}\int_0^{2\pi} u(x_0 + R\cos\varphi, y_0 + R\sin\varphi)R\,\mathrm{d}\varphi$$

(see also Example 3 below).

2. (*Converse of the Mean Value Theorem.*) Let u be continuous in Ω and such that its value at the centre of an arbitrary n-dimensional sphere $\overline{K} \in \Omega$ is equal to the mean value over the boundary of the sphere. Then u is harmonic in Ω.

3. (*The First Harnack Theorem.*) If a sequence of functions u_n, each of which is harmonic inside a bounded region Ω and continuous in $\overline{\Omega}$, converges uniformly on the boundary of that region, then the sequence u_n converges uniformly in the entire region Ω and the limiting function is harmonic in Ω.

4. (*The Second Harnack Theorem.*) If a series $\sum_{n=1}^{\infty} u_n$ of functions u_n, each of which is harmonic and non-negative in a region Ω, converges at an interior point of that region, then it converges everywhere in Ω and the limiting function is harmonic in Ω. The convergence is uniform in any closed bounded part of the region Ω.

5. (*Theorem on a Removable Singularity.*) If a function u is harmonic and bounded in a neighbourhood of a point P, with the exception of the point P, then the function u may be defined at the point P in such a way that u will be harmonic in the entire neigbourhood of the point P.

6. A function harmonic and bounded outside an n-dimensional sphere has a finite limit at infinity.

7. (*Liouville's Theorem.*) A function, harmonic and bounded in the entire n-dimensional space, is a constant. (Hence, if a harmonic function is not constant in the entire n-dimensional space, then it cannot be bounded.)

8. A function harmonic in the region Ω is *analytic* in that region, i.e. it can be expanded into a power series (in n variables) in the neigbourhood of any point of the region Ω.

9. A harmonic function bounded in a circle K is angular extensible almost everywhere on the boundary. This means that the following assertion is true for every point P of the boundary with the possible exception of points constituting a set of measure zero: If a sequence of points (x_n, y_n) converges to the point P and if all these points (x_n, y_n) lie in an angle $\alpha < 180°$ the arms of which lie, in a certain

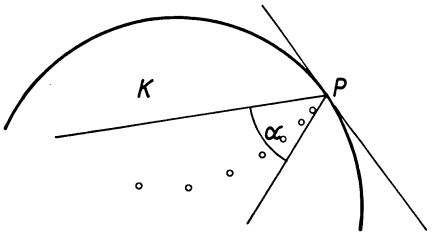


Fig. 18.4.

neighbourhood of the point P, in the circle under consideration (see Fig. 18.4), then there exists a finite limit $\lim_{n\to\infty} u(x_n, y_n)$ (which is the same for every sequence with the above mentioned properties).

Theorem 3. *Let two points*

$$P(x_1, x_2, \ldots, x_n), \quad Q(\xi_1, \xi_2, \ldots, \xi_n)$$

be given in n-dimensional space. Let us write

$$r = \sqrt{[(x_1 - \xi_1)^2 + (x_2 - \xi_2)^2 + \ldots + (x_n - \xi_n)^2]} \, .$$

Then, if Q is regarded as fixed, the functions

$$\frac{1}{r^{n-2}} \quad \text{for} \quad n > 2, \quad \ln \frac{1}{r} \quad \text{for} \quad n = 2 \tag{21}$$

constitute (as functions of the variables $x_1, x_2, \ldots, x_n$) solutions of Laplace's equation in the entire space, provided $P \neq Q$. The same holds for these functions considered as functions of $\xi_1, \xi_2, \ldots, \xi_n$ with P fixed , provided $Q \neq P$.

Example 1. For $n = 3$ and $P \neq Q$ the function $u = 1/r$ is a solution of the equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0.$$

Definition 8. The function

$$\gamma(r) = \frac{\Gamma\left(\frac{n}{2}\right)}{2(n-2)(\sqrt{\pi})^n} \frac{1}{r^{n-2}} \quad \text{for} \quad n > 2$$

or

$$\gamma(r) = \frac{1}{2\pi} \ln \frac{1}{r} \quad \text{for} \quad n = 2$$

is called the *fundamental solution of the Laplace equation in n-dimensional space* (Γ denoting the gamma function, § 13.11).

Example 2. The fundamental solution in three-dimensional space is thus

$$\gamma(r) = \frac{1}{4\pi} \frac{1}{r}. \tag{22}$$

Definition 9. Let us consider the interior Dirichlet problem (Definition 6) for Laplace's or Poisson's equation, respectively. The function

$$G(P, Q) = \gamma(r) + v(P, Q)$$

is called the *Green function* for the problem considered provided that $\gamma(r)$ is the fundamental solution of the Laplace equation (see Definition 8) and that, moreover, $v(P, Q)$ is (for a fixed Q) a harmonic function of the point P in the entire region Ω in question (including the point Q) and $\gamma(r) + v(P, Q)$ vanishes on the boundary of the region Ω.

REMARK 10. The Green function for the problem mentioned is therefore a harmonic function of the point P in the entire region Ω with the exception of the point Q. At this point it has a singularity given by the singularity of the fundamental solution (for example, in three-dimensional space by the singularity of the function (22)). Moreover, the Green function is zero if P is a point on the boundary.

Theorem 4. *The Green function is a symmetric function of the points P and Q, i.e. $G(P, Q) = G(Q, P)$.*

REMARK 11. To ensure existence of the Green function in a plane or in three-dimensional space it is sufficient to assume that the boundary of the region Ω is smooth or piecewise smooth (see also Remark 23).

Theorem 5. *Let the Dirichlet problem be given for the Poisson equation with a continuously differentiable right-hand side $f(Q)$ (or for the Laplace equation, if $f = 0$) and let the region Ω under consideration have a piecewise smooth boundary S. Let the boundary function $g(S)$ be continuous on S. If $G(P, Q)$ is the Green function for this problem, then the solution is given by the formula*

$$u(P) = -\int_{\Omega} G(P, Q) f(Q)\,\mathrm{d}Q - \int_{S} g(S)\frac{\partial G(P, Q)}{\partial n}\,\mathrm{d}S, \tag{23}$$

where $\partial G/\partial n$ denotes the outward normal derivative (Remark 6) *of the function G (taken as a function of Q with P fixed).*

REMARK 12. In the particular case of a plane problem, let the boundary function on the boundary curve k be given as a function of arc length s, i.e. $g = g(s)$. Let us denote by $a_1(s)$, $a_2(s)$ the direction cosines of the outward normal at the point (ξ, η) corresponding to the parameter s. Then (see equation (12)) formula (23) takes the form

$$u(x, y) = -\iint_{\Omega} G(x, y, \xi, \eta) f(\xi, \eta)\,\mathrm{d}\xi\,\mathrm{d}\eta -$$
$$-\int_{k} g(s)\left[a_1(s)\frac{\partial G(x, y, \xi, \eta)}{\partial \xi} + a_2(s)\frac{\partial G(x, y, \xi, \eta)}{\partial \eta}\right]\mathrm{d}s. \tag{24}$$

REMARK 13. To find explicitly the Green function for a given region is in general a difficult task. We mention some particular cases:

REMARK 14. Green function for the interior of the circle Ω with radius R and centre at the origin of the coordinate system is of the form

$$G(x, y, \xi, \eta) = \frac{1}{2\pi}\ln\frac{1}{r} - \frac{1}{2\pi}\ln\frac{R}{r_1 r_2}, \tag{25}$$

where

$$r = \sqrt{\left[(x-\xi)^2 + (y-\eta)^2\right]}\,, \quad r_1 = \sqrt{(\xi^2 + \eta^2)}\,,$$
$$r_2 = \sqrt{\left[\left(x - \frac{R^2}{r_1^2}\xi\right)^2 + \left(y - \frac{R^2}{r_1^2}\eta\right)^2\right]}\,.$$

REMARK 15. For the interior of a sphere Ω with radius R and centre at the origin we have

$$G(x, y, z, \xi, \eta, \zeta) = \frac{1}{4\pi r} - \frac{1}{4\pi}\frac{R}{r_1 r_2}, \tag{26}$$

where

$$r = \sqrt{[(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2]}\,, \quad r_1 = \sqrt{(\xi^2 + \eta^2 + \zeta^2)}\,,$$
$$r_2 = \sqrt{\left[\left(x - \frac{R^2}{r_1^2}\xi\right)^2 + \left(y - \frac{R^2}{r_1^2}\eta\right)^2 + \left(z - \frac{R^2}{r_1^2}\zeta\right)^2\right]}\,.$$

REMARK 16. For $f = 0$ we obtain from (23), (25) and (26) the solution of the Dirichlet problem for the Laplace equation in the interior of a circle or a sphere in the form of the so-called *Poisson integral*:

$$u(x, y) = \frac{1}{2\pi R}\int_0^{2\pi R} \frac{R^2 - r^2}{R^2 + r^2 - 2Rr\cos\varphi}\, g(s)\,\mathrm{d}s, \tag{27}$$

or

$$u(x, y, z) = \frac{1}{4\pi R^2}\iint_S \frac{R^2 - r^2}{(R^2 + r^2 - 2Rr\cos\varphi)^{3/2}}\, g(Q)\,\mathrm{d}S, \tag{28}$$

R standing for the radius of the circle or of the sphere (with centre at the origin), and

$$r = \sqrt{(x^2 + y^2)}\,, \quad \text{or} \quad r = \sqrt{(x^2 + y^2 + z^2)}\,, \tag{29}$$

respectively, φ being the angle between the radii drawn to the points (x, y) and (ξ, η), or (x, y, z) and (ξ, η, ζ), respectively. The point (x, y) or (x, y, z) is a fixed point of the circle (or sphere) considered, the point (ξ, η) (or (ξ, η, ζ)) traces out the circumference of the circle, or the boundary of the sphere, respectively. The function (27) (or (28)) is a solution of the interior Dirichlet problem for an arbitrary continuous g. Actual evaluation of the integral may be difficult and is usually carried out approximately.

REMARK 17. We shall now investigate the so-called *potential of a single layer and potential of a double layer* in a plane and in three-dimensional space. We shall assume, without repeating it explicitly, that the curves or surfaces in question are *simple* and *smooth* (see, however, Remark 22). Moreover, the curves will be required to have a continuous curvature, and the surfaces will be assumed to be of so-called *Liapunov type*. We do not attempt to give the precise definition of such surfaces here (the reader is referred e.g. to [438]); all smooth surfaces we meet in practice are of Liapunov type. The curve k or the surface S is supposed to be closed even though this assumption is not necessary for the definition of potentials. The derivative $\partial u/\partial n$ will always mean the derivative in the direction of the *outward*

normal to the closed curve or surface under consideration. The fixed point (x, y) or (x, y, z) at which the potential is evaluated will be denoted by P, the variable point of integration which traces out the curve k or surface S will be denoted by A; $\overrightarrow{AP}$ is the vector with initial point A and the end point P, and $\boldsymbol{n}_A$ and $\boldsymbol{n}_P$ the vectors of the outward normals at the points A, P, if $P \in k$, or $P \in S$, respectively. (See Fig. 18.5 below.) The given functions f_1, F_1, or f_2, F_2 (the so-called *densities*), are assumed to be continuous on k or on S, respectively.

Definition 10. The integrals

$$v(P) = \int_k f_1(A) \ln \frac{1}{r} \, \mathrm{d}s \tag{30}$$

and

$$V(P) = \iint_S F_1(A) \frac{1}{r} \, \mathrm{d}S \tag{31}$$

are called the *potentials of a single layer in the plane* and *in the space*, respectively.

The integrals

$$w(P) = \int_k f_2(A) \frac{\cos(\boldsymbol{n}_A, \overrightarrow{AP})}{r} \, \mathrm{d}s \tag{32}$$

and

$$W(P) = -\iint_S F_2(A) \frac{\cos(\boldsymbol{n}_A, \overrightarrow{AP})}{r^2} \, \mathrm{d}S \tag{33}$$

are called the *potentials of a double layer in the plane* and *in the space*, respectively (for notation and assumptions see Remark 17); r stands for the distance between the points A, P.

Theorem 6. *The functions* (30) – (33) *are harmonic (as functions of P) in the interior as well as in the exterior of the curve k and the surface S, respectively.*

Theorem 7. *The integrals* (30) *and* (31) *are convergent for $P \in k$ and $P \in S$, respectively. The function v and V are continuous functions of the point P in the entire plane and in the entire space, respectively (including the curve k and the surface S).*

Theorem 8. *The integrals* (32), (33) *converge for $P \in k$ and $P \in S$; however, the functions w, W have a jump on k and on S, respectively. If $P_0 \in k$, or $P_0 \in S$, then the function w, or W, is continuously extensible to the point P_0 from the interior as well as from the exterior of the curve k, or of the surface S, respectively. Let us write w_e, or W_e, for the value of the continuous extension from the exterior and w_i, or W_i, for the continuous extension from the interior of the curve k, or of the surface S, respectively. Further, let us denote by w_0, or W_0, the value at $P = P_0$ of the integral* (32), *or* (33), *respectively. Then the following relations hold at P_0:*

$$w_e = w_0 + \pi f_2(P_0), \qquad w_i = w_0 - \pi f_2(P_0), \tag{34}$$

$$W_e = W_0 - 2\pi F_2(P_0), \quad W_i = W_0 + 2\pi F_2(P_0). \tag{35}$$

Theorem 9. *Let $\boldsymbol{n}_{P_0}$ be the vector of the outward normal at $P_0 \in k$, or $P_0 \in S$. The functions v and V, given by relations* (30) *and* (31), *have derivatives in the direction of $\boldsymbol{n}_{P_0}$ from the exterior as well as from the interior of the curve k, or of the surface S, respectively. If the derivatives from the interior are denoted by $(\partial v/\partial n)_i$, $(\partial V/\partial n)_i$ and from the exterior by $(\partial v/\partial n)_e$, $(\partial V/\partial n)_e$, then the following relations hold at P_0:*

$$\left(\frac{\partial v}{\partial n}\right)_e = \left(\frac{\partial v}{\partial n}\right)_0 - \pi f_1(P_0), \qquad \left(\frac{\partial v}{\partial n}\right)_i = \left(\frac{\partial v}{\partial n}\right)_0 + \pi f_1(P_0), \tag{36}$$

$$\left(\frac{\partial V}{\partial n}\right)_e = \left(\frac{\partial V}{\partial n}\right)_0 - 2\pi F_1(P_0), \quad \left(\frac{\partial V}{\partial n}\right)_i = \left(\frac{\partial V}{\partial n}\right)_0 + 2\pi F_1(P_0), \tag{37}$$

where

$$\left(\frac{\partial v}{\partial n}\right)_0 = -\int_k f_1(A)\frac{\cos(\boldsymbol{n}_{P_0}, \overrightarrow{AP_0})}{r}\,\mathrm{d}s, \quad \left(\frac{\partial V}{\partial n}\right)_0 = -\iint_S F_1(A)\frac{\cos(\boldsymbol{n}_{P_0}, \overrightarrow{AP_0})}{r^2}\,\mathrm{d}S. \tag{38}$$

REMARK 18 (*Solution of the Dirichlet and the Neumann Problems for the Laplace Equation by making Use of Potentials; Reduction to Integral Equations*). The interior of the curve k (of the surface S) will be denoted by Ω, the exterior by Ω'. The Dirichlet problem involves finding a function $u(x, y)$, or $U(x, y, z)$ in the three-dimensional case, harmonic in Ω or in Ω', respectively, which assumes on the boundary the prescribed values g (or G in the three-dimensional case). When solving the Neumann problem, the values h or H of the outward normal derivative with respect to the region considered (see Definition 7 and Remark 6) are prescribed. In the case of exterior problems the function u should be bounded, while the function U should vanish at infinity.

REMARK 19. Notice that in the above equations as well as in the following equations the normal $\boldsymbol{n}_A$, or $\boldsymbol{n}_{P_0}$ points into the exterior of the curve, or of the surface in question, respectively, so that in the case of the exterior Neumann problem it is *not* an outward normal with respect to Ω' in the sense of Remark 6.

REMARK 20.

A. THE INTERIOR DIRICHLET PROBLEM. The functions u, U are assumed to be of the form (32), (33) ($P \in \Omega$), respectively, where f_2 and F_2 are the functions to be found. Making use of the second equations in (34) and (35), we obtain from the condition $u(P) \to g(P_0)$, or $U(P) \to G(P_0)$ ($P_0 \in k$, or $P_0 \in S$) the following integral equations for f_2, F_2, respectively:

$$f_2(P_0) - \frac{1}{\pi}\int_k \frac{\cos(\boldsymbol{n}_A, \overrightarrow{AP_0})}{r} f_2(A)\,\mathrm{d}s = -\frac{g(P_0)}{\pi} \tag{39}$$

and

$$F_2(P_0) - \frac{1}{2\pi} \iint_S \frac{\cos(\boldsymbol{n}_A, \overrightarrow{AP_0})}{r^2} F_2(A)\,\mathrm{d}S = \frac{G(P_0)}{2\pi}. \tag{40}$$

B. THE EXTERIOR DIRICHLET PROBLEM. The functions u and U are again assumed in the form (32), (33) with unknown functions f_2, F_2, respectively. Now, of course, we have $P \in \Omega'$. Making use of the first of equations (34), (35), we obtain for the unknown functions f_2 and F_2 the equations

$$f_2(P_0) + \frac{1}{\pi} \int_k \frac{\cos(\boldsymbol{n}_A, \overrightarrow{AP_0})}{r} f_2(A)\,\mathrm{d}s = \frac{g(P_0)}{\pi}, \tag{41}$$

$$F_2(P_0) + \frac{1}{2\pi} \iint_S \frac{\cos(\boldsymbol{n}_A, \overrightarrow{AP_0})}{r^2} F_2(A)\,\mathrm{d}S = -\frac{G(P_0)}{2\pi}. \tag{42}$$

(The vector $\boldsymbol{n}_A$ points into the exterior of the curve k or of the surface S.)

C. THE INTERIOR NEUMANN PROBLEM. The functions u, U are assumed in the form (30), (31). By making use of the second of equations (36) and (37), the conditions $\partial u/\partial n = h(P_0)$ and $\partial u/\partial n = H(P_0)$ (from the interior) yield the following integral equations for the unknown functions f_1 and F_1:

$$f_1(P_0) - \frac{1}{\pi} \int_k \frac{\cos(\boldsymbol{n}_{P_0}, \overrightarrow{AP_0})}{r} f_1(A)\,\mathrm{d}s = \frac{h(P_0)}{\pi} \tag{43}$$

or

$$F_1(P_0) - \frac{1}{2\pi} \iint_S \frac{\cos(\boldsymbol{n}_{P_0}, \overrightarrow{AP_0})}{r^2} F_1(A)\,\mathrm{d}S = \frac{H(P_0)}{2\pi}. \tag{44}$$

D. THE EXTERIOR NEUMANN PROBLEM. The functions u, U are assumed in the form (30), (31) (where $P \in \Omega'$). The first of equations (36) and (37) imply that

$$f_1(P_0) + \frac{1}{\pi} \int_k \frac{\cos(\boldsymbol{n}_{P_0}, \overrightarrow{AP_0})}{r} f_1(A)\,\mathrm{d}s = \frac{h(P_0)}{\pi}, \tag{45}$$

$$F_1(P_0) + \frac{1}{2\pi} \iint_S \frac{\cos(\boldsymbol{n}_{P_0}, \overrightarrow{AP_0})}{r^2} F_1(A)\,\mathrm{d}S = \frac{H(P_0)}{2\pi}. \tag{46}$$

In (43) – (46), the vector $\boldsymbol{n}_{P_0}$ points to the exterior of the curve k, or of the surface S, respectively (hence, in the case of the exterior problem it points into the interior of the region Ω' under consideration). In the case of the exterior problem, h or H, respectively is the prescribed derivative of u or U in the direction of the exterior normal to Ω', i.e. of the inward normal to the curve h, or to the surface S.

REMARK 21 (*solvability of equations* (39) – (46)). Equations (39), (40), (45), (46) are uniquely solvable for every continuous right-hand side. (In practice, to

find the solution may be difficult and numerical methods are usually employed.) Having solved these equations, the functions u, U are given by the integrals (32), (33) and (30), (31), respectively. Of course, we have $P \in \Omega$ for the interior Dirichlet problem and $P \in \Omega'$ for the exterior Neumann problem. The integral (30) (for the solution of the exterior Neumann problem) defines a function which is bounded in Ω' if and only if $\int_k h(s)\,ds = 0$, consequently (30) is a solution in the sense of Remark 7 if and only if this condition is satisfied.

Equations (43), (44) are solvable if and only if

$$\int_k h(s)\,ds = 0 \quad \text{or} \quad \iint_S H(Q)\,dS = 0,$$

respectively, in agreement with (18), (19).

Equations (41), (42) are not, in general, solvable (for an arbitrary continuous right-hand side). By Remark 8, the exterior Dirichlet problem has a solution, this solution, however, need not be of the form (32) or (33), for we do not require that the solution be, for large r, of the order $1/r$ in the plane or $1/r^2$ in the space. How to overcome this difficulty, see e.g. [369] (for the plane problem) or [438] (for the three-dimensional problem).

Example 3. Consider the Dirichlet problem for the unit circle with centre at the origin and with the boundary condition given by a continuous function $g(s)$. Making use of (32) and (39) we shall evaluate the required harmonic function at the origin. It is readily seen from Fig. 18.5 that:

$$\cos(\boldsymbol{n}_A, \overrightarrow{AP_0}) = \cos\left(\frac{\pi}{2} + \frac{s - s_0}{2}\right) = -\sin\frac{s - s_0}{2}, \quad r = 2\sin\frac{s - s_0}{2}.$$

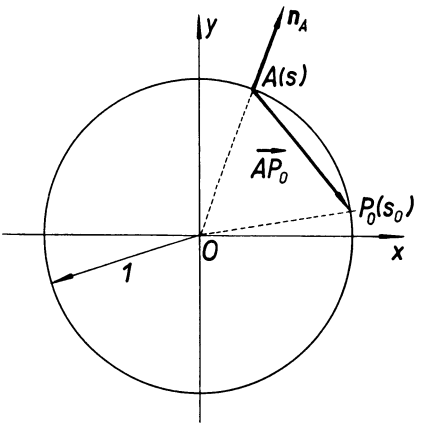


Fig. 18.5.

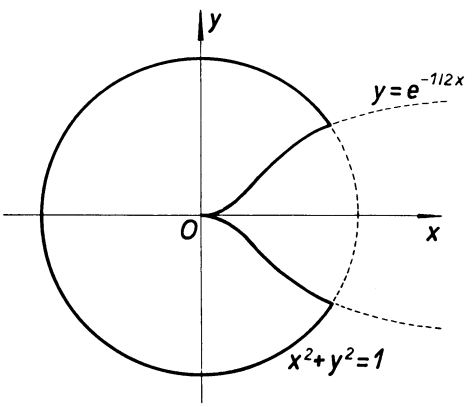


Fig. 18.6.

Equation (39) becomes:

$$f_2(s_0) + \frac{1}{2\pi}\int_0^{2\pi} f_2(s)\,\mathrm{d}s = -\frac{g(s_0)}{\pi}, \tag{47}$$

and its solution is of the form (see § 19.2)

$$f_2(s_0) = -\frac{g(s_0)}{\pi} + k,$$

where k is a constant. Substituting this in (47), we get

$$-\frac{g(s_0)}{\pi} + k - \frac{1}{2\pi^2}\int_0^{2\pi} g(s)\,\mathrm{d}s + \frac{1}{2\pi}\int_0^{2\pi} k\,\mathrm{d}s = -\frac{g(s_0)}{\pi}$$

whence

$$k = \frac{1}{4\pi^2}\int_0^{2\pi} g(s)\,\mathrm{d}s.$$

If we substitute the expression obtained for $f_2(s)$ into (32), where P denotes the origin, we get (because $r = 1$, $\cos(\boldsymbol{n}_A, \overrightarrow{AP}) = -1$)

$$w(0,0) = \int_0^{2\pi}\left(-\frac{g(s)}{\pi} + k\right).(-1)\,\mathrm{d}s = \frac{1}{\pi}\int_0^{2\pi} g(s)\,\mathrm{d}s - 2\pi k = \frac{1}{2\pi}\int_0^{2\pi} g(s)\,\mathrm{d}s.$$

(The above example is, of course, only an illustrative one; see Mean Value Theorem, Remark 9, point 1.)

REMARK 22. The assumption of smoothness of the boundary in the theory of potentials is rather restrictive for applications. In recent years, J. Král and his school have produced a modern theory which makes it possible to apply potentials even in the case of nonsmooth boundaries. (J. Král, Berlin, Springer 1980.)

Let us note that the theory of potentials is applicable also to other equations than to that of Laplace.

REMARK 23. Existence of solution of problems investigated in this paragraph can be proved under more general conditions than are those given in Remark 8. For instance, the Dirichlet problem for the Laplace equation in the plane is (uniquely) solvable (for a given continuous function on the coundary) if the boundary is a Jordan curve (Remark 14.1.3) only.

In three-dimensional space we have the following situation: Let S be the surface which is formed by revolution about the x-axis of the curve shown in Fig. 18.6. The point O is a cusp of this curve. In cases of this type, we encounter the following fact: There exists a function G continuous on S such that the interior Dirichlet problem for the Laplace equation has no solution in the above-defined classical

sense (that is, there exists no function U, which is harmonic in the interior V of S, continuous in $V + S$ and assumes the values G on S). However, there does exist a bounded function U_1, harmonic in V, continuous in $V + S$ with the exception of the point O and assuming the prescribed values G on S (except at the point O). It can be shown that such a situation cannot occur, if the point O may be taken as the vertex of a cone K having a non-zero solid angle at the point O and such that K has only the point O in common with $V + S$. (This was not possible in the foregoing example, since the cross-section passing through the x-axis had a cusp at O.) In this case, we say that the *exterior cone condition* is satisfied at the point O.

A region V is said to be *regular with respect to the interior Dirichlet problem for the Laplace equation* if the interior Dirichlet problem is solvable for every continuous function prescribed on the boundary of V. In space, every region whose boundary consists entirely of points which satisfy the above-mentioned exterior cone condition is regular with respect to the interior Dirichlet problem; in particular, a region with a smooth boundary is regular. In the plane, a Jordan region is a regular region. In every regular region, there exists a Green function for the interior Dirichlet problem.

Many results which we have presented for the Laplace and Poisson equation here may be generalized to other elliptic equations.

In applications, the prescribed boundary function is frequently discontinuous at some points of the boundary. In this case we must define what is meant by a solution: In applications, when solving the Dirichlet problem, a "classical" solution is, generally, understood to be a bounded function which satisfies the given differential equation and is continuously extensible to the given boundary function at all points where it is continuous.

REMARK 24. A reader, who is interested in being briefly acquainted with fundamental ideas of functional-analytical methods of solving sufficiently general problems in elliptic equations, is referred to § 18.8 (theorem on minimum of functional of energy, generalized solutions) and § 18.9 (the Lax-Milgram theorem, weak solutions, nonlinear problems).

REMARK 25 (*Methods of Solution*). As mentioned in introduction to this chapter, solution in a closed form can be obtained in very special cases only. In this section, the solution of the Poisson equation in the whole space (Theorem 1) and of the Dirichlet problem for the Laplace equation for a circle and for a sphere have been given (Remark 16). See also Chap. 26 concerning the Fourier method which makes it possible to get the solution in the form of an infinite series in some cases. See also Remarks 20 and 21 on the method of potentials.

In a majority of cases the solution is to be found approximately. For elliptic problems the most suitable methods are variational methods (Chap. 24), especially the finite element method, from among classical methods the method of finite differences.

18.5. Hyperbolic Equations. Wave Equation, the Cauchy Problem, the Mixed Problem. Generalized Solutions of Hyperbolic Equations

REMARK 1 (*informative*). We shall deal only with problems relating to the so-called wave equation, first with the Cauchy problem and then with the mixed problem (involving boundary conditions). When investigating hyperbolic equations, we encounter the following phenomenon which does not occur in the case of elliptic and parabolic equations with sufficiently smooth coefficients: If we understand, as usual, by a solution of the given problem a function which satisfies the given differential equation in the region considered and the prescribed initial and boundary conditions, then a solution need not exist (Remark 7) if the initial and boundary functions are not sufficiently smooth. Consequently, we introduce so-called *generalized solutions* (Definition 3).

Definition 1. The equation of the form

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \ldots + \frac{\partial^2 u}{\partial x_n^2} \tag{1}$$

is called the *wave equation.* In paricular, for $n = 1$ we obtain the so-called *equation of the vibrating string*

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \tag{2}$$

for $n = 2$ the equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \tag{3}$$

and for $n = 3$ the equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}. \tag{4}$$

By a (*classical*) *solution of equation* (1) *in a region* Ω we understand such a function $u(t, x_1, x_2, \ldots, x_n)$ which has, in Ω, continuous derivatives of the second order with respect to all variables and satisfies , in Ω, equation (1).

Definition 2. The *Cauchy problem* for equation (1) is to find, for $t > 0$, such a solution u of equation (1) that u and $\partial u/\partial t$ are continuously extensible for $t \to 0$ and the relations

$$u(0, x_1, \ldots, x_n) = \varphi_0(x_1, \ldots, x_n), \tag{5}$$

$$\frac{\partial u}{\partial t}(0, x_1, \ldots, x_n) = \varphi_1(x_1, \ldots, x_n) \tag{6}$$

hold.

REMARK 2. The functions (5), (6) are usually prescribed in the entire hyperplane $x_1, x_2, \ldots, x_n$ (for $n = 1$, for instance, on the entire x-axis). However, they may be prescribed only on a part of that hyperplane.

REMARK 3. The Cauchy problem for the equation

$$\frac{1}{a^2}\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \ldots + \frac{\partial^2 u}{\partial x_n^2} \tag{7}$$

(with conditions (5), (6)) may be reduced to the Cauchy problem for equation (1) by the substitution $at = \tau$. We have

$$\frac{\partial^2 u}{\partial \tau^2} = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \ldots + \frac{\partial^2 u}{\partial x_n^2} \tag{8}$$

with initial conditions for $\tau \to 0$:

$$u(0, x_1, \ldots, x_n) = \varphi_0(x_1, \ldots, x_n), \tag{9}$$

$$\frac{\partial u}{\partial \tau}(0, x_1, \ldots, x_n) = \frac{1}{a}\varphi_1(x_1, \ldots, x_n). \tag{10}$$

Theorem 1 (*Kirchhoff's Formula for $n = 3$*). *If the functions φ_0, φ_1 have continuous partial derivatives of the second order, then the solution of the Cauchy problem* (Definition 2) *for $n = 3$ is*

$$u = \frac{\partial u_{\varphi_0}}{\partial t} + u_{\varphi_1}, \tag{11}$$

where

$$u_\varphi(t, x_1, x_2, x_3) = \frac{1}{4\pi}\iint_{S_t(x_1,x_2,x_3)} \frac{\varphi(\alpha_1, \alpha_2, \alpha_3)}{t}\, \mathrm{d}S_t. \tag{12}$$

Integration is carried out over the surface S_t of the sphere of radius t with centre at the point (x_1, x_2, x_3).

To obtain u_{φ_0} and u_{φ_1} and to be able to use (11), we must, of course, evaluate (12) first for $\varphi = \varphi_0$ and then for $\varphi = \varphi_1$.

REMARK 4. Formula (11) is also applicable for $n = 2$ and $n = 1$. For $n = 2$ (retaining the assumption that φ_0 and φ_1 have continuous derivatives of the second order), the expression

$$u_\varphi(t, x_1, x_2) = \frac{1}{2\pi}\iint_{K_t} \frac{\varphi(\alpha_1, \alpha_2)}{\sqrt{[t^2 - (\alpha_1 - x_1)^2 - (\alpha_2 - x_2)^2]}}\, \mathrm{d}\alpha_1\, \mathrm{d}\alpha_2 \tag{13}$$

is to be substituted in the formula for u (*Poisson's formula*). Here K_t stands for the circle of radius t with centre at the point (x_1, x_2). For $n = 1$ we have

$$u_\varphi(t, x) = \frac{1}{2}\int_{x-t}^{x+t} \varphi(\alpha)\,d\alpha \tag{14}$$

so that

$$u(t, x) = \frac{\varphi_0(x+t) + \varphi_0(x-t)}{2} + \frac{1}{2}\int_{x-t}^{x+t} \varphi_1(\alpha)\,d\alpha \tag{15}$$

(*d'Alembert's formula*).

REMARK 5. For the equation

$$\frac{1}{a^2}\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2},$$

equation (15) together with (8) and (10) implies that

$$u(t, x) = \frac{\varphi_0(x+at) + \varphi_0(x-at)}{2} + \frac{1}{2a}\int_{x-at}^{x+at} \varphi_1(\alpha)\,d\alpha. \tag{16}$$

If it is required to solve the *non-homogeneous* equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \ldots + \frac{\partial^2 u}{\partial x_n^2} + f(x_1, x_2, \ldots, x_n, t)$$

with initial conditions (5), (6), then the following expressions are to be added to (11):

for $n = 3$:

$$\frac{1}{4\pi}\iiint_{L_t} \frac{f(\alpha_1, \alpha_2, \alpha_3, t-\rho)}{\rho}\,d\alpha_1\,d\alpha_2\,d\alpha_3,$$

where L_t is the sphere with centre at the point (x_1, x_2, x_3) and radius t,

$$\rho = \sqrt{[(x_1-\alpha_1)^2 + (x_2-\alpha_2)^2 + (x_3-\alpha_3)^2]}\,;$$

for $n = 2$:

$$\frac{1}{2\pi}\int_0^t d\tau \iint_{L_{t-\tau}} \frac{f(\alpha_1, \alpha_2, \tau)}{\sqrt{[(t-\tau)^2 - \rho^2]}}\,d\alpha_1\,d\alpha_2,$$

where $L_{t-\tau}$ is the circle with centre (x_1, x_2) and radius $t-\tau$,

$$\rho = \sqrt{[(x_1-\alpha_1)^2 + (x_2-\alpha_2)^2]}\,;$$

for $n = 1$:

$$\frac{1}{2}\int_0^t d\tau \int_{x-t+\tau}^{x+t-\tau} f(\alpha, \tau)\,d\alpha.$$

REMARK 6 (*Uniqueness and Well-posed Nature of the Cauchy Problem*). Under the assumptions mentioned, the solutions presented in Theorem 1 and Remark 4 are *unique*. Further: the problem formulated in Definition 2 is *well-posed*, i.e. the solution depends continuously on the initial conditions. For $n = 1$, this fact may be readily seen from (15). In the general case (for n dimensions) the following assertion is valid: Consider the solution in the interval $[0, T]$, T being any (finite) positive number. Denote by the symbol $[n/2]$ the greatest integer which is less or equal to $n/2$. (For $n = 1$, $[n/2] = 0$, for $n = 2$ and $n = 3$ we have $[n/2] = 1$, etc.). Then for an arbitrary $\varepsilon > 0$ there exists a $\delta > 0$ such that the change in u (in absolute value) is less than ε whenever the change in φ_0 and φ_1 and their derivatives up to the order $[n/2]$ is less than δ. In particular, for $n = 1$ it suffices to consider only changes in the functions φ_0 and φ_1, as may be seen directly from (15), of course. For $n = 2$, however, it is not sufficient to consider only φ_0 and φ_1; the derivatives must also be taken into account. It can be shown that $|u|$ may become large even if $|\varphi_0|$ and $|\varphi_1|$ are small since thay may, nevertheless, have large derivatives.

REMARK 7 (*generalized solution*). It is readily seen from (15) that if the functions φ_0 and φ_1 are not smooth enough (for example, if φ_0 does not possess the second

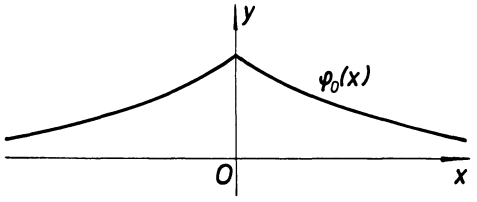


Fig. 18.7.

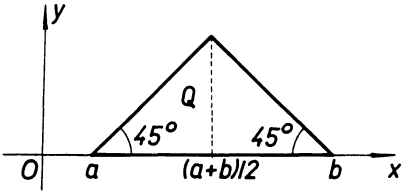


Fig. 18.8.

derivative, see Fig. 18.7, or φ_1 the first derivative), then the function u is not a solution (in the classical sense), since it does not possess derivatives of the second order. To come to a reasonable concept of the solution in such a case, let us proceed in the following way: Let, for example, the functions φ_0 and φ_1 in (15) be merely continuous. As shown above (Remark 6), the change of the function (15) will be arbitrarily small if the functions φ_0 and φ_1 are replaced (with a sufficient accuracy) by functions ψ_0, ψ_1 having two continuous derivatives. The function (15) corresponding to ψ_0 and ψ_1 will then constitute a solution in the usual sense. In this way we arrive at the concept of a generalized solution:

Definition 3. A function u is said to be a *generalized solution* of the given problem (in the given domain Ω) if there exists a sequence $u_1, u_2, \ldots, u_n, \ldots$ of solutions (in the usual sense) of this problem converging, for $n \to \infty$, uniformly to u in Ω.

REMARK 8. It is readily seen that if φ_0 and φ_1 are continuous, then (15) represents a generalized solution of the Cauchy problem for equation (2), for instance

in every rectangle $-a + T \leqq x \leqq a - T$, $0 \leqq t \leqq T$, while u is continuous in this rectangle. It is sufficient to make use of Remark 10 below and in virtue of the Weierstrass Theorem to replace the functions φ_0 and φ_1 by polynomials in the interval $[-a, a]$. (By a suitable choice of T and a, any point of the xt-plane can be included, so that in this sense (15) represents a generalized solution in every bounded part of the plane.)

REMARK 9. Generalized solutions may be introduced in various ways according to the purpose we wish to achieve. (See, for example [438], where existence and uniqueness of a generalized solution is proved for a wide class of problems.)

REMARK 10. It follows from (15) that the values of u in the entire triangle T drawn in Fig. 18.8 depend only on the values of φ_0 and φ_1 in the interval (a, b) and are independent of the values of these functions outside that interval. Similarly, it may be shown that the values of the solution (13) in a right circular cone K having its base P in the plane x_1x_2 and formed by generators which make an angle of $45°$ with the axis of the cone are uniquely determined by the functions φ_0 and φ_1 in P. The sides of the triangle T or the generators of the cone K are obviously characteristics of the equation (2), or (3), respectively. A similar result is valid also for $n > 2$.

On the Cauchy problem posed on a characteristic see § 18.1, especially Theorem 2, Examples 10 and 11.

REMARK 11 (*a mixed problem*). In some problems, the required solution u of the wave equation (1) is not only subject to the initial conditions (5), (6), in which $(x_1, x_2, \ldots, x_n)$ denotes a point in the given region Ω, but must also satisfy certain conditions on the boundary. Such a problem is said to be a *mixed problem*, and some typical examples follow.

Example 1. Find a solution of the equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \tag{17}$$

in the semi-infinite strip $\Omega(0 < x < l, t > 0)$ such that it is continuous and has a continuous derivative with respect to t for $0 \leqq x \leqq l$, $t \geqq 0$ and satisfies the following initial and boundary conditions:

$$u(0, x) = \varphi_0(x), \tag{18}$$

$$\frac{\partial u}{\partial t}(0, x) = \varphi_1(x), \tag{19}$$

$$u(t, 0) = 0, \tag{20}$$

$$u(t, l) = 0, \tag{21}$$

where $\varphi_0(0) = \varphi_0(l) = 0$, $\varphi_1(0) = \varphi_1(l) = 0$. (This is the problem of vibration of a string of length l, fixed at both ends; the initial position of the string $u(0, x)$ being given by the function $\varphi_0(x)$ and the initial velocity $\frac{\partial u}{\partial t}(0, x)$ by the function $\varphi_1(x)$; for the solution see § 26.1.)

Example 2. In a similar way, the mixed problem can be formulated in the case $n > 1$. For example, if $n = 2$, a solution of equation (3) is to be found which is continuous with its derivative with respect to t for $(x_1, x_2) \in \overline{\Omega}$ and $t \geqq 0$ (Ω being the given region in the x_1x_2-plane, with boundary S), satisfies conditions (5), (6) (for $n = 2$), and vanishes for $(x_1, x_2) \in S$, $t > 0$ (i.e. on the lateral surface of the cylinder in the interior of which the solution is to be found).

REMARK 12. Other conditions may be prescribed on the boundary S for $t > 0$; we may, for instance, have $\partial u/\partial n = 0$, or, more generally, $\partial u/\partial n + \sigma u = 0$, where $\partial u/\partial n$ stands for the outward normal derivative and σ is a non-negative constant (more generally a non-negative continuous function), or the corresponding non-homogeneous conditions. The usual requirement is that not only u but also the derivatives of the first order should be continuous for all $t \geqq 0$ and all points of the closed region $\overline{\Omega}$. These assumptions ensure uniqueness of the solution of the given problems and also continuous dependence on initial conditions in the following sense: Two solutions which satisfy the same boundary conditions on S for $t > 0$ (i.e. their difference satisfies $u = 0$ or $\partial u/\partial u = 0$ or $\partial u/\partial n + \sigma u = 0$) are arbitrarily close to each other if the difference of the functions φ_0 and φ_1 and their derivatives up to the order $[\frac{1}{2}n] + 1$ is sufficiently small. Here $[\frac{1}{2}n]$ denotes the greatest integer which is less than or equal to $\frac{1}{2}n$.

REMARK 13 (*Methods of Solution*). A typical mehod of solving mixed problems is the Fourier method. Details are given in Chap. 26, especially in § 26.1, where the problem for equation (17) under conditions (18)–(21) is solved. The Fourier method can also be applied to problems concerning more general equations of hyperbolic type and also in multidimensional cases.

Another typical method of finding solutions of mixed problems, especially useful for time-dependent boundary conditions, is the Laplace transformation (see Chap. 28). (Cf. a similar example for the heat-conduction equation presented in that chapter.)

The reader will find a variety of practical examples solved by Laplace transformation methods, for example in [77].

A typical numerical method is the method of finite differences (Chap. 27).

In recent years, the so-called *method of discretization in time* (the *Rothe method*, the *horizontal method of lines*) turned out to be a rather universal method of solution of sufficiently general hyperbolic problems (of "arbitrary" order in the

space variables), being both an effective theoretical tool (for obtaining sufficiently general existence theorems) and an efficient numerical method. See [390]; see also § 18.10.

On the other hand, the vertical method of lines (the Galerkin method) is frequently used, where space variables are discretized using the finite element method. See § 24.6.

The initial conditions are sometimes prescribed in a different way from that of the Cauchy problem defined in Definition 2. For instance, we may have to find a solution of equation (2) such that it assumes prescribed values on a curve $y = h(x)$, where $h'(x) < 0$,

$$u\Big|_{y=h(x)} = \varphi_0(x), \quad \frac{\partial u}{\partial t}\Big|_{y=h(x)} = \varphi_1(x).$$

For this case, a suitable method of solution has been proposed by Riemann (see [438], for example). A method due to Kirchhoff [438] is suitable for solutions of some multidimensional problems. On some questions concerning hyperbolic systems of equations see, for example, [369].

18.6. Parabolic Equations. The Heat-conduction Equation. The Cauchy Problem. Mixed Boundary Value Problems

REMARK 1. In this paragraph we deal with the *heat-conduction equation* which for $n = 3$, 2, or 1 is of the form

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} + f(x, y, z, t), \tag{1}$$

or

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + f(x, y, t), \tag{2}$$

or

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \tag{3}$$

respectively.

If $f \equiv 0$ (which means physically that in the region considered no sources of heat are present) the equation is said to be *homogeneous.*

The equation

$$\frac{1}{a^2}\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} + g(x, y, z, t)$$

is transformed by the substitution $a^2 t = \tau$ to the form (1) in the same way as in Remark 18.5.3.

Definition 1. The following problem is called the *Cauchy problem* for equation (1): To find a function $u(x, y, z, t)$ which is bounded and continuous at all points of the three-dimensional space xyz and for all $t \geqq 0$, for $t > 0$ satisfies equation (1) and for $t \to 0$ is continuously extensible to a continuous function $\varphi(x, y, z)$,

$$u(x, y, z, 0) = \varphi(x, y, z). \tag{4}$$

REMARK 2. The Cauchy problem for equation (2) (i.e. in the plane E_2) and for equation (3) (the one-dimensional problem) is defined in a similar way.

Theorem 1. *The function*

$$v(x, y, z, t, \xi, \eta, \zeta, \tau) = \frac{1}{8\pi^{3/2}(t-\tau)^{3/2}} e^{-r^2/[4(t-\tau)]}, \tag{5}$$

where

$$r = \sqrt{[(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2]}\,, \tag{6}$$

when regarded as a function of the variables x, y, z, t (for fixed ξ, η, ζ, τ) satisfies equation (1) whenever $t > \tau$. On the other hand, when regarded as a function of ξ, η, ζ, τ it satisfies the equation

$$-\frac{\partial u}{\partial \tau} = \frac{\partial^2 u}{\partial \xi^2} + \frac{\partial^2 u}{\partial \eta^2} + \frac{\partial^2 u}{\partial \zeta^2} \tag{7}$$

for $\tau < t$.

Theorem 2. *For $t > \tau$ the relation*

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} v \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = 1 \tag{8}$$

holds.

REMARK 3. The function (5) is called the *fundamental solution of the homogeneous equation* (1). The physical interpretation of this fundamental solution is, roughly speaking, that it decsribes the temperature field in space at time t due to a unit heat impulse at the point (ξ, η, ζ) at the instant τ.

REMARK 4. The fundamental solutions of the homogeneous equations (2) and (3), namely

$$v(x, y, t, \xi, \eta, \tau) = \frac{1}{4\pi(t-\tau)} e^{-r^2/[4(t-\tau)]} \quad \left(r = \sqrt{[(x-\xi)^2 + (y-\eta)^2]}\right), \tag{9}$$

$$v(x, t, \xi, \tau) = \frac{1}{2\sqrt{(\pi)}\,\sqrt{(t-\tau)}} e^{-(x-\xi)^2/[4(t-\tau)]}, \tag{10}$$

have quite similar properties.

Theorem 3. *If the functions* $f(x, y, z, t)$, $\varphi(x, y, z)$ *and their derivatives of the first order are bounded and continuous in the entire space* xyz *and for all* $t > 0$, *then the solution of the Cauchy problem* (Definition 1) *is*

$$u(x, y, z, t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(\xi, \eta, \zeta)\, v(x, y, z, t, \xi, \eta, \zeta, 0)\, d\xi\, d\eta\, d\zeta + \\ + \int_0^t \left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\xi, \eta, \zeta, \tau)\, v(x, y, z, t, \xi, \eta, \zeta, \tau)\, d\xi\, d\eta\, d\zeta\right] d\tau. \quad (11)$$

REMARK 5. This solution is *unique* (in Definition 1 we require the function u to be *bounded*). In addition, Cauchy's problem is well-posed in the sense that small changes of the functions f and φ cause small changes of the function u.

REMARK 6. It can be shown that if $f \equiv 0$ (i.e. if equation (1) is homogeneous), the function u given by formula (11) is, for $t > 0$, continuously differentiable infinitely many times no matter whether φ has derivatives or not. The heat-conduction equation differs essentially in this property from the wave equation (see Remark 18.5.7).

On the other hand, it should be noted that the Cauchy problem need not possess a solution for $t < 0$.

REMARK 7. The results for equations (2) and (3) are quite similar. The solution of the Cauchy problem is as follows:

$$u(x, y, t) = \frac{1}{4\pi t}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(\xi, \eta)\, e^{-[(x-\xi)^2+(y-\eta)^2]/4t} d\xi\, d\eta + \\ + \frac{1}{4\pi}\int_0^t \frac{1}{t-\tau}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\xi, \eta, \tau)\, e^{-[(x-\xi)^2+(y-\eta)^2]/[4(t-\tau)]} d\xi\, d\eta\right] d\tau, \quad (12)$$

$$u(x, t) = \frac{1}{2\sqrt{(\pi)}\,\sqrt{(t)}}\int_{-\infty}^{\infty} \varphi(\xi)\, e^{-(x-\xi)^2/4t} d\xi + \\ + \frac{1}{2\sqrt{(\pi)}}\int_0^t \frac{1}{\sqrt{(t-\tau)}}\left[\int_{-\infty}^{\infty} f(\xi, \tau)\, e^{-(x-\xi)^2/[4(t-\tau)]} d\xi\right] d\tau. \quad (13)$$

REMARK 8. Very frequently *mixed* (boundary value) problems are encountered. The following problem is typical: Find a solution of equation (3), continuous and bounded for $a \leqq x \leqq b$, $t \geqq 0$, such that on the segment $a \leqq x \leqq b$, $t = 0$ it assumes the values of a prescribed function $\varphi(x)$, i.e.

$$u(x, 0) = \varphi(x), \quad (14)$$

while for $x = a$, $x = b$ it takes the values of other prescribed functions

$$u(a, t) = \psi_1(t), \quad u(b, t) = \psi_2(t) \quad (t > 0). \tag{15}$$

(This is the problem of heat conduction in an insulated bar of length $b - a$ having initial temperature $u(x, 0) = \varphi(x)$ whose end points are kept at the temperatures $\psi_1(t)$ and $\psi_2(t)$, respectively. The inner sources of heat are characterized by the function $f(x, t)$. See Example 28.2.2 and § 26.3.)

Theorem 4 (*The Maximum Principle*). *Let* $f(x, t) \equiv 0$ (*i.e. equation* (3) *is homogeneous*). *Then the solution of the problem mentioned in Remark* 8 *has the following property: For any rectangle* $a \leqq x \leqq b$, $0 \leqq t \leqq T$ (T *being a positive number*) *the solution takes its maximum and minimum values either on the lower base* (*for* $t = 0$), *or on one of its lateral sides* ($x = a$, $x = b$).

REMARK 9. Theorem 4 implies immediately *uniqueness* of the solution. Moreover, the problem is well posed (for the non-homogeneous equation (3) as well) in the following sense: In every rectangle $a \leqq x \leqq b$, $0 \leqq t \leqq T$ the solution u changes only little if the function f and the functions (14) and (15) change only little. If the function f remains unchanged and if the change of the functions (14), (15) is smaller than ε, then the change of u is also smaller than ε.

REMARK 10. In the same way as in Remark 8 the problem may be formulated for equations (1) and (2). In the case of equation (2) a closed region $\overline{\Omega}$ of the plane xy is given instead of the segment $a \leqq x \leqq b$, $t = 0$, in the case of equation (1) a (closed) three-dimensional region is given. Boundary functions corresponding to those of (15) are not only functions of time but also, in general, functions of position on the boundary. Theorem 4 and Remark 9 are again valid.

REMARK 11. It is possible to pose some other problems different from those formulated in Remark 8. For example, it may be prescribed that for $x = b$, $\partial u/\partial n = 0$ should hold (physical interpretation: the end $x = b$ of the bar $a \leqq x \leqq b$ is insulated) or the condition $\partial u/\partial n + \alpha u = 0$ ($\alpha > 0$; Newton's condition of heat transfer) may be prescribed, or the corresponding non-homogeneous conditions are given, etc. For a wide class of such problems existence and uniqueness theorems can be proved (see e.g. [369]). As far as existence and uniqueness of the solution are concerned it does not matter if the boundary conditions are discontinuous at a finite number of points (if, for example, in the problem of Remark 8 the relations $\psi_1(0) = \varphi(a)$ or $\psi_2(0) = \varphi(b)$ do not hold, etc.) provided that we require the solution to be bounded and to satisfy the bundary conditions everywhere except at those points.

Homogeneous equations (1)–(3) may be solved by transformation into integral equations using the so-called *heat potentials* (see e.g. [323]) which are similar to those introduced in Definition 18.4.10.

For generalization of some results see [369].

REMARK 12 (*methods of solution*). The method of the Laplace transform is widely used. The problem of Remark 8 and problems of a similar nature are typical cases in which this method can be used (see Example 28.2.2). A variety of examples may be found, e.g., in [77]. In the case where several space coordinates are involved, the use of Laplace transforms is complicated by the fact that the transformed equation is again a *partial* differential equation.

Another efficient method is the Fourier method. If, for instance, we have $\psi_1 \equiv 0$, $\psi_2 \equiv 0$ and also $f \equiv 0$ (the homogeneous equation) in the problem of Remark 8, then using the Fourier method (§ 26.3), we arrive at the result (here $a = 0$, $b = l$)

$$u(x, t) = \sum_{n=1}^{\infty} a_n \sin \frac{n\pi x}{l} e^{-n^2\pi^2 t/l^2} \tag{16}$$

where

$$a_n = \frac{2}{l} \int_0^l \varphi(x) \sin \frac{n\pi x}{l} \, dx. \tag{17}$$

If $\varphi(x)$ is continuous in $[0, l]$ and $\varphi(0) = 0$, $\varphi(l) = 0$, then (16) is the desired solution. If $\varphi(x)$ is bounded and continuous except at a finite number of points or if $\varphi(0) \neq 0$ or $\varphi(l) \neq 0$, then the condition (14) is satisfied with the exception of those points (Remark 11).

An efficient and frequently used numerical method is the method of finite differences (Chap. 27). For solving more complicated problems, the method of discretization in time (the Rothe method, the horizontal method of lines) is rather universal, from the theoretical point of view (to obtain sufficienty general existence theorems) as well as from the numerical point of view (as an efficient numerical method, in particular in combination with the finite element method). See e.g. [390]. See also § 18.10. On the other hand, the "vertical" method of lines (cf. Remark 18.5.13) is very frequently used (see § 24.6).

18.7. Some Other Problems of Partial Differential Equations. Systems of Equations. Pfaffian Equation. Equations of Higher Order, Biharmonic Equation. Potential Flow, the Navier-Stokes Equations

REMARK 1. Solvability of the so-called *Pfaffian equation* is studied in geometry (Remark 2). The question of solvability of the system of two equations

$$\frac{\partial z}{\partial x} = A(x, y, z), \tag{1}$$

$$\frac{\partial z}{\partial y} = B(x, y, z) \tag{2}$$

for one unknown function $z(x, y)$ is connected with it. Let us assume that the functions A, B have continuous partial derivatives of the first order in a region Ω where the system (1), (2) is investigated. Then it can be shown that in order that this system should be solvable (or, as we say, that equations (1) and (2) should be *consistent*) a necessary condition is:

$$\frac{\partial A}{\partial y} + \frac{\partial A}{\partial z}B = \frac{\partial B}{\partial x} + \frac{\partial B}{\partial z}A. \tag{3}$$

If Ω is simply connected and if (3) is satisfied identically (i.e. for all x, y, z of the region Ω), then there exists a system of surfaces $z(x, y)$ satisfying (1), (2). (One and only one integral surface of the system (1), (2) passes then through each point $(x_0, y_0, z_0) \in \Omega$.)

If (3) is not satisfied identically, it is possible, in general, to express z as a function of x, y, thus $z = \varphi(x, y)$. If the system (1), (2) has a solution, then it is given by this relation. Whether $z = \varphi(x, y)$ is a solution or not should be verified by inspection.

REMARK 2. The equation of the form

$$P(x, y, z)\,\mathrm{d}x + Q(x, y, z)\,\mathrm{d}y + R(x, y, z)\,\mathrm{d}z = 0 \tag{4}$$

is called the *Pfaffian equation.*

Geometric interpretation: To every point $(x_0, y_0, z_0) \in \Omega$ (Ω is the simply connected region in question; P, Q, R and their derivatives of the first order are assumed to be continuous in Ω) there corresponds a vector with components P, Q, R. To solve equation (4) means, geometrically, to find a system of surfaces that are orthogonal to the field of those vectors (in Ω). The necessary and sufficient condition for such a system to exist is that the condition of integrability

$$P\left(\frac{\partial Q}{\partial z} - \frac{\partial R}{\partial y}\right) + Q\left(\frac{\partial R}{\partial x} - \frac{\partial P}{\partial z}\right) + R\left(\frac{\partial P}{\partial y} - \frac{\partial Q}{\partial x}\right) = 0 \tag{5}$$

be satisfied identically in Ω.

If condition (5) is not satisfied, then it can be shown that such a system of surfaces does not exist. However, it is possible to find one-dimensional integral manifolds (curves) of equation (4),

$$y = y(x), \quad z = z(x). \tag{6}$$

One of the functions (6) may be chosen arbitrarily; the other is then determined by solving the ordinary differential equation

$$P(x, y, z) + Q(x, y, z)\frac{\mathrm{d}y}{\mathrm{d}x} + R(x, y, z)\frac{\mathrm{d}z}{\mathrm{d}x} = 0.$$

REMARK 3. Systems of partial differential equations which are encountered in problems of physics and engineering are mostly of a different character. The solution of these systems is often reduced to the solution of a single equation of higher order. A typical example of such systems is the system of equations

$$\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} = 0, \quad \frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_y}{\partial y} = 0, \quad \Delta(\sigma_x + \sigma_y) = 0 \tag{7}$$

appearing in the so-called *plane problem of elasticity*, where the components σ_x, τ_{xy} and σ_y of the so-called *stress-tensor* are to be found (Δ being the Laplace operator). It can be shown that, in a simply connected region Ω, the solution of this system is equivalent, in the following sense, to the solution of the *biharmonic equation*

$$\frac{\partial^4 U}{\partial x^4} + 2\frac{\partial^4 U}{\partial x^2 \partial y^2} + \frac{\partial^4 U}{\partial y^4} = 0 \quad \text{(briefly } \Delta^2 U = 0\text{)} \tag{8}$$

for the so-called *Airy stress function* U: Every *biharmonic function* $U(x, y)$ (thus every function satisfying equation (8)) has the property that the functions

$$\sigma_x = \frac{\partial^2 U}{\partial y^2}, \quad \tau_{xy} = -\frac{\partial^2 U}{\partial x \partial y}, \quad \sigma_y = \frac{\partial^2 U}{\partial x^2} \tag{9}$$

satisfy the system (7), while, conversely, to every triple of functions σ_x, τ_{xy}, σ_y satisfying (7) there exists a biharmonic function U connected with these functions by the relations (9). Boundary conditions corresponding to the system (7) can be transformed into those for the function U, and the so-called *biharmonic problem* is then to be solved. For details see e.g. [22]. A very universal method for solving the biharmonic problem is the finite element method. Plane (and also three-dimensional) problems of elasticity can be solved as well in components of the displacement. Also here the finite element method is frequently applied.

Solution of systems of equations in the *theory of shells* can be reduced, similarly as in the case of equations (7), to the solution of a single equation of the eighth order.

In a similar way, equations of *hydrodynamics* and *electromagnetic field* are transformed into those for scalar and vector potentials, respectively. Two-dimensional problems for the flow of nonviscous incompressible fluids are solved fairly simply, as usual, using the theory of functions of a complex variable (see Example 21.3.3).

The *flow of viscous incompressible fluids* is governed by the system of equations

$$\frac{\partial \boldsymbol{v}}{\partial t} + (\nabla \times \boldsymbol{v}) \times \boldsymbol{v} + \frac{1}{2}\nabla(v^2) = -\frac{\nabla p}{\rho} - \nabla U + \frac{\eta}{\rho}\Delta \boldsymbol{v} \tag{10}$$

(the *Navier-Stokes equation*),

$$\operatorname{div} \boldsymbol{v} = 0 \tag{11}$$

(the *continuity equation*), with corresponding initial and boundary conditions. Here $\boldsymbol{v}$, or p, or U, or ρ, or η is *vector of velocity* of the fluid, or *pressure*, or *gravitational potential*, or *density*, or *dynamic viscosity* of the fluid, respectively. To the solution of such problems, the *method of discretization in time* (§ 18.10) combined with the *finite element method* (Chap. 24) have been applied with success. Recently, these methods were successfully applied even in the case of compressible fluids.

18.8. Elliptic Boundary Value Problems of Arbitrary Order. Generalized Solutions. Eigenvalue Problems

REMARK 1. Results, obtained for unbounded positive definite operators by means of functional analysis and summarized in Theorems 22.6.9 and 22.6.10, make it possible to prove existence of (generalized) solutions for a relatively very broad class of elliptic boundary value problems and to apply current variational methods to obtain these solutions, or their sufficiently close approximation. Let us remind these results, in brief:

Let us investigate equations of the form

$$Au = f, \tag{1}$$

where f is an element of a Hilbert space H and A is an operator with his domain of definition $D(A)$ dense in H. (For concrete equations see Example 1 below.) Let A be *positive definite on* $D(A)$ (Definition 22.6.4) and let

$$Fu = (Au, u) - 2(f, u) \tag{2}$$

be corresponding *functional of energy* (22.6.24). (Here, $(.,.)$ is the scalar product in H.) Then (see the quoted Theorem 22.6.9) an element $u_0 \in D(A)$ is a solution of equation (1) exactly if it minimizes functional (2) on $D(A)$. However, (see Remark 22.6.10), neither an element $u_0 \in D(A)$ satisfying (1) nor an element minimizing (2) need exist. To ensure existence of a solution of (1) (in a generalized sense), let us introduce, following Remark 22.6.10, a new scalar product

$$(u, v)_A = (Au, v) \tag{3}$$

on $D(A)$ and, on its base, the norm and distance (22.6.26). In this way, $D(A)$ is converted into a metric space. Under the assumption of positive definiteness of the operator A, this space can be completed, in a relatively simple way, adding certain elements of H to $D(A)$ and extending the scalar product (3), defined originally only for elements of $D(A)$, onto those new elements. In this way a complete, and

thus *Hilbert* space is obtained, called the *energetic space* and denoted by H_A. The functional (2) is then extended onto this space by

$$Fu = (u, u)_A - 2(f, u) \tag{4}$$

and can be shown to attain actually its minimum on H_A, for an element

$$u_0 \in H_A, \tag{5}$$

uniquely determined by the right-hand side f of equation (1). This element u_0 is called the *generalized solution* of that equation. Thus, *if the operator A is positive definite on D_A, then equation* (1) *actually has a solution – in the just explained generalized sense.* This generalized solution can then be obtained by minimizing, in H_A, the functional (4), thus using current variational methods (Chap. 24). If $u_0 \in D(A)$, then u_0 is the solution of equation (1) in the usual sense. (In boundary value problems for differential equations this case occurs if the given data of the problem are sufficiently smooth.)

For details see, e.g., [389], in particular Chaps. 10 and 11.

Usefulness of results, recalled in Remark 1, can be well shown in the following example:

Example 1. Let us consider the following boundary value problems:

$$\Delta^2 u = f \quad \text{in } \Omega = (0, a) \times (0, b), \tag{6}$$

$$u = 0, \quad \frac{\partial u}{\partial \nu} = 0 \quad \text{on the boundary } S. \tag{7}$$

(*Deflexion of a clamped rectangular plate* loaded vertically by a load proportional to the function $f(x, y)$.)

$$-u'' + (1 + \sin^2 x)\, u = f, \tag{8}$$

$$u(0) = 0, \quad u(\pi) = 0. \tag{9}$$

(A boundary value problem for an ordinary differential equation with non-constant coefficients.)

Each of these problems is a problem of the form

$$Au = f, \tag{10}$$

where in the first case we have

$$Au = \Delta^2 u = \frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4}, \tag{11}$$

in the second one

$$Au = -u'' + (1 + \sin^2 x)\, u. \tag{12}$$

If we choose, in the first case, $H = L_2(\Omega)$ and

$$D(A) = \left\{ u\,;\, u \in C^{(4)}(\overline{\Omega}),\ u = 0,\ \frac{\partial u}{\partial \nu} = 0 \text{ on } S \right\} \tag{13}$$

(i.e. $D(A)$ is the set of all functions u continuous with their partial derivatives up to the fourth order inclusive in the closed region $\overline{\Omega}$ and satisfying the boundary conditions (7)), then the operator A, given by (11), can be shown to be positive definite on $D(A)$. (See, for example, Tab. 24.1.) Thus, if $f \in L_2(\Omega)$, there exists exactly one generalized solution u_0 of the problem (6), (7), and this solution minimizes the corresponding energetic functional, given as well in Tab. 24.1.

The same results follow for the problem (8), (9), if we choose $H = L_2(0, \pi)$ and

$$D(A) = \left\{ u\,;\, u \in C^{(2)}\big([0, \pi]\big),\ u(0) = 0,\ u(\pi) = 0 \right\}. \tag{14}$$

REMARK 2. For how to establish positive definiteness of an operator A on its domain of definition see, in details, in [389]. See also Example 22.6.6 and, in particular, Chap. 24 of the present book, devoted to the application of variational methods to the solution of problems of the just discussed types.

REMARK 3. It is evident that if the right-hand side f in (8) is "sufficiently" discontinuous, the problem (8), (9) cannot have a classical solution, and consequently, a generalization of that concept – e.g. in the sense of Remark 1 – is necessary. A similar remark concerns the problem (6), (7). Here, the loading of the plate is very often a discontinuous function. To introduce the concept of a generalized solution is then quite natural.

REMARK 4 (*Eigenvalue Problems*). As stated in Theorem 22.6.11, a positive definite operator A (or, in other words, the corresponding equation $Au - \lambda u = 0$) has a countable set of *eigenvalues*

$$0 < \lambda_1 \leqq \lambda_2 \leqq \lambda_3 \leqq \dots, \quad \lim_{n \to \infty} \lambda_n = +\infty, \tag{15}$$

while the orthonormal (in H_A) system of corresponding (linearly independent) *eigenelements* (*eigenfunctions*) is complete in both the spaces H_A and H. These eigenelements (eigenfunctions) minimize, subsenquently, in the space H_A and in its subspaces, in the sense of the quoted theorem, the *Rayleigh quotient*

$$R(v) = \frac{(v, v)_A}{(v, v)}, \tag{16}$$

giving successively the eigenvalues $\lambda_1, \lambda_2, \ldots$.

In particular, the operator A given by (11), being positive definite on its domain of definition (13), all just mentioned results given in Theorem 22.6.11 can be applied to the eigenvalue problem

$$\Delta^2 u - \lambda u = 0 \quad \text{in } \Omega, \tag{17}$$

$$u = 0, \quad \frac{\partial u}{\partial \nu} = 0 \quad \text{on } S. \tag{18}$$

Thus, this problem has a countable set of eigenvalues (15), etc. The same assertion holds for the problem

$$-u'' + (1 + \sin^2 x)\, u - \lambda u = 0, \tag{19}$$

$$u(0) = 0, \quad u(\pi) = 0. \tag{20}$$

REMARK 5. Similar results can be obtained for eigenvalue problems of the form

$$Au - \lambda Bu = 0 \tag{21}$$

with operators A, B positive definite on their domains of definition $D(A)$, $D(B)$. A typical problem of this form is the eigenvalue problem

$$\Delta^2 u + \lambda \Delta u = 0 \quad \text{in } \Omega, \tag{22}$$

$$u = 0, \quad \frac{\partial u}{\partial \nu} = 0 \quad \text{on } S. \tag{23}$$

For details see e.g. [389], Chap. 39. See also Chap. 24 of the present book.

18.9. Weak Solutions of Boundary Value Problems. Nonlinear Problems

The theory of generalized solutions, discussed in the preceding paragraph, is sufficiently general and, being based on minimalization of the functional of energy, relatively familiar to "consumers" of mathematics, especially to engineers. However, its disadvantage lies in the fact that, positive definite operators being symmetric, only symmetric problems can be investigated. In this paragraph, we mention a more general theory based on the concept of the so-called weak solution of the given problem and of the Lax-Milgram theorem. The idea of that theory will be shown in the following simple example:

Example 1. Let us consider the boundary value problem

$$-\Delta u \equiv -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f(x, y) \quad \text{in } \Omega, \tag{1}$$

$$u = 0 \quad \text{on } S, \tag{2}$$

where Ω is a bounded region in E_2 with a Lipschitz boundary S (Remark 22.4.10), and $f \in L_2(\Omega)$. Denote

$$V = \{ v \,;\, v \in W_2^{(1)}(\Omega),\ u = 0 \text{ on } S \text{ in the sense of traces} \} = \mathring{W}_2^{(1)}(\Omega). \tag{3}$$

Thus, V is the subspace of such functions from the Sobolev space $W_2^{(1)}(\Omega)$ (Remark 22.4.10) for which $v = 0$ on S is satisfied in the sense of traces (Remark 22.4.11). This subspace is denoted by $\mathring{W}_2^{(1)}(\Omega)$, as usual. Let, first, u be a classical solution of the problem (1), (2). (Such a solution need not exist, of course.) Let us choose a fixed function $v \in V$, multiply equation (1) by this function and integrate over the region Ω. Using the Green theorem and the fact that, v belonging to V, $v = 0$ on S, we obtain

$$-\iint_\Omega \frac{\partial^2 u}{\partial x^2} v \,\mathrm{d}x\,\mathrm{d}y = \iint_\Omega \frac{\partial u}{\partial x}\frac{\partial v}{\partial x}\,\mathrm{d}x\,\mathrm{d}y. \tag{4}$$

A similar result holds for $\partial^2 u/\partial y^2$. Equation (1) thus yields

$$\iint_\Omega \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right) \mathrm{d}x\,\mathrm{d}y = \iint_\Omega f v \,\mathrm{d}x\,\mathrm{d}y. \tag{5}$$

This result has been obtained for *every* $v \in V$. Therefore, (5) is often called an *integral identity.* Let us rewrite it in the form

$$((u, v)) = (f, v) \quad \text{for all } v \in V, \tag{6}$$

where the expression

$$((u, v)) = \iint_\Omega \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right) \mathrm{d}x\,\mathrm{d}y, \tag{7}$$

linear both in u and v, is called the *bilinear form* corresponding to the Laplace operator $-\Delta$ (appearing on the left-hand side of (1)) and to the boundary condition (2), and

$$(f, v) = \iint_\Omega f v \,\mathrm{d}x\,\mathrm{d}y. \tag{8}$$

The integral identity (5), or (6), has been derived under the condition that u is a classical solution of the problem (1), (2). (Thus assuming $u \in C^{(2)}(\overline{\Omega})$, etc.) However, the integral (7) has sense if $\partial u/\partial x$ and $\partial u/\partial y$ belong to $L_2(\Omega)$ only. This

is the case, for example, if $u \in W_2^{(1)}(\Omega)$. Condition (2) can then be considered in the sense of traces. These facts establish the reason why to look for the solution u of the problem (1), (2) among the functions from the space $\mathring{W}_2^{(1)}(\Omega)$, and lead to the following rather natural generalization of the concept of a classical solution:

Definition 1. Under a *weak solution of the problem* (1), (2) we understand a function

$$u \in \mathring{W}_2^{(1)}(\Omega), \tag{9}$$

for which the integral identity

$$((u, v)) = (f, v) \quad \text{for all } v \in \mathring{W}_2^{(1)}(\Omega) \tag{10}$$

is satisfied, with $((u, v))$ and (f, v) given by (7), (8).

Existence of exactly one weak solution of the problem (1), (2) (with $f \in L_2(\Omega)$ only) is ensured by the following *Lax-Milgram Theorem:*

Theorem 1. *Let in a Hilbert space V a bilinear form $((u, v))$ be given. Let there exist two constants $K > 0$ and $\alpha > 0$* (independent of u, v) *such that for every $u, v \in V$ we have*

$$|((u, v))| \leqq K \|u\|_V \|v\|_V \quad (V\text{-boundedness of the form } ((u, v))), \tag{11}$$

$$((v, v)) \geqq \alpha \|v\|_V^2 \quad (V\text{-ellipticity of the form } ((u, v))). \tag{12}$$

Then every bounded linear functional F on V can be expressed in the form

$$Fv = ((u_0, v)) \quad \text{for all } v \in V \tag{13}$$

with u_0 uniquely determined by this functional.

The right-hand side (f, v) of the integral identity (6) is evidently a bounded linear functional in $V = \mathring{W}_2^{(1)}(\Omega)$, for due to the Schwarz inequality, we have

$$|(f, v)| \leqq \|f\|_{L_2(\Omega)} \cdot \|v\|_{L_2(\Omega)} \leqq \|f\|_{L_2(\Omega)} \cdot \|v\|_{\mathring{W}_2^{(1)}(\Omega)}, \tag{14}$$

because (see Remark 22.4.10)

$$\|v\|_{W_2^{(1)}(\Omega)}^2 = \|v\|_{L_2(\Omega)}^2 + \left\|\frac{\partial v}{\partial x}\right\|_{L_2(\Omega)}^2 + \left\|\frac{\partial v}{\partial y}\right\|_{L_2(\Omega)}^2. \tag{15}$$

In a similar way, we get

$$\left|\iint_\Omega \frac{\partial u}{\partial x}\frac{\partial v}{\partial x}\,\mathrm{d}x\,\mathrm{d}y\right| \leqq \left\|\frac{\partial u}{\partial x}\right\|_{L_2(\Omega)} \left\|\frac{\partial v}{\partial x}\right\|_{L_2(\Omega)} \leqq \|u\|_{\mathring{W}_2^{(1)}(\Omega)} \|v\|_{\mathring{W}_2^{(1)}(\Omega)}$$

and similarly for the derivatives with respect to y. Thus

$$|((u, v))| \leqq 2 \|u\|_{\mathring{W}_2^{(1)}(\Omega)} \cdot \|v\|_{\mathring{W}_2^{(1)}(\Omega)}. \tag{16}$$

Moreover the so-called *Friedrichs inequality* ([389], Chap. 30) yields

$$((v, v)) = \iint_\Omega \left[\left(\frac{\partial v}{\partial x} \right)^2 + \left(\frac{\partial v}{\partial y} \right)^2 \right] \mathrm{d}x \, \mathrm{d}y \geqq \alpha \|v\|^2_{\mathring{W}_2^{(1)}(\Omega)} \tag{17}$$

for all $v \in \mathring{W}_2^{(1)}(\Omega)$, with $\alpha > 0$ independent of v.

All assumptions of the Lax-Milgram theorem being fulfilled, for every $f \in L_2(\Omega)$ there exists exactly one weak solution $u_0 \in W_2^{(1)}(\Omega)$ of the problem (1), (2). (While existence of a classical solution is not ensured – if f is discontinuous, for example, classical solution does not exist.) Moreover (see e.g. [389], Chap. 34), the weak solution u_0 minimizes, on V, the functional

$$((u, u)) - 2(f, u),$$

and current variational methods to this minimalization can be applied.

REMARK 1. In Example 1 only the idea of the theory of weak solutions of boundary value problems has been shown. For the whole theory see, e.g., [389]. The problem (1), (2) is so simple (and moreover symmetric) that the concept of a generalized solution, discussed in the preceding paragraph, would have been quite sufficient for its treating. On the other hand, this problem being so simple, it was possible to demonstrate, in a very lucid way, that no symmetry is needed in the case of the weak formulation of a problem. Just this fact represents a very advantage of this approach, in the general case. Moreover, the weak formulation makes it possible to investigate non-homogeneous boundary conditions in a much more unified and surveyable way than in the case of generalized solutions.

REMARK 2 (*Nonlinear Problems*). The concept of the *Gâteau differential*, discussed in § 22.8, enables us to treat a broad class of nonlinear problems in a proper way. The idea how to do it is shown in the following example:

Example 2. Let us consider the nonlinear problem

$$-u'' + 20u^3 = \sin \pi x, \tag{18}$$

$$u(0) = 0, \quad u(1) = 0. \tag{19}$$

Similarly as in Example 1, denote

$$V = \mathring{W}_2^{(1)}(0, 1) = \{ v;\ v \in W_2^{(1)}(0, 1),\ v(0) = 0,\ v(1) = 0 \}. \tag{20}$$

In contrast to the quoted example, it is not necessary to add "in the sense of traces" in (20) because (see (33) below), for functions from $\mathring{W}_2^{(1)}(0, 1)$, the boundary conditions (19) are fulfilled in the ordinary sense.

Following the way of Example 1 and multiplying (18) by an arbitrary function $v \in \mathring{W}_2^{(1)}(\Omega)$, integrating over the interval $[0, 1]$ and using integration by parts in the first term, we come to the weak formulation of the problem (18), (19): To find such a function

$$u \in \mathring{W}_2^{(1)}(0, 1) \tag{21}$$

that the integral identity

$$\int_0^1 u'v' \, dx + 20 \int_0^1 u^3 v \, dx = \int_0^1 v \sin \pi x \, dx \quad \text{for all } v \in \mathring{W}_2^{(1)}(\Omega) \tag{22}$$

be satisfied, i.e.,

$$((u, v)) = (v, \sin \pi x) \quad \text{for all } v \in \mathring{W}_2^{(1)}(\Omega) \tag{23}$$

with

$$((u, v)) = \int_0^1 u'v' \, dx + 20 \int_0^1 u^3 v \, dx, \tag{24}$$

$$(v, \sin \pi x) = \int_0^1 v \sin \pi x \, dx. \tag{25}$$

To ensure existence of such a weak solution u, it is sufficient – following § 22.8 – to construct on $\mathring{W}_2^{(1)}(0, 1)$ a functional G for which

$$G'(u, v) = \int_0^1 u'v' \, dx + 20 \int_0^1 u^3 v \, dx - \int_0^1 v \sin \pi x \, dx \tag{26}$$

holds and to prove that there exists an $u_0 \in \mathring{W}_2^{(1)}(0, 1)$ for which that functional attains its minimum on $\mathring{W}_2^{(1)}(0, 1)$. Because, by Theorem 22.8.1, we then have, at that point, $G'(u_0, v) = 0$ for all $v \in \mathring{W}_2^{(1)}(0, 1)$, i.e.

$$\int_0^1 u_0'v' \, dx + 20 \int_0^1 u_0^3 v \, dx - \int_0^1 v \sin \pi x \, dx = 0 \quad \text{for all } v \in \mathring{W}_2^{(1)}(0, 1), \tag{27}$$

so that for that u_0 the integral identity (22) is satisfied.

In our case, the functional G is easily obtained by (22.8.21) from Remark 22.8.6:

$$\begin{aligned} Gu &= \int_0^1 \left[\int_0^1 tu'u' \, dx + 20 \int_0^1 (tu)^3 u \, dx - \int_0^1 u \sin \pi x \, dx \right] dt = \\ &= \frac{1}{2} \int_0^1 u'^2 \, dx + 5 \int_0^1 u^4 \, dx - \int_0^1 u \sin \pi x \, dx. \end{aligned} \tag{28}$$

Evidently, we have

$$Gu = Fu - gu, \tag{29}$$

where

$$gu = \int_0^1 u \sin \pi x \, dx \tag{30}$$

and

$$Fu = \frac{1}{2} \int_0^1 u'^2 \, dx + 5 \int_0^1 u^4 \, dx \tag{31}$$

is the functional discussed in details in Example 22.8.1.

Using the well-known formula for the norm in $W_2^{(1)}(0, 1)$,

$$\|u\|^2_{W_2^{(1)}(0,1)} = \int_0^1 u^2 \, dx + \int_0^1 u'^2 \, dx \tag{32}$$

and an embedding result (Example 22.4.7),

$$u \in \mathring{W}_2^{(1)}(0, 1) \Rightarrow u \text{ is continuous in } [0, 1], \tag{33}$$

while a constant c (independent of u) exists such that

$$\|u\|_{C[0,1]} = \max_{0 \leqq x \leqq 1} |u(x)| \leqq c \, \|u\|_{W_2^{(1)}(0,1)}, \tag{34}$$

the functional (31) has been proved, in the quoted Example 22.8.1, to be well defined on the entire space $W_2^{(1)}(0, 1)$. Then its first and second Gâteaux differentials have been determined,

$$F'(u, v) = \int_0^1 u'v' \, dx + 20 \int_0^1 u^3 v \, dx, \tag{35}$$

$$F''(u, v, w) = \int_0^1 u'w' \, dx + 60 \int_0^1 u^2 vw \, dx. \tag{36}$$

Now, because of $u \in W_2^{(1)}(0, 1) \Rightarrow u \in L_2(0, 1)$, the functional g is also defined for all $u \in W_2^{(1)}(0, 1)$. Moreover,

$$g(u + tv) = \int_0^1 (u + tv) \sin \pi x \, dx,$$

so that

$$g'(u, v) = \frac{d}{dt} \int_0^1 g(u + tv)\Big|_{t=0} = \int_0^1 v \sin \pi x \, dx; \tag{37}$$

further,

$$g'(u + tw, v) = \int_0^1 v \sin \pi x \, dx,$$

whence

$$g''(u, v, w) = \frac{\mathrm{d}}{\mathrm{d}t} g'(u + tw, v)\Big|_{t=0} = 0. \tag{38}$$

Thus, the functional G is also defined on the whole space $W_2^{(1)}(0, 1)$, and, consequently, on the whole space $\mathring{W}_2^{(1)}(0, 1)$, because $\mathring{W}_2^{(1)}(0, 1)$ is a subspace of $W_2^{(1)}(0, 1)$, and

$$G'(u, v) = F'(u, v) - g'(u, v) =$$
$$= \int_0^1 u'v' \,\mathrm{d}x + 20 \int_0^1 u^3 v \,\mathrm{d}x - \int_0^1 v \sin \pi x \,\mathrm{d}x \tag{39}$$

(as known from (26) already),

$$G''(u, v, w) = F''(u, v, w) = \int_0^1 v'w' \,\mathrm{d}x + 60 \int_0^1 u^2 vw \,\mathrm{d}x. \tag{40}$$

To establish that a functional G attains its (only) minumum on $\mathring{W}_2^{(1)}(0, 1)$, it is sufficient (Theorem 22.8.3) to prove that

(i) G is defined on the whole space $\mathring{W}_2^{(1)}(0, 1)$ and has everywhere its first and second Gâteaux differentials;

(ii) $G'(u, v)$ is bounded in the following sense: Let M be the set of all $u \in$ $\in \mathring{W}_2^{(1)}(0, 1)$ such that $\|u\|_{\mathring{W}_2^{(1)}(0,1)} \leqq r$. Then a constant K (dependent on r, but indepedent of $u \in M$) exists such that

$$\left|G'(u, v)\right| \leqq K \|v\|_{\mathring{W}_2^{(1)}(0,1)} \tag{41}$$

holds for all $u \in M$ and $v \in \mathring{W}_2^{(1)}(0, 1)$;

(iii) a constant $k > 0$ exists (independent of v, u) such that

$$G''(u, v, v) \geqq k \|v\|^2_{\mathring{W}_2^{(1)}(0,1)}. \tag{42}$$

For our functional (28) the requirements (i) have already been established.

Ad (ii): We have

$$\left|G'(u, v)\right| \leqq \left|\int_0^1 u'v' \,\mathrm{d}x\right| + 20 \left|\int_0^1 u^3 v \,\mathrm{d}x\right| + \left|\int_0^1 v \sin \pi x \,\mathrm{d}x\right| \leqq$$
$$\leqq \left(r + 20r^3c^3 + \frac{1}{\sqrt{2}}\right) \|v\|_{\mathring{W}_2^{(1)}(0,1)} \tag{43}$$

for every $u \in M$ and $v \in \mathring{W}_2^{(1)}(0, 1)$, because by the Schwarz inequality and by (32) we have, first,

$$\left|\int_0^1 u'v' \,\mathrm{d}x\right| = \left|(u', v')_{L_2(0,1)}\right| \leqq \|u'\|_{L_2(0,1)} \|v'\|_{L_2(0,1)} \leqq$$
$$\leqq \|u\|_{\mathring{W}_2^{(1)}(0,1)} \|v\|_{\mathring{W}_2^{(1)}(0,1)} \leqq r \|v\|_{\mathring{W}_2^{(1)}(0,1)};$$

further, by (34) and the Schwarz inequality,

$$20\left|\int_0^1 u^3 v\,\mathrm{d}x\right| \leqq 20\int_0^1 |u|^3\,|v|\,\mathrm{d}x = 20\,r^3c^3\int_0^1 |v|\,\mathrm{d}x \leqq$$
$$\leqq 20\,r^3c^3\,\|v\|_{L_2(0,1)} \leqq 20\,r^3c^3\,\|v\|_{\mathring{W}_2^{(1)}(0,1)}$$

and, finally,

$$\left|\int_0^1 v\sin\pi x\,\mathrm{d}x\right| = \left|(v,\sin\pi x)_{L_2(0,1)}\right| \leqq$$
$$\leqq \left\|\sin\pi x\right\|_{L_2(0,1)}\|v\|_{L_2(0,1)} \leqq$$
$$\leqq \frac{1}{\sqrt{2}}\,\|v\|_{\mathring{W}_2^{(1)}(0,1)}\,.$$

Thus, r being fixed, the value $r + 20r^3c^3 + \frac{1}{\sqrt{2}}$ for K in (41) can be taken.

Ad (iii): Let us note, first, that the Friedrichs inequality (see, e.g. [389], Chap. 30) yields for functions from $\mathring{W}_2^{(1)}(0,1)$

$$\int_0^1 u'^2\,\mathrm{d}x \geqq m\int_0^1 u^2\,\mathrm{d}x \tag{44}$$

with $m > 0$ independent of u. So we have

$$G''(u,v,v) = \int_0^1 v'^2\,\mathrm{d}x + \int_0^1 u^2v^2\,\mathrm{d}x \geqq \int_0^1 v'^2\,\mathrm{d}x =$$
$$= \frac{1}{2}\int_0^1 v'^2\,\mathrm{d}x + \frac{1}{2}\int_0^1 v'^2\,\mathrm{d}x \geqq$$
$$\geqq \frac{1}{2}\left(m\int_0^1 v^2\,\mathrm{d}x + \int_0^1 v'^2\,\mathrm{d}x\right) \geqq k\,\|v\|^2_{\mathring{W}_2^{(1)}(0,1)} \tag{45}$$

with

$$k = \min\left(\frac{m}{2}, \frac{1}{2}\right).$$

All the requirements (i), (ii), (iii) being fulfilled, Theorem 22.8.3 ensures existence of exactly one $u_0 \in \mathring{W}_2^{(1)}(0,1)$ for which the functional G assumes its minimum on $\mathring{W}_2^{(1)}(0,1)$ and for which, consequently, the integral identity (27) is fulfilled.

In this way, existence of a (unique) weak solution of the problem (18), (19) has been established.

REMARK 3. A relatively simple problem (18), (19) has been choosen to show the ideas. The Hilbert space $\mathring{W}_2^{(1)}(0,1)$ could be taken for the space V in (20), because,

thanks to the embedding theorems, the "nonlinear" integral $\int_0^1 u^3 v \, dx$ exists here. In a similar way we can go on when solving, for example, the problem

$$-\Delta u + u^7 = f \quad \text{in } \Omega,$$
$$u = 0 \quad \text{on } S,$$

where G is a bounded region in E_2 with a Lipschitz boundary S. Also here the Hilbert space $V = \mathring{W}_2^{(1)}(\Omega)$ can be choosen, because, thanks to (22.4.22), the integral $\int_\Omega u^q v \, dx$ exists with an arbitrary $q \in [1, +\infty)$, thus with $q = 7$, for example. However, in general, Banach spaces should be applied when nolinear problems are to be solved. For a detailed theory see, e.g., [160], where also "nonpotential" problems are investigated (thus such where the theory based on the concept of the Gâteuax differential of a functional cannot be applied).

18.10. Application of Variational Methods to the Solution of Partial Differential Equations Containing Time. The Method of Discretization in Time (the Rothe Method, the "Horizontal" Method of Lines)

Variational methods, developed originally for solution of elliptic boundary value problems, can be as well applied to the solution of parabolic and hyperbolic problems. One of the ways how to achieve it is to apply the Galerkin semidiscretization method (the "vertical" method of lines) mentioned in § 24.6. Another possibility gives the *method of discretizaton in time* (the *Rothe method*, the *"horizontal" method of lines*) investigated extensively in [390]. Its very simple idea becomes clear from the following example:

Example 1. Consider the parabolic equation

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = \sin x \tag{1}$$

in the region $Q = (0, \pi) \times (0, 1)$, with the initial and boundary conditions

$$u(x, 0) = 0, \tag{2}$$
$$u(0, t) = 0, \quad u(\pi, t) = 0. \tag{3}$$

Here, the solution can be easily found:

$$u(x, t) = (1 - e^{-t}) \sin x. \tag{4}$$

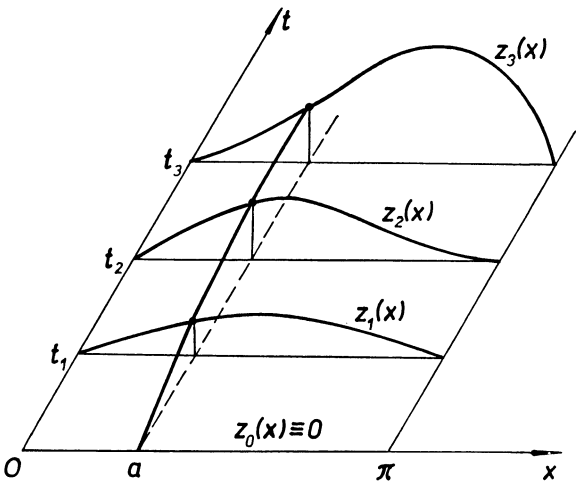


Fig. 18.9.

Thus, the example is only illustrative. We have chosen it because it is simple enough to show the idea and, moreover, it gives the possibility of comparing results, obtained by the method considered, with the exact solution.

Let us divide the interval $[0, 1]$ into p subintervals of the length $h = 1/p$ by the points

$$t_1 = h, \quad t_2 = 2h, \quad t_3 = 3h, \quad \ldots$$

(Fig. 18.9), with $t_0 = 0$, $t_p = ph = 1$. Choose $z_0(x) \equiv 0$ in accordance with (2) and find succesively for $t = t_1$, $t = t_2$, $t = t_3$, ..., the functions $z_1(x)$, $z_2(x)$, $z_3(x)$, ... as solutions of the problems

$$\frac{z_1 - z_0}{h} - z_1'' = \sin x, \quad z_1(0) = 0, \quad z_1(\pi) = 0, \tag{5}$$

$$\frac{z_2 - z_1}{h} - z_2'' = \sin x, \quad z_2(0) = 0, \quad z_2(\pi) = 0, \tag{6}$$

$$\frac{z_3 - z_2}{h} - z_3'' = \sin x, \quad z_3(0) = 0, \quad z_3(\pi) = 0, \tag{7}$$

.....

We thus have replaced, for every t_k, the derivative $\partial u/\partial t$ in (1) by the corresponding difference quotient and the derivative $\partial^2 u/\partial x^2$ by the ordinary second derivative. The given problem has been reduced, in this way, to successive solution of ordinary differential equations with the boundary conditions $z_k(0) = 0$, $z_k(\pi) = 0$ (by (3)).

Solution of these problems is very simple in our case. Because of $z_0(x) \equiv 0$ and of the boundary conditions given, the solution of the problem (5) can be evidently assumed in the form

$$z_1 = a_1 \sin x. \tag{8}$$

An easy computation yields

$$z_1 = \left(1 - \frac{1}{1+h}\right) \sin x; \tag{9}$$

substituting this result into (6) for z_1 and assuming z_2 in the form $a_2 \sin x$ again, we get

$$z_2 = \left(1 - \frac{1}{(1+h)^2}\right) \sin x;$$

in the same way we obtain

$$z_3 = \left(1 - \frac{1}{(1+h)^3}\right) \sin x$$

and, generally,

$$z_k = \left(1 - \frac{1}{(1+h)^k}\right) \sin x. \tag{10}$$

If we choose, for example, $h = 0{\cdot}01$ and $k = 20$, or $k = 40$, or $k = 100$ (which correspond to the values $t = 0{\cdot}2$, or $t = 0{\cdot}4$, or $t = 1$), we get

$$z_{20} = \left(1 - \frac{1}{(1+0{\cdot}01)^{20}}\right) \sin x = 0{\cdot}1805 \sin x,$$

or

$$z_{40} = \left(1 - \frac{1}{(1+0{\cdot}01)^{40}}\right) \sin x = 0{\cdot}3284 \sin x,$$

or

$$z_{100} = \left(1 - \frac{1}{(1+0{\cdot}01)^{100}}\right) \sin x = 0{\cdot}6303 \sin x,$$

respectively. Corresponding values of the exact solution (4) are

$$z(x, 0{\cdot}2) = 0{\cdot}1813 \sin x,$$
$$z(x, 0{\cdot}4) = 0{\cdot}3297 \sin x,$$
$$z(x, 1) = 0{\cdot}6321 \sin x.$$

Thus, the values of the approximate solution, obtained by the method of discretization in time at these points, and the values of the exact solution are in a very good harmony.

We see that this method gives an approximate solution at certain discrete points $t_1, t_2, \ldots, t_p$ of the given interval only. An approximate solution defined on the whole region Q can be obtained, if required, by constructing the so-called *Rothe function*, defined in every subinterval $I_k = [t_{k-1}, t_k]$ by

$$u_1(x, t) = z_{k-1}(x) + \frac{t - t_{k-1}}{h}\big(z_k(x) - z_{k-1}(x)\big), \tag{11}$$

$k = 1, 2, \ldots, p$. So for every fixed $x = a$ from the interval $[0, \pi]$, this function is a piecewise linear function of t in the interval $[0, 1]$ and for $t = t_1$, $t = t_2$, ... it assumes the values $z_1(a)$, $z_2(a)$, etc. (Fig. 18.9).

If we are interested in *convergence* of the method, let us consider the divisions $d_1, d_2, d_3, \ldots, d_n, \ldots$ of the interval $[0, 1]$ into subintervals of the length h, $h/2$, $h/4, \ldots, h/2^{n-1}, \ldots$, respectively, and for each of these divisions construct the corresponding Rothe function. In our case it is easy to prove that the sequence

$$u_1(x, t), \quad u_2(x, t), \quad u_3(x, t), \quad \ldots, \quad u_n(x, t), \ldots$$

of these Rothe functions converges in $\overline{Q}$ to the solution (4) of the problem (1), (2), (3). (Roughly speaking, the approximate solution, obtained by the method of discretization in time, is the "better", the finer is the division of the given time interval.)

As said above, Example 1 is only illustrative, because the problem (1)–(3) was so simple that there was no need to apply any approximate method. However, it has well shown characteristic features of the method and, at the same time, what can be expected when using it.

In the case of boundary value problems with one space variable we solve, successively, *ordinary* differential equations with corresponding boundary conditions. However, the main importance of the method lies in its application to approximate solution of partial differential equations in several space variables, first of all of equations of the form

$$\frac{\partial u}{\partial t} + Au = f \quad \text{in } Q = \Omega \times (0, T) \tag{12}$$

with corresponding initial and boundary conditions. Here A is an elliptic operator (e.g. the operator $-\Delta$, or the operator Δ^2, or an elliptic operator of an even order with variable coefficients) and f is the given function. The method of discretization in time then yields *elliptic equations* of the form

$$\frac{z_1 - z_0}{h} + Az_1 = f \quad \text{in } \Omega, \tag{13}$$

$$\frac{z_2 - z_1}{h} + Az_2 = f \quad \text{in } \Omega, \tag{14}$$

$$\ldots\ldots\ldots\ldots\ldots$$

with corresponding boundary conditions, to be solved, successively, to obtain the functions z_1, z_2, etc. The function z_0 is known from the given initial condition (in Example 1 we had $z_0 \equiv 0$). (For the sake of simplicity, we assumed, when writing down equations (13), (14), the function f and the coefficients of the operator A not to depend on time; if it is *not* the case, equation (13) is to be replaced by the equation

$$\frac{z_1 - z_0}{h} + A_1 z_1 = f_1 \quad \text{in } \Omega, \tag{15}$$

where f_1, or A_1, is the function f, or the operator A, respectively, taken for $t = t_1$, etc. Also boundary conditions may depend on time.) The elliptic problems (13), (14) can then be solved by current variational methods (the finite element method, etc.). If the boundary conditions and the operator A are independent of time (and this is a very frequent case in applications), the successive solution of problems (13), (14), ... goes forward very quickly, because the matrix of corresponding systems of equations remains unchanged. Moreover, *the method of discretization in time is numerically stable.*

As concerns *theoretical* questions (existence, convergence), let us assume, first, that the initial and boundary conditions are homogeneous. Let the region Ω be bounded and its boundary Lipschitzian (see Remark 22.4.10; thus not necessarily smooth). If $f \in L_2(\Omega)$ and the operator A is positive definite or the corresponding bilinear form V-elliptic (see §§ 18.8, 18.9), then each of the problems (13), (14), ... is uniquely solvable (it has exactly one generalized, or weak solution in the sense of the quoted paragraphs), and the sequence of corresponding Rothe functions, constructed similarly as in (11), converges in a certain sense (in a certain functional space not characterized exactly here, see [390]) to the so-called *weak solution* of the given problem. This result can be extended, without difficulties, to the case of non-homogeneous initial and boundary conditions. For details, the reader is referred to [390], where also the cases are analysed when the operator A is nonlinear, or when the problem is hyperbolic (thus when an equation of the form

$$\frac{\partial^2 u}{\partial t^2} + Au = f$$

with corresponding initial and boundary conditions is given), etc. In that book, also a lot of numerical examples can be found.

19. INTEGRAL EQUATIONS

By KAREL REKTORYS

References: [22], [99], [109], [143], [232], [241], [252], [264], [273], [274], [276], [323], [342], [415], [418], [432], [434], [438], [469], [471].

19.1. Integral Equations of Fredholm's Type. Solvability, Fredholm's Theorems. Systems of Integral Equations

REMARK 1. Many problems encountered in applications, as well as in mathematics itself, lead to the solution of integral equations, in particular of the so-called Fredholm equations (see below). As shown in Example 22.5.3, the integral operator, given by

$$Au = \int_a^b K(x, s)u(s)\,\mathrm{d}s,$$

occurring in these equations, is a completely continuous operator. Hence, the theory of integral equations can be taken as a special case of the theory of operator equations with completely continuous operators. However, we are not going to treat integral equations in such an abstract way in this chapter. We give here the "classical" theory of integral equations in the space $L_2(a, b)$ which is a very natural space for these equations to be treated in.

Basic concepts, concerning the space $L_2(a, b)$ of square Lebesgue integrable functions, have been given in § 16.1. For a brief orientation of the reader, let us remind:

Every (real) function which is continuous, or piecewise continuous in $[a, b]$, is Lebesgue integrable in this interval. A (Lebesgue) measurable function f (unbounded in general) is called *square integrable* on (or in) $[a, b]$ if the integral

$$\int_a^b f^2(x)\,\mathrm{d}x$$

is convergent (= of finite value). We write

$$f \in L_2(a, b).$$

By a *scalar product* of two (real) functions $f, g \in L_2(a, b)$ we understand the number

$$(f, g) = \int_a^b f(x)g(x)\,\mathrm{d}x,$$

by the *norm* of a function f the number

$$\|f\| = \sqrt{(f, f)} = \sqrt{\left[\int_a^b f^2(x)\,\mathrm{d}x\right]}$$

and by the *distance* of functions f, g the number

$$\varrho(f, g) = \|f - g\| = \sqrt{\left\{\int_a^b \left[g(x) - f(x)\right]^2 \mathrm{d}x\right\}}.$$

The set of all square integrable functions on the interval $[a, b]$, with the just mentioned operations of scalar product, norm and distance, is called the (*real*) *space* $L_2(a, b)$. (See, however, the following text.)

Two functions f, g are called *equivalent* in the space $L_2(a, b)$, if their distance is equal to zero, i.e. if

$$\int_a^b \left[g(x) - f(x)\right]^2 \mathrm{d}x = 0.$$

We write

$$f = g \quad \text{in} \quad L_2(a, b).$$

Two equivalent functions can differ, in $[a, b]$, on a set of measure zero, e.g. at a finite number of points. We say also that they are *equal almost everywhere* in $[a, b]$. In the space $L_2(a, b)$, they are taken for equal, or, in other words, they represent the same element of this space. (In this sense, one speaks about the space L_2 as about the set of *classes* of mutually equivalent square integrable functions, with the operations of scalar product, norm and distance defined as corresponding operations with arbitrary representants of these classes.) If we say that f is a *zero function* in $L_2(a, b)$ (we write

$$f = 0 \quad \text{in} \quad L_2(a, b)),$$

then this function is a function which is either equal to zero in the whole interval $[a, b]$, or is different from zero at points constituting a set of measure zero, e.g. at a finite number of points. If we say that f is a *non-zero function* in $L_2(a, b)$ (we write

$$f \neq 0 \quad \text{in} \quad L_2(a, b)),$$

then this function is different from zero on a set of positive measure, e.g. on the whole interval $[a, b]$ or on a subinterval of that interval. For example, the function $\sin 5x$ is a non-zero function in $L_2(0, \pi)$.

In this sense we speak about zero or non-zero functions in the whole chapter.

Let us note that if two equivalent functions f, g are *continuous* in $[a, b]$, then they are equal at all points of that interval.

In the theory of integral equations we often deal with *complex functions of a real variable*, i.e. with functions of the form

$$f(x) = f_1(x) + \mathrm{i} f_2(x),$$

where f_1 and f_2 are real functions. A (measurable) function f is then called square integrable in $[a, b]$ if

$$\int_a^b |f(x)|^2 \, \mathrm{d}x < +\infty$$

(what happens if and only if both the functions f_1 and f_2 are square integrable in that interval). All what has been said about the real space $L_2(a, b)$ in the preceding text, remains valid for the *complex space* $L_2(a, b)$, except that the scalar product is defined here by the relation

$$(f, g) = \int_a^b f(x)\overline{g(x)} \, \mathrm{d}x$$

($\overline{g(x)}$ being the complex conjugate to $g(x)$) and that the integrals

$$\int_a^b f^2(x) \, \mathrm{d}x, \quad \text{or} \quad \int_a^b [g(x) - f(x)]^2 \, \mathrm{d}x,$$

appearing in the definitions of the norm and distance, are to be replaced by the integrals

$$\int_a^b |f(x)|^2 \, \mathrm{d}x, \quad \text{or} \quad \int_a^b |g(x) - f(x)|^2 \, \mathrm{d}x,$$

respectively.

In a similar way the *real*, or *complex space* $L_2(\Omega)$ is defined, where Ω is a bounded region in E_2 (or, more generally, in E_n), with the scalar product (for $n = 2$)

$$(f, g) = \iint_\Omega f(x, y) g(x, y) \, \mathrm{d}x \, \mathrm{d}y,$$

or

$$(f, g) = \iint_\Omega f(x, y)\overline{g(x, y)} \, \mathrm{d}x \, \mathrm{d}y,$$

respectively.

In this chapter, under $L_2(a, b)$, or $L_2(\Omega)$, complex spaces L_2 are to be understood, in general.

Definition 1. The equation

$$f(x) - \int_a^b K(x, s) f(s)\, ds = g(x) \tag{1}$$

is called a *linear integral equation of the second kind.* (On integral equations of the first kind see in § 19.7. A brief analysis of a typical *nonlinear* equation can be found in Example 22.5.1.) Here $x, s \in [a, b]$ are real variables, the function $K(x, s)$, called the *kernel* of equation (1), is defined in the closed square $\overline{Q} = [a, b] \times [a, b]$, $g(x)$ is a given function defined in $[a, b]$, $f(x)$ is the unknown function.

In general, we have

$$K(x, s) = K_1(x, s) + iK_2(x, s), \quad g(x) = g_1(x) + ig_2(x),$$

with K_1, K_2, g_1, g_2 real. *In what follows, we assume that* $K \in L_2(Q)$, $g \in L_2(a, b)$, i.e. (see Remark 1) that

$$\iint_Q |K(x, s)|^2\, dx\, ds, \quad \int_a^b |g(x)|^2\, dx$$

are finite numbers. In this case, (1) is called the *Fredholm equation.* (Let us note that there is no uniformity in terminology. For example, many authors understand under a Fredholm equation every equation of the form (1), thus every integral equation of the second kind. Also other definitions are in use.)

Definition 2. By a *solution of equation* (1) we understand such a function $f \in$ $\in L_2(a, b)$, for which equation (1) is satisfied in the interval $[a, b]$ almost everywhere, i.e. for all $x \in [a, b]$ with the possible exception of points constituting a set of measure zero (e.g. with the possible exception of a finite number of points).

Example 1. The function

$$f(x) = x^2 - \tfrac{5}{8}x - \tfrac{1}{8}$$

(see Example 19.2.1) is a solution of the equation

$$f(x) - \int_0^1 6(x + s) f(s)\, ds = x^2,$$

because for $x \in [0, 1]$ the equality

$$x^2 - \tfrac{5}{8}x - \tfrac{1}{8} - \int_0^1 6(x + s)\left(s^2 - \tfrac{5}{8}s - \tfrac{1}{8}\right) ds = x^2$$

is satisfied (even for *all* x of that interval here).

REMARK 2. In the case of two variables, the equation of the second kind is of the form

$$f(x_1, x_2) - \iint_{\Omega} K(x_1, x_2, x_3, x_4) f(x_3, x_4)\, \mathrm{d}x_3\, \mathrm{d}x_4 = g(x_1, x_2),$$

where Ω is a given domain; very frequently $\overline{\Omega}$ is the square $a \leqq x_3 \leqq b$, $a \leqq x_4 \leqq b$. The variables x_1, x_2, x_3, x_4 run through the set $\overline{Q} = \overline{\Omega} \times \overline{\Omega}$ (thus $(x_3, x_4) \in \overline{\Omega}$, $(x_1, x_2) \in \overline{\Omega}$). In the special case, where Ω is a square, Q is the four-dimensional interval $a \leqq x_i \leqq b$, $i = 1, 2, 3, 4$. If the kernel $K(x_1, x_2, x_3, x_4)$ is square integrable in Q, the equation is called *Fredholm's equation*. The (Fredholm) integral equation for an unknown function of several variables is defined similarly. The theory and computing methods are very much alike for cases both of one and several variables. Therefore, in the following text, we shall deal with the one-dimensional case only. The results obtained may easily be generalized to the case of functions of several variables.

REMARK 3. A parameter λ (generally complex) is often introduced into equation (1),

$$f(x) - \lambda \int_a^b K(x, s) f(x)\, \mathrm{d}s = g(x). \tag{2}$$

For $\lambda = 1$ we obtain equation (1) as a special case. If $g(x) \equiv 0$, we obtain the equation

$$f(x) - \lambda \int_a^b K(x, s) f(x)\, \mathrm{d}s = 0, \tag{3}$$

the so-called *homogeneous equation corresponding to the equation* (2).

Definition 3. The equation

$$F(x) - \overline{\lambda} \int_a^b \overline{K(s, x)} F(s)\, \mathrm{d}s = 0 \tag{4}$$

is called the *adjoint equation to equation* (3). Recall that the bar above λ and $K(s, x)$ denotes the complex conjugate value (i.e. if $\lambda = \lambda_1 + \mathrm{i}\lambda_2$ and $K(x, s) = K_1(x, s) + \mathrm{i}K_2(x, s)$, $\lambda_1, \lambda_2, K_1, K_2$ real, then $\overline{\lambda} = \lambda_1 - \mathrm{i}\lambda_2$ and $\overline{K(s, x)} = K_1(s, x) - \mathrm{i}K_2(s, x)$, respectively). The kernel $\overline{K(s, x)}$ is called the *adjoint kernel to the kernel* $K(x, s)$.

Example 2. The equation

$$f(x) - \lambda \int_0^2 (x^3 + s) f(s)\, \mathrm{d}s = 0$$

is a Fredholm equation (for any value of λ). The equation

$$F(x) - \overline{\lambda} \int_0^2 (s^3 + x) F(s)\, \mathrm{d}s = 0$$

is the corresponding adjoint equation. Note the interchange of variables x and s in the kernels of the original and the adjoint equations.

Definition 4. Any number $\lambda = \lambda_0$ for which equation (3) possesses a *non-zero* (Remark 1) solution $\varphi(x)$ is called a *characteristic value* or a *characteristic number* or an *eigenvalue* of equation (3) (or of the kernel $K(x, s)$).The function $\varphi(x)$ is called the *characteristic function* or the *eigenfunction* associated with the number λ_0.

(The terms *characteristic value* and *characteristic number* are preferred to the term *eigenvalue* in the literature, bacause, in the operator theory, the latter is used in the case of equations of the form $Af - \lambda f = 0$, while we consider equations of the form $f - \lambda Af = 0$ here.)

FREDHOLM'S THEOREMS. *Let* (3) *be a Fredholm equation, i.e. let* $K \in L_2(Q)$. *Then:*

Theorem 1. *In any bounded part of the complex* λ*-plane, there exist only a finite number of characteristic values.*

Thus the only possible point of accumulation of the characteristic values is the point $\lambda = \infty$.

Theorem 2. *At least one characteristic function is associated with each characteristic value. The number of linearly independent* (Definition 12.8.2) *characteristic functions associated with a fixed characteristic value is finite.*

Theorem 3. *If* λ_0 *is a characteristic value of equation* (3) (*with kernel* $K(x, s)$), *then the complex conjugate number* $\overline{\lambda_0}$ *is a characteristic value of the adjoint equation*

$$F(x) - \overline{\lambda} \int_a^b \overline{K(s, x)} F(s)\, ds = 0. \tag{5}$$

Equations (3) *and* (5) *have the same number of linearly independent characteristic functions* (*corresponding to* λ_0 *and* $\overline{\lambda_0}$, *respectively*).

Theorem 4 (*Fredholm's Alternative*). *For the equation*

$$f(x) - \lambda \int_a^b K(x, s) f(s)\, ds = g(x) \tag{6}$$

there are only two possibilities:

either this equation possesses one and only one solution $f \in L_2(a, b)$ *for any* $g \in L_2(a, b)$ (*in particular,* $f(x) \equiv 0$ *is the only continuous solution for* $g(x) \equiv 0$),

or the corresponding homogeneous equation (3) *has a non-zero solution.*

REMARK 4. In the second case, λ is a characteristic value of the kernel $K(x, s)$ and hence (Theorem 3) $\overline{\lambda}$ is a characteristic value of the kernel $\overline{K(s, x)}$. Consequently, the equation

$$\psi(x) - \overline{\lambda} \int_a^b \overline{K(s, x)} \psi(s)\, ds = 0 \tag{7}$$

has a finite number (Theorem 3) of linearly independent characteristic functions. Let us denote them by

$$\psi_1(x), \psi_2(x), \ldots, \psi_k(x). \tag{8}$$

Then equation (6) is solvable precisely for those functions $g \in L_2(a, b)$ which are orthogonal to all functions (8), i.e. for which the relations

$$\int_a^b g(x)\overline{\psi_i(x)}\, dx = 0, \quad i = 1, 2, \ldots, k \tag{9}$$

hold.

In this case, equation (6) has obviously more than one solution: For if $f(x)$ is a solution of equation (6), corresponding to $g(x)$ ($g(x)$ satisfying conditions (9)) and if

$$\varphi_1(x), \varphi_2(x), \ldots, \varphi_k(x) \tag{10}$$

are linearly independent solutions of the homogeneous equation

$$\varphi(x) - \lambda \int_a^b K(x, s)\varphi(s)\, ds = 0, \tag{11}$$

then, obviously, the function

$$f(x) + c_1\varphi_1(x) + \ldots + c_k\varphi_k(x),$$

where $c_1, \ldots, c_k$ are arbitrary constants, is again a solution of equation (6).

REMARK 5. The Fredholm alternative is often applied: If it is known that equation (11) has (in the domain of square integrable functions) only the zero solution (as can often be expected because of the nature of the technical problem in question), then equation (6) has one and only one solution $f \in L_2(a, b)$ for each function $g \in L_2(a, b)$.

Theorem 5. *If $K(x, s)$ is continuous in $\overline{Q}$ and $g(x)$ is continuous in $[a, b]$ and if equation (6) is required to be satisfied at all points of the interval $[a, b]$, then its solutions (if they exist) are continuous functions in $[a, b]$.*

REMARK 6. A *system* of two integral equations

$$f_1(x) - \int_a^b K_{11}(x, s) f_1(s)\,ds - \int_a^b K_{12}(x, s) f_2(s)\,ds = g_1(x), \tag{12}$$

$$f_2(x) - \int_a^b K_{21}(x, s) f_1(s)\,ds - \int_a^b K_{22}(x, s) f_2(s)\,ds = g_2(x) \tag{13}$$

may be reduced to a single integral equation as follows:

Instead of the interval $[a, b]$, consider the interval $[a, 2b-a]$ whose length is double

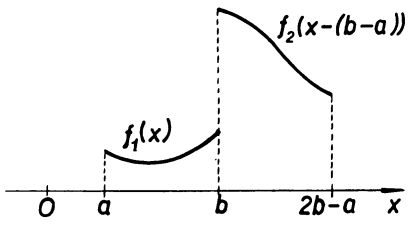


Fig. 19.1.

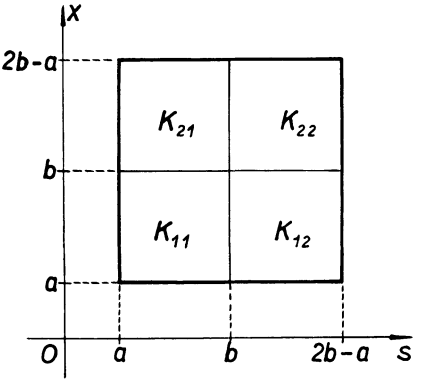


Fig. 19.2.

that of $[a, b]$ (Fig. 19.1) and define functions $F(x)$ and $G(x)$ by the formulae

$$F(x) = \begin{cases} f_1(x) & \text{for} \quad a \leqq x \leqq b, \\ f_2\big(x - (b-a)\big) & \text{for} \quad b < x \leqq 2b - a \end{cases}$$

($f_2\big(x - (b-a)\big)$ is the function $f_2(x)$ "shifted" by the distance $b - a$ to the right),

$$G(x) = \begin{cases} g_1(x) & \text{for} \quad a \leqq x \leqq b, \\ g_2\big(x - (b-a)\big) & \text{for} \quad b < x \leqq 2b - a. \end{cases}$$

Similarly (see Fig. 19.2):

$$K(x, s) = \begin{cases} K_{11}(x, s) & \text{for } a \leqq x \leqq b, \quad a \leqq s \leqq b, \\ K_{12}\big(x, s - (b-a)\big) & \text{for } a \leqq x \leqq b, \quad b < s \leqq 2b - a, \\ K_{21}\big(x - (b-a), s\big) & \text{for } b < x \leqq 2b - a, \ a \leqq s \leqq b, \\ K_{22}\big(x - (b-a), s - (b-a)\big) & \text{for } b < x \leqq 2b - a, \ b < s \leqq 2b - a. \end{cases}$$

In this notation, we may rewrite equations (12) and (13) in the form

$$F(x) - \int_a^b K(x, s) F(s)\,ds - \int_b^{2b-a} K(x, s) F(s)\,ds = G(x) \quad (a \leqq x \leqq b),$$

$$F(x) - \int_a^b K(x, s) F(s)\,ds - \int_b^{2b-a} K(x, s) F(s)\,ds = G(x) \quad (b < x \leqq 2b - a)$$

or

$$F(x) - \int_a^{2b-a} K(x, s)F(s)\,\mathrm{d}s = G(x) \quad (a \leqq x \leqq 2b - a). \tag{14}$$

The equation (14) and the system of equations (12) and (13) are equivalent. If $K_{ij} \in L_2(Q)$ $(i = 1, 2;\ j = 1, 2)$, then $K \in L_2(R)$ (R is the square $a \leqq x \leqq 2b - a$, $a \leqq s \leqq 2b - a$) and equation (14) is a Fredholm equation, so that Theorems 1 – 4 may be applied. Coming back to the system (12), (13) and using, in particular, Theorem 4 we conclude that:

If $K_{ij} \in L_2(Q)$ $(i = 1, 2;\ j = 1, 2)$, then either the system (12), (13) possesses a non-zero solution for zero right-hand sides, or to each pair of functions $g_1 \in \in L_2(a, b)$, $g_2 \in L_2(a, b)$ there exists one and only one pair of functions $f_1 \in \in L_2(a, b)$, $f_2 \in L_2(a, b)$ which constitute the solution of the system (12), (13).

19.2. Equations with Degenerate Kernels

The equation of the form

$$f(x) - \int_a^b \Big[\sum_{k=1}^{n} a_k(x)b_k(s)\Big] f(s)\,\mathrm{d}s = g(x) \tag{1}$$

is said to have a *degenerate kernel.* Thus, such a kernel is a finite sum of products, the factors of which are functions of only one variable x or s, respectively. The functions $a_k(x)$ may be assumed to be linearly independent (otherwise the number of terms of the kernel could be reduced).

Theorem 1. *Any solution of equation* (1) *is of the form*

$$f(x) = g(x) + \sum_{k=1}^{n} c_k a_k(x). \tag{2}$$

In fact,

$$\sum_{k=1}^{n} \int_a^b a_k(x)b_k(s)f(s)\,\mathrm{d}s = \sum_{k=1}^{n} a_k(x) \int_a^b b_k(s)f(s)\,\mathrm{d}s = \sum_{k=1}^{n} c_k a_k(x).$$

Substituting (2) into (1) and comparing coefficients of linearly independent functions $a_k(x)$, we obtain a system of n linear algebraic equations for the n unknown coefficients c_k. *This system is solvable if and only if the corresponding integral equation is solvable.*

Example 1.

$$f(x) - \int_0^1 6(x+s)f(s)\,\mathrm{d}s = x^2. \tag{3}$$

According to (2), the solution is of the form

$$f(x) = x^2 + c_1 x + c_2 \tag{4}$$

(because $a_1(x) = x$, $a_2(x) = 1$). Substituting (4) into (3) we obtain

$$x^2 + c_1 x + c_2 - \int_0^1 6(x+s)(s^2 + c_1 s + c_2)\,\mathrm{d}s = x^2,$$
$$(c_1 - 2 - 3c_1 - 6c_2)x + (c_2 - \tfrac{3}{2} - 2c_1 - 3c_2) = 0.$$

Comparing the coefficients of corresponding powers of x, we find that

$$-2c_1 - 6c_2 - 2 = 0, \quad -2c_1 - 2c_2 - \tfrac{3}{2} = 0,$$

from which it follows that

$$c_1 = -\tfrac{5}{8}, \quad c_2 = -\tfrac{1}{8},$$

hence

$$f(x) = x^2 - \tfrac{5}{8}x - \tfrac{1}{8}.$$

Example 2.

$$f(x) - \lambda \int_0^1 x f(s)\,\mathrm{d}s = 0. \tag{5}$$

Here $a_1(x) = x$, $b_1(s) = \lambda$. So by (2) the solution is of the form

$$f(x) = cx. \tag{6}$$

Substituting (6) into (5), we get

$$cx - \lambda \int_0^1 x \,.\, cs\,\mathrm{d}s = 0,$$

or

$$cx - \tfrac{1}{2}\lambda cx = 0. \tag{7}$$

If $\lambda \neq 2$, then $c = 0$ and (5) has as its only solution zero. (Thus by Fredholm alternative (Theorem 19.1.4) the corresponding *non-homogeneous* equation has for $\lambda \neq 2$ exactly one solution for an arbitrary right-hand side $g \in L_2(0, 1)$).

If $\lambda = 2$, then (7) is satisfied for any c; $\lambda = 2$ is the (only) characteristic value of equation (5). The corresponding characteristic functions are of the form

$$f(x) = cx \quad (c = \text{const.} \neq 0),$$

each of them being thus a multiple of x; so there is only one linearly independent characteristic function corresponding to the value $\lambda = 2$.

The equation adjoint to equation (5) is of the form (Definition 19.1.3)

$$\psi(x) - \overline{\lambda} \int_0^1 s \,.\, \psi(s)\, \mathrm{d}s = 0. \tag{8}$$

According to Theorem 19.1.3, $\lambda = 2$ is the (only) characteristic value of equation (8). The solution of the equation

$$\psi(x) - 2 \int_0^1 s \,.\, \psi(s)\, \mathrm{d}s = 0 \tag{9}$$

is of the form

$$\psi(x) = a \quad (a = \text{const.}) \tag{10}$$

and, as can easily be verified by substitution, is the solution of (9) for arbitrary a. Solutions of another type do not exist (for, by Theorem 1, any solution of (9) must be of the form (10)). According to Remark 19.1.4, the equation

$$f(x) - 2 \int_0^1 x f(s)\, \mathrm{d}s = g(x) \tag{11}$$

has a solution if and only if the function $g(x)$ is orthogonal to the functions (10), i.e. if and only if

$$\int_0^1 a \,.\, g(x)\, \mathrm{d}x = 0$$

holds for any a, i.e. if and only if

$$\int_0^1 g(x)\, \mathrm{d}x = 0. \tag{12}$$

If we find such a solution, we obtain all other solutions by adding an arbitrary multiple of the solution cx of the corresponding homogeneous equation

$$f(x) - 2 \int_0^1 x f(s)\, \mathrm{d}s = 0$$

(Remark 19.1.4).

Let us choose $g(x) = 1 - 2x$, satisfying evidently the condition (12). For this function, equation (11), i.e. the equation

$$f(x) - 2 \int_0^1 x f(s)\, \mathrm{d}s = 1 - 2x,$$

has obviously a solution

$$f(x) = 1 - 2x,$$

because

$$\int_0^1 xf(s)\,\mathrm{d}s = x\int_0^1 (1-2s)\,\mathrm{d}s = 0.$$

All solutions of the equation (11) (with $g(x) = 1 - 2x$) will then be of the form

$$f(x) = 1 - 2x + c_1 x, \tag{13}$$

where c_1 is an arbitrary constant.

19.3. Equations with Symmetric Kernels

The kernel $K(x, s)$ (or the corresponding integral equation) is said to be *symmetric* if

$$K(x, s) = \overline{K(s, x)}, \tag{1}$$

i.e. if the kernel equals the function which is obtained by interchanging the variables in $K(x, s)$ and by taking the complex conjugate value. In particular, if the kernel is real (the most frequent case in applications), then the symmetry is expressed by the relation

$$K(x, s) = K(s, x). \tag{2}$$

Example 1. The kernels $x + s$, $\mathrm{i}(x - s)$ are symmetric, the kernels $x^2 + s$, $\mathrm{i}(x + s)$ are not symmetric (for example, in the last case we have

$$K(x, s) = -\overline{K(s, x)}$$

since $\bar{\mathrm{i}} = -\mathrm{i}$).

For a *Fredholm equation*

$$f(x) - \lambda\int_a^b K(x, s)f(s)\,\mathrm{d}s = 0 \tag{3}$$

with *symmetric* kernel the following assertions are true:

Theorem 1. *If $K(x, s)$ is not a zero function* (Remark 19.1.1), *then there exists at least one characteristic value of equation* (3).

Theorem 2. *All characteristic values of equation* (3) *are real* (*even if $K(x, s)$ is not real*).

Theorem 3. *The maximum of the expression*

$$\left| \iint_Q K(x, s)\varphi(x)\overline{\varphi(s)}\,dx\,ds \right|, \tag{4}$$

taken over all functions φ, which are square integrable and normalized, i.e. for which the relation

$$\int_a^b |\varphi^2(x)|\,dx = 1 \tag{5}$$

holds, is equal to $|1/\lambda_1|$, where λ_1 is the characteristic value with smallest modulus. The function $\varphi_1(x)$ which maximizes expression (4) is a characteristic function corresponding to this characteristic value λ_1.

Theorem 4. *The characteristic functions associated with different characteristic values are orthogonal,* i.e.

$$\int_a^b \varphi_m(x)\overline{\varphi_n(x)}\,dx = 0, \quad \text{if} \quad \lambda_m \neq \lambda_n. \tag{6}$$

REMARK 1. By Theorem 19.1.2, to each characteristic value there corresponds a finite number of linearly independent characteristic functions

$$\varphi_1(x), \varphi_2(x), \ldots, \varphi_k(x). \tag{7}$$

We shall assume in the sequel that the functions (7) are *normalized* (i.e. that they satisfy the relation (5)) and *orthogonal*, i.e. that they fulfil

$$\int_a^b \varphi_i(x)\overline{\varphi_j(x)}\,dx = 0 \quad \text{if} \quad i \neq j. \tag{8}$$

The orthogonalization process is described in Remark 16.2.15.

It is convenient to number the characteristic values in such a way that to every characteristic function (7) there corresponds exactly one characteristic value; then to the k linearly independent functions (7) there correspond k characteristic values $\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(k)}$, which of course are all equal, i.e. $\lambda^{(1)} = \lambda^{(2)} = \ldots = \lambda^{(k)}$. Let the characteristic values corresponding to all normalized and mutually orthogonal characteristic functions of equation (3) be numbered in such a manner that their absolute values form a non-decreasing sequence, i.e.

$$|\lambda_1| \leqq |\lambda_2| \leqq |\lambda_3| \leqq \ldots. \tag{9}$$

Theorem 5. *Let $\varphi_1(x), \ldots, \varphi_n(x)$ be characteristic functions corresponding to characteristic values $\lambda_1, \ldots, \lambda_n$. Let $\varphi(x)$ be any function satisfying the condition* (5) *and also the conditions*

$$\int_a^b \varphi(x)\overline{\varphi_1(x)}\,dx = 0, \ldots, \quad \int_a^b \varphi(x)\overline{\varphi_n(x)}\,dx = 0. \tag{10}$$

Then the absolute value of the characteristic value λ_{n+1} *is equal to the reciprocal of the maximum value of the integral* (4) *in the class of functions* $\varphi(x)$ *with the above properties.*

REMARK 2. To determine $|\lambda_1|$ we use (4), and we find also the characteristic functions $\varphi_1(x)$ corresponding to λ_1. To find $|\lambda_2|$ we again look for the maximum of (4), but among all functions φ which are normalized in $[a, b]$ we consider only those that are orthogonal to $\varphi_1(x)$. Proceeding in this way, we find successively $|\lambda_3|$, $|\lambda_4|$, etc.

Theorems 3 and 5 are of importance for the practical evaluation of characteristic values, because they make possible to apply variational methods (see e.g. § 29.5).

Theorem 6 (*The Hilbert–Schmidt Theorem*). *Equation* (3) *has a finite or countable set of normalized and mutually orthogonal characteristic functions*

$$\varphi_1(x), \varphi_2(x), \ldots, \varphi_n(x), \ldots, \tag{11}$$

which correspond to the characteristic values $\lambda_1, \lambda_2, \ldots, \lambda_n, \ldots$ $(|\lambda_1| \leqq |\lambda_2| \leqq \ldots \leqq |\lambda_n| \leqq \ldots)$. *For each function* $f(x)$ *which can be expressed in the form*

$$f(x) = \int_a^b K(x, s)h(s)\,\mathrm{d}s, \quad h(s) \in L_2(a, b), \tag{12}$$

the relation

$$\lim_{k\to\infty} \int_a^b \left[f(x) - \sum_{n=1}^{k} a_n\varphi_n(x)\right]^2 \mathrm{d}x = 0$$

holds, where

$$a_n = \int_a^b f(x)\overline{\varphi_n(x)}\,\mathrm{d}x, \tag{13}$$

i.e. the series

$$\sum_{n=1}^{\infty} a_n\varphi_n(x) \tag{14}$$

converges in the mean to the function $f(x)$.

If, in addition, the integral

$$\int_a^b |K(x, s)|^2\,\mathrm{d}s \tag{15}$$

is bounded by the same constant for all $x \in [a, b]$, *then the series* (15) *converges to* $f(x)$ *absolutely and uniformly in* $[a, b]$.

The coefficients a_n can be expressed in terms of the characteristic values λ_n and of the function $h(x)$ as follows:

$$a_n = \frac{h_n}{\lambda_n}, \quad \text{where} \quad h_n = \int_a^b h(x)\overline{\varphi_n(x)}\,dx. \tag{16}$$

19.4. The Resolvent

Definition 1. The number λ is said to be *regular* for the equation

$$f(x) - \lambda \int_a^b K(x, s)f(s)\,ds = 0 \tag{1}$$

if λ is not a characteristic value of that equation.

Definition 2. If for each regular λ and for an arbitrary function $g \in L_2(a, b)$ the solution of the equation

$$f(x) - \lambda \int_a^b K(x, s)f(s)\,ds = g(x) \tag{2}$$

can be written in the form

$$f(x) = g(x) + \lambda \int_a^b \Gamma(x, s, \lambda)g(s)\,ds, \tag{3}$$

then the function

$$\Gamma(x, s, \lambda) \tag{4}$$

is called the *resolvent* of equation (2).

Example 1. The solution of the equation with degenerate kernel

$$f(x) - \lambda \int_0^1 (x + s)f(s)\,ds = g(x) \tag{5}$$

is, in view of equation (19.2.2), of the form

$$f(x) = g(x) + c_1 x + c_2. \tag{6}$$

Substituting (6) into (5) and equating corresponding coefficients, we get

$$c_1 = \frac{\lambda\left(1 - \frac{1}{2}\lambda\right)\int_0^1 g(s)\,ds + \lambda^2 \int_0^1 sg(s)\,ds}{1 - \lambda - \frac{1}{12}\lambda^2},$$

$$c_2 = \frac{\frac{1}{3}\lambda^2 \int_0^1 g(s)\,ds + \lambda\left(1 - \frac{1}{2}\lambda\right)\int_0^1 sg(s)\,ds}{1 - \lambda - \frac{1}{12}\lambda^2}. \tag{7}$$

Substitution of these values of c_1 and c_2 into (6) yields

$$f(x) = g(x) + \int_0^1 \lambda \frac{(12 - 6\lambda)x + 4\lambda + s(12\lambda x + 12 - 6\lambda)}{12 - 12\lambda - \lambda^2} g(s)\,ds. \tag{8}$$

Hence

$$\Gamma(x, s, \lambda) = \frac{(12 - 6\lambda)x + 4\lambda + s(12\lambda x + 12 - 6\lambda)}{12 - 12\lambda - \lambda^2}. \tag{9}$$

(Cf. Example 19.2.1, where $\lambda = 6$, $g(x) = x^2$.)

For a Fredholm equation (2) the following theorems are valid:

Theorem 1. *For equation* (2) *there exists a resolvent for every regular* λ.

Theorem 2. *The resolvent is a meromorphic function* (see Remark 20.4.10) *of the complex variable* λ *in the entire* λ*-plane. The characteristic values are poles of the resolvent.*

Theorem 3. *If* λ *is regular, then*

$$\Gamma(x, s, \lambda) = \frac{D(x, s, \lambda)}{D(\lambda)}, \tag{10}$$

where

$$D(x, s, \lambda) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} B_n(x, s)\lambda^n, \tag{11}$$

$$D(\lambda) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} c_n \lambda^n \tag{12}$$

and

$$B_0(x, s) = K(x, s), \tag{13}$$

$$B_n(x, s) = \int_a^b \cdots \int_a^b \Delta_n\,dt_1\,dt_2 \ldots dt_n, \tag{14}$$

$$\Delta_n(x, s, t_1, t_2, \ldots, t_n) = \begin{vmatrix} K(x, s), & K(x, t_1), & K(x, t_2), & \ldots, & K(x, t_n) \\ K(t_1, s), & K(t_1, t_1), & K(t_1, t_2), & \ldots, & K(t_1, t_n) \\ K(t_2, s), & K(t_2, t_1), & K(t_2, t_2), & \ldots, & K(t_2, t_n) \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots \\ K(t_n, s), & K(t_n, t_1), & K(t_n, t_2), & \ldots, & K(t_n, t_n) \end{vmatrix}, \tag{15}$$

$$c_0 = 1, \tag{16}$$

$$c_n = \int_a^b \cdots \int_a^b \begin{vmatrix} K(t_1, t_1), & K(t_1, t_2), & \ldots, & K(t_1, t_n) \\ K(t_2, t_1), & K(t_2, t_2), & \ldots, & K(t_2, t_n) \\ \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots \\ K(t_n, t_1), & K(t_n, t_2), & \ldots, & K(t_n, t_n) \end{vmatrix} dt_1\,dt_2 \ldots dt_n. \tag{17}$$

The series (11), (12) *are convergent for all values of* λ.

REMARK 1. If

$$|\lambda| \leqq \frac{1}{C}, \quad \text{where} \quad C^2 = \int_a^b \int_a^b |K^2(x, s)| \, dx \, ds,$$

then there is another expression for the resolvent, namely

$$\Gamma(x, s, \lambda) = \sum_{m=1}^{\infty} K_m(x, s)\lambda^{m-1}, \tag{18}$$

where $K_m(x, s)$ is the so-called *m-th iterated kernel* (corresponding to the kernel $K(x, s)$) given by the recurrence formula

$$K_m(x, s) = \int_a^b K(x, t)K_{m-1}(t, s) \, dt \quad (m \geqq 2), \tag{19}$$
$$K_1(x, s) = K(x, s).$$

REMARK 2. The functions (14) and (17) can be evaluated by applying the following recurrence formulae:

$$c_{n+1} = \int_a^b B_n(s, s) \, ds, \tag{20}$$

$$B_n(x, s) = c_n K(x, s) - n \int_a^b K(x, t)B_{n-1}(t, s) \, dt. \tag{21}$$

Example 2. Let us determine the resolvent of equation (5). From (16) and (13) $c_0 = 1$, $B_0(x, s) = x + s$. Using (20) and (21) we find

$$c_1 = \int_0^1 2s \, ds = 1,$$

$$B_1(x, s) = x + s - \int_0^1 (x + t)(t + s) \, dt = \tfrac{1}{2}(x + s) - xs - \tfrac{1}{3},$$

$$c_2 = \int_0^1 \left(s - s^2 - \tfrac{1}{3}\right) ds = -\tfrac{1}{6},$$

$$B_2(x, s) = -\tfrac{1}{6}(x + s) - 2 \int_0^1 (x + t)\left[\tfrac{1}{2}(t + s) - ts - \tfrac{1}{3}\right] dt = 0.$$

It follows readily from (20) and (21) that $c_3 = 0$, $B_3(x, s) \equiv 0$, $c_4 = 0$, $B_4(x, s) \equiv$ $\equiv 0, \ldots$. Hence, according to (11), (12) and (10) the resolvent of equation (5) is

$$\Gamma(x, s, \lambda) = \frac{x + s - \left[\frac{1}{2}(x + s) - xs - \frac{1}{3}\right]\lambda}{1 - \lambda - \frac{1}{12}\lambda^2} \tag{22}$$

in agreement with (9). (This example is used only to illustrate the underlying idea; in fact, the equation has a degenerate kernel and can be solved by the procedure of § 19.2)

If λ is sufficiently small (more exactly, for $|\lambda| \leqq \sqrt{\frac{6}{7}}$ in view of Remark 1 and of the relation $C^2 = \int_0^1 \int_0^1 (x+s)^2 \, dx \, ds = \frac{7}{6}$), then the resolvent of equation (5) can be expressed in the form (18). We have

$$K_1(x, s) = K(x, s) = x + s,$$

$$K_2(x, s) = \int_0^1 (x+t)(t+s) \, dt = \tfrac{1}{2}(x+s) + xs + \tfrac{1}{3},$$

$$K_3(x, s) = \int_0^1 (x+t)\left[\tfrac{1}{2}(t+s) + ts + \tfrac{1}{3}\right] dt = xs + \tfrac{7}{12}x + \tfrac{7}{12}s + \tfrac{1}{3}$$

so that in view of (18)

$$\Gamma(x, s, \lambda) = x + s + \left[\tfrac{1}{2}(x+s) + xs + \tfrac{1}{3}\right]\lambda + \left[xs + \tfrac{7}{12}x + \tfrac{7}{12}s + \tfrac{1}{3}\right]\lambda^2 + \ldots. \quad (23)$$

We can obtain the same formula from (22) if we write

$$\frac{1}{1 - \lambda - \frac{1}{12}\lambda^2} = 1 + \left(\lambda + \tfrac{1}{12}\lambda^2\right) + \left(\lambda + \tfrac{1}{12}\lambda^2\right)^2 + \ldots,$$

and consider terms up to and including λ^2.

Having found the resolvent (22), we can easily find the solution of equation (5) for any given (regular) λ and any given $g(x)$; e.g. if $\lambda = 6$, $g(x) = x^2$, then by virtue of (3) we get

$$f(x) = x^2 + 6 \int_0^1 \frac{x + s - 6\left[\frac{1}{2}(x+s) - xs - \frac{1}{3}\right]}{1 - 6 - 3} s^2 \, ds = x^2 - \tfrac{5}{8}x - \tfrac{1}{8},$$

in agreement with example 19.2.1.

Theorem 4. *If $K(x, s)$ possesses continuous partial derivatives of the first order in $\overline{Q}$, then for every regular λ the resolvent $\Gamma(x, s, \lambda)$ has continuous partial derivatives of the first order with respect to x and s.*

REMARK 3. From the form of (3) the continuous dependence* of the solution on $g(x)$ clearly follows.

* The solution $f(x)$ of (2) is "continuously dependent" on $g(x)$ if, roughly speaking, a slight change in $g(x)$ causes a slight change in $f(x)$ for λ fixed and regular.

19.5. Equations Involving Weak Singularities. Singular Equations

Definition 1. Equations of the form

$$f(x) - \lambda \int_a^b \frac{H(x, s)}{|x - s|^{\alpha}} f(s)\, ds = g(x), \tag{1}$$

where $H(x, s)$ is a bounded (integrable) function, $0 < \alpha < 1$, are said to have a *weak singularity.* (If $\alpha < \frac{1}{2}$, we have Fredholm's equations.)

Theorem 1. *All iterated kernels* (Remark 19.4.1) *of equation* (1) *starting from a certain kernel are bounded.*

Theorem 2. *All four of Fredholm's theorems* (Theorems 19.1.1 – 19.1.4) *hold for equation* (1).

REMARK 1. It can be shown that Fredholm's theorems remain valid not only for equations involving a weak singularity but, more generally, for any equation whose iterated kernels, starting from a certain kernel, are bounded.

REMARK 2. Theorems 1 and 2 hold also for equations in several variables with a weak singularity. For two independent variables such equations are of the form

$$f(x_1, x_2) - \lambda \iint_{\Omega} \frac{H(x_1, x_2, x_3, x_4)}{r^{\alpha}} f(x_3, x_4)\, dx_3\, dx_4 = g(x_1, x_2), \tag{2}$$

where Ω is a bounded two-dimensional region, $H(x_1, x_2, x_3, x_4)$ is a bounded integrable function, r is the distance between the points (x_1, x_2) and (x_3, x_4), $0 < \alpha < 2$. The equation corresponding to (2) for the case of n variables is of similar form, with $0 < \alpha < n$. For details see e.g. [323].

Definition 2. Integral equations of the form

$$f(x) - \lambda \int_a^b \frac{A(x, s)}{x - s} f(s)\, ds = g(x) \tag{3}$$

where $A(x, s)$ is a differentiable function of the variables x and s, are called *singular integral equations.*

REMARK 3. The integral appearing in (3) may be divergent, in general. If it is considered in the sense of its principal value (Remark 13.8.3), an extensive theory for equation (3) can be established. For details see [342]. We shall merely consider two typical cases:

A. The *equation with the so-called Hilbert kernel,*

$$af(x) + \frac{b}{2\pi} \int_0^{2\pi} \cot \frac{1}{2}(x - s) f(s)\, ds = g(x), \tag{4}$$

possesses for $a \neq 0$, $a^2 + b^2 \neq 0$ (a, b may be complex) the following solution:

$$f(x) = \frac{a}{a^2 + b^2} g(x) - \frac{b}{2\pi(a^2 + b^2)} \int_0^{2\pi} g(s) \cot \frac{1}{2}(s - x)\, ds + \\ + \frac{b^2}{2\pi a(a^2 + b^2)} \int_0^{2\pi} g(s)\, ds. \tag{5}$$

If $a = 0$, $b \neq 0$, equation (4) becomes an equation of the first kind (see § 19.7) and has a solution if and only if

$$\int_0^{2\pi} g(s)\, ds = 0. \tag{6}$$

The solution is then

$$f(x) = -\frac{1}{2\pi b} \int_0^{2\pi} g(s) \cot \frac{1}{2}(s - x)\, ds + C \tag{7}$$

where C is an arbitrary constant.

When $a^2 + b^2 = 0$, then equation (4) cannot, in general, be solved.

B. The *equation with the so-called Cauchy kernel* is of the form

$$af(z) + \frac{b}{\pi i} \int_c \frac{f(t)}{t - z}\, dt = g(z), \tag{8}$$

where c is a simple curve in the complex plane, piecewise smooth, closed and positively oriented with respect to its interior V (Remark 14.7.1), while $g(z)$, or $f(z)$ is a function of the complex variable $z = x + iy$, given, or to be found as the function of the point z on c, respectively. If $a^2 - b^2 \neq 0$, then equation (8) has the solution

$$f(z) = \frac{a}{a^2 - b^2} g(z) - \frac{b}{(a^2 - b^2)\pi i} \int_c \frac{g(t)}{t - z}\, dt. \tag{9}$$

REMARK 4. Other kinds of singularities may occur when the given interval is infinite and is transformed to a finite one. For example, the interval $(0, \infty)$ is transformed by the substitution $x = t/(1 - t)$ into the interval $(0, 1)$; using this, simultaneously with the substitution $s = \sigma/(1 - \sigma)$, the new kernel (which is a function of t and σ) may become unbounded. In such cases, for existence theorems and also for construction of approximate solutions, it is often convenient to employ the method of successive approximations.

19.6. Equations of Volterra Type

These equations are of the form

$$f(x) - \lambda \int_a^x K(x, s) f(s)\, \mathrm{d}s = g(x) \quad (a \leqq x \leqq b), \tag{1}$$

where the kernel $K(x, s)$ is bounded and integrable. Thus these equations are special cases of equations of the form

$$f(x) - \lambda \int_a^b K(x, s) f(s)\, \mathrm{d}s = g(x)$$

if $K(x, s)$ is bounded and equal to zero for $x < s \leqq b$.

Theorem 1. *If $g(x)$ is absolutely integrable, then equation (1) possesses one and only one solution for every λ. The solution may be obtained as the limit of a uniformly convergent sequence of successive approximations*

$$\begin{aligned} f_0(x) &= g(x), \\ f_1(x) &= g(x) + \lambda \int_a^x K(x, s) f_0(s)\, \mathrm{d}s, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ f_{n+1}(x) &= g(x) + \lambda \int_a^x K(x, s) f_n(s)\, \mathrm{d}s, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{aligned} \tag{2}$$

Theorem 2. *Let $|K(x, s)| < M$. Then*

$$|f_{n+1}(x) - f_n(x)| \leqq \frac{|\lambda|^n M^n (b-a)^{n-1}}{(n-1)!} \int_a^b |g(s)|\, \mathrm{d}s. \tag{3}$$

REMARK 1. The inequality (3) is of importance for estimating the error, when using the method of successive approximations.

REMARK 2. Volterra's equation can often be reduced to a differential equation (and vice versa).

Example 1. The equation

$$f(x) - \lambda \int_0^x \mathrm{e}^{x-s} f(s)\, \mathrm{d}s = g(x) \tag{4}$$

differentiated with respect to x ($g(x)$ is assumed to be differentiable) gives

$$f'(x) - \lambda\, \mathrm{e}^{x-x} f(x) - \lambda \int_0^x \mathrm{e}^{x-s} f(s)\, \mathrm{d}s = g'(x). \tag{5}$$

Eliminating the integral between (4) and (5) we obtain a simple differential equation for $f(x)$

$$f'(x) - (\lambda + 1)f(x) = g'(x) - g(x). \tag{6}$$

If we prescribe for the solution of this equation the condition $f(0) = g(0)$, as follows from (4) by setting $x = 0$, then the solution of equation (6) is the solution of equation (4), and conversely.

Example 2. Let us transform the problem

$$y'' + a(x)y = f(x), \quad y(0) = y_0, \quad y'(0) = y_0', \tag{7}$$

into an integral equation of the Volterra type.

On integrating equation (7) twice we get

$$\int_0^x \mathrm{d}u \int_0^u y''(t)\,\mathrm{d}t + \int_0^x \mathrm{d}u \int_0^u a(t)y(t)\,\mathrm{d}t = \int_0^x \mathrm{d}u \int_0^u f(t)\,\mathrm{d}t. \tag{8}$$

Further

$$\int_0^u y''(t)\,\mathrm{d}t = y'(u) - y'(0), \quad \int_0^x [y'(u) - y'(0)]\,\mathrm{d}u = y(x) - y(0) - xy'(0). \tag{9}$$

According to Cauchy-Dirichlet's formula (17.7.4),

$$\int_0^x \mathrm{d}u \int_0^u a(t)y(t)\,\mathrm{d}t = \int_0^x (x - u)a(u)y(u)\,\mathrm{d}u, \tag{10}$$

$$\int_0^x \mathrm{d}u \int_0^u f(t)\,\mathrm{d}t = \int_0^x (x - u)f(u)\,\mathrm{d}u = F(x). \tag{11}$$

Putting (9), (10), (11) into (8), we get

$$y(x) - y(0) - xy'(0) + \int_0^x (x - u)a(u)y(u)\,\mathrm{d}u = F(x)$$

or

$$y(x) + \int_0^x (x - u)a(u)y(u)\,\mathrm{d}u = F(x) + y_0 + xy_0'. \tag{12}$$

19.7. Integral Equations of the First Kind

Integral equations of the form

$$\int_a^b K(x, s)f(s)\,\mathrm{d}s = g(x), \tag{1}$$

where $f(x)$ is the unknown function, are called *integral equations of the first kind.*

Generally speaking, equation (1) has no solution.

Example 1. The equation

$$\int_0^1 xf(s)\,\mathrm{d}s = x^2 \tag{2}$$

does not possess a solution; for any function $f(s)$, the left-hand side is of the form kx and this is evidently not identically equal to x^2 in the interval [0, 1] for any k. Clearly it is easy to decide whether an equation of the first kind with degenerate kernel is solvable or not.

Equations of the first kind are not encountered in practice so often as equations of the second kind. Equations of the first kind are extensively dealt with in [415].

The following equation, known as *Abel's integral equation*, is of importance:

$$\int_0^x \frac{f(s)}{(x-s)^\alpha}\,\mathrm{d}s = g(x) \quad (0 < \alpha < 1).$$

The solution is

$$f(x) = \frac{\sin \alpha\pi}{\pi}\left(\frac{g(0)}{x^{1-\alpha}} + \int_0^x \frac{g'(s)}{(x-s)^{1-\alpha}}\,\mathrm{d}s\right).$$

In particular, for the case $\alpha = \frac{1}{2}$ (which frequently occurs), this becomes

$$\int_0^x \frac{f(s)}{\sqrt{(x-s)}}\,\mathrm{d}s = g(x), \quad f(x) = \frac{1}{\pi}\left(\frac{g(0)}{\sqrt{x}} + \int_0^x \frac{g'(s)}{\sqrt{(x-s)}}\,\mathrm{d}s\right).$$

20. FUNCTIONS OF ONE AND MORE COMPLEX VARIABLES

References: [3], [4], [12], [22], [66], [70], [78], [91], [128], [146], [158], [163], [187], [197], [213], [221], [227], [253], [271], [296], [309], [313], [347], [381], [383], [408], [447], [477], [494].

A. FUNCTIONS OF ONE COMPLEX VARIABLE

By KAREL REKTORYS

20.1. Fundamental Concepts. Limit and Continuity. The Derivative. The Cauchy-Riemann Equations. Applications of the Theory of Functions of One Complex Variable.

REMARK 1. A complex number $z = x + iy$ is often represented by a point with coordinates x, y in the so-called complex (or Gaussian) plane (see § 1.6). We usually speak of the point z instead of the number z. If we speak of a region G of complex numbers, we mean the corresponding region of points (x, y) in the plane. A similar meaning is to be given to the statement "the point z lies on the curve c", etc.

On operations with complex numbers see § 1.6 and § 1.21.

REMARK 2. Complex numbers can also be represented on the so-called *Riemann sphere* touching the Gaussian plane at the origin (which is taken as the "south pole" of the sphere). If we join an arbitrary point of the Gaussian plane to the "north pole" of the sphere (as centre of projection), then there is a one-to-one correspondence between the points z and z' of the Gaussian plane and of the spherical surface, respectively (the so-called *stereographic projection*, Fig. 20.1). To the "north pole" there corresponds the point $z = \infty$. In this sense, there exists *exactly one* point at infinity, $z = \infty$.

The (Gaussian) plane of complex numbers together with the point $z = \infty$ is called the *closed* (or *completed* or *extended*) *plane of complex numbers* (the *closed plane*, in brief).

Definition 1. Let M be a set of complex numbers in the Gaussian plane (in the sense of Remark 1). If a relationship is given, by virtue of which to every point $z \in M$ there corresponds one and only one number w (complex, in general), we say that a *function*

$$w = f(z) \tag{1}$$

is defined on M. The set M is called the *domain of definition of the function* (1).

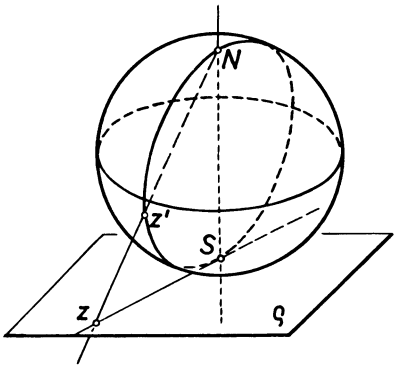


Fig. 20.1.

REMARK 3. The function (1) can also be interpreted in the following way: to every point $(x, y) \in M$ (where $z = x + iy$) there corresponds a complex number $w = u + iv$, i.e.

$$f(z) = u(x, y) + iv(x, y). \tag{2}$$

Thus, the investigation of functions of a complex variable can be reduced to the investigation of two functions u, v of the real variables x, y.

Example 1.

$$f(z) = z^2 = (x + iy)^2 = x^2 - y^2 + i \cdot 2xy. \tag{3}$$

Thus

$$u(x, y) = x^2 - y^2, \quad v(x, y) = 2xy. \tag{4}$$

REMARK 4. We write briefly

$$\operatorname{Re} f(z) = u \quad \text{or} \quad \mathrm{R}[f(z)] = u;$$
$$\operatorname{Im} f(z) = v \quad \text{or} \quad \mathrm{I}[f(z)] = v$$

(and call these functions the *real part* and the *imaginary part of the function* $f(z)$). In particular,

$$\operatorname{Re} z = x, \quad \operatorname{Im} z = y.$$

REMARK 5. A function $w = f(z)$ can be interpreted geometrically as a mapping from one complex plane into another (see also "conformal mapping", Chap. 21).

Definition 2. By a *δ-neighbourhood of a point z_0* we mean all the points of the complex plane whose distance from z_0 is less than δ, i.e. all the points z satisfying

$$|z - z_0| < \delta .$$

(Note that the relation $|z - z_0| = \delta$ characterizes all points whose distance from z_0 is equal to δ, i.e. the circumference of the circle with centre z_0 and radius δ.)

Definition 3. We say that $f(z)$ has a *limit* A (A being a complex number, in general) *at the point* z_0 (which need not belong to the domain of definition M), or that $f(z)$ *tends to the limit* A for $z \to z_0$, if for an arbitrary $\varepsilon > 0$, there exists a $\delta > 0$ such that the inequality

$$|f(z) - A| < \varepsilon$$

holds for all $z \neq z_0$ of the δ-neighbourhood of z_0, i.e. for all z satisfying

$$0 < |z - z_0| < \delta .$$

We write

$$\lim_{z \to z_0} f(z) = A .$$

REMARK 6. This definition is equivalent to the following one: $f(z)$ has a limit A at z_0 if the relation

$$\lim_{n \to \infty} f(z_n) = A$$

holds for *every* sequence of points $z_n \neq z_0$ converging to z_0.

Definition 4. If $z_0 \in M$ and if

$$\lim_{z \to z_0} f(z) = f(z_0) , \tag{5}$$

we say that $f(z)$ is *continuous at* z_0.

If $f(z)$ if continuous at every point of a region G, we say that $f(z)$ is *continuous in* G.

Definition 5. If $z_0 \in M$ is *not* an interior point (Definition 22.1.1) of the domain of definition M of the function $f(z)$ and if (5) holds for $z \in M$ (i.e. $|f(z) - f(z_0)| < \varepsilon$ for $|z - z_0| < \delta$ and $z \in M$), we say that $f(z)$ is *continuous at* z_0 *with respect to* M.

REMARK 7. This occurs especially in connection with continuity at boundary points of the domain of definition, or continuity on a given curve, if $f(z)$ is defined only at points of this curve, etc.

Definition 6. If $f(z)$ is continuous in a region G and continuous with respect to the closed region $\overline{G}$ ($\overline{G} = G \cup c$) at every point of the boundary c, we say that $f(z)$ is *continuous in* $\overline{G}$.

Definition 7. If $f(z)$ is defined and continuous only in G and if there exists a function $g(z)$ defined on the boundary c of the region G such that the function $h(z)$ defined by $h(z) = f(z)$ in G and by $h(z) = g(z)$ on c, is continuous in $\overline{G}$, then we say that $f(z)$ is *continuously extensible on the boundary c to the function* $g(z)$. (If $g(z)$ exists, then it is uniquely determined by $f(z)$.)

Definition 8. If the limit

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} \tag{6}$$

exists, we say that $f(z)$ has a *derivative at the point* z_0 (or that $f(z)$ is *differentiable* or *monogenic at* z_0). We denote the limit (6) by $f'(z_0)$.

Note (cf. Remark 6) that the limit in (6) should be independent of the manner in which $z \to z_0$. This fact leads to interesting consequences, among others to conditions (7) below.

Theorem 1. *In order that a function* $f(z) = u(x, y) + \mathrm{i}v(x, y)$ *should have a derivative at the point* $z_0 = x_0 + \mathrm{i}y_0$ *it is necessary and sufficient that* $u(x, y)$ *and* $v(x, y)$ *have total differentials* (see § 12.3) *and that their derivatives at this point satisfy the so-called Cauchy-Riemann equations*

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{7}$$

REMARK 8. Hence, if $f(z)$ has a derivative at z_0, then u and v have total differentials at (x_0, y_0) and satisfy (7). Conversely: If u and v have total differentials at (x_0, y_0) and satisfy (7), then the function $u + \mathrm{i}v$ has a derivative at $z_0 = x_0 + \mathrm{i}y_0$.

Theorem 2. *If the derivative of a function*

$$f(z) = u(x, y) + \mathrm{i}v(x, y)$$

exists, then it is given by the formulae

$$f'(z) = \frac{\partial u}{\partial x} + \mathrm{i}\frac{\partial v}{\partial x} = \frac{\partial f}{\partial x} \quad \textit{or} \quad f'(z) = -\mathrm{i}\left(\frac{\partial u}{\partial y} + \mathrm{i}\frac{\partial v}{\partial y}\right) = -\mathrm{i}\frac{\partial f}{\partial y}. \tag{8}$$

REMARK 9. For the evaluation of derivatives of functions of a complex variable, the same rules hold as for functions of a real variable. In particular, $(fg)' = f'g + fg'$, $(z^n)' = nz^{n-1}$ for every positive integer n, etc.

Theorem 3. *If $f(z)$ has a derivative at the point z_0, then $f(z)$ is continuous at z_0.*

Definition 9. If $f(z)$ has a derivative at every point of a region G, it is said to be *holomorphic (regular) in G.*

REMARK 10. A holomorphic function is often called an *analytic function.* However, this concept is also used for *multi-valued* functions (e.g. $\ln z$) which have a derivative (compare § 20.6).*

Theorem 4. *If the function $f(z) = u(x, y) + iv(x, y)$ is holomorphic in G, then u and v are harmonic* (Definition 18.4.4) *in G.*

Theorem 5. *If $f(z)$ is holomorphic in G, then it has derivatives of all orders in G.*

REMARK 11. We use a bar to denote the complex conjugate:

$$\overline{z} = x - iy\,, \quad \overline{f(z)} = u(x, y) - iv(x, y)\,.$$

Theorem 6. *If $f(z)$ is holomorphic in G, then $\overline{f(z)}$ is not holomorphic in G, in general.*

Example 2. The function

$$f(z) = z^2 = (x + iy)^2 = x^2 - y^2 + i\,.\,2xy \tag{9}$$

(where thus $u(x, y) = x^2 - y^2$, $v(x, y) = 2xy$) is holomorphic in the whole Gaussian plane, for u, v have everywhere continuous partial derivatives and, consequently, the total differential, and they obviously satisfy the Cauchy-Riemann conditions (7), since

$$\frac{\partial u}{\partial x} = 2x\,, \quad \frac{\partial v}{\partial y} = 2x\,, \quad \frac{\partial u}{\partial y} = -2y\,, \quad \frac{\partial v}{\partial x} = 2y\,.$$

Further, $f'(z) = 2z$. In fact, according to (8) we have

$$f'(z) = 2x + i\,.\,2y = -i(-2y + i\,.\,2x) = 2(x + iy) = 2z\,.$$

*In the English literature there are further variations; when consulting other works the reader should carefully examine the definitions used.

(See also Remark 9, of course.) The functions u and v are obviously harmonic, for

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 2 - 2 = 0\,, \quad \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0 + 0 = 0\,.$$

The complex conjugate function

$$\overline{f(z)} = \overline{z^2} = x^2 - y^2 - \mathrm{i}\,.\,2xy$$

is not holomorphic in any region of the Gaussian plane, because — with the exception of the point $z = 0$ — the Cauchy-Riemann conditions (7) are nowhere satisfied.

REMARK 12. Using equations (7), we can find, corresponding to a given function $u(x, y)$ (or $v(x, y)$) which is harmonic in a simply connected region G, a so-called *conjugate function* $v(x, y)$ (or $u(x, y)$, respectively), so that the function

$$f(z) = u(x, y) + \mathrm{i}v(x, y)$$

be holomorphic in G. This conjugate function is uniquely determined up to an additive constant.

Theorem 5 then implies that *every function harmonic in G has derivatives in G of all orders.*

REMARK 13. We call a function $f(z)$ *univalent* or *simple in a region G*, if for every pair of different points z_1, z_2 of this region the relation $f(z_1) \neq f(z_2)$ holds.

Hence a univalent function does not assume the same value at two different points of G. Univalent holomorphic functions are of great importance in conformal mappings where a one-to-one mapping of regions is concerned. The following assertion holds: If $f'(z_0) \neq 0$, then $f(z)$ is univalent in a certain neighbourhood of z_0.

REMARK 14. Functions of a complex variable have extensive applications. For example, in the study of two-dimensional irrotational sourceless flow we define the so-called *complex potential of the flow*, i.e. a holomorphic function $f(z) = u(x, y) + \mathrm{i}v(x, y)$. Equations $u = \text{const.}$ then represent equipotential curves, while equations $v = \text{const.}$ give the trajectories of the flow (cf. Example 21.3.3). In electrotechnics, functions of a complex variable are used on the one hand in elementary considerations where vectors of basic electrical quantities are expressed by complex numbers, and on the other in solving more complicated problems (e.g. in solving differential equations with the help of the Laplace transform, etc.). We also use functions of a complex variable to solve two-dimensional problems in elasticity,

where the so called Airy stress function is expressed by two holomorphic functions (see e.g. [22]). Properties of series, integral theorems (especially the Residue Theorem, see § 20.5) and conformal mappings are also widely applied.

20.2. Integral of a Function of a Complex Variable. The Cauchy Integral Theorem. Cauchy's Integral Formula

Integral of a function of a complex variable along a curve c is defined in a similar way as the line integral of a function of a real variable. Let us consider a simple finite piecewise smooth oriented curve c (Definition 14.1.1; in the following text, we speak briefly of a curve) with initial point z_0 (Fig. 20.2) and a function

$$w = f(z)$$

defined on this curve. Let us divide the curve at the points $z_1, z_2, \ldots, z_{n-1}$,

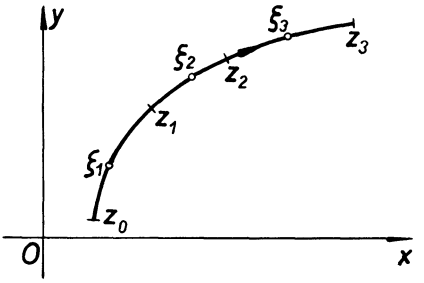


Fig. 20.2.

numbered in the sense of its positive orientation, into n arcs $c_1, c_2, \ldots, c_n$. Let us denote by $l_1, l_2, \ldots, l_n$ the lengths of these arcs. The *norm* $\nu(d)$ *of the chosen partition* d is defined as the greatest of the numbers $l_1, l_2, \ldots, l_n$. On each arc c_i let us choose an arbitrary point ξ_i and let us construct the sum

$$\sigma(d) = \sum_{i=1}^{n} f(\xi_i)(z_i - z_{i-1}) \tag{1}$$

(depending on the partition d and on the points ξ_i on c_i).

Definition 1. If there exists a (generally complex) number I such that

$$\lim_{n\to\infty} \sigma(d_n) = I$$

holds for every sequence of partitions d_n satisfying

$$\lim_{n\to\infty} \nu(d_n) = 0$$

and for every choice of points ξ_i on c_i, then we call this number the *integral of the function* $f(z)$ *along the* (*oriented*) *curve* c *and write*

$$\int_c f(z)\,\mathrm{d}z = I. \tag{2}$$

Roughly speaking, the integral (2) is the limit of integral sums (1) for the case that the norm of the partition of the curve c tends to zero.

Theorem 1. *If* $f(z)$ *is continuous on* c (cf. Remark 20.1.7), *then the integral* (2) *exists. In particular, if* $f(z)$ *is continuous* (*or even holomorphic*) *in a region* G, *then there exists the integral of* $f(z)$ *along every curve* c (*with the properties mentioned above*) *which lies in* G.

Theorem 2. *The inequality*

$$\left|\int_c f(z)\,\mathrm{d}z\right| \leqq Ml$$

holds, where l *is the length of the curve* c *and* M *is the maximum* (or l.u.b., see Definition 1.3.3) *of* $|f(z)|$ *on* c.

REMARK 1. The integral (2) has properties similar to those of line integrals. In particular,

$$\int_c [k_1 f_1(z) + k_2 f_2(z)]\,\mathrm{d}z = k_1 \int_c f_1(z)\,\mathrm{d}z + k_2 \int_c f_2(z)\,\mathrm{d}z,$$

$$\int_c f(z)\,\mathrm{d}z = -\int_{c'} f(z)\,\mathrm{d}z,$$

where c' is the curve c with opposite orientation, etc.

Definition 2. A function $F(z)$ satisfying

$$F'(z) = f(z)$$

in G is called a *primitive function of* $f(z)$ *in* G.

Theorem 3. *To every function* $f(z)$ *holomorphic in a simply connected region* G *there exists a primitive function.*

REMARK 2. For regions which are not simply connected the assertion of this theorem does not hold, in general.

Theorem 4. *If $F(z)$ is a primitive function of $f(z)$ in G and if c is an arbitrary curve lying in G* (with the properties mentioned above) *with initial point z_1 and end point z_2, then*

$$\int_c f(z)\,\mathrm{d}z = F(z_2) - F(z_1)\,. \tag{3}$$

REMARK 3. The fact that the value of the integral does not depend on the path of integration in this case but only on the initial and end points of the curve c is often expressed by writing it as

$$\int_{z_1}^{z_2} f(z)\,\mathrm{d}z\,. \tag{4}$$

Example 1. The primitive function of $f(z) = z^2$ is $F(z) = \frac{1}{3}z^3 + C$ in the whole plane. Hence (for every curve)

$$\int_{z_1}^{z_2} z^2\,\mathrm{d}z = \frac{z_2^3}{3} - \frac{z_1^3}{3}\,.$$

REMARK 4. If an integral cannot be evaluated with the aid of a primitive function (i.e. if either the primitive function does not exist, or is difficult to find), we represent the integral by line integrals (see § 14.7); we write $f(z) = u(x, y) + \mathrm{i}v(x, y)$, $\mathrm{d}z = \mathrm{d}x + \mathrm{i}\,\mathrm{d}y$, then multiplying formally we obtain

$$\int_c f(z)\,\mathrm{d}z = \int_c [u(x, y)\,\mathrm{d}x - v(x, y)\,\mathrm{d}y] + \mathrm{i}\int_c [v(x, y)\,\mathrm{d}x + u(x, y)\,\mathrm{d}y]\,.$$

We can also use other methods (see Example 3). In particular, the substitution method and method of integration by parts (under similar assumptions as in § 13.2) can be applied.

Example 2. Let us evaluate the integral

$$\int_c \frac{1}{z}\,\mathrm{d}z\,,$$

where c is the circumference of the circle with centre at the origin and radius a, oriented positively (Remark 14.7.1) with respect to its interior.

For every point on the circumference we have $z = a(\cos\varphi + \mathrm{i}\sin\varphi) = a\mathrm{e}^{\mathrm{i}\varphi}$. Differentiating, we get $\mathrm{d}z = a\mathrm{i}\mathrm{e}^{\mathrm{i}\varphi}\,\mathrm{d}\varphi$, so that

$$\int_c \frac{1}{z}\,\mathrm{d}z = \int_0^{2\pi} \frac{a\mathrm{i}\mathrm{e}^{\mathrm{i}\varphi}\,\mathrm{d}\varphi}{a\mathrm{e}^{\mathrm{i}\varphi}} = \mathrm{i}\int_0^{2\pi} \mathrm{d}\varphi = 2\pi\mathrm{i}\,.$$

The function $\ln z$ could not be used as a primitive function directly, since it is not a *single-valued* function in the region under consideration, and hence is not holomorphic.

This example also shows that the integral of a holomorphic function along a closed curve need not vanish. (Cf. Theorem 5 below. Here — by contrast with the assumptions of that theorem — the function $f(z) = 1/z$ has a singular point $z = 0$ in the interior of the given circle.)

REMARK 5. The method of integrating by parts leads to the formula

$$\int_{z_1}^{z_2} f_1(z)F_2(z)\,dz = [F_1(z)F_2(z)]_{z_1}^{z_2} - \int_{z_1}^{z_2} F_1(z)f_2(z)\,dz, \tag{5}$$

where $F_1(z)$, $F_2(z)$ are primitive functions (Definition 2) of $f_1(z)$, $f_2(z)$, respectively, in G and c is a curve lying in G with initial point z_1 and end-point z_2. This formula gives for a closed curve c the result

$$\int_c f_1(z)F_2(z)\,dz = -\int_c F_1(z)f_2(z)\,dz, \tag{6}$$

since for a closed curve we have $F_1(z_2)F_2(z_2) = F_1(z_1)F_2(z_1)$ and the first term of the right-hand side in (5) vanishes. Equation (6) finds frequent application.

As far as curves are investigated in what follows, we will always assume them to be simple, finite and piecewise smooth (as in the preceding text).

Theorem 5 (*The Caychy Integral Theorem*). *Let c be a closed oriented curve (with the properties mentioned above). Let $f(z)$ be a function holomorphic in the interior G of the curve c and continuous in $\overline{G} = G \cup c$. Then*

$$\int_c f(z)\,dz = 0. \tag{7}$$

REMARK 6. Theorem 5 holds the more true if c lies in a region G where $f(z)$ is holomorphic and if the whole interior of c belongs to G.

A similar remark holds for Theorems 6 and 7 below.

If $f(z)$ is not holomorphic in the whole interior of the curve c, then (7) is not valid, in general. See Example 2, where this condition was violated at the point $z = 0$.

Theorem 6. *Let c_0, c_1 be closed curves with the same orientation* (as shown in Fig. 20.3) *and with the above-mentioned properties, the curve c_1 lying in the interior of c_0. Let $f(z)$ be holomorphic in the doubly connected region G, the boundary of which is constituted by these curves* (Fig. 20.3, the shaded area), *and let it be*

continuous in $\overline{G} = G \cup c_0 \cup c_1$. (It is not necessary to assume that $f(z)$ be defined, or even holomorphic in the interior of the curve c_1.) *Then*

$$\int_{c_1} f(z)\,\mathrm{d}z = \int_{c_0} f(z)\,\mathrm{d}z. \tag{8}$$

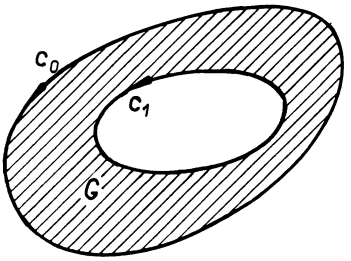


Fig. 20.3.

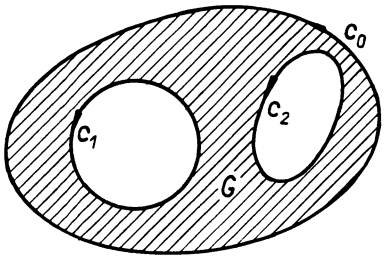


Fig. 20.4.

REMARK 7. Theorem 6 has frequent applications. For example, the integral of the function $f(z) = 1/z$ along any closed curve c_0 (of the above-mentioned properties), positively oriented with respect to its interior in which the point $z = 0$ is contained, has the value $2\pi\mathrm{i}$. Because — according to Theorem 6 — this integral is equal to that of Example 2, a being chosen sufficiently small in order that c be contained in the interior of c_0.

REMARK 8. Theorem 6 can be generalized: Let $c_0, c_1, c_2, \ldots, c_n$ be closed curves with above-mentioned properties and with the same orientation (see Fig. 20.4 sketched for $n = 2$). Let the curves $c_1, c_2, \ldots, c_n$ lie mutually in their exteriors (i.e. $c_2, c_3, \ldots, c_n$ outside of c_1, etc.) and let all of them lie in the interior of c_0. If $f(z)$ is holomorphic in the $(n+1)$-tuply connected region G, the boundary of which is constituted by the curves $c_0, c_1, c_2, \ldots, c_n$ (Fig. 20.4, the shaded area), and continuous in $\overline{G} = G \cup c_0 \cup c_1 \cup c_2 \cup \cdots \cup c_n$, then

$$\int_{c_0} f(z)\,\mathrm{d}z = \int_{c_1} f(z)\,\mathrm{d}z + \int_{c_2} f(z)\,\mathrm{d}z + \cdots + \int_{c_n} f(z)\,\mathrm{d}z. \tag{9}$$

Theorem 7 (*The Cauchy Integral Formula*). *Let c be a closed curve* (of properties mentioned above), *positively oriented with respect to its interior G, and let $f(z)$ be holomorphic in G and continuous in $\overline{G} = G \cup c$. Let $z_0 \in G$* (Fig. 20.5). *Then we have*

$$f(z_0) = \frac{1}{2\pi\mathrm{i}} \int_c \frac{f(z)\,\mathrm{d}z}{z - z_0}. \tag{10}$$

Moreover,

$$f^{(n)}(z_0) = \frac{n!}{2\pi\mathrm{i}} \int_c \frac{f(z)\,\mathrm{d}z}{(z - z_0)^{n+1}}, \quad n = 1, 2, \ldots, \tag{11}$$

holds, where $f^{(n)}(z_0)$ denotes the n-th derivative of the function $f(z)$ at the point z_0.

REMARK 9. If the point z_0 lies in the exterior of the curve c, then the functions $g(z) = f(z)/(z - z_0)$, $g_1(z) = f(z)/(z - z_0)^2$, ... are holomorphic in G and according to Cauchy's theorem 5 the integrals (10), (11) are equal to zero.

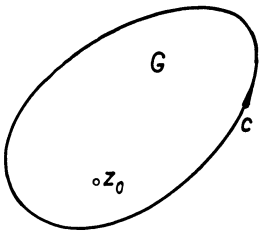


Fig. 20.5.

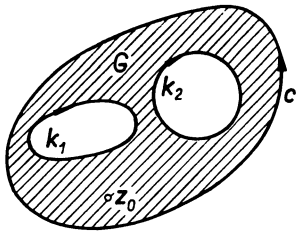


Fig. 20.6.

Example 3. Let us evaluate the integral

$$\int_c \frac{dz}{z^3}$$

along a closed curve positively oriented with respect to its interior G and let G contain the origin $z = 0$.

If we choose $f(z) \equiv 1$ (so that $f(z)$ is holomorphic even in the entire plane), and $z_0 = 0$, then (11) gives, for $n = 2$,

$$f''(0) = \frac{2}{2\pi i} \int_c \frac{1 \,.\, dz}{(z - 0)^3}$$

and hence, because $f''(z) \equiv 0$ and, consequently, $f''(0) = 0$,

$$\int_c \frac{dz}{z^3} = 0 \,.$$

REMARK 10. Theorem 7 can be generalized quite similarly as Theorem 6 (see Remark 8). In the situation shown in Fig. 20.6 (note the orientation of the curves k_1, k_2 !) we have

$$f(z_0) = \frac{1}{2\pi i} \left[\int_c \frac{f(z)\,dz}{z - z_0} + \int_{k_1} \frac{f(z)\,dz}{z - z_0} + \int_{k_2} \frac{f(z)\,dz}{z - z_0} \right]. \tag{12}$$

Formula (11) can be generalized in a quite similar way.

20.3. Integrals of Cauchy's Type. The Plemelj Formulae

REMARK 1. In this paragraph, by the word *curve* we understand a simple finite piecewise smooth closed curve, positively oriented with respect to its interior. We

denote its points by t instead of z, i.e. $t = x + iy$. The interior of the curve will be denoted by S^+, the exterior by S^-.

Definition 1. Let a function $f(t)$ be given on a curve c. We say that $f(t)$ has a *derivative with respect to the curve* c (briefly *on* c) at *a point* $t_0 \in c$, if there exists a finite (generally complex) limit

$$\lim \frac{f(t) - f(t_0)}{t - t_0} \tag{1}$$

for $t \to t_0$ along the curve c.

Definition 2. We say that $f(t)$ *satisfies the Hölder condition with the exponent* μ in the neighbourhood of a point $t_0 \in c$ if there exists an arc I on c such that t_0 is its interior point and that there exist two constants M and μ ($M > 0$, $0 < \mu \leqq 1$) such that for every two points t_1, t_2 of I the relation

$$|f(t_2) - f(t_1)| \leqq M|t_2 - t_1|^\mu \tag{2}$$

holds. In particular, if there exist constants M and μ ($M > 0$, $0 < \mu \leqq 1$) such that (2) holds for every pair of points t_1, t_2 of the curve c, we say that $f(t)$ *satisfies the Hölder condition* (*with exponent* μ) *on* c.

REMARK 2. We shall briefly say that $f(t)$ *fulfils the condition* H (or, in more detail, the *condition* $H(\mu)$).

Theorem 1. *If* $f(t)$ *has a continuous* (*or, more generally, a bounded*) *derivative on* c, *then it satisfies the condition* H *on* c (*and hence in the neighbourhood of every point of* c) *with the exponent* 1.

Definition 3. Let a function $f(t)$ be defined on c. The integral

$$F(z) = \frac{1}{2\pi i} \int_c \frac{f(t)\,dt}{t - z} \tag{3}$$

is called an *integral of Cauchy's type.*

Theorem 2. *If* $f(t)$ *is continuous on* c (*it is sufficient if* $f(t)$ *is integrable on* c), *then* $F(z)$ *is a holomorphic function of* z *both in* S^+ *and in* S^- (for the meaning of S^+ and S^- see Remark 1). *The derivatives of* $F(z)$ *in* S^+ *as well as in* S^- *are given by the formulae*

$$F'(z) = \frac{1}{2\pi i} \int_c \frac{f(t)\,dt}{(t - z)^2}, \ldots, F^{(n)}(z) = \frac{n!}{2\pi i} \int_c \frac{f(t)\,dt}{(t - z)^{n+1}} \tag{4}$$

which are obtained by formal differentiation with respect to z *under the integral sign in* (3).

REMARK 3. If we choose a point z on the curve c, then, in general, the integral (3) does not exist (it is divergent). However, the principal value (see Definition 4 below) of this integral can exist.

Definition 4. Let the point t_0 lie on c. Choose $\varepsilon > 0$ such that the circumference of the circle with centre t_0 and radius ε (and also all circumferences of such circles with radius less than ε) intersect the curve c exactly at two points (denote them by t_1, t_2; Fig. 20.7). (If the curve c has the properties mentioned in Remark 1, then such an $\varepsilon > 0$ exists for every point $t_0 \in c$.) If there exists a finite limit

$$\lim_{\varepsilon \to 0} \int_{c - \overset{\frown}{t_1 t_2}} \frac{f(t)\,\mathrm{d}t}{t - t_0}, \tag{5}$$

where $\overset{\frown}{t_1 t_2}$ is the arc of the curve c lying within the circumference of the circle with centre t_0 and radius ε, we say that the *integral*

$$\int_c \frac{f(t)\,\mathrm{d}t}{t - t_0} \tag{6}$$

exists in the sense of its principal value.

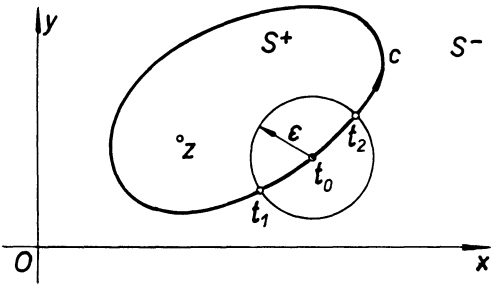


Fig. 20.7.

REMARK 4. *If the point t_0 lies on c, then, in what follows, we shall always interpret the integral* (6) *to mean its principal value. The function $f(t)$ will always be assumed to be integrable on c.*

Theorem 3. *If $f(t)$ has a continuous derivative in the neighbourhood of t_0* (Definition 1) (*it is sufficient to require only that $f(t)$ fulfils condition H in the neighbourhood of t_0*, cf. Definition 2 and Remark 2), *then the principal value of the integral* (6) *exists.*

REMARK 5. By the continuous extensibility (on the curve c) of functions defined in S^+ (or S^-) we mean the extensibility in the sense of Definition 20.1.7. Further,

we say that a function $g(z)$ defined in S^+ (or S^-) is *continuously extensible on c at the point* $t_0 \in c$ if there exists a number A such that

$$\lim_{z \to t_0} g(z) = A, \quad \text{where } z \in S^+ \text{ (or } z \in S^-, \text{ respectively)}.$$

REMARK 6. Hence, $g(z)$ is continuously extensible to a value A at the point $t_0 \in c$ if, for every sequence of points $z_n \in S^+$ (or $z_n \in S^-$, respectively) satisfying $\lim_{n \to \infty} z_n = t_0$, we have

$$\lim_{n \to \infty} g(z_n) = A.$$

Theorem 4. *Let $f(t)$ satisfy the condition H in the neighbourhood of the point $t_0 \in c$. (A sufficient condition for this is that $f(t)$ have a continuous derivative in a neighbourhood of t_0.) Then the function*

$$F(z) = \frac{1}{2\pi i} \int_c \frac{f(t)\,dt}{t - z} \tag{7}$$

is continuously extensible on the curve c from S^+ as well as from S^- at that point and we have

$$F^+(t_0) = \tfrac{1}{2} f(t_0) + \frac{1}{2\pi i} \int_c \frac{f(t)\,dt}{t - t_0}, \tag{8}$$

$$F^-(t_0) = -\tfrac{1}{2} f(t_0) + \frac{1}{2\pi i} \int_c \frac{f(t)\,dt}{t - t_0}. \tag{9}$$

Here $F^+(t_0)$ and $F^-(t_0)$ stand for the values of the continuous extension of $F(z)$ from S^+ and from S^-, respectively; by integrals (8), (9) *we mean their principal values.*

Formulae (8) and (9) are called the *Plemelj formulae.*

REMARK 7. Formulae (8) and (9) show that if we cross the curve c from S^- to S^+ at the point t_0, then the function (7) has, at that point, a jump $f(t_0)$.

REMARK 8. It can be shown that if $f'(t)$ (Definition 1) satisfies the condition H in the neighbourhood of the point t_0, then the function

$$F'(z) = \frac{1}{2\pi i} \int_c \frac{f(t)\,dt}{(t - z)^2}$$

is continuously extensible at t_0 on the curve c from S^+ as well as from S^-. An analogous assertion is valid for derivatives of higher order.

Theorem 5 (*Cauchy's Theorem*). *Let $\varphi(z)$ be holomorphic in S^+ and continuous in $\overline{S}^+$ (i.e. in $S^+ \cup c$). Then* (cf. Theorem 20.2.7 and Remark 20.2.9)

$$\frac{1}{2\pi i}\int_c \frac{\varphi(t)\,dt}{t-z} = \varphi(z) \quad \text{if } z \in S^+, \tag{10}$$

$$\frac{1}{2\pi i}\int_c \frac{\varphi(t)\,dt}{t-z} = 0 \quad \text{if } z \in S^-. \tag{11}$$

If $\varphi(z)$ is holomorphic in S^- and at infinity (i.e. if $\varphi(1/z)$ has a removable singularity at the origin, cf. Remark 20.4.11 below) *and continuous in $\overline{S}^-$ (i.e. in $S^- \cup c$), then*

$$\frac{1}{2\pi i}\int_c \frac{\varphi(t)\,dt}{t-z} = \varphi(\infty)\ (= \lim_{z\to\infty}\varphi(z)) \quad \text{if } z \in S^+, \tag{12}$$

$$\frac{1}{2\pi i}\int_c \frac{\varphi(t)\,dt}{t-z} = -\varphi(z) + \varphi(\infty) \quad \text{if } z \in S^-. \tag{13}$$

Example 1. Let the origin belong to S^+. Then

$$\int_c \frac{dt}{t^2} = 0, \tag{14}$$

because the function $\varphi(z) = 1/z$ is holomorphic in S^- and at infinity and is continuous in $\overline{S}^-$: According to (12), we have

$$\frac{1}{2\pi i}\int_c \frac{1/t}{t-z}\,dt = 0 \tag{15}$$

for every $z \in S^+$. Putting $z = 0$ in (15) we obtain (14).

Theorem 6. *A function $\varphi(t)$ continuous on c is a continuous extension of some function $\varphi(z)$ holomorphic in S^+ if and only if*

$$\frac{1}{2\pi i}\int_c \frac{\varphi(t)\,dt}{t-z} = 0 \quad \text{for every } z \in S^-; \tag{16}$$

$\varphi(t)$ is a continuous extension of some function $\psi(z)$ holomorphic in S^- and at infinity if and only if

$$\frac{1}{2\pi i}\int_c \frac{\varphi(t)\,dt}{t-z} = a \quad \text{for every } z \in S^+, \tag{17}$$

where a is a constant. (This constant is equal to the value $\psi(\infty)$.)

REMARK 9. The relation

$$F(z) = \frac{1}{2\pi i} \int_c \frac{f(t)\,dt}{t - z}, \quad z \in S^+, \tag{18}$$

does *not* imply (even if $f(t)$ has a derivative on c) that $f(t)$ is a continuous extension of $F(z)$ from S^+ on c ($F^+(t) = f(t)$ need not hold). According to Theorem 6 we can have $f(t) - F^+(t) = g(t)$, where $g(t)$ is a continuous extension of a function holomorphic in S^- and vanishing at infinity.

REMARK 10. The results of this paragraph can be generalized, for instance, to the case of multiply connected regions. Similar results also hold for the case where c is not a closed curve but, for example, the x-axis (the axis of real numbers) in the Gaussian plane. S^+ and S^- are then half-planes. The formulation of these results can be found in [22], § 5.7 and § 5.11.

20.4. Series. Taylor's Series, Laurent's Series. Singular Points of Holomorphic Functions

In this paragraph we use the notation $a = \alpha + i\beta$, $a_n = \alpha_n + i\beta_n$, where α, β, α_n, β_n are real numbers.

Definition 1. We say that a sequence of complex numbers

$$a_1, a_2, \ldots, a_n, \ldots$$

has the *limit* a (or tends to the limit a) if to every $\varepsilon > 0$ there corresponds a number n_0 such that for every $n > n_0$ we have

$$|a - a_n| < \varepsilon, \quad \text{i.e.} \quad \sqrt{[(\alpha - \alpha_n)^2 + (\beta - \beta_n)^2]} < \varepsilon.$$

We write

$$\lim_{n\to\infty} a_n = a.$$

Definition 2. Let us write

$$s_n = a_1 + a_2 + \cdots + a_n.$$

We say that the *series*

$$a_1 + a_2 + \cdots + a_n + \cdots \tag{1}$$

is *convergent* and *has the sum* s if

$$\lim_{n\to\infty} s_n = s.$$

Theorem 1. *The series* (1) *is convergent if and only if both series of real numbers*

$$\alpha_1 + \alpha_2 + \cdots + \alpha_n + \cdots, \quad \beta_1 + \beta_2 + \cdots + \beta_n + \cdots$$

are convergent. If the first of them has the sum α *and the second the sum* β*, then*

$$s = \alpha + \mathrm{i}\beta .$$

Theorem 2. *The convergence of the series*

$$|a_1| + |a_2| + \cdots + |a_n| + \cdots$$

implies the convergence of the series (1). *The series* (1) *is then said to be absolutely convergent.*

In the theory of functions of a complex variable, *series of functions* (*function series*) play an important role. Many functions can be expressed (or directly defined) by infinite series.

Basic definitions and results are very similar to those discussed for the case of functions of a real variable in Chap. 15. They will be formulated for regions here, although they can be easily generalized for the case of more general domains.

Definition 3. Let a sequence of functions

$$f_1(z), f_2(z), \ldots, f_n(z), \ldots,$$

defined in a region G, be given. We say, that this sequence is *convergent in* G and that $f(z)$ is its *limit*, if the sequence of numbers

$$f_1(z_0), f_2(z_0), \ldots, f_n(z_0), \ldots$$

is convergent for every fixed $z_0 \in G$ and has the limit $f(z_0)$.

Denote

$$f_1(z) + f_2(z) + \cdots + f_n(z) = s_n(z) .$$

We say that the series

$$f_1(z) + f_2(z) + \cdots + f_n(z) + \cdots \tag{2}$$

is *convergent in* G and has the *sum* $s(z)$, if the sequence $\{s_n(z)\}$ (the so-called *sequence of partial sums of the series* (2)) is convergent in G and has the limit $s(z)$.

REMARK 1. (i) It may happen that the functions $f_k(z)$ $(k = 1, 2, \dots)$ are defined in a region G, however, the series (2) does not converge for all $z \in G$, but converges in a "smaller" domain, say D. Then D is called *the domain of convergence of the series* (2).

(ii) (*Uniform Convergence*). According to Definitions 3 and 1, the series (2) is convergent in G if to an arbitrary $\varepsilon > 0$ and an arbitrary point $z_0 \in G$ such an n_0 can be found, dependent on the choice of ε and z_0, in general, that

$$|s(z_0) - s_n(z_0)| < \varepsilon$$

holds for all $n > n_0$. If to an arbitrary $\varepsilon > 0$ the same n_0 can be found for the whole region G (thus dependent on ε, but independent of $z_0 \in G$), the series (2) is said to be *uniformly convergent in* G. A simple criterion for uniform convergence gives the so-called *Weierstrass M-Test*: *Let such non-negative numbers* $A_1, A_2, \dots, A_n, \dots$ *exist that*

$$|f_1(z)| \leqq A_1\,, \quad |f_2(z)| \leqq A_2\,, \quad \dots\,, \quad |f_n(z)| \leqq A_n\,, \quad \dots$$

holds for every $z \in G$ *and let, at the same time, the series* (of numbers)

$$A_1 + A_2 + \cdots + A_n + \cdots$$

be convergent. Then the series (2) *is uniformly convergent in* G.

Often the case is encountered that the series (2) is not uniformly convergent in the whole region G, but that it converges uniformly in *every* bounded *closed* region (say $\overline{B}$) contained in G. (Thus, corresponding to a given $\varepsilon > 0$, the above-mentioned n_0 may be different for different regions $\overline{B}$.) Then we say that the series (2) converges *almost* (or *locally*) *uniformly in* G.

If (2) converges uniformly in G, then it converges almost uniformly in G, of course.

Theorem 3 (*The Weierstrass Theorem*). *Let the functions*

$$f_1(z), f_2(z), \dots, f_n(z), \dots$$

be holomorphic in a region G *and let the series* (2) *be almost uniformly convergent in* G (Remark 1). (*The second assumption is satisfied, for example, if the series* (2) *is uniformly convergent in* G.) *Then*

1. *The function* $s(z)$ *defined by the sum of the series* (2) *is holomorphic in* G.

2. *If we differentiate the series* (2) n *times term by term, we obtain a series which is again almost uniformly convergent in* G *and whose sum in* G *is equal to the* n*-th derivative of the function* $s(z)$.

3. *If* $z_0 \in G$, $z \in G$, *then*

$$\int_{z_0}^{z} s(t)\,dt = \int_{z_0}^{z} f_1(t)\,dt + \int_{z_0}^{z} f_2(t)\,dt + \cdots + \int_{z_0}^{z} f_n(t)\,dt + \cdots .$$

Definition 4. By a *power series* (in complex variable) we mean the series

$$a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \dots , \tag{3}$$

where the a_n are constants (generally complex).

REMARK 2. For $z_0 = 0$, a power series has the form

$$a_0 + a_1 z + a_2 z^2 + \dots . \tag{4}$$

Since the series (3) can be transformed into a series of the form (4) by the substitution $z - z_0 = z'$ (translation in the Gaussian plane), it is sufficient to consider the series (4). Theorems valid for the series (4) hold for the series (3) as well, if we write $z - z_0$ instead of z.

Theorem 4. *To every series* (4) *there corresponds a number* $r \geqq 0$ ($r = +\infty$ *is also admitted*) *such that* (4) *is convergent for all* z *satisfying* $|z| < r$ *and divergent for all* z *satisfying* $|z| > r$.

The number r is called the *radius of convergence of the series* (4). The circle with centre at the point z_0 and with radius r is called the *circle of convergence.*

REMARK 3. In order to find the radius of convergence of the series (4), we can use rules analogous to those governing real series. In particular: The series (4) and the series

$$|a_0| + |a_1||z| + |a_2||z|^2 + \dots \tag{5}$$

have the same radius of convergence.

Since (5) is a series with non-negative terms, the criteria of § 10.2 may be applied to determine its radius of convergence. In particular, they imply:

Theorem 5. *If there exists a limit*

$$\lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} = l \quad \text{or} \quad \lim_{n\to\infty} \sqrt[n]{|a_n|} = l ,$$

then

$$r = \frac{1}{l} .$$

(*For* $l = 0$ *we have* $r = +\infty$, *for* $l = +\infty$ *we have* $r = 0$.)

Theorem 6 (*Abel's Theorem*). *Let a power series*

$$a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \cdots = s(z)$$

having a radius of convergence r, converge at a point t on the circumference of its circle of convergence (i.e. at a point t such that $|t - z_0| = r$). Let S be its sum at this point. Then there exists the angular extension S of $s(z)$ from the interior K of the circle of convergence to the point t. In details: For every sequence of points $z_n \in K$ which converges to t and which lies in the interior of some angle $\alpha < \pi$ with vertex at t whose arms lie, in a neighbourhood or t, inside of K (Fig. 20.8), *we have*

$$\lim_{n\to\infty} s(z_n) = S.$$

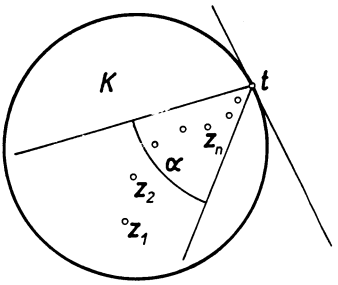


Fig. 20.8.

Theorem 7. *A power series converges absolutely in its circle of convergence K ($|z - z_0| < r$). In addition, it converges uniformly in every closed region contained in K (i.e. almost uniformly in K).*

Hence Theorem 3 applies to power series in K. It implies especially: *The sum $s(z)$ of a power series is a holomorphic function in K. Its derivatives can be calculated by differentiating the given series term by term. All these series have the same radius of convergence.*

REMARK 4. Some functions in complex variable may be defined by power series (naturally, in their domain of convergence). The most important are the following:

$$e^z = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \ldots \quad (r = +\infty),$$

$$\sin z = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \ldots \quad (r = +\infty),$$

$$\cos z = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \ldots \quad (r = +\infty).$$

For real arguments ($z = x$), these functions coincide with the well-known functions e^x, $\sin x$, $\cos x$. (We say that they are an *extension of these functions into the complex plane* (cf. § 20.6).) For the functions e^z, $\sin z$, $\cos z$ all formulae valid for functions of a real argument still hold. For example,

$$e^{z_1}e^{z_2} = e^{z_1+z_2}, \quad \sin(z_1 + z_2) = \sin z_1 \cos z_2 + \cos z_1 \sin z_2,$$
$$(e^z)' = e^z, \quad (\sin z)' = \cos z, \quad (\cos z)' = -\sin z,$$

etc. From the expansions in series we easily get the so-called *Euler relation*

$$e^{iz} = \cos z + i \sin z$$

which is often used for $z = x$ when solving linear differential equations with constant coefficients. Further we have

$$\cosh z = \frac{e^z + e^{-z}}{2} = \cos iz, \quad \sinh z = \frac{e^z - e^{-z}}{2} = -i \sin iz,$$
$$\cos z = \cosh iz, \quad \sin z = -i \sinh iz,$$
$$\sin z = \sin x \cosh y + i \cos x \sinh y, \quad \cos z = \cos x \cosh y - i \sin x \sinh y,$$
$$\cos(z + 2k\pi) = \cos z, \quad \sin(z + 2k\pi) = \sin z, \quad e^{z+2k\pi i} = e^z$$

(k being an integer), etc.

Definition 5. We say that a point z_0 ($z_0 \neq \infty$) is a *regular point* (*ordinary point*) of a function $f(z)$ if there exists a (circular) neighbourhood of z_0 such that $f(z)$ is holomorphic in this neighbourhood. A point which is not regular is called a *singular point* of $f(z)$.

Example 1. The function $f(z) = 1/z$ has only one singular point, namely $z = 0$. Every other point of the Gaussian plane is a regular point of this function.

Theorem 8 (*The Taylor Series*). *Let $z = z_0$ be a regular point of a function $f(z)$. Then, in the neighbourhood of z_0, we have*

$$f(z) = a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \dots, \tag{6}$$

where

$$a_n = \frac{f^{(n)}(z_0)}{n!}. \tag{7}$$

The radius of convergence r of the series (6) *is equal to the distance of the point z_0 from the nearest singular point of $f(z)$.* (*More exactly: The radius r is equal to the* g. l. b. *of the distances of z_0 from all singular points of $f(z)$.*) *The series* (6) *is uniquely determined by the function $f(z)$.*

REMARK 5. It follows from Cauchy's integral formula that the coefficients of the series (6) can also be evaluated as the integrals

$$a_n = \frac{1}{2\pi i}\int_c \frac{f(z)\,dz}{(z-z_0)^{n+1}} \tag{8}$$

along an arbitrary simple closed curve c positively oriented with respect to its interior, lying in a neighbourhood of z_0 in which $f(z)$ is holomorphic and containing the point z_0 in its interior.

Example 2. For $f(z) = 1/z$ and $z_0 = 2$, we have

$$f^{(n)}(2) = (-1)^n \frac{n!}{2^{n+1}}$$

and, according to (6) and (7),

$$\frac{1}{z} = \frac{1}{2} - \frac{1}{2^2}(z-2) + \frac{1}{2^3}(z-2)^2 - \frac{1}{2^4}(z-2)^3 + \dots .$$

The radius of convergence r is equal to the distance of the point $z = 2$ from the singular point $z = 0$, hence $r = 2$.

REMARK 6. Here (and often for other simple rational functions) it is also possible to get the Taylor expansion in such a way that we first rearrange the formula for the functional relation to be of the form of a sum of a geometric series; from the condition for the convergence of a geometric series (the absolute value of the common ratio is to be less than 1), we find the radius of convergence r:

$$\frac{1}{z} = \frac{1}{2+(z-2)} = \frac{1}{2}\cdot\frac{1}{1+\dfrac{z-2}{2}} = \frac{1}{2}\left[1 - \frac{z-2}{2} + \frac{(z-2)^2}{2^2} - \frac{(z-2)^3}{2^3} + \dots\right];$$

$$\left|\frac{z-2}{2}\right| < 1, \text{ i.e. } |z-2| < 2, \text{ hence } r = 2.$$

Theorem 9 (*The Laurent Series*). *Let $f(z)$ be holomorphic in the annulus M with centre z_0, inner radius r_1 and outer radius r_2. Then for $z \in M$ we have*

$$f(z) = \sum_{n=-\infty}^{\infty} a_n (z-z_0)^n, \tag{9}$$

where

$$a_n = \frac{1}{2\pi i}\int_c \frac{f(z)\,dz}{(z-z_0)^{n+1}} \quad (n = 0, \pm 1, \pm 2, \dots), \tag{10}$$

c being the circumference of an arbitrary circle with centre z_0, lying in M and positively oriented with respect to its interior.

REMARK 7. By the convergence of the series (9) we mean the convergence of *both* series

$$\sum_{n=0}^{\infty} a_n(z-z_0)^n, \quad \sum_{n=1}^{\infty} \frac{a_{-n}}{(z-z_0)^n}. \tag{11}$$

The first series is called the *regular part*, the second series the *principal part of the Laurent series*. The domain of convergence of series (11) can be a set greater than M. The regular part (as a power series) converges *inside* a certain circle (with centre z_0), the principal part *outside* a certain circle. The series (9) then converges in the common annulus. If $f(z)$ is holomorphic everywhere inside the smaller circle, then the principal part of the Laurent series vanishes (which also follows from Cauchy's integral theorem, since the integrand in (10) is then a holomorphic function for $n = -1, -2, \dots$) and the Laurent series coincides with the Taylor series.

Theorem 10. *The series* (9) *is uniquely determined by the holomorphic function* $f(z)$.

REMARK 8. It is not always necessary, if we expand a function into its Laurent series, to calculate the coefficients of this series according to (10). We can often — especially in the case of simple rational functions — use a similar method as in the case of the Taylor series (Remark 6), i.e. write the given function in a form involving fractions each having the form of the sum of a geometric series.

Example 3. Let us expand the function

$$f(z) = \frac{1}{(z-1)(z-3)}$$

in Laurent series with centre at the point $z = 0$ (i.e. in powers of z) and converging in the annulus $1 < |z| < 3$. We have

$$\frac{1}{(z-1)(z-3)} = -\frac{1}{2}\left(\frac{1}{z-1} - \frac{1}{z-3}\right) = -\frac{1}{6}\cdot\frac{1}{1-\frac{z}{3}} - \frac{1}{2z}\cdot\frac{1}{1-\frac{1}{z}} =$$

$$= -\frac{1}{6}\sum_{n=0}^{\infty}\left(\frac{z}{3}\right)^n - \frac{1}{2z}\sum_{n=0}^{\infty}\left(\frac{1}{z}\right)^n = -\left[\dots\frac{1}{2z^2} + \frac{1}{2z} + \frac{1}{6} + \frac{z}{18} + \dots\right].$$

We must be careful to ensure that the common ratio of each of the geometric series be of absolute value less than 1. Therefore we first put the function in the form shown above, for in this example we are considering values of z for which $1 < |z| < 3$. In fact, after this arrangement is made, the common ratios of both geometric series will be of absolute value less than 1, since

$$\left|\frac{z}{3}\right| < 1, \quad \left|\frac{1}{z}\right| < 1.$$

Example 4. Let us develop the same function in a Laurent series with centre $z = 1$ (i.e. in powers of $z - 1$), converging for $0 < |z - 1| < 2$. We have

$$\frac{1}{(z-1)(z-3)} = \frac{1}{(z-1)[(z-1)-2]} = -\frac{1}{2(z-1)} \cdot \frac{1}{1-\frac{1}{2}(z-1)} =$$
$$= -\frac{1}{2(z-1)} \sum_{n=0}^{\infty} \left(\frac{z-1}{2}\right)^n = -\frac{1}{2(z-1)} - \frac{1}{4} - \frac{1}{8}(z-1) - \dots .$$

REMARK 9. Example 4 illustrates an important case of the development of a function $f(z)$ in a Laurent series in the neighbourhood of an *isolated singular point* (i.e. of a singular point such that there is no other singular point in a sufficiently small neighbourhood of it). In this case, the inner circle reduces to a point.

REMARK 10. Let (9) be a Laurent expansion of a (holomorphic) function $f(z)$, converging in a neighbourhood G of an *isolated* singular point z_0 of $f(z)$ (we naturally consider the neighbourhood G without the point z_0). Exactly three cases can then arise:

1. If there exists a $k > 0$ such that we have $a_{-k} \neq 0$ in (9) but $a_{-l} = 0$ for all $l > k$ (hence the principal part of the Laurent series has only a finite number of terms), we say that $f(z)$ has *a pole of the k-th order* or *a pole of order k at* z_0.

A pole is characterized by the relation

$$\lim_{z \to z_0} f(z) = \infty .$$

In details: To an arbitrary $K > 0$ such a $\delta > 0$ can be found that in the δ-neighbourhood of the point z_0 — except for the point z_0 itself — we have $|f(z)| > K$.

A pole of the first order is often called a *simple pole* and a pole of the second order a *double pole.*

A (holomorphic) function which has no singular points in the complex plane other than poles is called a *meromorphic function.* It can be shown that if a function $f(z)$ is single-valued in the closed plane (Remark 20.1.2) and has at infinity at most a pole (not an essential singularity, see Remark 11 below), then it is meromorphic if and only if it is a rational function, i.e. a function which can be expressed in the form

$$f(z) = \frac{a_0 + a_1 z + \dots + a_n z^n}{b_0 + b_1 z + \dots + b_m z^m} .$$

2. If the principal part of the Laurent series has an infinite number of terms, we say that $f(z)$ has an *essential singularity at* z_0.

In this case, in an arbitrary small neighbourhood of z_0 the difference between $f(z)$ and an arbitrarily chosen number A can be made arbitrarily small. More precisely:

Given an arbitrary (generally complex) number A we can find a sequence of points $z_1, z_2, z_3, \ldots$ converging to the point z_0 such that

$$\lim_{n\to\infty} f(z_n) = A.$$

This is true even for $A = \infty$.

3. If the Laurent series has only the regular part, we say that $f(z)$ has a *removable singularity at* z_0. For example, the function

$$\frac{\sin z}{z} = 1 - \frac{z^2}{3!} + \frac{z^4}{5!} - \ldots$$

(see Remark 4) has a removable singularity at $z = 0$. If we define $f(z)$ at z_0 by the value a_0 (in this example by the value 1), then $f(z)$ is holomorphic in a neighbourhood of z_0. A (holomorphic) function $f(z)$ has a removable singularity at an isolated singular point if and only if it is bounded in the neighbourhood of this point.

These three cases are the only possible types of singularities of (single-valued) holomorphic functions at isolated singular points.

Multi-valued functions (§ 20.6) can have other types of singularities.

REMARK 11. If $f(z)$ is defined for all sufficiently large z (briefly: *in a neighbourhood of infinity*), we speak of a *Laurent series in the neighbourhood of infinity.* By the substitution

$$z - z_0 = \frac{1}{t} \tag{12}$$

we reduce the investigation of a Laurent series in the neighbourhood of infinity to the investigation of another Laurent series at $t = 0$. If this series has a pole, an essential singularity or a removable singularity at the point $t = 0$, we say that the original series (in powers of $z - z_0$) has a *pole*, an *essential singularity* or a *removable singularity at infinity*, respectively. For example, if the regular part of the original Laurent series has an infinite number of terms, then the corresponding Laurent series in the variable t has its *principal* part with an infinite number of terms and we say that $f(z)$ has an essential singularity at infinity. (Cf. also Remark 10, point 2.)

REMARK 12. A holomorphic function given by a power series with an infinite number of terms and radius of convergence $r = +\infty$ (i.e. converging for every finite z) is called an *entire* or *integral transcendental function.* As examples we have the functions $\sin z$, $\cos z$, e^z (Remark 4). An entire transcendental function has its only singular point at $z = \infty$ and this point is an essential singularity. Conversely, every function holomorphic in the entire plane and having an essential singularity at infinity is an entire transcendental function.

Theorem 11 (*Liouville's Theorem*). *If a function $f(z)$ is holomorphic and bounded in the whole plane, then it is merely a constant.*

20.5. The Residue of a Function. Residue Theorem and its Applications

Definition 1. Let $f(z)$ be a holomorphic function and z_0 its isolated singular point. Then we can develop $f(z)$ in the neighbourhood of z_0 (for $z \neq z_0$) in its Laurent series (Theorem 20.4.9),

$$f(z) = \sum_{n=-\infty}^{\infty} a_n (z - z_0)^n = \cdots + \frac{a_{-2}}{(z - z_0)^2} + \frac{a_{-1}}{z - z_0} + a_0 + a_1(z - z_0) + \ldots . \quad (1)$$

The number a_{-1} is called *the residue of the function $f(z)$ at the point z_0*.

Example 1. The residue of the function

$$f(z) = \frac{1}{(z-1)(z-3)}$$

at the point $z = 1$ is $-\frac{1}{2}$ (Example 20.4.4).

REMARK 1. According to (20.4.10) we have (since $n = -1$)

$$a_{-1} = \frac{1}{2\pi i} \int_c f(z)\,dz .$$

In some cases we can find the residue of a function more easily:

If $f(z)$ has a pole of the *first* order at z_0, then

$$a_{-1} = \lim_{z \to z_0} (z - z_0) f(z),$$

while if $f(z)$ has a pole of the k-th order ($k > 1$) at z_0, then

$$a_{-1} = \frac{1}{(k-1)!} \lim_{z \to z_0} \frac{d^{k-1}}{dz^{k-1}} [(z - z_0)^k f(z)] .$$

If $f(z)$ has, in the neighbourhood of z_0, the form

$$f(z) = \frac{\varphi(z)}{\psi(z)}$$

($\varphi(z)$, $\psi(z)$ being holomorphic in the neighbourhood of the point z_0) and if $\varphi(z_0) \neq \neq 0$, $\psi(z_0) = 0$, $\psi'(z_0) \neq 0$, then $f(z)$ has a pole of the first order at z_0. In this case, the residue is equal to

$$a_{-1} = \frac{\varphi(z_0)}{\psi'(z_0)}. \tag{2}$$

Example 2. The function

$$f(z) = \frac{1}{\sin z}$$

has at $z = 0$ a pole of the first order and the residue is $a_{-1} = 1$, since

$$\sin 0 = 0, \quad \cos 0 \neq 0, \quad a_{-1} = \frac{1}{\cos 0} = \frac{1}{1} = 1.$$

Theorem 1 (*Residue Theorem*). *Let c be a simple piecewise smooth closed curve, positively oriented with respect to its interior V. Let $f(z)$ be a function holomorphic in V with the exception of a finite number of singular points $z_1, z_2, \ldots, z_n$* (Fig. 20.9, where $n = 3$) *and continuous in $\overline{V} = V \cup c$ with the exception of these points. Then the integral*

$$\frac{1}{2\pi i} \int_c f(z)\,dz$$

is equal to the sum of residues at the points $z_1, z_2, \ldots, z_n$. In symbols:

$$\frac{1}{2\pi i} \int_c f(z)\,dz = \sum_{k=1}^{n} \operatorname{res}[f(z)]_{z=z_k}. \tag{3}$$

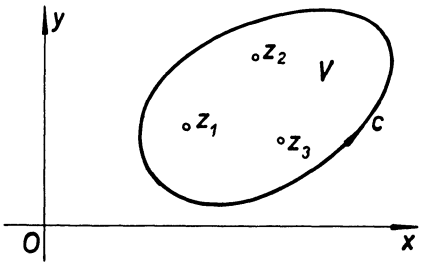


Fig. 20.9.

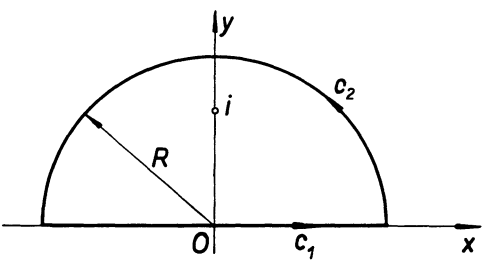


Fig. 20.10.

REMARK 2. This theorem has many applications, one of the most important of which is its use in evaluating integrals which cannot be calculated with the help of elementary primitive functions, e.g.

$$\int_0^\infty \frac{\sin x}{x}\,dx, \quad \int_0^\infty \frac{x^\alpha}{1+x}\,dx \quad (-1 < \alpha < 0),$$

etc. Here we shall give an illustrative example showing the fundamental idea of the method:

Example 3. Let us evaluate the integral

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}x}{1+x^2}. \tag{4}$$

(In this case, of course, the integral can be computed by means of the primitive function $\arctan x$; we know that its value is π.)

The integral (4) is convergent and therefore

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}x}{1+x^2} = \lim_{R\to+\infty} \int_{-R}^{R} \frac{\mathrm{d}x}{1+x^2}. \tag{5}$$

Let us draw a semi-circle c_2 with centre at the origin and radius $R > 1$ and write $c = c_1 \cup c_2$, where c_1 is the segment $-R \leqq x \leqq R$ of the x-axis. The orientation is evident from Fig. 20.10. Within the curve c the function

$$f(z) = \frac{1}{1+z^2}$$

has only one singular point, namely $z = +\mathrm{i}$, where it has a pole of the first order. The residue can be calculated, for example, by means of (2) (putting $\varphi(z) \equiv 1$, $\psi(z) = 1 + z^2$),

$$a_{-1} = \left[\frac{1}{2z}\right]_{z=\mathrm{i}} = \frac{1}{2\mathrm{i}}.$$

Then from (3), we obtain

$$\int_c f(z)\,\mathrm{d}z = 2\pi\mathrm{i}\frac{1}{2\mathrm{i}} = \pi. \tag{6}$$

The relation (6) holds for every $R > 1$, hence

$$\lim_{R\to+\infty} \int_c f(z)\,\mathrm{d}z = \pi.$$

Consequently (by (5))

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}x}{1+x^2} = \lim_{R\to+\infty} \int_{-R}^{R} \frac{\mathrm{d}x}{1+x^2} = \lim_{R\to+\infty} \int_{c_1} f(z)\,\mathrm{d}z =$$

$$= \lim_{R\to+\infty} \int_c f(z)\,\mathrm{d}z - \lim_{R\to+\infty} \int_{c_2} f(z)\,\mathrm{d}z = \pi - \lim_{R\to+\infty} \int_{c_2} f(z)\,\mathrm{d}z. \tag{7}$$

However, for every $R > 1$, we have (according to Theorem 20.2.2)

$$\left|\int_{c_2} f(z)\,\mathrm{d}z\right| = \left|\int_{c_2} \frac{\mathrm{d}z}{1+z^2}\right| \leqq \pi R \,.\, \max\left|\frac{1}{1+z^2}\right|_{c_2} \leqq \pi R \frac{1}{R^2-1}, \tag{8}$$

because $|z^2 + 1| \geqq |z^2| - 1$ and on c_2 we have $|z| = R$. Since

$$\lim_{R \to +\infty} \frac{R}{R^2 - 1} = 0$$

we have, using (8),

$$\lim_{R \to +\infty} \left| \int_{c_2} f(z)\, \mathrm{d}z \right| = 0$$

and hence also

$$\lim_{R \to +\infty} \int_{c_2} f(z)\, \mathrm{d}z = 0$$

whence, using (7),

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}x}{1 + x^2} = \pi\,.$$

20.6. Logarithm, Power. Analytic Continuation. Analytic Functions

The so-called *principal branch of the function "logarithm z"* is defined by the relation

$$\ln_0 z = \ln r + \mathrm{i}\varphi\,, \quad r > 0\,, \quad -\pi < \varphi \leqq \pi\,, \tag{1}$$

where r is the absolute value and φ the argument of the number z, i.e.

$$z = r(\cos\varphi + \mathrm{i}\sin\varphi)\,, \quad -\pi < \varphi \leqq \pi\,. \tag{2}$$

The function (1) is holomorphic ($(\ln_0 z)' = 1/z$) in the entire (open) plane with the exception of the point $z = 0$ and all the points on the negative real axis; on this axis it is discontinuous, because its imaginary part has a jump of -2π. In order to eliminate this discontinuity, we define the so-called *second branch of the logarithmic function* by the relation

$$\ln_1 z = \ln_0 z + 2\pi\mathrm{i} = \ln r + \mathrm{i}(\varphi + 2\pi)\,.$$

Let U be a neighbourhood of an arbitrary point z_0 on the negative real axis which does not contain the point $z = 0$. Let us define, in this neighbourhood,

$$\begin{aligned} f(z) &= \ln_0 z \quad \text{if } \operatorname{Im} z \geqq 0 \quad \text{(i.e. for points of the upper half-plane)}\,, \\ f(z) &= \ln_1 z \quad \text{if } \operatorname{Im} z < 0 \quad \text{(i.e. for points of the lower half-plane)}\,. \end{aligned} \tag{3}$$

Then $f(z)$ is continuous and holomorphic in U. We say that $\ln_1 z$ is the *analytic continuation of the function* $\ln_0 z$ *from the upper half-plane into the lower half-plane* through the negative real axis.

Similarly, we define other branches of the logarithmic function,

$$\ln_n z = \ln_0 z + 2n\pi i = \ln r + i(\varphi + 2n\pi) \quad (n \text{ an integer}).$$

Definition 1. The set of all these branches is called the *multi-valued function* $\ln z$ and the function $\ln_0 z$ the *principal branch* of the function $\ln z$.

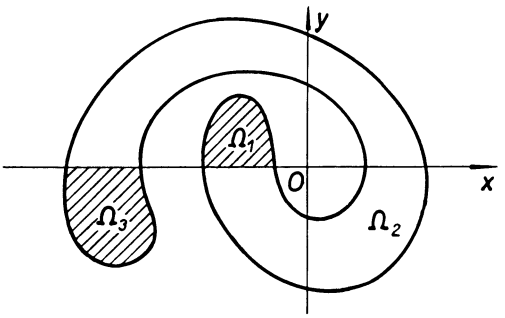


Fig. 20.11.

REMARK 1. If G is an arbitrary *simply connected* region which does not contain the point $z = 0$, then we can assign to each point $z \in G$ a value of the function $\ln z$ (i.e. the value of a certain branch of this function) such that the function $\ln z$ is holomorphic (hence single-valued) in G.

Example 1. In the region G illustrated in Fig. 20.11, we can choose the branches of the function $\ln z$ in the following way:

$$\ln z = \ln_0 z \quad \text{in } G_1,$$
$$\ln z = \ln_1 z \quad \text{in } G_2,$$
$$\ln z = \ln_2 z \quad \text{in } G_3.$$

REMARK 2. Each of the branches of the function $\ln z$ is single-valued in the Gaussian plane. Let us consider an infinite number of Gaussian planes $\ldots, R_{-2}, R_{-1}, R_0, R_1, R_2, \ldots$ and let us assign to each branch of $\ln_n z$ the plane R_n. Let us imagine that these planes are made of stiff paper and that they are cut along the negative real axis. We have obtained the function $\ln z$ by "joining" the function $\ln_1 z$ to the function $\ln_0 z$, etc. Similarly, let us "attach" the plane R_1 to the plane R_0 so that we join the left lower half-plane R_1 to the left upper half-plane R_0 along the negative real axis. In the same way, let us "attach" the plane R_2 to R_1, etc., and proceed similarly with the planes $R_{-1}, R_{-2}, \ldots$. Thus we get a surface consisting of an infinity of sheets which is called the *Riemann surface* of the function $\ln z$. To each sheet of this surface, there corresponds a certain branch of the function $\ln z$

(and conversely). If z moves round the point $z = 0$ in the positive sense and does not leave the Riemann surface (i.e. if z, after moving once round the point $z = 0$, passes from R_n to R_{n+1}), then the values of the function $\ln z$ change continuously with z. Since during this movement the function $\ln z$ never returns to the initial value (the imaginary part of $\ln z$ continually increases), the singular point $z = 0$ of $\ln z$ is called a *branch point of infinite order* or a *transcendental branch point* of this function. A singularity of this kind is often called a *logarithmic singularity.*

REMARK 3. The notation $\ln z$, $\ln_n z$ is not uniformly used in the literature. The multi-valued function here denoted by $\ln z$ is often denoted by $\log z$ or $\operatorname{Log} z$. For the argument φ of the function $\ln_0 z$, the interval $[0, 2\pi)$ is often chosen. We therefore recommend the reader to study the notation individually in each publication.

REMARK 4. The co-called *general power of* z is a (generally multivalued) function defined by the relation

$$z^n = e^{n \ln z} = e^{n[\ln r + i(\varphi + 2k\pi)]} \quad (k \text{ an integer}). \tag{4}$$

For real *irrational* n, this function is infinitely multi-valued, the point $z = 0$ being a transcendental branch point, and its Riemann surface has an infinity of sheets. If n is rational, then, since $e^{i.2\pi l} = 1$ if l is an integer ($e^{i.2\pi l} = \cos 2\pi l + i \sin 2\pi l = 1$), we come back after a finite number of rotations of the point z round the origin (i.e. for a certain k in equation (4)) to the same value. The function z^n then has only a finite number of different branches, the point $z = 0$ is a so-called *algebraic branch point* (*branch point of finite order*). The corresponding Riemann surface has then only a finite number of sheets.

For example, the function $\sqrt{z} = z^{1/2}$ has two branches (in (4) we choose $k = 0$ and $k = 1$), each of which is a (-1)-multiple of the other (since $e^{(1/2).2\pi i} = -1$). To each of them, there corresponds one sheet of a two-sheet Riemann surface. Let us denote them by R_0, R_1. We obtain the Riemann surface (cf. Remark 2) by "attaching" the left lower half-plane of the plane R_1 to the left upper half-plane of the plane R_0 and then (after a further rotation of the point z round the origin) the left upper half-plane of the plane R_1 to the left lower half-plane of the plane R_0. (All these operations we naturally perform only mentally.)

Generally, for a natural number m, the function $\sqrt[m]{z} = z^{1/m}$ has m (different) branches, which can be defined (for example) by the following relations:

$$z_1 = e^{(1/m) \ln_0 z}, \quad z_2 = e^{(1/m) \ln_1 z}, \quad \ldots, \quad z_m = e^{(1/m) \ln_{m-1} z}.$$

REMARK 5. Everything which has been said concerning the functions $\ln z$, z^n is also valid for the functions $\ln(z - z_0)$, $(z - z_0)^n$; it is sufficient to substitute $z - z_0 = z'$. Of course, the branch point will now be the point z_0.

Definition 2. Let a region Ω be the intersection of the regions O_1 and O_2 (Fig. 20.12). Let a holomorphic function $f_1(z)$ be defined in O_1. If there exists a holomorphic function $f_2(z)$ in O_2 such that $f_1(z) = f_2(z)$ in Ω, we say that $f_2(z)$ is an *analytic continuation (extension) of the function* $f_1(z)$ *from* O_1 *onto* O_2 *through* Ω.

REMARK 6. If there exists such a function, then it is the only one, as the following theorem implies:

Theorem 1. *Let* M *be an infinite set of points lying in a region* B *and having in* B *at least one point of accumulation* (Definition 22.1.3). (For example, a segment or a region lying in B satisfy these conditions.) *Let a function* $g(z)$ *be defined on* M. *If there exists such a holomorphic function* $G(z)$ *in* B *that* $G(z) = g(z)$ *at the points of the set* M, *then this function is unique.*

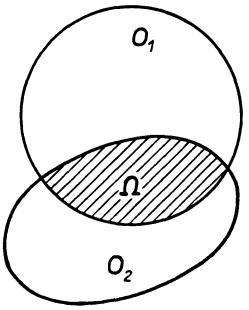


Fig. 20.12.

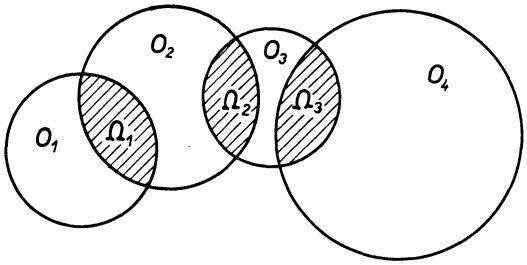


Fig. 20.13.

REMARK 7. Theorem 1 implies, for example, that the function $\sin z$ defined in Remark 20.4.4 is the only holomorphic function in the Gaussian plane which coincides, for real z, with the function $\sin x$ defined in the calculus of functions of a real variable.

REMARK 8. Let us have a so-called *chain of regions* $O_1, O_2, \ldots, O_n$, i.e. a system of regions O_k such that each region O_k $(k = 2, 3, \ldots, n-1)$ has non-empty intersections Ω_{k-1}, Ω_k precisely with the regions O_{k-1} and O_{k+1} and that these intersections Ω_{k-1}, Ω_k are simply connected regions, all mutually disjoint (Fig. 20.13).

Let $f_k(z)$ be holomorphic functions defined in O_k, such that $f_k(z) = f_{k+1}(z)$ in Ω_k. Then the function $f_n(z)$ is called an *analytic continuation of the function* f_1 *onto the region* O_n *through the given chain of regions.*

REMARK 9. If we make two chains of regions, both leading to the region O_n, it may happen that we reach O_n in each case with a different analytic continuation of the function $f_1(z)$. Or, in other words, if the chain under consideration is closed

($O_n = O_1$), then, after passing through the chain, we come to the region O_n with a holomorphic function different from the one we started with. If, for instance, we start from the circle K_1 (Fig. 20.14) with the value of the function $\ln_0 z$, we return to K_1 after passing through the chain with values of the function $\ln_1 z$.

REMARK 10. We also often denote the function $f_2(z)$ of Definition 2, which is an analytic continuation of the function $f_1(z)$, by $f_1(z)$ again and say that we *have continued the function* $f_1(z)$ (analytically) *from the region* O_1 *onto the region* $O_1 \cup O_2$. We similarly speak of a continuation (extension) of the function $f_1(z)$ in the case of a chain of regions. A function (generally multi-valued) which arises as an analytic continuation of a holomorphic function onto a region D is called an *analytic function in* D. If we carry out all possible analytic continuations of the function $f_1(z)$ we come to a so-called *complete analytic function* (generally multivalued) which cannot be further continued. Its domain is called the *natural domain of the analytic function* in question.

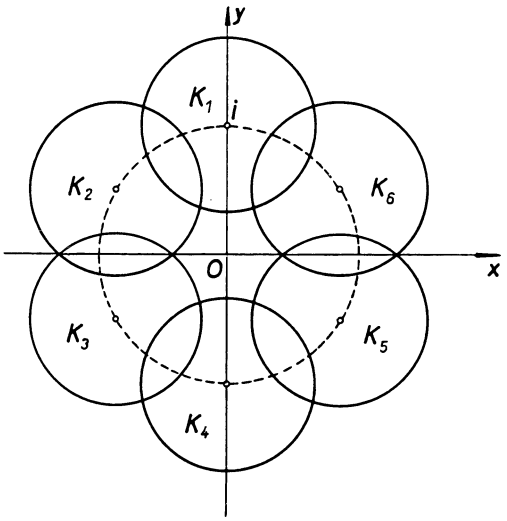


Fig. 20.14.

For example, the natural domain of the analytic function $\ln z$ is the whole (open) plane except the point $z = 0$. In the same way (as in the case of the function $\ln z$) we speak also in the general case of *branches* of a (multi-valued) analytic function.

REMARK 11. The idea of a simple method of constructing the analytic continuation of a given function was suggested by Weierstrass: Let $f(z)$ be a holomorphic function in a region G. Let us choose $z_1 \in G$ and expand $f(z)$ in a power series with centre at z_1. This series converges at least in the circle K_1 which lies in G and has

its radius equal to the distance of the point z_1 from the boundary of the region G. It may happen that the circle of convergence is greater. Then $f(z)$ is continued onto the region $G \cup K_1$. Further, we choose $z_2 \in G$, and proceed similarly at all the points of G. We then expand the function $f(z)$ thus continued into a power series at the points of the new region, and so on. In this way, we obtain a complete analytic function.

REMARK 12. Let us consider the function $f(z)$ defined by the series

$$f(z) = z + z^2 + z^6 + \cdots + z^{n!} + \ldots . \tag{5}$$

The series (5) converges in the circle $|z| < 1$. It can be shown that the function (5) cannot be continued to any region greater than this circle. We say that no point of the circumference of the circle $|z| = 1$ is a *point of continuability* of the function (5) or that this circumference is a *natural boundary* of the function. The function $f(z)$ defined by the series

$$1 - z + z^2 - z^3 + \ldots \tag{6}$$

also converges in the circle $|z| < 1$. Its sum equals $1/(1 + z)$. The function (6) can be analytically continued to the entire plane with the exception of the point $z = -1$. The analytic function so obtained is actually single-valued and is the function $1/(1 + z)$. Every point of the circumference $|z| = 1$ is a point of continuability of the function under consideration, with the exception of the point $z = -1$ which is its only singular point.

Theorem 2. *Let $f(z)$ be an analytic (generally multi-valued) function in a region G. Let D be a simply connected region lying in G. Then we can assign to the function $f(z)$ at every point $z \in D$ such a value (i.e. the value of one of its branches) that the function thus defined is holomorphic (hence single-valued) in D.*

REMARK 13. Compare Remark 1 and Example 1. Theorem 2 is a special case of a more general Monodromy Theorem, see e.g. [408].

B. FUNCTIONS OF SEVERAL COMPLEX VARIABLES

By JAROSLAV FUKA

INTRODUCTORY REMARK. The theory of functions of several complex variables is not a straightforward generalization of the theory of functions of one complex variable: in some questions concerning holomorphy, integral formulae, Taylor expansion etc., there is a certain similarity, but many essential properties of functions

of several complex variables have no analogue in the one-dimensional case. It is very important to have a good geometrical idea of the regions in which the functions of several complex variables are investigated. Therefore, in § 20.7 some regions important in the theory of these functions as well as in applications are analysed.

Recently, functions of several complex variables have been applied in many branches of theoretical physics, especially in the axiomatic field theory (holomorphic relativistic fields).

In the following paragraphs 20.7–20.11 we write $z = (z_1, z_2, \ldots, z_n)$, where $z_j = x_j + \mathrm{i}y_j$ $(j = 1, 2, \ldots, n)$, x_j, y_j real, and speak simply of the *point* z. The space $\mathbb{C}^n$ of such points is identified with the space $\mathbb{R}^{2n}$ of the corresponding points $(x_1, y_1, x_2, y_2, \ldots, x_n, y_n)$ and the distance $d(z, w)$ of two points $z = (z_1, z_2, \ldots, z_n)$, $w = (w_1, w_2, \ldots, w_n)$, $(z_j = x_j + \mathrm{i}y_j$, $w_j = u_j + \mathrm{i}v_j)$, is then defined by the formula

$$d(z, w) = \|w - z\| = \sqrt{\sum_{j=1}^{n} [(u_j - x_j)^2 + (v_j - y_j)^2]} = \sqrt{\sum_{j=1}^{n} |w_j - z_j|^2} \,.$$

In this way, all concepts such as the *δ-neighbourhood of a point*, the *open* or *closed set*, the *boundary of a set, bounded sets, regions* (= open connected sets), *compact sets* (= closed bounded sets), etc., can be defined in $\mathbb{C}^n$ word by word similarly as in § 22.1.

20.7. Important Regions in $\mathbb{C}^n$

We give some important examples of regions and their boundaries in $\mathbb{C}^n$ and point out how one visualizes them for the purpose of studying holomorphic functions. The reader is recommended to compare the case $n = 1$ with $n \geqq 2$.

Example 1. The *ball* with radius $r > 0$ and centre at the point $a = (a_1, a_2, \ldots, a_n) \in \mathbb{C}^n$:

$$B^n(a, r) = \{z \in \mathbb{C}^n : \|z - a\| < r\} \,.$$

It is the ordinary Euclidean ball of dimension $2n$; its boundary is the $(2n - 1)$-dimensional Euclidean sphere

$$S^{2n-1}(a, r) = \{z \in \mathbb{C}^n : \|z - a\| = r\} \,.$$

Example 2. The *polydisk* (*polycylinder*) with radius $r > 0$ and centre at the point $a \in \mathbb{C}^n$:

$$U^n(a, r) = \{z \in \mathbb{C}^n : |z_j - a_j| < r,\ j = 1, 2, \ldots, n\} \,.$$

Evidently $U^n(a, r) = U(a_1, r) \times U(a_2, r) \times \cdots \times U(a_n, r)$, i.e. $U^n(a, r)$ is the Cartesian product of n plane disks $U(a_j, r)$ of radius r centred at the points a_j. The boundary of $U^n(a, r)$ is the set of all points z, for which at least one coordinate, say z_k, lies on the boundary of the k-th disk $U(a_k, r)$, i.e. $|z_k - a_k| = r$, and the other coordinates z_j, $j \neq k$, change arbitrarily in the closed disks $\overline{U(a_j, r)}$.

We describe in more detail the *bidisk*

$$U^2 = \{z \in \mathbb{C}^2 : z = (z_1, z_2), |z_1| < 1, |z_2| < 1\}$$

of radius 1 centred at the origin. This four-dimensional body is the intersection of two unbounded four-dimensional cylinders $x_1^2 + y_1^2 < 1$ (x_2, y_2 arbitrary) and $x_2^2 + y_2^2 < 1$ (x_1, y_1 arbitrary). The boundary $\partial U^2 = \Gamma_1 \cup \Gamma_2$ is three-dimensional. Here

$$\Gamma_1 = \bigcup_{0 \leqq \theta \leqq 2\pi} \{z_1 = e^{i\theta}, |z_2| \leqq 1\}, \quad \Gamma_2 = \bigcup_{0 \leqq \theta \leqq 2\pi} \{|z_1| \leqq 1, z_2 = e^{i\theta}\}.$$

The *frame* (the *distinguished boundary*) is the intersection

$$\Gamma = \Gamma_1 \cap \Gamma_2 = \{z \in \mathbb{C}^2 : |z_1| = 1, |z_2| = 1\},$$

the Cartesian product of two circles, hence geometrically the torus (see § 3.2 and Fig. 20.15).

More generally, one can consider the *polydisk*

$$U^n(a, \boldsymbol{r}) = \{z \in \mathbb{C}^n : |z_j - a_j| < r_j, j = 1, 2, \ldots, n\}$$

with the vectorial radius $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)$.

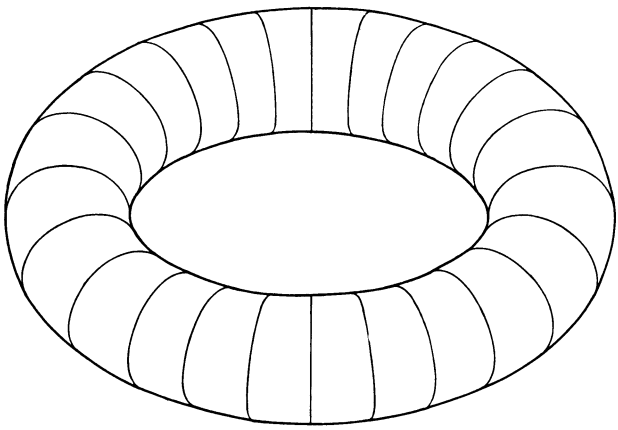


Fig. 20.15.

Example 3. The *Reinhardt domain* with centre at the point $a \in \mathbb{C}^n$ is a region G satisfying the following condition: If the point $z = (z_1, \ldots, z_n)$ belongs to G, then every point

$$w = \left(a_1 + (z_1 - a_1)e^{i\theta_1}, \ldots, a_n + (z_n - a_n)e^{i\theta_n}\right), \quad 0 \leqq \theta_j < 2\pi, \quad j = 1, 2, \ldots, n,$$

belongs to G. Without loss of generality one can put $a = O = (0, 0, \ldots, 0)$. Such a Reinhardt domain (thus with centre at the point $a = O$) contains, with every point $z = (z_1, \ldots, z_n)$, also every point with the same moduli $|z_j|$ and all possible arguments θ_j, $0 \leqq \theta_j < 2\pi$, $j = 1, 2, \ldots, n$, i.e., the n-dimensional torus

$$T_n = \{\zeta_1 \in \mathbb{C};\ |\zeta_1| = |z_1|\} \times \cdots \times \{\zeta_n \in \mathbb{C};\ |\zeta_n| = |z_n|\},$$

the Cartesian product of n circles. Hence it is possible to represent the $2n$-dimensional Reinhardt domain in an n-dimensional diagram, whose construction is clear from Fig. 20.16, where the diagrams of the unit ball and unit polydisk in $\mathbb{C}^2$ and $\mathbb{C}^3$ are presented. In the two-dimensional diagram, every point $(|z_1|, |z_2|)$ represents a torus

$$\{\zeta \in \mathbb{C}^2;\ |\zeta_1| = |z_1|,\ |\zeta_2| = |z_2|\},$$

and analogously in the three- or n-dimensional diagram every point represents a three- or n-dimensional torus, i.e. the Cartesian product of three or n circles.

The Reinhardt domain is called *complete* if it contains, with every point z, also every point $\zeta = (\zeta_1, \ldots, \zeta_n)$ such that $|\zeta_j - a_j| \leqq |z_j - a_j|$, $j = 1, 2, \ldots, n$, i.e. if it contains, with every point z, the closed polydisk centred at the point a with the vectorial radius $\boldsymbol{r} = (|z_1 - a_1|, \ldots, |z_n - a_n|)$. Reinhardt domains sketched in Figs. 20.16 a–d are complete, the domain in Fig. 20.16 e is not complete. For $n = 1$, every complete Reinhardt domain is a disk $\{z \in \mathbb{C}:\ |z - a| < R\}$, every non-complete Reinhardt domain is an annulus $\{z \in \mathbb{C}:\ r < |z - a| < R\}$. The balls (Example 1) and the polydisks (Example 2) are complete Reinhardt domains. Complete Reinhardt domains play, for the Taylor expansion of holomorphic functions in $\mathbb{C}^n$, $n > 1$, a similar role as the disks do in the case $n = 1$ (see e.g. [187], [381]).

Example 4. Given a region B in $\mathbb{R}^n$, the *tube domain with the base* B is a region in $\mathbb{C}^n$ of the form $T_B^n = B + i\mathbb{R}^n$ (or $\mathbb{R}^n + iB$), i.e. $(x_1 + iy_1, \ldots, x_n + iy_n) \in T_B^n$ if and only if $(x_1, \ldots, x_n) \in B$ (or $(y_1, \ldots, y_n) \in B$). For $n = 1$, the tube domains are strips $\alpha < x < \beta$ (or $\alpha < y < \beta$) and halfplanes $\alpha < x$ (or $\alpha < y$), $x < \alpha$ (or $y < \alpha$).

Example 5. The following tube domains M_+ and M_-, the co-called *complexified light* (*absolute*) *cones*, play a considerable role in mathematical physics, especially in the axiomatic quantum field theory, being natural domains for defining *holomorphic*

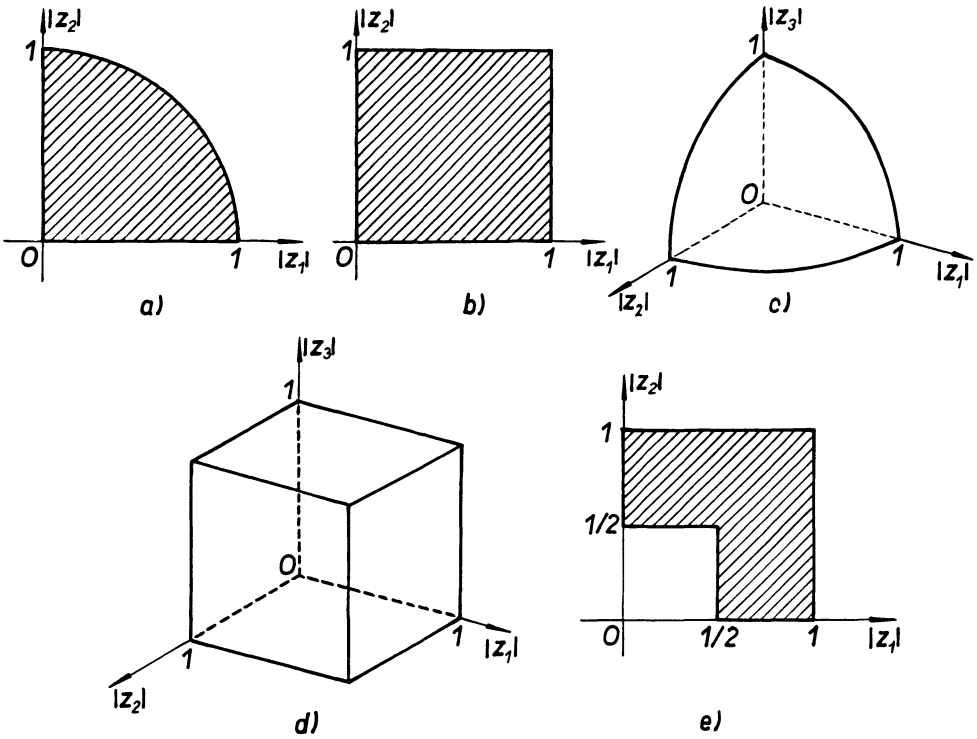


Fig. 20.16 a, b, c, d, e. a) Fourdimensional sphere $\|z\| < 1$. b) Sixdimensional sphere $\|z\| < 1$. c) Bidisc U^2. d) Tridisc U^3. e) $\Omega = U^2 \setminus \overline{U^2(\frac{1}{2})}$.

relativistic fields. $M_+ = \mathbb{R}^4 + \mathrm{i}C_+$, where

$$C_+ = \{(y_0, y_1, y_2, y_3);\ y_0^2 - y_1^2 - y_2^2 - y_3^2 > 0,\ y_0 > 0\},$$

is the *forward light cone* representing physically the propagation of a light signal sent in all directions from the origin $y_1 = y_2 = y_3 = 0$ at the time $y_0 = 0$, and similarly $M_- = \mathbb{R}^4 + \mathrm{i}C_-$, where

$$C_- = \{(y_0, y_1, y_2, y_3);\ y_0^2 - y_1^2 - y_2^2 - y_3^2 > 0,\ y_0 < 0\},$$

is the *backward light cone* representing physically the points, from which the light signal may reach the origin at the time $y_0 = 0$.

The boundary of M_+ (M_-) is the seven-dimensional set $\mathbb{R}^4 + \mathrm{i}\partial C_+$ ($\mathbb{R}^4 + \mathrm{i}\partial C_-$), where

$$\partial C_+ = \{(y_0, y_1, y_2, y_3);\ y_0^2 - y_1^2 - y_2^2 - y_3^2 = 0,\ y_0 \geqq 0\},$$
$$\partial C_- = \{(y_0, y_1, y_2, y_3);\ y_0^2 - y_1^2 - y_2^2 - y_3^2 = 0,\ y_0 \leqq 0\}.$$

The common part of $\mathbb{R}^4 + i\partial C_+$ and $\mathbb{R}^4 + i\partial C_-$ is the four-dimensional "edge" $\mathbb{R}^4 + iO$, where $O = (0, 0, 0, 0)$.

20.8. Functions of Several Complex Variables. Complex Derivative, Complex Differential, Holomorphic Functions

In what follows we will be concerned (until otherwise stated) with functions $f(z_1, \ldots, z_n)$ (with complex values, in general) defined in a region $G \subset \mathbb{C}^n$. For such functions the concepts of *limit, continuity, derivative* etc. are defined in a quite similar way as for $n = 1$ (§ 20.1). The following definitions are of fundamental importance.

Definition 1. Let U_δ be a neighbourhood of a given point $a = (a_1, \ldots, a_n)$. We say that

(i) f has a *complex derivative* with respect to the variable z_j at the point a (we write $\dfrac{\partial f}{\partial z_j}(a)$) if the limit

$$\lim_{h \to 0} \frac{f(a_1, \ldots, a_{j-1}, a_j + h, a_{j+1}, \ldots, a_n) - f(a_1, \ldots, a_{j-1}, a_j, a_{j+1}, \ldots, a_n)}{h}$$

exists (i.e. if f has, at the point a, the complex derivative in the sense of Definition 20.1.8 as a function of the j-th variable z_j only, the other variables being kept fixed).

(ii) f is *complex differentiable* at the point a, if there exists a complex linear function

$$l(z) = l_1 z_1 + \cdots + l_n z_n \quad (l_i \in \mathbb{C}) \tag{1}$$

so that

$$\lim_{\|h\| \to 0} \frac{\|f(a + h) - f(a) - l(h)\|}{\|h\|} = 0 .$$

Here $h = (h_1, \ldots, h_n)$, $\|h\| = \sqrt{(h_1^2 + \cdots + h_n^2)}$, $O = (0, \ldots, 0)$, $a + h = (a_1 + h_1, \ldots, a_n + h_n)$. $l(z)$ is called the *complex differential of the function f at the point a.*

REMARK 1. From (ii) continuity of f at the point a easily follows and, moreover, existence of $\dfrac{\partial f}{\partial z_j}(a)$ for every $j = 1, 2, \ldots, n$, with $l_j = \dfrac{\partial f}{\partial z_j}(a)$; hence

$$l(z) = \sum_{j=1}^{n} \frac{\partial f}{\partial z_j}(a) z_j .$$

Definition 2. Let G be a region in $\mathbb{C}^n$. The function f is said to be *holomorphic in* G, if it has a complex differential at every point $a \in G$.

Example 1. For the function of two complex variables $f(z) = z_1\overline{z_2}$ we calculate $f(a+h) - f(a) = a_1\overline{h}_2 + \overline{a}_2 h_1 + h_1\overline{h}_2$. From $\left|\dfrac{h_j}{\sqrt{(h_1^2+h_2^2)}}\right| \leqq 1$ for $j = 1, 2$ it follows

$$\left|\frac{h_1 h_2}{\sqrt{(h_1^2+h_2^2)}}\right| \leqq \max(|h_1|, |h_2|), \text{ so that } \lim_{\|h\|\to 0} \frac{h_1\overline{h}_2}{\sqrt{(h_1^2+h_2^2)}} = 0.$$

Thus $f(z)$ is complex differentiable at every point $a = (0, a_2)$, with the complex differential $l(z) = \overline{a}_2 z_1$. At every point $a = (a_1, a_2)$, $a_1 \neq 0$, f is only real differentiable (as a function of four real variables x_j, y_j, $j = 1, 2$), not complex differentiable. The differential at such points is of the form $l(z) = \overline{a}_2 z_1 + a_1\overline{z}_2$, and this is not a complex linear function (i.e. a function of the form (1)), because of the "bad" second term $a_1\overline{z}_2$. The function f is not holomorphic in any domain $G \subset \mathbb{C}^2$, because every open set in $\mathbb{C}^2$ contains some point (a_1, a_2), $a_1 \neq 0$.

Theorem 1. *The following conditions are equivalent:*

(i) *f is holomorphic in the region $G \subset \mathbb{C}^n$;*

(ii) $\dfrac{\partial f}{\partial z_j}(a)$, $j = 1, 2, \ldots, n$, *exist at every point $a \in G$.*

REMARK 2. The assertion (ii)⇒(i) is a very deep theorem by Hartogs (see [271]). The finest step of the proof is the proof of continuity of f, following from (ii).

20.9. Cauchy-Riemann Equations. Pluriharmonic Functions

Let $f(z) = u(x_1, y_1, \ldots, x_n, y_n) + iv(x_1, y_1, \ldots, x_n, y_n)$ be holomorphic in $G \subset \mathbb{C}^n$. Because it is holomorphic in every variable z_j (see Remark 20.8.1), u and v satisfy in G, for every $j = 1, 2, \ldots, n$, the Cauchy-Riemann equations (see (20.1.7))

$$\frac{\partial u}{\partial x_j} = \frac{\partial v}{\partial y_j}, \quad \frac{\partial u}{\partial y_j} = -\frac{\partial v}{\partial x_j}. \tag{1}$$

Fix μ, ν, $1 \leqq \mu \leqq n$, $1 \leqq \nu \leqq n$. Differentiating the first of equations (1) for $j = \mu$ with respect to x_ν and the second one for $j = \nu$ with respect to y_μ and then the first of these equations for $j = \mu$ ($j = \nu$) with respect to y_ν (y_μ) and eliminating the corresponding derivatives of the function v, we obtain for the function u the system of n^2 equations

$$\frac{\partial^2 u}{\partial x_\mu \partial x_\nu} + \frac{\partial^2 u}{\partial y_\mu \partial y_\nu} = 0, \quad \frac{\partial^2 u}{\partial x_\mu \partial y_\nu} - \frac{\partial^2 u}{\partial x_\nu \partial y_\mu} = 0. \tag{2}$$

The same system is obtained for the function v.

Definition 1. Let G be a region in $\mathbb{C}^n$. The real function u with continuous second order derivatives in G, satisfying in G the system (2), is called *pluriharmonic.*

Example 1. Summing equations (2) for $\mu = \nu = 1, 2, \ldots, n$, we see, that every pluriharmonic function is harmonic (Definition 18.4.4). The following example shows that for $n > 1$ the class of pluriharmonic functions is a proper subclass of the class of harmonic functions. The function $u(z) = u(z_1, z_2, \ldots, z_n) = x_1x_2 + y_1y_2$ $(= \operatorname{Re} z_1\overline{z}_2)$ is clearly harmonic (it is linear in each variable) but not pluriharmonic:

$$\frac{\partial^2 u}{\partial x_1 \partial x_2} + \frac{\partial^2 u}{\partial y_1 \partial y_2} = 1 + 1 = 2 \neq 0 .$$

Theorem 1. *Let u be a pluriharmonic function in $G \subset \mathbb{C}^n$, $a \in G$. Then there exists a polydisk $U^n(a, r) \subset \overline{U^n(a, r)} \subset G$ and a function v pluriharmonic in $U^n(a, r)$ so that $f = u + \mathrm{i}v$ is holomorphic in $U^n(a, r)$.*

20.10. Local Properties of Holomorphic Functions. The Cauchy Integral Formula. The Taylor Expansion

By a repeated application of the one-dimensional Cauchy formula (Theorem 20.2.7) one obtains:

Theorem 1. *Let f be holomorphic in a region $G \subset \mathbb{C}^n$ and let $a \in G$. Then there exists a polydisk $U^n(a, \boldsymbol{r})$ with $\overline{U^n(a, \boldsymbol{r})} \subset G$, $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)$, so that for every $z \in U^n(a, \boldsymbol{r})$, $z = (z_1, z_2, \ldots, z_n)$, the following formula holds:*

$$f(z) = \frac{1}{(2\pi\mathrm{i})^n} \int_{C_{r_1}} \int_{C_{r_2}} \cdots \int_{C_{r_n}} \frac{f(\zeta)}{(\zeta_1 - z_1)(\zeta_2 - z_2) \ldots (\zeta_n - z_n)} \mathrm{d}\zeta_1 \mathrm{d}\zeta_2 \ldots \mathrm{d}\zeta_n , \quad (1)$$

where C_{r_j} is the circumference $|\zeta_j - a_j| = r_j$.

This formula is in a certain sense an analogue of the Cauchy integral formula (20.2.10), to which it reduces in the case $n = 1$. But there is an essential difference here: In one-dimensional case one integrates over the full topological boundary of the disk $U(a, r)$, while for $n > 1$ one integrates only over its part, namely over the frame (the distinguished boundary, Example 20.7.2) of the polydisk.

From (1), the following local properties of holomorphic functions can be derived in the same way as for $n = 1$:

Theorem 2 (*Existence of derivatives of all orders*). *Every function f holomorphic in a region G has, at every point $a \in G$, derivatives of all orders; they may be written in the form*

$$\frac{\partial^{\alpha_1+\cdots+\alpha_n} f}{\partial z_1^{\alpha_1} \dots \partial z_n^{\alpha_n}}(a) = \frac{\alpha_1! \dots \alpha_n!}{(2\pi\mathrm{i})^n} \int_{C_{r_1}} \cdots \int_{C_{r_n}} \frac{f(\zeta)}{(\zeta_1 - a_1)^{\alpha_1+1} \dots (\zeta_n - a_n)^{\alpha_n+1}} \mathrm{d}\zeta_1 \dots \mathrm{d}\zeta_n. \tag{2}$$

(Cf. (20.2.11). Notation as in Theorem 1.)

Theorem 3 (*Local Taylor's Expansion; cf. Theorem* 20.4.8). *Let f be holomorphic in G and let $a \in G$. Then for every $U^n(a, \boldsymbol{r})$ with $\overline{U^n(a, \boldsymbol{r})} \subset G$ the following formula holds:*

$$f(z) = \sum_{k=0}^{\infty} \sum_{\nu_1+\nu_2+\cdots+\nu_n=k} a_{\nu_1\nu_2\dots\nu_n} (z_1 - a_1)^{\nu_1} (z_2 - a_2)^{\nu_2} \dots (z_n - a_n)^{\nu_n}, \tag{3}$$

where

$$a_{\nu_1\nu_2\dots\nu_n} = \frac{1}{\nu_1!\ \nu_2!\ \dots \nu_n!} \frac{\partial^{\nu_1+\nu_2+\cdots+\nu_n} f}{\partial z_1^{\nu_1} \partial z_2^{\nu_2} \dots \partial z_n^{\nu_n}}(a) =$$

$$= \frac{1}{(2\pi\mathrm{i})^n} \int_{C_{r_1}} \cdots \int_{C_{r_n}} \frac{f(\zeta)}{(\zeta_1 - a_1)^{\nu_1+1} (\zeta_2 - a_2)^{\nu_2+1} \dots (\zeta_n - a_n)^{\nu_n+1}} \mathrm{d}\zeta_1 \mathrm{d}\zeta_2 \dots \mathrm{d}\zeta_n.$$

The series converges uniformly on compact subsets of $U^n(a, \boldsymbol{r})$.

Theorem 4 (*The Uniqueness Theorem*). *Let a function f be holomorphic in a region $G \subset \mathbb{C}^n$. If f and all its derivatives vanish at some point $a \in G$, then f is identically zero in G.*

REMARK 1. Here it is essential that G is connected. If G consists of two (or more) components, e.g. if $G = U^n(O, 1) \cup U^n(a, 1)$, where $a = (0, 0, \dots, 2)$, then the function f equal to 0 in $U^n(0, 1)$ and 1 in $U^n(a, 1)$ fulfils the conditions but not the assertion of Theorem 4.

Theorem 5 (*The Identity Theorem*). *Let $G \subset \mathbb{C}^n$ be a region and f_1, f_2 be holomorphic functions in G. Let $B \subset G$ be a non-empty open set on which f_1 and f_2 are equal to each other. Then $f_1 = f_2$ in G.*

REMARK 2. For $n \geqq 2$, Theorem 20.6.1 does not hold. The functions $f_1(z_1, z_2) = z_2$ and $f_2(z_1, z_2) = z_1 z_2$ are not identical, although $f_1(z_1, 0) = f_2(z_1, 0)$ holds for every $z_1 \in \mathbb{C}$.

20.11. On Some Different Properties of Functions of One and Several Complex Variables. Analytic Continuation. Domain of Holomorphy. Biholomorphic mapping

The most important part of the theory of holomorphic functions of several variables ($n > 1$) is that one which is in some sense radically different from the complex functions theory of one variable. This concerns, first of all, global properties of holomorphic functions. We give only a few examples of such rather surprising phenomena here. We suppose $n \geqq 2$ all the time.

Example 1. *Analytic continuation.* Let $G = U^2 \setminus \overline{U^2(\frac{1}{2})}$ be the region from Fig. 20.16 e. Then every function f holomorphic in G can be holomorphically extended to the whole U^2, i.e. to every function f, holomorphic in G, there exists a unique F holomorphic in U^2 so that $F = f$ holds in G. This situation is thus different from that in one variable, where in every domain $G \subset \mathbb{C}$ it is possible to construct a function f, which has no holomorphic extension to a greater domain.

Theorem 1 (*the so-called Kugelsatz*). *Let G be a region in $\mathbb{C}^n$, $K \subset G$ a compact set, $G \setminus K$ connected. Then every function holomorphic in $G \setminus K$ can be uniquely extended to a function holomorphic in G.*

A specific theorem on analytic continuation in tube domains with conic bases is the following one:

Theorem 2. *Let M_+, M_- be the light cones from Example* 20.7.5. *Let the complex function f, defined in the set $W = M_+ \cup M_- \cup (\mathbb{R}^4 + \mathrm{i}O)$, be continuous in W and holomorphic in $M_+ \cup M_-$. Then there is a region G in $\mathbb{C}^4$ containing W and a function F holomorphic in G such that $F = f$ on W.*

REMARK 1. This theorem is a very special case of the so-called *Edge of the Wedge Theorem*, which has fundamental application in the scattering theory (see [477]). As we have seen in Example 20.7.5, $\mathbb{R}^4 + \mathrm{i}O$ is the common boundary of M_+ and M_-, so it is the "edge" of the "wedge" W. Every function, holomorphic on the open set $M_+ \cup M_-$ and continuous up to its four-dimensional "edge", can be extended holomorphically to the eight-dimensional neighbourhood of the "edge" — a truly surprising fact; cf. also the relevant assertion about light cones in Example 2 below.

The *domains of holomorphy* in $\mathbb{C}^n$ are, roughly speaking, regions, in which there exists a holomorphic function that can be no more extended holomorphically to a greater domain. The following exact definition (seemingly complicated) takes into account the possibility that the boundary of a domain can "intersect" itself and that the phenomenon of the multivalence, already known from the construction of the logarithmic function (see § 20.6), could occur.

Definition 1. A region $G \subset \mathbb{C}^n$ is called the *domain of holomorphy*, if to every pair (U_1, U_2) of non-empty sets in $\mathbb{C}^n$, $U_1 \subset U_2 \cap G$, U_2 a region not contained in G, there exists a function f holomorphic in G with the following property: for every function F holomorphic in U_2 there is a point $z_1 \in U_1$ such that $f(z_1) \neq F(z_1)$, i.e. f cannot be the restriction on U_1 of any function holomorphic in U_2 (see Fig. 20.17).

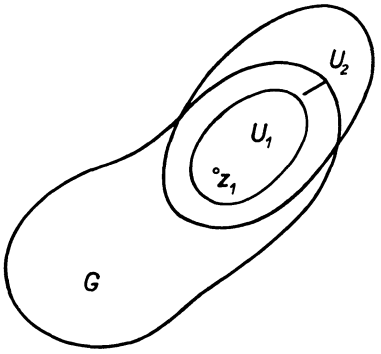


Fig. 20.17.

Example 2. The bidisk U^2 is a domain of holomorphy: take $f(z_1, z_2) = \varphi(z_1).\varphi(z_2)$, where φ is the function (20.6.5). More generally, every polydisk is a domain of holomorphy. Every convex domain in $\mathbb{C}^n$ (especially the ball) is a domain of holomorphy. The tube domain (see Example 20.7.4) is a domain of holomorphy if and only if its base is convex, thus especially the complexified forward and backward light cones from Example 20.7.5 are domains of holomorphy. The region from Example 1 and the regions of the form $G \setminus K$, $K \subset G$, K compact, are not domains of holomorphy (Theorem 1).

There arises a fundamental problem: How to characterize geometrically the domains of holomorphy. For its solution the reader is referred to the books ([187], [197], [271], [381]).

An easy consequence of the Kugelsatz (Theorem 1) are the following facts with no analogues in the case $n = 1$:

Theorem 3. *Let f be a holomorphic function in a region $G \subset \mathbb{C}^n$, $Z_a = \{z \in G; f(z) = a\}$ the level set of f. Then*

(i) *no point of Z_a is isolated,*

(ii) *Z_a is either empty or reaches up to the boundary of G.*

REMARK 2. For $n = 1$, every point of Z_a is isolated, except that $f(z) = \text{const.}$ identically. The set Z_a can evidently be finite and, consequently, cannot reach up

to the boundary, in such a case — e.g.

$$G = \{z;\ |z| < 1\}, \quad f(z) = z(z - \tfrac{1}{2}), \quad a = 0.$$

REMARK 3. In contrast to functions of one complex variable, the singularities of a holomorphic function cannot be isolated for $n \geqq 2$ (as stated in Theorem 3). Therefore the quotients of two holomorphic functions, the so called *meromorphic functions*, can be interpreted as functions only in a very broad sense. For example, the function $f(z) = z_2/z_1$ has a pole, i.e.

$$\lim_{(z_1,z_2)\to(0,a)} f(z_1, z_2) = \infty$$

(cf. Remark 20.4.10) at every point $z = (0, a)$, $a \neq 0$. The point $z = (0, 0)$ is the *point of indetermination* of function f: One cannot assign any reasonable value to f at that point, because, e.g., for every $k \in \mathbb{C}$ we have

$$\lim_{z_1\to 0} f(z_1, kz_1) = k.$$

Definition 2. Let G be a region in $\mathbb{C}^n$, $\varphi_1, \ldots, \varphi_n$ complex functions in G. A function $F = (\varphi_1, \ldots, \varphi_n)$ from G into $\mathbb{C}^n$ is called a *holomorphic mapping*, if each function φ_j, $j = 1, 2, \ldots, n$, is holomorphic in G. A *biholomorphic mapping* is defined as a one-to-one holomorphic mapping.

REMARK 4. For $n = 1$, the biholomorphic mapping is just conformal, i.e. conserves locally the angles of curves (§ 21.1). For $n \geqq 2$ the concepts of biholomorphicity and conformality do not agree. For example, for $n = 2$ the mapping $F = (z_1, 2z_2)$ is biholomorphic but not conformal, and the conformal mapping $F = \left(\dfrac{z_1}{|z_1|^2 + |z_2|^2}, \dfrac{z_2}{|z_1|^2 + |z_2|^2}\right)$ is not biholomorphic.

In contrast to the Riemann Mapping Theorem for $n = 1$ (Theorem 21.2.1) we have

Theorem 4. *There exists no biholomorphic mapping of the ball $B^2 \subset \mathbb{C}^2$ onto the bidisk $U^2 \subset \mathbb{C}^2$.*

21. CONFORMAL MAPPING

By JAROSLAV FUKA

References: [3], [66], [213], [227], [252], [262], [267], [296], [313], [352], [357].

The book [262] which is a dictionary of common conformal mappings is particularly recommended to the reader.

21.1. The Concept of Conformal Mapping

The theory of conformal mapping is of great importance in many branches of technology and physical sciences (for example, in the theories of elasticity, of flow around aerofoils, of two-dimensional stationary vector fields, etc.).

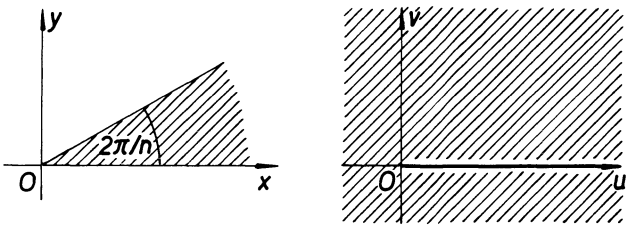


Fig. 21.1. The mapping by the function $w = z^n$.

Definition 1. Let a region G be mapped on a region B with one-to-one correspondence by a function $w = f(z) = u(x, y) + iv(x, y)$. If, moreover, the function $f(z)$ is holomorphic on G (Definition 20.1.9), then the mapping given by the function $f(z)$ is called a *conformal mapping* (or *transformation*) *of the region G on* or *on to the region B*. We also say that $f(z)$ *maps the region G conformally on the region B*.

REMARK 1. The function $z = f^{-1}(w)$, inverse to the function $w = f(z)$, maps B conformally on G.

Example 1. The function $w = z^n$ (n is a positive integer) maps the sector $0 < \arg z < 2\pi/n$ of the z-plane conformally on the w-plane, from which the positive part of the real axis (including the origin) is excluded (Fig. 21.1).

Example 2. The function $w = e^z$ maps the strip $h_1 < \operatorname{Im} z < h_2$, $0 < h_2 - h_1 \leqq \leqq 2\pi$, of the z-plane on a sector in the w-plane with vertex at the origin and arms making angles h_1, h_2 with the positive real axis (Fig. 21.2).

REMARK 2. A conformal mapping of a region G on B has the following two basic properties:

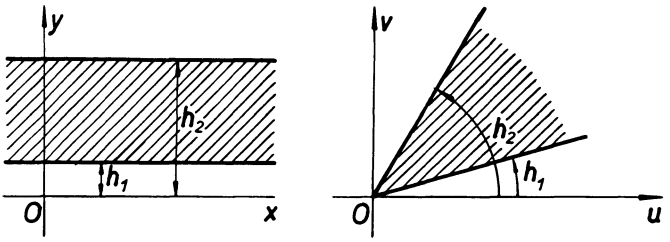


Fig. 21.2. The mapping by the function $w = e^z$.

1. It transforms every circle (we write "circle", in this chapter, instead of "circumference of a circle", for brevity) of infinitely small diameter into a circle of infinitely small diameter. The exact meaning of this assertion is as follows: Let a curve β_r be the image of a circle c_r, the equation of which is $|z - z_0| = r$ (cf. Definition 20.1.2) and which lies in G (Fig. 21.3a). Let us construct a circle k_r in B:

$$|\zeta - f(z_0)| = |f'(z_0)|\, r\,. \tag{1}$$

Let $\varrho(r)$ denotes the maximum of the distances of points $w = f(z)$ lying on β_r from the circumference of the circle k_r. Then

$$\frac{\varrho(r)}{r} \to 0 \quad \text{as} \quad r \to 0\,. \tag{2}$$

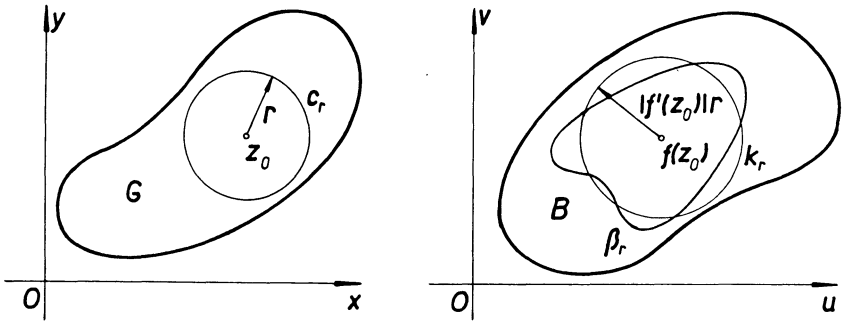


Fig. 21.3a.

2. It preserves angles. More precisely: Let γ_1, γ_2 be smooth curves lying in G and let α be the oriented angle between γ_1 and γ_2 at their point of intersection z_0.

Then the images β_1, β_2 of the curves γ_1, γ_2 are also smooth curves and the oriented angle between them (at the point $f(z_0)$) is also α (Fig. 21.3b).

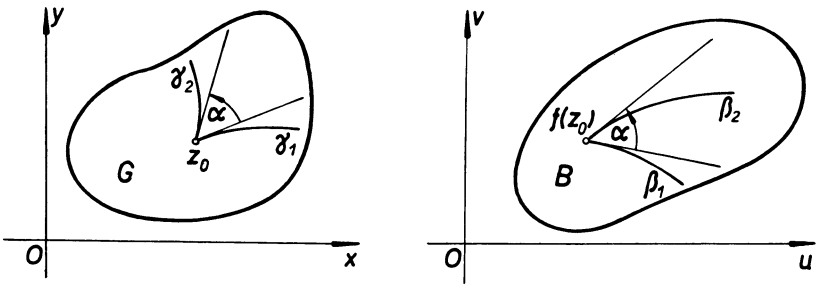


Fig. 21.3b.

REMARK 3. From (1) and (2), we deduce the geometrical significance of the absolute value of the derivative $f'(z)$: $|f'(z_0)|$ determines the "change of scale" at the point z_0. The number $\arg f'(z_0)$ has the following geometrical significance: Let γ denote an arbitrary smooth curve in the z-plane, passing through the point z_0, and β its image in the w-plane, where $w = f(z)$. If we superimpose the z- and w-planes so that the coordinate axes remain parallel and the points z_0 and $w_0 = f(z_0)$ coincide, then $\arg f'(z_0)$ is the angle between the curves γ and β. We say briefly that $\arg f'(z_0)$ is the angle of rotation at the point z_0.

REMARK 4. As an important example of conformal mapping, let us mention the mapping of the form $w = (az+b)/(cz+d)$, $ad-bc \neq 0$, which is called *homographic* and has the following properties:

The mapping composed of two homographic mappings is a homographic mapping. The mapping inverse to a homographic mapping is also homographic. The image of a straight line or of the circumference of a circle is again a straight line or the circumference of a circle and the image of a circle or of a half-plane is again a circle or a half-plane. Conversely, every conformal mapping of a circle or a half-plane on a circle or a half-plane is homographic. A homographic mapping maps the closed plane, or the closed plane from which one point is excluded, conformally on the closed plane or on the closed plane excluding one point. Conversely, if $f(z)$ is a conformal mapping of the closed plane, or the closed plane excluding one point, on a region B, then $f(z)$ is a homographic mapping and B is the closed plane or the closed plane excluding one point.

Further: Three mutually different points z_1, z_2, z_3 either determine the circumference k of a circle or lie on a straight line p. Let the circle k (or the straight line p) be oriented curves. Let G denote the interior of k (in this case let the circumference k be oriented in such a way that G lies on the left-hand side of k if we move along k in the positive sense of its orientation). Alternatively, let G denote

that one of the two half-planes determined by the straight line p which lies on the left-hand side if we move along p in the positive sense. Let the points z_1, z_2, z_3 follow each other in the order mentioned according to the positive orientation of k or p. Similarly, let the points w_1, w_2, w_3 of the w-plane lie either on a circumference k' or on a straight line p', and let us have the same convention concerning the orientation of k' or p' with respect to the corresponding circle B or the half-plane B, respectively, and concerning the ordering of the points w_1, w_2, w_3 on k' or p', respectively (Fig. 21.4). Every conformal mapping of the region G on B is a homographic one. If this mapping (let us denote it by $w = f(z)$) maps the points z_1, z_2, z_3 on the points w_1, w_2, w_3 (in this order), then it is uniquely determined (see Remark 21.2.1) by the relation

$$\left(\frac{w - w_1}{w - w_2}\right) \Big/ \left(\frac{w_3 - w_1}{w_3 - w_2}\right) = \left(\frac{z - z_1}{z - z_2}\right) \Big/ \left(\frac{z_3 - z_1}{z_3 - z_2}\right). \tag{3}$$

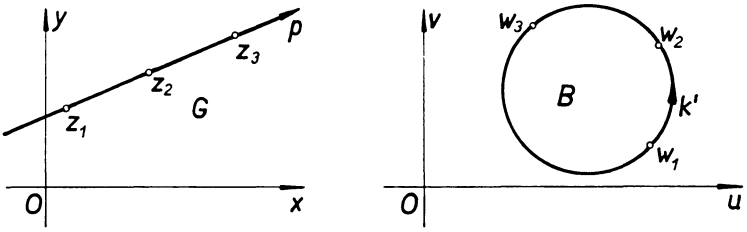


Fig. 21.4.

Example 3. Let us find a conformal mapping of the upper half-plane G of the z-plane on the unit circle B with centre at the origin of the w-plane such that the points $z_1 = -1$, $z_2 = 0$, $z_3 = 1$ of the x-axis are mapped on the points $w_1 = 1$, $w_2 = \mathrm{i}$, $w_3 = -1$ lying on the boundary of the circle B; Fig. 21.5.

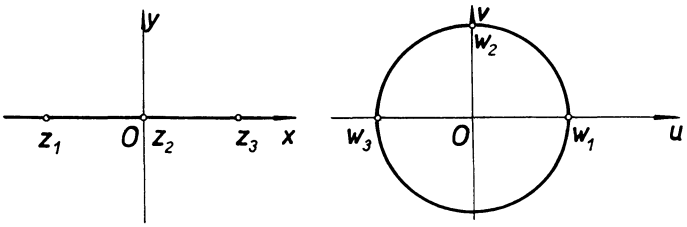


Fig. 21.5.

According to (3), we have

$$\left(\frac{w - 1}{w - \mathrm{i}}\right) \Big/ \left(\frac{-1 - 1}{-1 - \mathrm{i}}\right) = \left(\frac{z - (-1)}{z - 0}\right) \Big/ \left(\frac{1 - (-1)}{1 - 0}\right)$$

and consequently

$$w = \frac{z - \mathrm{i}}{\mathrm{i}(z + \mathrm{i})}. \tag{4}$$

REMARK 5. The formula (3) can be used even if one of the points under consideration is the point at infinity. If e.g. $z_3 = \infty$, then in (3) we put $(z_3 - z_1)/(z_3 - z_2) = 1$, so that the right-hand side of equation (3) is $(z - z_1)/(z - z_2)$. If, e.g., $w_1 = \infty$, then we write the left-hand side of equation (3) in the form

$$\left(\frac{w - w_1}{w_3 - w_1}\right) \Big/ \left(\frac{w - w_2}{w_3 - w_2}\right)$$

and put $(w - w_1)/(w_3 - w_1) = 1$, so that the left-hand side of equation (3) is $(w_3 - w_2)/(w - w_2)$, etc.

In our example it follows from (4) that to the point $z = \infty$ there corresponds the point $w = -\mathrm{i}$. If we now choose the mutually corresponding points

$$z_1 = -1\,,\ z_2 = 1\,,\ z_3 = \infty\,,\quad w_1 = 1\,,\ w_2 = -1\,,\ w_3 = -\mathrm{i}$$

we again obtain, by the method just described, the mapping (4).

21.2. Existence and Uniqueness of Conformal Mapping

If two regions G and B are given in the plane, a fundamental question arises as to whether there exists a conformal mapping of the region G on B. In the case of simply connected regions, the answer is given by the following theorem:

Theorem 1 (*The Riemann Theorem*). *Let G be a simply connected region with a boundary containing at least two points* (see Remark 2) *and let $z_0 \in G$. Then there exists a conformal mapping $w = f(z)$ which maps the region G on the unit circle $|w| < 1$ such that*

$$f(z_0) = 0\,,\quad f'(z_0) > 0\,. \tag{1}$$

The function $f(z)$ is uniquely determined by the conditions (1).

REMARK 1. Conditions (1) have the following geometrical meaning (Fig. 21.6): Two mutually corresponding points $z = z_0$ and $w = 0$ are given and the angle of rotation at z_0 is zero (see Remark 21.1.3). The conditions determining uniquely the function $f(z)$ can also be chosen in other ways; for example, if G and B are Jordan regions (see Remark 14.1.3), then $f(z)$ is uniquely determined if three pairs of mutually corresponding points on the boundaries of the regions G and B are given (Fig. 21.7: see Example 21.1.3). (Cf. also Theorem 21.4.1.)

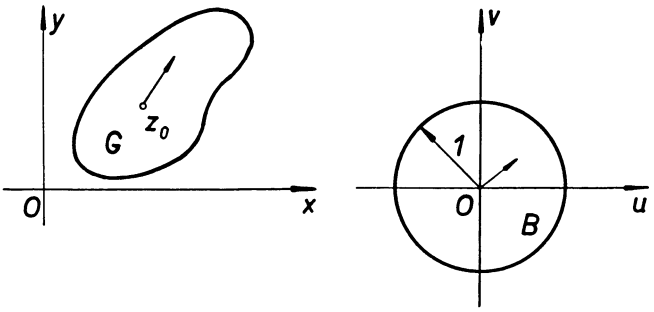


Fig. 21.6.

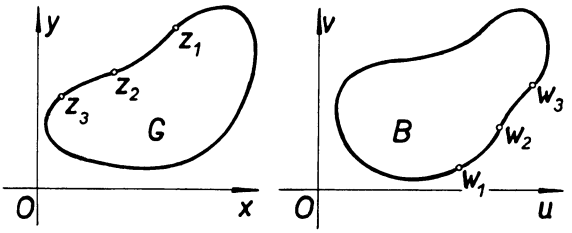


Fig. 21.7.

REMARK 2. Theorem 1 gives no information about regions, the boundary of which consists of only one point or no point at all, e.g. the closed plane or the closed plane excluding one point (this point can also be the point $z = \infty$). These cases have been discussed in Remark 21.1.4. Neither the closed plane, nor the closed plane excluding one point, can be conformally mapped on a circle. The region G of Fig. 21.8 (the closed plane with a circular hole with centre at the origin and

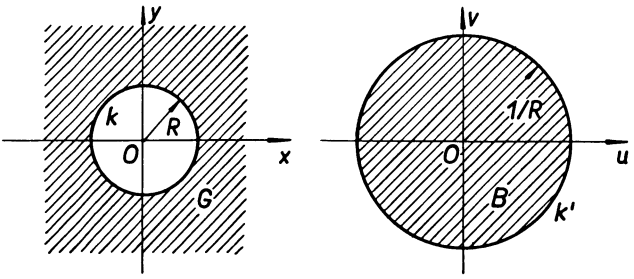


Fig. 21.8.

radius R) can be mapped, by the so-called "inversion" $w = 1/z$, on the circle with centre at the origin and radius $a = 1/R$. (If the centre is at a point z_0, then the corresponding mapping will be $w = 1/(z - z_0)$; if the hole is not circular and, at

the same time, z_0 is an interior point of it, then we can first map the region G by the function $w = 1/(z - z_0)$ on a bounded simply connected region, and then map this region on a circle.) If we exclude one point (e.g. the point $z = \infty$) from the above-mentioned region G (the closed plane with a circular hole and with one point excluded), we can map G in a similar way on a circle excluding one point.

In the case of multiply connected regions, the following theorem holds:

Theorem 2. *Every n-tuply connected region G ($n \geqq 2$) can be conformally mapped on one of the following regions B_i ($i = 1, 2, \ldots, 5$), where*

B_1 *is an annulus with centre at the origin excluding $n - 2$ concentric circular arcs;*

B_2 *is a circle with centre at the origin excluding $n - 1$ concentric circular arcs;*

B_3 *is the plane excluding n concentric circular arcs;*

B_4 *is the plane excluding n segments lying on rays starting from the origin;*

B_5 *is the plane excluding n parallel segments with the same angle ϑ between their direction and the x-axis* (Fig. 21.9a, b, c, d, e).

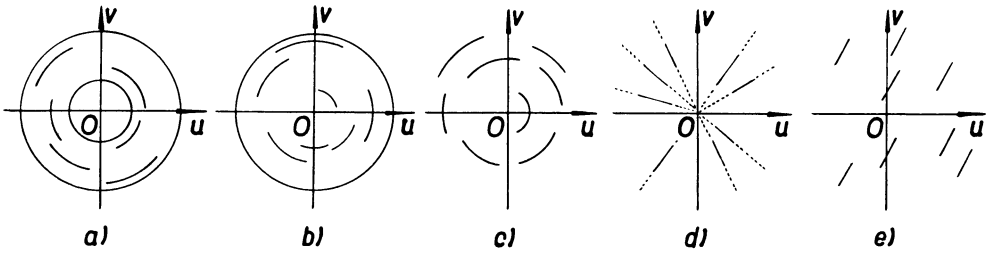


Fig. 21.9.

REMARK 3. The difference between Theorem 2 and Theorem 1 consists in the fact that the shape of the region B_i (e.g. the ratio of the radii of the annulus B_1) cannot be arbitrarily chosen a priori, as distinct from the case of simply connected regions. For example, a conformal mapping of an annulus $r_1 < |z| < r_2$ on an annulus $\varrho_1 < |w| < \varrho_2$ exists if and only if $r_1/r_2 = \varrho_1/\varrho_2$.

The closed plane with two circular holes (Fig. 21.10) can be conformally mapped by the inversion $w = 1/z$ (or $w = 1/(z - z_0)$, see Remark 2) on an eccentric annulus (which can then eventually be mapped on a concentric annulus as in Example 21.3.2). In general, the closed plane with two holes can be mapped conformally on an annulus. The open plane (or the closed plane excluding one point) with two holes cannot be conformally mapped on an annulus. Obviously, an annulus cannot be a conformal image of the closed plane with more than two holes.

Theorem 3. *There exists only one function $w = f(z)$ which maps an n-tuply connected region G on a region of type B_5* (cf. Theorem 2) *so that to a given point*

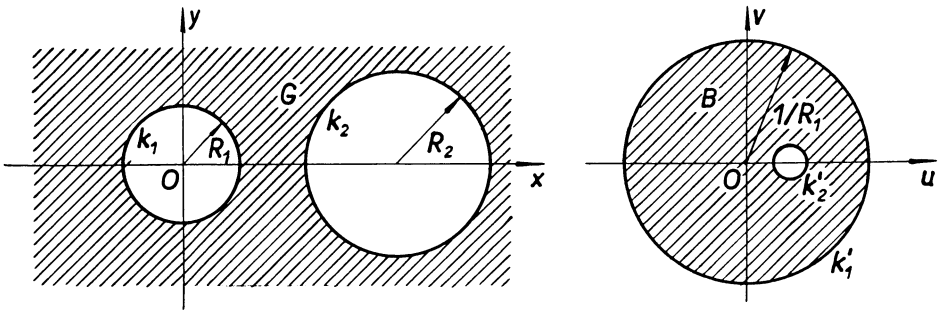


Fig. 21.10.

$z_0 \in G$ *there corresponds the point* $w = \infty$ *and that the expansion of* $f(z)$ *in the neighbourhood of the point* z_0 *has the form*

$$f(z) = \frac{1}{z - z_0} + \alpha_1(z - z_0) + \alpha_2(z - z_0)^2 + \dots$$

if z_0 *is a finite point and*

$$f(z) = z + \frac{\beta_1}{z} + \frac{\beta_2}{z^2} + \dots$$

if $z_0 = \infty$.

REMARK 4. Analogous theorems hold for regions of type B_i $(i = 1, \dots, 4)$.

21.3. Methods of Performing Conformal Mappings

Example 1. Let us find the conformal mapping of the exterior of the circumference c of a circle on the plane excluding a circular arc γ (Fig. 21.11). The circumference c has its centre at the point $z = ih$, $h > 0$, and passes through the point a, $a > 0$; the arc γ is given by the points $-a$, ih, a ($-a$, a being its end-points).

If φ stands for the angle shown in Fig. 21.11, then the angle between c and the positive x-axis is $\alpha = \frac{1}{2}\pi - \varphi$ and the angle, at the point $w = a$, between γ and the positive u-axis is $\beta = \pi - 2\varphi$. The mapping

$$z_1 = \frac{z - a}{z + a}$$

maps the circumference of the circle c on a straight line passing through the origin of the z_1-plane. (According to Remark 21.1.4, the image of the circumference of a circle is either a straight line or a circle, if a homographic mapping is used.

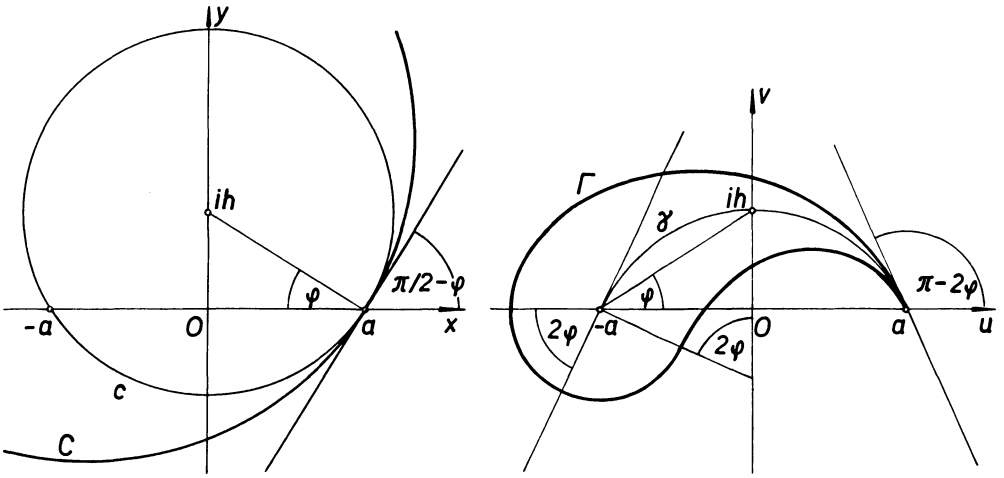


Fig. 21.11.

However, for $z = a$ we have $z_1 = 0$, for $z = -a$ we have $z_1 = \infty$, so that a circle is out of the question.) This straight line makes an angle α with the x_1-axis, since the x-axis is mapped on the x_1-axis and conformal mappings preserve angles. Hence, the function

$$z_2 = z_1^2 = \left(\frac{z-a}{z+a}\right)^2$$

maps (see Example 21.1.1) the exterior of the circumference of the circle c on the z_2-plane excluding the ray from the origin making an angle $2\alpha = \beta$ with the positive x_2-axis. By the mapping

$$w_1 = \frac{w-a}{w+a}$$

(see Remark 21.1.4), the complement of the arc γ is mapped on the w_1-plane excluding the ray from the origin making an angle β with the positive u_1-axis, since the u-axis is transformed on to the u_1-axis and conformal mappings preserve angles. Hence, putting $z_2 = w_1$, we get

$$w = \frac{1}{2}\left(z + \frac{a^2}{z}\right). \tag{1}$$

REMARK 1. Let us draw the circumference of a circle (denoted by C) touching c at the point a and lying in the exterior of c (Fig. 21.11). The mapping (1) maps C on a curve Γ containing the arc γ in its interior and having a cusp at a. The function (1) maps the exterior of C on the exterior of Γ. This provides a basis for the study of aerofoils (the so-called *Joukowski aerofoils*).

In Example 1, we have used a simple *method of combining elementary conformal mappings* which is often very effective since simple mappings (such as the

homographic mapping) have simple properties which can be well illustrated geometrically.

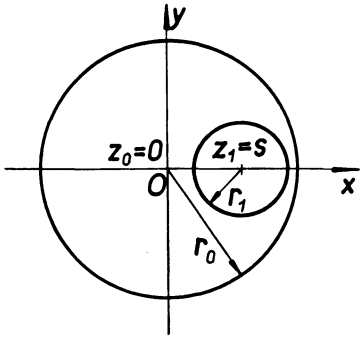


Fig. 21.12.

The following example shows a further simple application of homographic mappings.

Example 2. In electrostatic field theory (when determining the field of an eccentric cylindrical condenser), the conformal mapping of an eccentric annulus on a concentric one is frequently used. If we use notation of Fig. 21.12 ($r_0 > s + r_1$, $r_1 > 0$, $s \geqq 0$) and write

$$t = \sqrt{\frac{(r_0 + s)^2 - r_1^2}{(r_0 - s)^2 - r_1^2}}\,,$$

$$R_0 = \frac{r_0(t+1) - (s + r_1)(t-1)}{(s + r_1)(t+1) - r_0(t-1)}\,,$$

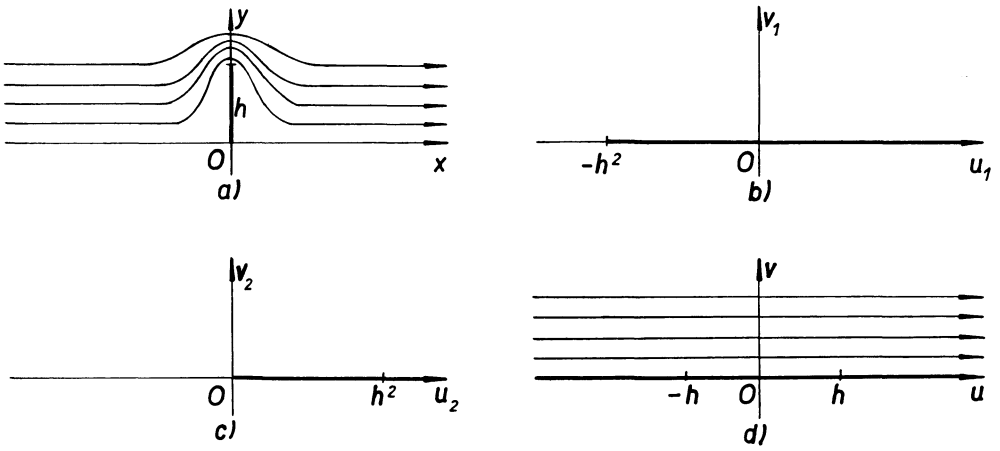


Fig. 21.13 a, b, c, d.

then the mapping

$$w = R_0 \frac{(t+1)z - r_0(t-1)}{-(t-1)z + r_0(t+1)}$$

maps the given eccentric annulus on a concentric annulus with centre at $w = 0$ and inner radius R_0, outer radius $R_1 = 1$.

Example 3 (*Flow Round an Obstacle*). Let us consider a steady irrotational flow in the upper u-half-plane with an obstacle of height h (Fig. 21.13a). Without this obstacle the flow would be uniform and the streamlines would be parallel to the x-axis. Let the velocity at infinity be unity, $v_\infty = 1$.

We have to find the corresponding complex potential of the flow, i.e. a holomorphic function $w = f(z)$ defined on the upper half-plane with the necessary slit of length h along the imaginary axis (let us denote this region by G) and such that $v = \operatorname{Im} f(z) = \text{const.}$ are the streamlines, $u = \operatorname{Re} f(z) = \text{const.}$ are the equipotential lines, perpendicular to the streamlines. At the same time, the complex number $\overline{f'(z)}$ at each point z will determine the velocity vector of the flow, not only in direction, but also in absolute value. Obviously, in the limit the boundary of the region G (i.e. the x-axis + the slit) is to be a streamline. Further, the condition $v_\infty = 1$ implies $f'(z) \to 1$ for $z \to \infty$.

We are going to map the region G on the upper half-plane Ω. We shall see that this mapping $w = f(z)$ is the required function. (For more detailed treatment of these problems see, e.g., [296]). We shall apply a combination of elementary mappings. First, the function $w_1 = z^2$ maps the region G on a region G_1 of the w_1-plane with a slit $[-h^2, +\infty)$ on the real axis, Fig. 21.13b (since the mapping $w_1 = z^2$ doubles angles with vertex at the origin, Example 21.1.1). Next, the function $w_2 = w_1 + h^2$ maps the region G_1 on a region G_2, with a slit $[0, +\infty)$ on the real axis (Fig. 21.13c; this is a translation in the direction of the x-axis). Finally the mapping $w = \sqrt{w_2}$ maps the region G_2 on the upper half-plane Ω of the w-plane (Fig. 21.13d). Hence, the required function is

$$w = \sqrt{(z^2 + h^2)} \ .$$

The flow in the region G corresponds, according to this equation, to the flow in the half-plane Ω. The streamlines $v = \operatorname{Im} \sqrt{(z^2 + h^2)} = \text{const.}$ in G correspond to the streamlines $v = \text{const.}$ in the half-plane Ω. Obviously

$$f'(z) = \frac{z}{\sqrt{(z^2 + h^2)}} \ ,$$

so that $f'(z) \to 1$ as $z \to \infty$. For $z \to ih$ we have $f'(z) \to \infty$ (the so-called *defect* on the edge).

In the case where $v_\infty = a$ is specified, the required function is

$$w = a\sqrt{(z^2 + h^2)} \ .$$

Theorem 1 (*Boundary-Correspondence Principle*). *Let $w = f(z)$ be a function continuous on a closed region $\overline{G}$, bounded by a Jordan curve Γ* (Remark 14.1.3) *and holomorphic on G.*

If the function $f(z)$ is uniquely invertible on Γ and maps Γ on a Jordan curve β, then $f(z)$ is uniquely invertible on $\overline{G}$ and maps G conformally on the interior of the curve β.

REMARK 2. Theorem 1 holds even if G is the exterior of Γ or if Γ is a generalized Jordan curve (this is a curve which is a stereographic projection (i.e. a projection from the "north pole") of a Jordan curve lying on the Riemann sphere (see Remark 20.1.2) on to the complex plane) and G is one of the regions with the boundary Γ (e.g. if Γ is a straight line, and G is a half-plane). (See e.g. [313], [296].)

Example 4. Let us study the mapping of the upper half-plane, given by the following function (for notation, see Remark 20.2.3):

$$w = f(z) = \int_0^z \frac{\mathrm{d}t}{\sqrt{[(1-t^2)(1-k^2t^2)]}} \ , \quad 0 < k^2 < 1 \tag{2}$$

(the "elliptic integral of the first kind" in Legendre's normal form).

We take that branch (§ 20.6) of the double-valued function $\sqrt{[(1-t^2)(1-k^2t^2)]}$ which, for $t \in (0, 1)$, assumes positive values. (To be precise we denote the positive root of a ($a > 0$) by the symbol $\sqrt[+]{a}$.) Then the function (2) is holomorphic on the upper half-plane and continuous on the closed upper half-plane.

Let us find the image of the whole real axis under the mapping (2) (Fig. 21.14). If $z = x$, $0 < x < 1$, then the value of $f(x)$ lies in the interval $(0, \omega_1)$, where

$$\omega_1 = \int_0^1 \frac{\mathrm{d}t}{\sqrt[+]{[(1-t^2)(1-k^2t^2)]}} \ .$$

In the interval $(1, 1/k)$, the integrand is of the form $1/\pm \mathrm{i}\sqrt[+]{[(t^2-1)(1-k^2t^2)]}$. We must choose the sign so that the above-mentioned branch is continuous in the upper half-plane. If we pass from the point $1-\delta$ to the point $1+\delta$ along a semicircle with centre at the point 1 and lying in the upper half-plane (suppose $0 < 1-\delta < 1$

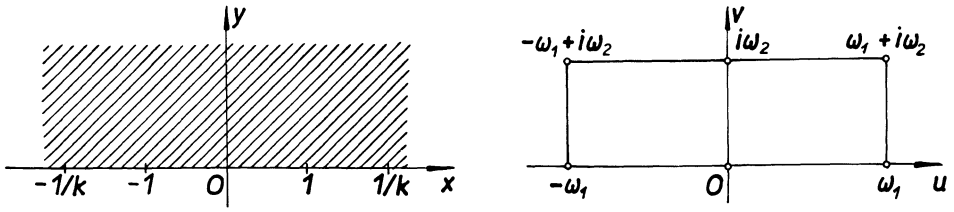


Fig. 21.14.

and $1 < 1 + \delta < 1/k$), the value of the expression $\varphi(t) = (1 - t^2)(1 - k^2t^2)$ changes from + to −, while $\arg \varphi(t)$ changes from 0 to $-\pi$, since if t passes along the given semi-circle, the values of $\varphi(t)$ lie in the *lower* half-plane. Thus, $\arg \sqrt{[\varphi(t)]}$ becomes equal to $-\pi/2$, and therefore the *minus* sign is to be chosen.

Hence, for $1 < x < 1/k$ we have

$$f(x) = \int_0^x \frac{dt}{\sqrt{[(1-t^2)(1-k^2t^2)]}} = \int_0^1 \frac{dt}{\sqrt{[(1-t^2)(1-k^2t^2)]}} + \\ + i\int_1^x \frac{dt}{\sqrt{[(t^2-1)(1-k^2t^2)]}} = \omega_1 + i\int_1^x \frac{dt}{\sqrt{[(t^2-1)(1-k^2t^2)]}}.$$

The points $f(x)$ therefore lie on the segment parallel to the imaginary axis, with endpoints ω_1, $\omega_1 + i\omega_2$, where

$$\omega_2 = \int_1^{1/k} \frac{dt}{\sqrt{[(t^2-1)(1-k^2t^2)]}}.$$

Similarly, it can be seen that $f(x)$ describes the segments from $\omega_1 + i\omega_2$ to $i\omega_2$, from $i\omega_2$ to $-\omega_1 + i\omega_2$, from $-\omega_1 + i\omega_2$ to $-\omega_1$ and from $-\omega_1$ to 0, when x describes the segments of the real axis from $1/k$ to $+\infty$, from $-\infty$ to $-1/k$, from $-1/k$ to -1 and from -1 to 0, respectively.

Hence, from Theorem 1 and Remark 2, we may conclude that the function (2) maps the upper half of the z-plane conformally on the rectangle with vertices $-\omega_1$, ω_1, $\omega_1 + i\omega_2$, $-\omega_1 + i\omega_2$ in the w-plane, without being obliged to verify the one-to-one correspondence of the mapping, which would be a rather complicated operation.

Theorem 2 (*The Riemann–Schwarz Reflection Principle*). *Let G be a region bounded by a Jordan curve Γ and let the curve Γ contain a segment δ. Let $w = f(z)$ be continuous on $G + \delta$ and holomorphic on G and let the segment δ be mapped by the function $f(z)$ on a segment λ* (Fig. 21.15). *Let G^* be the region symmetric to*

G with respect to the straight line containing δ and let us define the function $f^*(z)$ on G^* in the following way: $f^*(z) = f(z)$ if $z \in \delta$ and $f^*(z^*)$ is the point in the w-plane symmetric to the point $f(z)$ with respect to the straight line containing λ, if the point z^* is symmetric to the point z in the z-plane with respect to the straight line containing δ. Then the function $f^*(z)$ is the analytic continuation of the function $f(z)$ from $G + \delta$ into the region $G + G^* + \delta$.

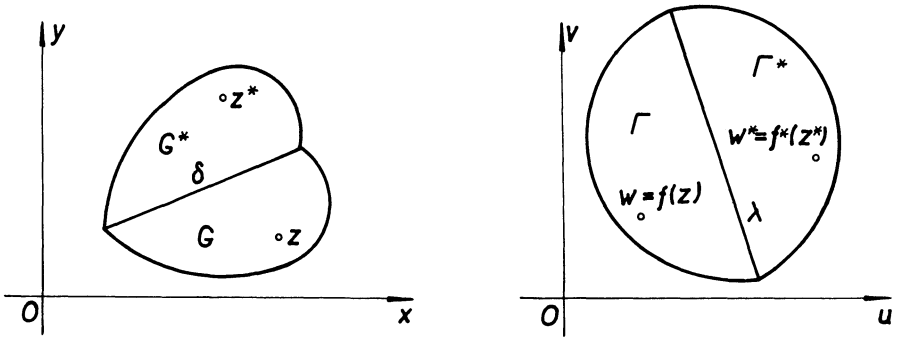


Fig. 21.15.

REMARK 3. The segments δ and λ very often lie on the real axis. The values $f(z)$ and $f^*(z^*)$ are then conjugate complex numbers.

The principle of reflection has numerous applications in conformal mapping and can be easily generalized (see, e.g. [313], [296]). For example, by its help the following (the so-called *Schwarz–Christoffel theorem*) may be proved:

Theorem 3. *If a function $w = f(z)$ maps the upper half-plane $\operatorname{Im} z > 0$ conformally on the interior of a bounded polygon G with angles $\alpha_k \pi$ $(0 < \alpha_k \leqq 2$, $k = 1, 2, \ldots, n$, $\sum_{k=1}^{n} \alpha_k = (n-2))$ so that the vertices of the polygon correspond to the points a_k of the real axis $(-\infty < a_1 < a_2 < \cdots < a_n < +\infty)$, then*

$$f(z) = C \int_{z_0}^{z} (z - a_1)^{\alpha_1 - 1} (z - a_2)^{\alpha_2 - 1} \ldots (z - a_n)^{\alpha_n - 1} \, dz + C_1 , \tag{3}$$

where z_0, C, C_1 are certain constants.

REMARK 4. If e.g. $a_n = \infty$, i.e. if one of the vertices of the polygon corresponds to the point at infinity, then in formula (3) the factor containing a_n is excluded.

REMARK 5. Formula (3) holds even for a polygon with a vertex (or several vertices) lying at the point ∞, if we define the angle between two straight lines at the point ∞ as equal to the angle at their intersection, multiplied by -1.

REMARK 6. Theorem 3 may be converted in the following way: The function (3) with α_k and a_k $(k = 1, 2, \ldots, n)$ satisfying

$$-2 \leqq \alpha_k \leqq 2, \quad \sum_{k=1}^{n} \alpha_k = n - 2 \quad \text{and} \quad -\infty < a_1 < a_2 < \cdots < a_n < \infty,$$

maps the upper half-plane conformally on some polygon with n sides. However, in practice this polygon is usually given and we wish to map it conformally on the upper half-plane, i.e. to find the points a_k, $k = 1, 2, \ldots, n$, and the constants C and C_1. For more detailed treatment see § 21.8.

An example of the mapping (3) is the mapping (2), where we have $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \frac{1}{2}$ and the points a_k are ± 1, $\pm 1/k$.

Theorem 4 (*The Green Function*). *Let a simply connected region G with a boundary Γ and a point $z_0 \in G$ be given. If there exists a function $U(z)$ $(= U(x, y)$, where $z = x + \mathrm{i}y)$, harmonic on G and continuous on $G + \Gamma$, which assumes the value $\ln(1/|\xi - z_0|)$ at each point $\xi \in \Gamma$, then the function*

$$f(z) = \mathrm{e}^{\mathrm{i}\alpha}(z - z_0)\mathrm{e}^{U(z)+\mathrm{i}V(z)},$$

where $V(z)$ is a harmonic function conjugate to $U(z)$ on G (cf. Remark 20.1.12) *and α is an arbitrary real number, maps the region G conformally on the unit circle so that $f(z_0) = 0$.*

Conversely, if $f(z)$ is a conformal mapping of the region G on the unit circle such that $f(z_0) = 0$, then $g(z, z_0) = (1/2\pi).\ln(1/|f(z)|)$ is the Green function (cf. Definition 18.4.9) *of the region G with a pole at z_0.*

Hence, under certain assumptions concerning the smoothness of the boundary of the region G, for every function $u(z)$, harmonic on G and continuous on $G + \Gamma$, the following relation holds:

$$u(z) = -\int_{\Gamma} u(\zeta)\frac{\partial g(\zeta, z)}{\partial n}\,\mathrm{d}s,$$

where Γ is the boundary of the region G and $\partial g/\partial n$ the derivative of the function g with respect to the outward normal.

REMARK 7. Theorem 4 reduces the search for a conformal mapping to the search for the solution of a Dirichlet problem (§ 18.4) together with the problem of finding a harmonically conjugate function, and conversely.

Theorem 5 (*Extremal Properties of Conformal Mappings*). *Let G be a simply connected domain containing the origin and let its boundary contain at least two*

points. Then from among all the functions $f(z)$, holomorphic on G and such that $f(0) = 0$, $f'(0) = 1$, only the function $f(z)$ mapping the region G conformally on a circle yields

(a) *the minimum of the value* $M(f) = \sup_{z \in G} |f(z)|$,

(b) *the minimum of the value* $P(f) = \iint_G |f'(z)|^2 \, dx \, dy$ (i.e. the minimum of the area of the image of G),

(c) *the minimum of the value* $D(f) = \int_\Gamma |f'(z)| \, ds$ (i.e. the minimum of the length of the image of the curve Γ; we suppose that Γ is of finite length).

REMARK 8. Also in the case of multiply connected regions, it is possible to seek conformal mappings by means of variational problems.

21.4. Boundary Properties of Conformal Mappings

Theorem 1. *Let $f(z)$ be a conformal mapping of a Jordan region G on a Jordan region B* (Remark 14.1.3). *Then the function $f(z)$ is continuously extensible on the boundary of the region G, i.e. a function $g(z)$ exists which is equal to the function $f(z)$ in G and is continuous on the closed region G (i.e. including the boundary). Moreover, g is one-to-one.*

REMARK 1. The theorem holds even for n-tuply connected regions bounded by Jordan curves.

REMARK 2. The study of the correspondence between boundaries of general simply connected regions led C. Carathéodory to the so-called *theory of prime ends.* (See, e.g. [313].)

In the theory of conformal mappings an important question arises, whether a "small" change of the mapped region implies a "small" change of the region into which it is mapped.

Theorem 2. *Let G_i $(i = 1, 2, \dots)$ be a sequence of regions containing the point $z = 0$ and lying in a certain circle K. Let $G_i \subset G_{i+1}$ for every i (i.e. the G_i form an "increasing" sequence of regions, each of them being contained in all the following ones). Let us write $G = \lim_{n \to \infty} G_n$ (i.e. G is the union of all these regions). Let $\{f_n(z)\}$ be a sequence of functions mapping conformally the regions G_n on the unit circle $|w| < 1$ and normed by the conditions $f(0) = 0$, $f'(0) > 0$* (compare Theorem 21.2.1); *let $\{\varphi_n(w)\}$ be functions inverse to the functions $f_n(z)$ (i.e. mapping $|w| < 1$ on G_n). Let $f(z)$ map the region G conformally (with the conditions mentioned above) on $|w| < 1$ and let $\varphi(z)$ be its inverse function. Then the sequence $f_n(z)$ converges to $f(z)$ almost uniformly on G* (Remark 20.4.1) *and the sequence $\varphi_n(w)$ converges to $\varphi(w)$ almost uniformly on $|w| < 1$.*

Hence, if the G_n converge to G, then the $f_n(z)$ converge to $f(z)$. Theorem 2 is a special case of a much more general theorem of C. Carathéodory (see [313]).

NUMERICAL METHODS IN CONFORMAL MAPPINGS

21.5. Variational Methods

REMARK 1. Let G be a bounded simply connected region containing the origin. Let $H_2(G)$ denote the space of all functions $f(z)$ holomorphic on G and such that

$$P(f) = \iint_G f(z)\overline{f(z)}\,dx\,dy < +\infty. \tag{1}$$

We define the scalar product of the functions $f(z)$, $g(z)$ in $H_2(G)$ in the following way:

$$(f, g) = \iint_G f(z)\overline{g(z)}\,dx\,dy.$$

By Theorem 21.3.5, the derivative of the function $\varphi(z)$ which maps G conformally on a circle and fulfills the conditions $\varphi(0) = 0$, $\varphi'(0) = 1$ is the solution of the variational problem

$$P(\varphi') = \iint_G |\varphi'(z)|^2\,dx\,dy = \iint_G \varphi'(z)\overline{\varphi'(z)}\,dx\,dy = \min. \tag{2}$$

For the solution of this variational problem, we shall use the *Ritz method*.

Let us consider an arbitrary system of linearly independent functions belonging to $H_2(G)$: $u_0(z), u_1(z), \ldots, u_i(z), \ldots$, with $u_0(0) \neq 0$ and let us seek an approximate solution in the form

$$\varphi_n'(z) = \sum_{i=0}^{n} c_i u_i(z), \tag{3}$$

with the condition

$$\varphi_n'(0) = 1, \tag{4}$$

so that the integral $P(\varphi_n')$ be minimal. For this it is necessary and sufficient that the relation

$$\iint_G \varphi_n'(z)\overline{\eta(z)}\,dx\,dy = 0 \tag{5}$$

be satisfied for every function of the form (3) satisfying the condition $\eta(0) = 0$. (In fact, if $\psi(z) = \varphi_n'(z) + \eta(z)$ is another function satisfying condition (4), i.e. $\eta(0) = 0$, we have

$$P(\psi) - P(\varphi_n') = \iint_G \varphi_n' \overline{\eta}\, dx\, dy + \iint_G \overline{\varphi_n'} \eta\, dx\, dy + \iint_G \eta\overline{\eta}\, dx\, dy,$$

whence it easily follows that $P(\psi) - P(\varphi_n') \geqq 0$ for all "admissible" functions ψ if and only if the condition (5) is satisfied.) If we choose, for $\eta(z)$, the functions $v_i(z) = u_i(z) - [u_i(0)/u_0(0)]u_0(z)$, $i = 1, 2, \ldots, n$ (cf. Example 1 below), we get from (5) a system of equations to determine the coefficients c_i:

$$\sum_{i=0}^{n} \alpha_{ij} c_i = 0, \tag{6}$$

where $\alpha_{ij} = \iint_G u_i \bar{v}_j\, dx\, dy \quad (j = 1, 2, \ldots, n)$.

Simultaneously with condition (4),

$$\sum_{i=0}^{n} u_i(0) c_i = 1, \tag{7}$$

we obtain a system of $n + 1$ equations which has a unique solution.

As usual, it is advantageous to use an orthonormal system $\{u_n(z)\}$. In this case the radius R of the circle on which the region G is mapped is

$$R = \sqrt{\frac{S}{\pi \sum_{n=0}^{\infty} |u_n(0)|^2}},$$

where S is the area of the region G.

In numerical calculations, the evaluation of the integrals α_{ij} and the solution of the above-mentioned system of equations are problems of importance. The calculations may be very laborious even in the case of very simple regions. Calculation of the α_{ij} can be simplified by choosing the $u_i(z)$ so that $u_i(0) = 0$ if $i \neq 0$, $u_0(0) = 1$. (This can evidently interfere with the orthonormality of the system considered.) Then (7) implies $c_0 = 1$ so that the number of unknown quantities is decreased by one; further $v_k(z) = u_k(z)$, and hence

$$\alpha_{ij} = \iint_G u_i(z)\overline{u_j(z)}\, dx\, dy.$$

Example 1. Let us apply the method of Remark 1 in order to find the conformal mapping of the square $G(-1 \leqq x \leqq 1, \ -1 \leqq y \leqq 1)$ on a circle. Let us choose

$u_n = z^n$, $n = 0, 1, 2, \ldots$. Then

$$\alpha_{ij} = \iint_G z^i \bar{z}^j \, dx \, dy, \quad i = 0, 1, \ldots, n, \quad j = 1, 2, \ldots, n.$$

Let us choose $n = 4$. We calculate easily that

$$\alpha_{1,1} = \frac{8}{3}, \ \alpha_{2,2} = \frac{232}{45}, \ \alpha_{3,3} = \frac{264}{35}, \ \alpha_{0,4} = -\frac{16}{15}, \ \alpha_{4,4} = \frac{8 \cdot 2131}{25 \cdot 7 \cdot 9}.$$

The remaining coefficients vanish. Hence, the system of equations (6), (7) for the unknown quantities c_i has the form

$$\alpha_{0,j} c_0 + \alpha_{j,j} c_j = 0, \quad j = 1, \ldots, 4, \quad c_0 = 1,$$

and therefore $c_0 = 1$, $c_1 = c_2 = c_3 = 0$, $c_4 = -\alpha_{0,4}/\alpha_{4,4} = 210/2131 = 0 \cdot 098\,55$. Thus the approximate solution is $\varphi_4'(z) = 1 + c_4 z^4$, $\varphi_4(z) = z + 0 \cdot 019\,71 z^5$.

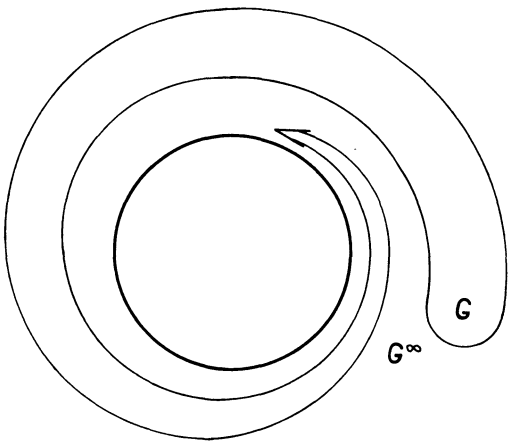


Fig. 21.16.

Theorem 1. *If the system* $u_i(z)$, $i = 0, 1, 2, \ldots$, *is complete* (cf. § 16.2 and Remark 22.4.9) *in* $H_2(G)$ (see Theorem 2 below), *then*

$$\lim_{n \to \infty} \iint_G \varphi_n'(z) \overline{\varphi_n'(z)} \, dx \, dy = \iint_G \varphi'(z) \overline{\varphi'(z)} \, dx \, dy,$$

$\varphi_n'(z) \to \varphi'(z)$ *and* $\varphi_n(z) \to \varphi(z)$ *almost uniformly on* G. *Here we have*

$$\varphi_n(z) = \int_o^z \varphi_n'(z) \, dz$$

(*according to our assumptions, the origin lies in* G *and* $\varphi(0) = 0$).

Definition 1. By a *Carathéodory region*, or briefly by a C-*region*, we mean a simply connected bounded region whose boundary is at the same time the boundary of that region lying in the complement of $\overline{G}$ which contains the point ∞.

REMARK 2. A Jordan region is, therefore, a C-region. The region in Fig. 21.16 (the "interior" of a spiral winding round a circle) is a C-region, but the regions in Figs. 21.17 and 21.18 are not C-regions.

Theorem 2. *In a* C-*region, there exists a complete orthonormal system of polynomials.*

REMARK 3. In the regions of Figs. 21.17 and 21.18, neither the polynomials, nor even an arbitrary system of entire functions form a system complete in $H_2(G)$.

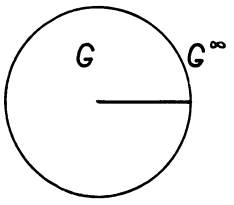


Fig. 21.17.

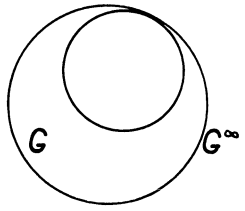


Fig. 21.18.

REMARK 4. As we have seen, in a C-region we can seek the function which maps G on a circle by means of the Ritz method; as a complete system we can use, for instance, an orthonormal system of polynomials.

REMARK 5. We can use analogous methods also in case (c) of Theorem 21.3.5.

21.6. The Method of Integral Equations

REMARK 1. In the case of a simply connected region, we can reduce our problem with the aid of Theorem 21.3.4 to a Dirichlet problem which can be solved by the method of integral equations (§ 18.4; see also § 24.6).

Example 1. We shall show the procedure for the construction of a system of integral equations for the real part of the function $w = f(z)$ which maps the exterior of a system of oriented curves $\Gamma_1, \Gamma_2, \ldots, \Gamma_n$ with continuous curvature on the plane $w = u + iv$ with slits parallel to the real axis,

$$v = v_k \quad (k = 1, 2, \ldots, n)$$

(Fig. 21.19) so that $f(\infty) = \infty$.

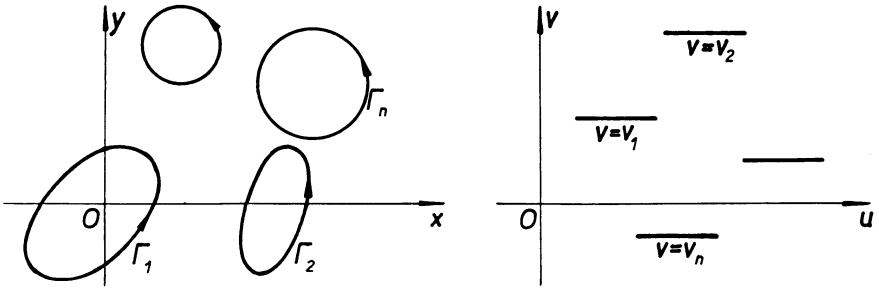


Fig. 21.19.

Let us draw a circle C with centre at the origin, sufficiently large so that all the curves Γ_k lie in the interior of C, and let us apply the Cauchy formula (Remark 20.2.10). We have

$$f(z) = \frac{1}{2\pi i}\left[\int_C \frac{f(\zeta)\,d\zeta}{\zeta - z} - \sum_{k=1}^{n}\int_{\Gamma_k} \frac{f(\zeta)\,d\zeta}{\zeta - z}\right]$$

(the point $z \in G$ lying in the interior of C). Since the expansion of $f(z)$ in the neighbourhood of infinity is

$$f(z) = cz + c_0 + \frac{c_1}{z} + \dots,$$

we have

$$\frac{1}{2\pi i}\int_C \frac{f(\zeta)\,d\zeta}{\zeta - z} = cz + c_0\,.$$

Hence, the function $f(z)$ will be of the form

$$f(z) = cz + c_0 - \sum_{k=1}^{n} \frac{1}{2\pi i}\int_{\Gamma_k} \frac{\mu_k(\zeta)\,d\zeta}{\zeta - z}\,, \tag{1}$$

where $\mu_k(\zeta) = \xi_k(\zeta) + i\eta_k(\zeta) = f(\zeta)$ for ζ lying on Γ_k. Since $\eta_k(\zeta) = v_k$ on Γ_k and the point z lies in the exterior of Γ_k, we have

$$\int_{\Gamma_k} \frac{\eta_k(\zeta)\,d\zeta}{\zeta - z} = 0\,.$$

Therefore, if we separate the real parts in (1), we obtain

$$\operatorname{Re} f(z) = \operatorname{Re}(cz + c_0) - \sum_{k=1}^{n} \frac{1}{2\pi}\int_{\Gamma_k} \frac{\xi_k(\sigma)\cos(\boldsymbol{\nu}, \boldsymbol{r})}{r}\,d\sigma\,,$$

where $z - \zeta = \boldsymbol{r}$, $(\boldsymbol{\nu}, \boldsymbol{r})$ is the angle between the outward normal $\boldsymbol{\nu}$ and the radius-vector $\boldsymbol{r}$ at the point ζ, $r = |\boldsymbol{r}|$ and σ is the parameter of the length of arc.

If we let the point z tend to a point of the curve Γ_k from the exterior of Γ_k, we have

$$\int_{\Gamma_k} \frac{\zeta_k(\sigma)\cos(\boldsymbol{\nu}, \boldsymbol{r})}{r} \mathrm{d}\sigma \to \int_{\Gamma_k} \frac{\zeta_k(\sigma)\cos(\boldsymbol{\nu}, \boldsymbol{r})}{r} \mathrm{d}\sigma - \pi\zeta_k(\sigma)$$

(cf. equation (18.4.34); in the first integral the point z does not lie on Γ_k, in the second integral it does). Consequently, we get

$$\xi_k(\sigma) = 2\,\mathrm{Re}(cz + c_0)_{\Gamma_k} - \sum_{j=1}^{n} \frac{1}{\pi} \int_{\Gamma_j} \frac{\xi_j(\sigma)\cos(\boldsymbol{\nu}, \boldsymbol{r})}{r} \mathrm{d}\sigma\,. \tag{2}$$

Under the assumption of continuity of the curvature of the curve Γ_k, the kernel $\cos(\boldsymbol{\nu}, \boldsymbol{r})/r$ is continuous on Γ_k. Hence, the Fredholm Alternative (Theorem 19.1.4) holds for the system (2). It can be proved that the homogeneous system (2) (i.e. that obtained by setting $c = 0$, $c_0 = 0$ in (2)) has a unique solution $\xi_k = 0$, $k = 1, 2, \ldots, n$. This implies, by the Fredholm Alternative, that the system (2) is solvable for an arbitrary choice of $c \neq 0$, c_0. If we solve the system (2) by any of the approximate methods (cf. Chaps. 29 or 24), we get $\xi_k(\sigma)$ and then calculate $f(z)$ by (1), where it is evidently sufficient to write $\xi_k(\zeta)$ instead of $\mu_k(\zeta)$.

21.7. Mapping of "Adjacent" Regions

REMARK 1 (*The Method of a Small Parameter*). Let us have s system of Jordan curves Γ_λ depending on a real parameter λ and containing the origin in their interior. Let the curve Γ_λ be given in the parametric form $z = z(t, \lambda)$. Let the function $w = f(z, \lambda)$ satisfying $f(0, \lambda) = 0$, $f'_z(0, \lambda) = 1$ map the interior G_λ of the curve Γ_λ on the circle $|w| < R_\lambda$. If $z(t, \lambda)$ is an analytic function of the parameter λ in the neighbourhood of the point $\lambda = 0$, we can expect, at least in some cases, the function $f(z, \lambda)$ also to be analytic in the neighbourhood of the point $\lambda = 0$ and hence expressible in a Taylor series

$$f(z, \lambda) = f_0(z) + \lambda f_1(z) + \cdots + \lambda^n f_n(z) + \ldots \tag{1}$$

If we know how to find the functions $f_n(z)$, then we can also find the function $f(z, \lambda)$. Let $\{u_n(z)\}$ be a complete system of functions defined on a region B containing all the G_λ for sufficiently small λ, $u_n(0) = 0$ $(n = 1, 2, \ldots)$, $u'_1(0) = 1$, $u'_n(0) = 0$ for $n > 1$. Then $f(z, \lambda) = \sum_{n=1}^{\infty} \alpha_n(\lambda) u_n(z)$. On the boundary, we have $|f(z, \lambda)|^2 = R_\lambda^2$. Hence if we expand the function $|\sum_{n=1}^{\infty} \alpha_n(\lambda) u_n(z(t, \lambda))|^2$ into

a Fourier series $C_0 + \sum_{n=1}^{\infty} (C_n \cos nt + C_n' \sin nt)$ on the boundary, we obtain, by comparing coefficients, $C_n = 0$, $C_n' = 0$, $C_0 = R_\lambda^2$. This is an infinite system of quadratic equations for the coefficients $\alpha_n(\lambda)$ which can be solved by successive approximations, taking, as a rule, only a few first coefficients $\alpha_n(\lambda)$. The convergence of the approximation process has been proved only under very special assumptions (see [252]).

REMARK 2. Let Γ be a curve given in polar coordinates by the equation $r = r(\varphi) = 1 - \delta(\varphi)$ and let $|\delta(\varphi)| < \varepsilon$, $|\delta'(\varphi)| < \varepsilon$, $|\delta''(\varphi)| < \varepsilon$. Then the function

$$w = f^*(z) = z\left(1 + \frac{1}{2\pi}\int_0^{2\pi} \frac{1 + ze^{-it}}{1 - ze^{-it}}\delta(t)\,dt\right)$$

differs from the function $w = f(z)$, $f(0) = 0$ (which maps the interior of the curve Γ on $|w| < 1$) by quantities of at least the second order in ε.

REMARK 3. A similar expression for the principal part of the function $f(z)$ holds also for a region adjacent to a half-plane or a region adjacent to another region (see [296]).

21.8. Mapping of the Upper Half-plane on a Polygon

Let K be a polygon with n sides in the w-plane. We wish to find the mapping of the upper half-plane on K, i.e. to find the constants $\alpha_1, \ldots, \alpha_n, a_1, \ldots, a_n, C, C_1$ in formula (21.3.3).

We solve the problem in the following way:

1. We take $\alpha_i = \beta_i$, where $\pi\beta_i$ are the magnitudes of the angles of the polygon K.

2. We determine the a_k from the relations

$$\lambda_1 : \lambda_2 : \cdots : \lambda_n = l_1 : l_2 : \cdots : l_n$$

where l_i are the lengths of the sides of K, and

$$\lambda_i = \int_{a_i}^{a_{i+1}} (z - a_1)^{\alpha_1 - 1} \ldots (z - a_n)^{\alpha_n - 1}\,dz;$$

we choose three of the points a_i arbitrarily, e.g. $a_1 = p_1$, $a_2 = p_2$, $a_n = p_n$.

3. The function

$$f(z) = \int_{z_0}^{z} (z - a_1)^{\alpha_1 - 1} \ldots (z - a_n)^{\alpha_n - 1}\,dz$$

maps the upper half-plane on a polygon K^*, similar to K. The constants C and C_1 are then determined by a translation, a rotation and a homothetic transformation so that K^* is transformed into K.

The constants $a_3, \ldots, a_{n-1}$ can be determined by the Newton–Fourier method which we proceed to describe.

The equations for $a_3, \ldots, a_{n-1}$ have the form

$$\lambda_2 = \frac{l_2}{l_1}\lambda_1, \ldots, \lambda_{n-2} = \frac{l_{n-2}}{l_1}\lambda_1 . \tag{1}$$

Let us take, for initial values, numbers $a_3^{(0)}, \ldots, a_{n-1}^{(0)}$ which differ little from the numbers $a_3, \ldots, a_{n-1}$, then expand the expressions in (1) into a Taylor series in $\delta_3^{(1)} = a_3 - a_3^{(0)}, \ldots, \delta_{n-1}^{(1)} = a_{n-1} - a_{n-1}^{(0)}$ taking their first terms. This gives

$$I_k^{(0)} + \delta_3^{(1)}\frac{\partial I_k^{(0)}}{\partial a_3^{(0)}} + \cdots + \delta_{n-1}^{(1)}\frac{\partial I_k^{(0}}{\partial a_{n-1}^{(0)}} = \frac{l_k}{l_1}\left[I_1^{(0)} + \delta_3^{(1)}\frac{\partial I_1^{(0)}}{\partial a_3^{(0)}} + \cdots + \delta_{n-1}^{(1)}\frac{\partial I_1^{(0)}}{\partial a_{n-1}^{(0)}}\right]$$

$(k = 2, 3, \ldots, n-2)$, where

$$I_k^{(0)} = \int_{a_k^{(0)}}^{a_{k+1}^{(0)}} (z-p_1)^{\alpha_1-1}(z-p_2)^{\alpha_2-1}(z-a_3^{(0)})^{\alpha_3-1}\ldots$$
$$\ldots(z-a_{n-1}^{(0)})^{\alpha_{n-1}-1}(z-p_n)^{\alpha_n-1}\,dz .$$

This is a system of linear equations in $\delta_3^{(1)}, \ldots, \delta_{n-1}^{(1)}$ which we solve and then repeat the same process for $a_3^{(1)} = a_3^{(0)} + \delta_3^{(1)}, \ldots, a_{n-1}^{(1)} = a_{n-1}^{(0)} + \delta_{n-1}^{(1)}$, etc. It can be proved that the system of equations for $\delta_3^{(k)}, \ldots, \delta_{n-1}^{(k)}$ are always solvable and that this process converges for a certain class of initial values. The reader will find a detailed treatment of this method in [252].

21.9. A Small Dictionary of Conformal Mappings

REMARK 1. In this paragraph, we present several examples of regions, whose conformal mapping on a canonical region (i.e. on the unit disk or its exterior, the upper half-plane or the strip in the case of simply connected region, on some of regions of type $B_1 - B_5$ from Theorem 21.2.2 in the case of multiply connected regions, etc.) can be explicitly written. The explicit formula is of use, indeed, only if one succeeds in describing the functional dependence of the parameters of the conformal mapping on the constants characterizing the geometric shape of the

region (cf. e.g. formula (21.3.3) and § 21.8); however, this dependence is essentially known only in very special cases (see [267]), and in general it is necessary to use numerical methods, e.g. some of those described in §§ 21.5–21.8.

REMARK 2. (a) The function $w = \operatorname{sn} z = \operatorname{sn}(z, k)$ occurring in the texts to Figs. 21.28, 21.29 is one of the so-called *Jacobi elliptic functions*. It is defined, for example, as the inverse function of the elliptic integral (21.3.2), i.e. as the unique solution of the equation

$$z = \int_0^w \frac{\mathrm{d}t}{\sqrt{[(1-t^2)(1-k^2t^2)]}}$$

for given z and k (cf. also § 13.12). The functions $\theta_1(z, \tau)$, $\theta_2(z, \tau)$ from the same texts are the so-called *Jacobi theta-functions*. For basic information about elliptic and theta-functions see [227].

(b) Observe that one function can realize conformal mapping of different regions; cf. the texts to Figs. 21.22 and 21.30.

(c) In the book [267], the reader can find a detailed information on the dependence of the conformal mapping on the geometry of the region.

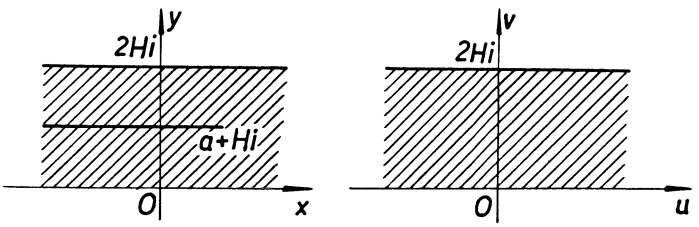


Fig. 21.20. The mapping of an infinite strip with a cut by the function

$$w = (H/\pi)\ln(\mathrm{e}^{\pi z/H} + \mathrm{e}^{\pi a/H}), \quad H > 0, \ a \text{ real},$$

on an infinite strip.

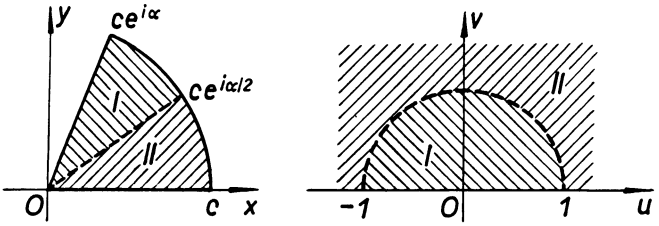


Fig. 21.21. The mapping of the sector of a circle by the function

$$w = \left(\frac{\left(\frac{z}{c}\right)^{\pi/\alpha} + 1}{\left(\frac{z}{c}\right)^{\pi/\alpha} - 1} \right)^2, \quad c > 0,\ 0 < \alpha \leqq 2\pi,$$

on the upper half-plane.

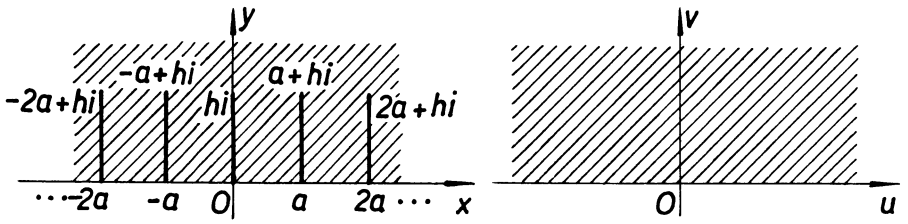


Fig. 21.22. The mapping of the upper half-plane, with segments $x = ka$ $(k = 0, \pm 1, \pm 2, \ldots)$, $0 \leqq y \leqq h$ excluded, by the function

$$w = (a/\pi) \arccos\left(\frac{\cosh \frac{\pi z}{a}}{\cosh \frac{\pi h}{a}} \right), \quad h \geqq 0,\ a > 0,$$

on the upper half-plane.

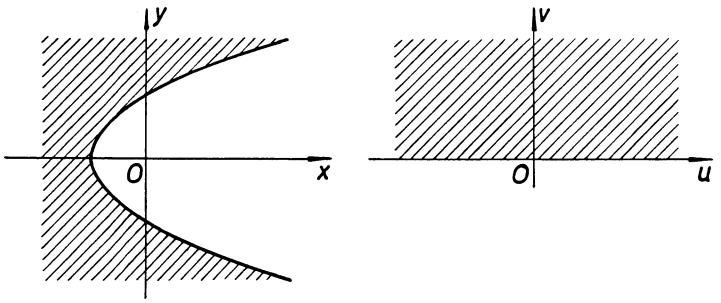


Fig. 21.23. The mapping of the exterior of the parabola $y^2 = 2p\left(x + \frac{p}{2}\right)$ by the function $w = \sqrt{z} - \sqrt{\left(\frac{p}{2}\right)}\,\mathrm{i}$ on the upper half-plane.

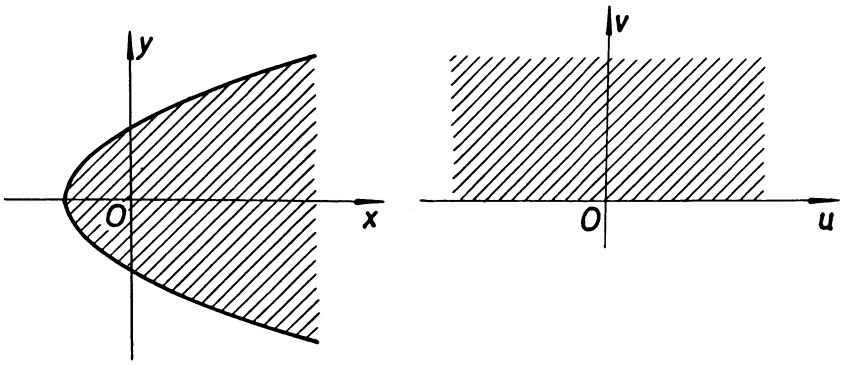


Fig. 21.24. The mapping of the interior of the parabola $y^2 = 2p\left(x + \frac{p}{2}\right)$ by the function

$$w = \mathrm{i}\sqrt{(2)}\cosh\left(\pi\sqrt{\left(\frac{z}{2p}\right)}\right)$$

on the upper half-plane.

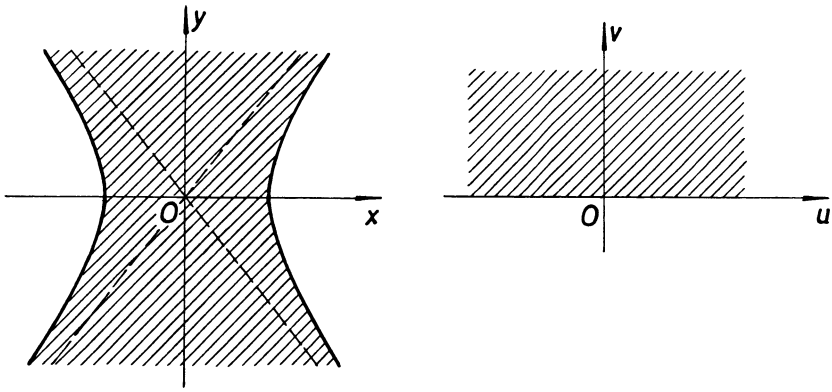


Fig. 21.25. The mapping of the exterior of the hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

by the function

$$w = \left(\frac{z + \sqrt{(z^2 - c^2)}}{c\mathrm{e}^{\mathrm{i}\vartheta}}\right)^{\pi/(\pi - 2\vartheta)}, \quad c = \sqrt{(a^2 + b^2)}\,,\ \vartheta = \arcsin\frac{a}{c}\,,$$

on the upper half-plane.

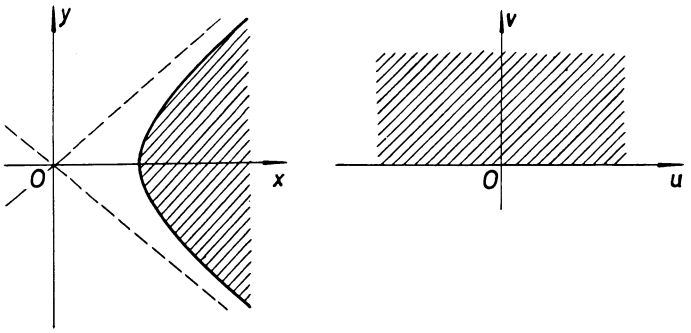


Fig. 21.26. The mapping of the interior of the right-hand branch of the hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

by the function

$$w = \mathrm{i}\sqrt{(2)}\cosh\left(\frac{\pi}{2\vartheta}\operatorname{arcosh}\frac{z}{c}\right), \quad c = \sqrt{(a^2 + b^2)}\,, \quad \vartheta = \arcsin\frac{a}{c}\,,$$

on the upper-half plane.

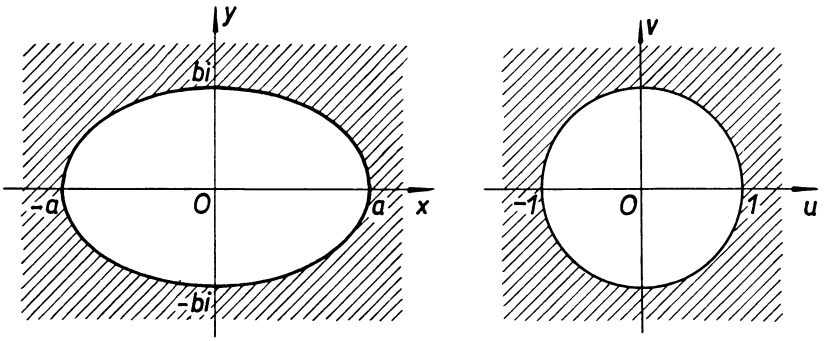


Fig. 21.27. The mapping of the exterior of the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

by the function

$$w = \frac{z + \sqrt{(z^2 - c^2)}}{a + b}, \quad a \geqq b, \quad c = \sqrt{(a^2 - b^2)}\,,$$

on the exterior of the unit circle with centre at the origin.

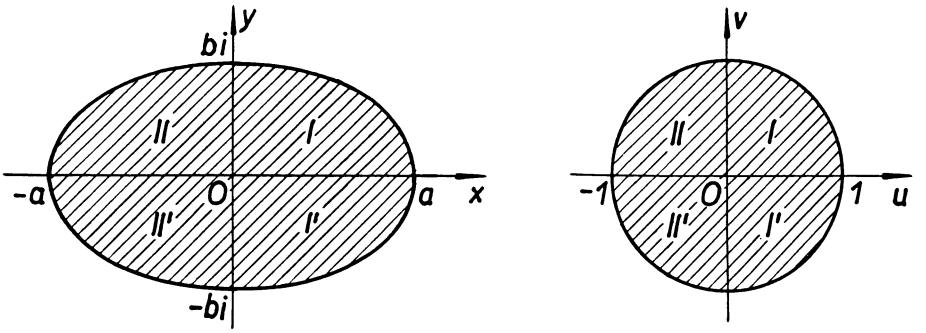


Fig. 21.28. The mapping of the interior of the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ by the function

$$w = \sqrt{(k)}\ \mathrm{sn}\left(\frac{2K}{\pi} \arcsin \frac{z}{\sqrt{(z^2 - b^2)}},\ k\right), \quad a > b > 0,$$

on the interior of the unit circle with centre at the origin. Here,

$$K = \int_0^1 \frac{\mathrm{d}x}{\sqrt{[(1-x^2)(1-k^2x^2)]}}, \quad k = \left(\frac{\theta_2(0,\tau)}{\theta_3(0,\tau)}\right)^2, \quad \tau = (2\mathrm{i}/\pi) \ln \frac{a+b}{a-b}.$$

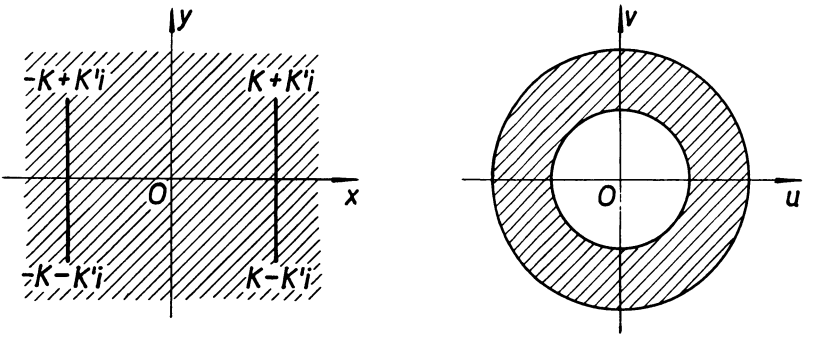


Fig. 21.29. The mapping of the plane, with segments excluded as shown in the figure, by the function

$$w = \mathrm{e}^{(\pi/K')\,\mathrm{sn}(z,k)} \text{ with } k = \left(\frac{\theta_2(0,\tau)}{\theta_3(0,\tau)}\right)^2, \quad \tau = \frac{\mathrm{i}K'}{K},$$

on the annulus $\mathrm{e}^{-\pi K/K'} < |w| < \mathrm{e}^{\pi K/K'}$.

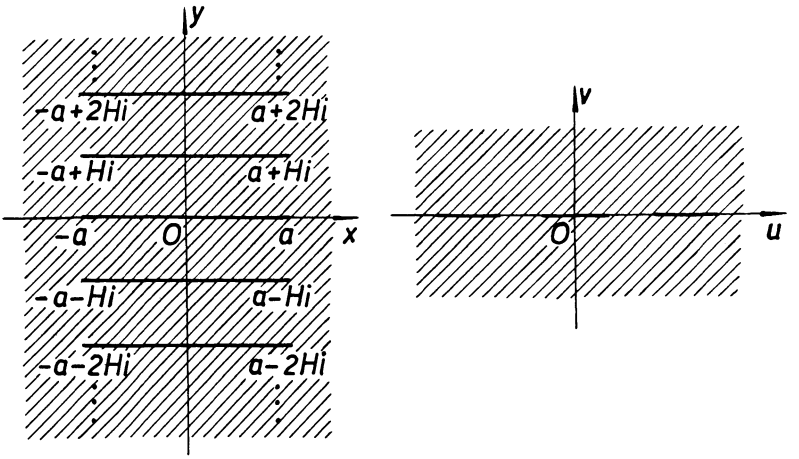


Fig. 21.30. The mapping of the plane, with segments $-a \leqq x \leqq a,\ y = kH$ $(k = 0, \pm 1, \pm 2, \dots)$ excluded, by the function

$$w = \arccos\left(\frac{\cosh\frac{\pi z}{H}}{\cosh\frac{\pi a}{H}}\right)$$

on the plane with segments $k\pi - b \leqq u \leqq k\pi + b \quad (k = 0, \pm 1, \pm 2, \dots),\ v = 0$ excluded. Here

$$b = \arccos\frac{1}{\cosh\frac{\pi a}{H}}.$$

22. FUNDAMENTALS OF THE THEORY OF SETS AND FUNCTIONAL ANALYSIS

By KAREL REKTORYS

References: [8], [27], [32], [46], [90], [93], [94], [97], [99], [111], [135], [162], [172], [203], [236], [263], [273], [274], [283], [284], [285], [305], [307], [312], [316], [317], [328], [329], [340], [346], [365], [389], [390], [393], [394], [397], [426], [429], [432], [457], [479], [507], [508], [510], [515].

22.1. Open and Closed Sets of Points in E_n. Regions

In many branches of mathematics, we come across the concepts of region, boundary of a set under consideration, etc. Let us clarify these concepts first for a plane.

Denote by d the *distance* between two points A and B in the plane. If, in this plane, a cartesian system of coordinates x, y is given (§ 5.1), then

$$d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}\,, \tag{1}$$

where x_1, y_1 and x_2, y_2 are coordinates of the points A and B, respectively. The set of all points of the plane with distance defined by (1) is called the *Euclidean space* E_2. (An elementary definition of this space has been presented here.)

All points in the plane, the distance of which from a given point P is less than a positive number δ, constitute the so-called *δ-neighbourhood of the point P*. (The point P belongs to this δ-neighbourhood.)

Definition 1. Let a set M of points in the plane be given. A point $P \in M$ is called an *interior point* of the set M if a (sufficiently small) δ-neighbourhood of P can be found which belongs entirely to M (i.e. all its points belong to M).

Definition 2. A point $P \in M$ is called an *isolated point* of a set M if a (sufficiently small) δ-neighbourhood of P can be found such that from among all the points of this δ-neighbourhood only the point P belongs to the set M.

Definition 3. P is called a *point of accumulation* (*accumulation point*, *cluster point*, *limit point*) of a set M if every δ-neighbourhood of P contains infinitely many points of M.

Definition 4. A point P is called a *boundary point* of a set M if every δ-neighborhood of P contains at least one point belonging to M and at least one point which does not belong to M. All boundary points of a set M constitute the so-called *boundary* of the set M.

Example 1. Let M_1 be the set of all points of a circle (= circular disc) K_1; let the circumference h_1 of this circle does not belong to M_1 (Fig. 22.1). Each point of

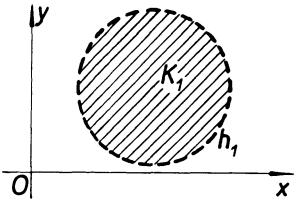


Fig. 22.1.

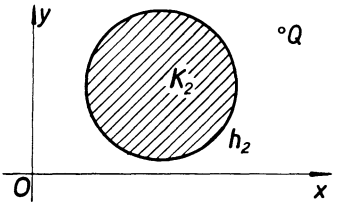


Fig. 22.2.

the set M_1 is an interior point of M_1. Points of the circumference h_1 are boundary points of the set M_1.

Example 2. Let the set M_2 consist of the point Q and of all points of a circle K_2 including its circumference h_2 (Fig. 22.2). The point Q is an isolated point and, at the same time, a boundary point of the set M_2. Points of h_2 are boundary points and points lying inside of h_2 are interior points of the set M_2.

REMARK 1. From Examples 1 and 2 it is clear that a boundary point of a set M may, but need not, belong to M. The same is true for a point of accumulation of M.

REMARK 2. A boundary point need not be a point of accumulation, and vice versa. Each interior point is a point of accumulation but not a boundary point. An isolated point is a boundary point but not a point of accumulation.

Definition 5. A set M is called *open* if each point of the set M is an interior point of this set (Example 1).

Definition 6. An open set M is called *connected* if every two points of M can be joined by a polygonal line (i.e. by a curve consisting of a finite number of straight line segments) which lies entirely in M (i.e., each point of it belongs to M). An open connected set is called a *region.*

Example 3. Examples of a region are: the set M_1 of Example 1 (the so-called *open circle* (= *open disc*)); an annulus with both boundary circumferences excluded; the plane xy.

REMARK 3. If every two points of a given set M can be joined by a segment lying entirely in M, then the set M is said to be *convex*. The set M_1 from Example 1 furnishes an example of such a set.

Definition 7. A set N obtained by completing a set M by all points of accumulation of M is called the *closure* of M. Notation: $N = \overline{M}$. (The symbol $[M]$ is also used.)

Definition 8. If $\overline{M} = M$, then the set M is called *closed*.

REMARK 4. If M is a region, then it is customary to call its closure $\overline{M}$ a *closed region*. An example of a closed region is a circle with its boundary included (the so-called *closed circle* or *closed disc*).

Definition 9. A set M is called *bounded* if a circle with a finite radius R can be found that M lies inside of that circle.

Definition 10. A bounded region is called *k-tuply connected* if its boundary consists of k closed curves.

Instead of an 1-tuply, or 2-tuply connected region we speak of a *simply*, or *doubly connected* region, respectively.

REMARK 5. We have presented an intuitive but not quite exact definition here. An exact definition can be found e.g. in [408]. (See also Remark 9.)

Example 4. The set M_1 from Example 1 (the open circle) is a simply connected region. An annulus is a doubly connected region. (Here we speak of a doubly connected region even if the inner circumference degenerates into a single point – which, of course, does not belong to the annulus.)

REMARK 6. In three-dimensional (or n-dimensional) space with a Cartesian coordinate system (in the *Euclidean space* E_3 (or E_n)), the *distance* d between two points $A(a_1, a_2, a_3)$, $B(b_1, b_2, b_3)$ (or $A(a_1, a_2, \ldots, a_n)$, $B(b_1, b_2, \ldots, b_n)$) is defined by the formula

$$d = \sqrt{[(b_1 - a_1)^2 + (b_2 - a_2)^2 + (b_3 - a_3)^2]}$$

(or

$$d = \sqrt{[(b_1 - a_1)^2 + (b_2 - a_2)^2 + \ldots + (b_n - a_n)^2]}\,). \tag{2}$$

(In E_1, the formula (2) reduces to $d = \sqrt{(b_1 - a_1)^2} = |b_1 - a_1|$.) A *$\delta$-neighbourhood* of a point P in E_n is defined in the same manner as in the two-dimensional case (in E_2). It is often called an *n-dimensional sphere* with centre P and radius δ and denoted by $S(P, \delta)$. In E_3 a δ-neighbourhood of a point P is a sphere in the ordinary sense, with centre at P and radius δ (and with its surface excluded), while in E_1 it is the open interval $(x_0 - \delta, x_0 + \delta)$ with centre at the point $P(x_0)$.

All definitions given above for the case of the space E_2 (concerning the concepts of interior, isolated, accumulation and boundary points, open and connected sets, regions, etc.) remain the same for E_n (see, however, Remark 9). Here, a *polygonal line in* E_n consists of a finite number of segments, while by a segment, joining the points $A(a_1, a_2, \ldots, a_n)$, $B(b_1, b_2, \ldots, b_n)$, the set of points with coordinates

$$a_k + (b_k - a_k)t, \quad 0 \leqq t \leqq 1, \quad k = 1, 2, \ldots, n, \tag{3}$$

is understood. An example of a region in E_3 is an (*open*) *sphere* (= the interior of a spherical surface), a cube without its boundary faces, etc. In E_1, to the concept of a region there corresponds an open interval, to the concept of a closed one a closed interval.

A set M in E_n is called *bounded*, if a sphere S with a finite radius R exists such that M lies inside of that sphere. If M is a region, we speak about a *bounded region* in E_n.

REMARK 7. An analogue of the real Euclidean space E_n is the *complex n-dimensional space* C_n, whose elements are points with complex coordinates. The distance between two points $A(a_1 + \mathrm{i}c_1, a_2 + \mathrm{i}c_2, \ldots, a_n + \mathrm{i}c_n)$, $B(b_1 + \mathrm{i}d_1, b_2 + \mathrm{i}d_2, \ldots, b_n + \mathrm{i}d_n)$ (a_k, b_k, c_k, d_k being real numbers) is defined by the formula

$$d = \sqrt{\Big[\sum_{k=1}^{n}(b_k - a_k)^2 + \sum_{k=1}^{n}(d_k - c_k)^2\Big]}\,.$$

REMARK 8. The concept of a *complement* of a set $M \subset N$ in the set N is often used: the complement is the set $N \setminus M$, i.e., the set N with the points belonging to M removed. For example, the complement of a closed circle in E_2 is an (unbounded) region obtained by removing this closed circle from the xy plane.

REMARK 9. Using the concept of a complement, a simply connected region in E_2 (in the xy plane) may be defined as follows: A bounded region $M \subset E_2$ is called *simply connected*, if its complement is a connected set.

For example, the interior of an ellipse is a simply connected region, since its complement is a connected set.

In the space E_3, however, the above-stated definition is not adequate; it fails to express characteristic properties of simple connectivity (ensuring, e.g., validity of some important theorems in integral calculus; for example, a torus would be a simply connected region, in the sense of the above definition). The reader familiar with the fundamentals of topology knows that a bounded region M in E_3 is simply connected if it is the so-called homeomorphic image of a sphere. (Very roughly speaking: if M can be obtained from a sphere by a "continuous deformation".) For an other definition see [350].

Definition 11. A set (of points) is called *countable* if there is a one-to-one correspondence between all its elements (points) and the positive integers 1, 2, 3, ... (i.e. if its elements (points) can be ordered in a sequence.) It is called *at most countable*, if it is countable or has only a finite number of elements.

It can be shown, for instance, that the set of points in E_1 with rational coordinates (the set of rational numbers) is countable. The same can be shown to hold in E_n.

22.2. Metric Spaces

In the last paragraph, we introduced the Euclidean space E_n, the elements of which were points and in which the distance was defined by formula (22.1.2). In a similar way, spaces of a more general nature can be introduced. In Chap. 16 (§ 16.1) the so-called space $L_2(a, b)$ has been defined. In the real case its elements are real functions which are square integrable (in the Lebesgue sense) in the interval $[a, b]$. The distance $d(f, g)$ between two elements f and g of $L_2(a, b)$ is defined by the formula

$$d(f, g) = \sqrt{\left\{\int_a^b [g(x) - f(x)]^2 \, dx\right\}}. \tag{1}$$

The space E_n and $L_2(a, b)$ are examples of so-called metric spaces.

Definition 1. A set M is called a *metric space* X if, for each pair of elements u, v belonging to M, the *distance* $d(u, v)$ is defined, having the following properties:

$$d(u, v) \geqq 0, \quad \text{while} \quad d(u, v) = 0 \quad \text{if and only if} \quad u = v, \tag{2}$$

$$d(u, v) = d(v, u), \tag{3}$$

$$d(u, z) \leqq d(u, v) + d(v, z) \tag{4}$$

for every u, v, $z \in M$.

We also say that a *metric* is given in (or on) the set M. Instead of elements, we often speak of *points* of the set M, or of the space X.

Example 1. For E_n and $L_2(a, b)$ it can be easily verified that the distance, defined by (22.1.2), or (1), respectively, satisfies the above three requirements (the so-called *axioms of the metric*). (For the space L_2 see, of course, Remark 1 below.)

REMARK 1. Let us remind (see Chap. 16) that two functions $f(x)$, $g(x)$ are called *equivalent* in the space $L_2(a, b)$, if their distance is equal to zero, i.e. if

$$\sqrt{\left\{\int_a^b [g(x) - f(x)]^2 \, dx\right\}} = 0$$

(or, what is the same, if

$$\int_a^b [g(x) - f(x)]^2 \, dx = 0), \tag{5}$$

or, in other words, if these functions are different, in the interval $[a, b]$, at most on a set of (Lebesgue) measure zero (for example, at a finite number of points of that interval). We write

$$f = g \quad \text{in } L_2(a, b). \tag{6}$$

All mutually equivalent functions are taken for equal in the space $L_2(a, b)$, they form a single element of this space; an arbitrary of these functions can be taken as a representant of this element. A similar remark holds for spaces $L_p(a, b)$, $L_2(\Omega)$ and $L_p(\Omega)$ defined below.

REMARK 2 (*complex space* $L_2(a, b)$). In applications, most often real functions are encountered. Then the just mentioned *real space* $L_2(a, b)$ with the metric (1) is introduced. Its generalization is the complex space $L_2(a, b)$, the elements of which are functions of the form $f(x) = f_1(x) + if_2(x)$, where f_1 and f_2 are real functions square integrable (in the Lebesgue sense) in the interval $[a, b]$. (Then also the function f is square integrable in $[a, b]$, as follows from the relation $|f(x)|^2 = f_1^2(x) + f_2^2(x)$.) The distance $d(f, g)$ is then defined by the formula

$$d(f, g) = \sqrt{\left[\int_a^b |g(x) - f(x)|^2 \, dx\right]}. \tag{7}$$

(See also § 16.1.)

REMARK 3 (generalizations). A more general metric space than $L_2(a, b)$ is the metric space $L_p(a, b)$ consisting of *functions integrable in* $[a, b]$ (in the Lebesgue sense) *with the p-th power* $(1 \leqq p < \infty)$; the distance is defined by the formula

$$d(f, g) = \left[\int_a^b |g(x) - f(x)|^p \, dx\right]^{1/p}.$$

For $p = 2$ we get the space $L_2(a, b)$.

Let Ω be a bounded region in E_n (Remark 22.1.6). Analogously to the space $L_2(a, b)$, or $L_p(a, b)$, the space $L_2(\Omega)$, or $L_p(\Omega)$, can be defined, respectively, as the space of functions square integrable, or integrable with the p-th power $(1 \leqq p < \infty)$ in Ω (in the Lebesgue sense), with the distance defined by

$$d(u, v) = \sqrt{\left\{\int_\Omega [v(x) - u(x)]^2 \, dx\right\}} \tag{8}$$

(by

$$d(u, v) = \sqrt{\left[\int_{\Omega} |v(x) - u(x)|^2 \, dx\right]} \tag{9}$$

in the complex case), or by

$$d(u, v) = \left[\int_{\Omega} |v(x) - u(x)|^p \, dx\right]^{1/p} \tag{10}$$

(in the real as well as complex case), respectively. Here, the brief notation

$$\int_{\Omega} u(x) \, dx$$

is used instead of

$$\underbrace{\int\int \dots \int_{\Omega}}_{n\text{-times}} u(x_1, \dots, x_n) \, dx_1 \dots dx_n.$$

For $n = 1$, the space $L_2(a, b)$, or $L_p(a, b)$ is obtained.

The space $L_2(\Omega)$ and $L_p(\Omega)$ can be defined for more general sets than for bounded regions.

REMARK 4. Further important metric spaces are the spaces $C([a, b])$ and $C(\overline{\Omega})$:

The *space* $C([a, b])$ (often denoted by $C[a, b]$, sometimes also by $C(a, b)$) is a space whose elements are functions continuous in $[a, b]$, with the distance defined by

$$d(u, v) = \max_{a \leqq x \leqq b} |v(x) - u(x)|. \tag{11}$$

Similarly, *the space* $C(\overline{\Omega})$ is defined as the space of functions continuous in a bounded closed region $\overline{\Omega}$ in E_n, with the distance defined by

$$d(u, v) = \max_{x \in \overline{\Omega}} |v(x) - u(x)|, \tag{12}$$

where $x = (x_1, \dots, x_n)$.

In what follows, the brief notation L_2, L_p, C will often be used instead of $L_2(a, b)$, $L_p(a, b)$, $C([a, b])$, $L_2(\Omega)$, $L_p(\Omega)$, $C(\overline{\Omega})$, if no misunderstanding may arise.

Definition 2. An element u of a metric space X is called the *limit* of a sequence $u_1, u_2, u_3, \dots$ of elements from X (or the sequence $u_1, u_2, u_3, \dots$ is said to be *convergent* in the space X to the element u) if

$$\lim_{n \to \infty} d(u, u_n) = 0. \tag{13}$$

This fact is denoted by $u_n \to u$ (in X).

REMARK 5. Instead of convergence in $L_2(a, b)$, the term *convergence in the mean* (see § 16.1) is often used. The relation (13) reads in this case (for the real case)

$$\lim_{n\to\infty} \sqrt{\left\{ \int_a^b \left[u(x) - u_n(x)\right]^2 \mathrm{d}x \right\}} = 0, \tag{14}$$

or, what is the same,

$$\lim_{n\to\infty} \int_a^b \left[u(x) - u_n(x)\right]^2 \mathrm{d}x = 0. \tag{15}$$

A similar remark holds for the space $L_2(\Omega)$.

Theorem 1. *A sequence $u_1, u_2, u_3, \ldots$ of elements of a metric space X has at most one limit u.* (For equality of functions in the space L_2 (or L_p) see Remark 1.) *If it has a limit, i.e. if it is convergent, then any subsequence of it is also convergent, with the same limit.*

Definition 3. The set of all elements (points) of a metric space X with distances from a given element (point) $u \in X$ less than δ ($\delta > 0$) is called a *δ-neighbourhood of u.* (Instead of δ-neighbourhood, the term a *sphere with centre at the point u and radius δ* is often used; in symbols $S(u, \delta)$.)

Definition 4. Let M be a set of elements of X. The element $u \in X$ is called a *point of accumulation* of the set M if any δ-neighbourhood of the element u contains infinitely many elements of the set M. (Obviously, a point of accumulation of a set M *need not belong to* M.)

Definition 5. The set $\overline{M}$ obtained by adjoining, to M, all its points of accumulation is called the *closure* of the set M (in the space X).

Definition 6. A set M is called *closed* (in X) if $M = \overline{M}$. A set M is called *open*, if the complement $X \setminus M$ (Remark 22.1.8) is a closed set.

Let us note that the open set can also be defined as a set consisting of interior points only. Here, a point of the given set is called its *interior point*, if such a (sufficiently small) δ-neighbourhood of that point exists which belongs entirely to that set.

Definition 7. The set M (equipped with the metric of the space X) is called *dense in X*, if $\overline{M} = X$. (That is, if to every element $u \in X$ and to every $\varepsilon > 0$ an element $v \in M$ can be found such that $d(u, v) < \varepsilon$, or, in other words, if every point of the space X is either an element of the set M, or a point of accumulation of this set.)

Example 2. It can be shown that the set M of all polynomials is dense in the space $C([a, b])$, i.e. that for every function $u \in C([a, b])$ and every $\varepsilon > 0$ a polynomial P can be found such that

$$\max_{a \leqq x \leqq b} \left|u(x) - P(x)\right| < \varepsilon$$

(the *Weierstrass theorem*).

The set of all polynomials is dense as well in the space $L_2(a, b)$, or, more generally, in $L_p(a, b)$ with $1 \leqq p < \infty$.

Similar assertions are true for the spaces $C(\overline{\Omega})$ and $L_2(\Omega)$ (or $L_p(\Omega)$). Here, polynomials in n variables are in question, of course.

22.3. Complete, Separable, Compact Spaces

Definition 1. A sequence $\{u_n\}$ of elements of a metric space X is called a *Cauchy sequence* (or a *fundamental sequence*) if for every $\varepsilon > 0$ a number n_0 (depending on the choice of ε) can be found such that

$$d(u_m, u_n) < \varepsilon \tag{1}$$

holds whenever both numbers m and n exceed n_0.

Theorem 1. *Every convergent sequence in X is a Cauchy sequence.*

Definition 2. A metric space X is called *complete* if any Cauchy sequence $\{u_n\}$ of elements from X has a limit belonging to the space X.

Theorem 2. *If M is a closed set in a complete space X, then M (equipped with the metric of this space) itself is a complete metric space.*

Theorem 3. *The spaces E_n, C, L_p (in particular, L_2) are complete.*

Not every space is complete:

Example 1. It is well known that the number $\sqrt{2}$ is not a rational number. At the same time, a sequence of rational numbers

$$x_1, x_2, x_3, \ldots \tag{2}$$

exists in E_1 such that

$$\lim_{n\to\infty} x_n = \sqrt{2}\,. \tag{3}$$

The sequence (2) is a Cauchy sequence in E_1 (by Theorem 1). Let X be the metric space elements of which are rational numbers, with the same metric $d(x_1, x_2) = |x_2 - x_1|$. Thus the sequence (2) is as well a Cauchy sequence in X. This sequence has no limit in the space X ($\sqrt{2}$ does not belong to X), so that X is *not* complete.

REMARK 1 (*Completion of a metric space*). If a metric space X is not complete, then it can be completed by "adding the so-called *ideal elements*". In details:

Two metric spaces X, Y with metrics ρ, or σ, respectively, are called *isometric*, if a one-to-one mapping F of the space X onto the space Y exists that preserves the distance: $\sigma(Fu, Fv) = \rho(u, v)$ holds for each couple of elements u, v from X.

From the point of view of metric properties only (convergence, completeness, etc.), such two spaces are equivalent.

Let X be not complete. Then it can be shown that a complete space Z can be constructed, containing a subspace Y, dense in Z, which is isometric with the space X.

The idea of construction of the space Z is the following: All Cauchy sequences in the space X are divided into two groups A, B of (Cauchy) sequences which have, or do not have a limit in the space X, respectively. It can be shown that for any two Cauchy sequences $\{u_n\}$, $\{v_n\}$ a finite limit

$$\lim_{n\to\infty} \rho(u_n, v_n) = a \tag{4}$$

always exists, without regard to which of the groups A, B they belong. All Cauchy sequences for which $a = 0$ are said to belong to the same *class*. All these classes are then taken as elements of a new space Z in which a metric is introduced on base of (4), and the so constructed space is shown to be complete. Also the so-called *stationary sequences* $\{u, u, u, \ldots\}$ are considered, belonging obviously to the group A. The class of Cauchy sequences, corresponding to such a stationary sequence, corresponds then to the element $u \in X$. All those classes constitute the space Y mentioned above. The space Y is shown to be dense in Z. The "remaining" classes then represent the above mentioned "ideal elements". In this sense the completion of an uncomplete metric space is to be understood. In general, it is difficult to characterize the nature of the "ideal elements". However, in many important cases, in particular in the case of many functional spaces, current in applications, the given space X can be completed simply by adding elements of a well-known character. For example, the Sobolev space $W_2^{(k)}(\Omega)$ is obtained by completing the space $S_2^{(k)}(\Omega)$ with such functions from the space $L_2(\Omega)$ which have square integrable generalized derivatives up to the order k in Ω (Remark 22.4.10). A similar situation is encountered when constructing the space H_A in Remark 22.6.10.

Definition 3. A metric space is called *separable* if it contains an at most countable (i.e. finite or countable, Definition 22.1.11) set M which is *dense* in X.

Theorem 4. *Spaces E_n, C, L_p with $1 \leqq p < \infty$ (in particular, the space L_2) are separable.*

REMARK 2. In E_n, all points with rational coordinates constitute a countable set M considered above. So also do all polynomials with rational coefficients in the spaces C and L_p $(1 \leqq p < \infty)$. (When spaces $C(\overline{\Omega})$ and $L_p(\Omega)$ are in question,

polynomials in n variables are to be considered, of course; in the case of complex spaces, coefficients of the polynomials are of the form $r+is$, where r, s are rational numbers.)

Definition 4. A set M of a metric space X is said to be *precompact* (or *relatively compact*) in that space, if any sequence of elements from M contains a subsequence convergent in X. If, in addition, the limits of all these subsequences belong to M, we say that M is *compact* (in X).

REMARK 3. In particular, a metric space X is called *compact*, if any sequence of elements of X contains a subsequence converging to an element of X.

Theorem 5. *Every bounded set M in E_n is relatively compact. If, in addition, M is closed, then it is compact* (in E_n).

REMARK 4. Spaces E_n, C, L_p (in particular, L_2) are *not* compact. To show it, for example, for the space E_1, it is sufficient to consider the sequence of points $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, ...

The interval $M = [3, 5]$ is a compact set in E_1 (because it is bounded and closed).

Theorem 6. *If X is complete, then a necessary and sufficient condition for a set $M \subset X$ to be relatively compact in X is that to every $\varepsilon > 0$ a finite ε-net N_ε for the set M exists, i.e. a finite set $N_\varepsilon \subset X$ such that to each $u \in M$ an $u_\varepsilon \in N_\varepsilon$ with $d(u, u_\varepsilon) < \varepsilon$ can be found.*

Theorem 7. *A necessary and sufficient condition for a set $M \subset C([a, b])$ to be compact* (in $C([a, b])$) *is that all functions $u \in M$ be equicontinuous and uniformly bounded* (see the Arzelà-Ascoli theorem, § 15.1).

A similar theorem holds in the space $C(\overline{\Omega})$.

Theorem 8. *A compact metric space is separable.*

Theorem 9. *Let a* (real) *function $u(x_1, x_2, \ldots, x_n)$ be continuous on a set $M \subset$ $\subset E_n$. Then, if M is compact, u assumes its maximum and minimum on M.*

REMARK 5. Theorem 9 is a generalization of well-known theorems relating to extrema of functions continuous in a closed finite interval or in a closed bounded region. Further, we have:

Theorem 10. *Let $f(u)$ be a real continuous functional* (Remark 22.5.3 and Definition 22.5.2) *on $M \subset X$. If M is compact, then $f(u)$ assumes its maximum and minimum on M.*

22.4. Linear Spaces. Normed Spaces. Banach and Hilbert Spaces. Orthogonal Systems. Generalized Derivatives, Sobolev Spaces, Embedding Theorems. Distributions

Definition 1. A set M of elements $u, v, z, \ldots$ is called a *linear set* (*linear space, vector space*) if it has the following properties:

If u, v are any two elements of M and a is a number (real or complex), then the sum $u + v$ and the product au are defined, and both $u + v$ and au belong to M; moreover, these operations obey the usual rules of linear algebra, i.e.

$$u + v = v + u, \qquad u + (v + z) = (u + v) + z,$$
$$a(u + v) = au + av, \qquad (a + b)u = au + bu,$$
$$a(bu) = (ab)u, \qquad 1 \cdot u = u,$$
$$\text{if} \quad u + v = u + z, \quad \text{then} \quad v = z.$$

REMARK 1. From these rules it follows that for any two elements u, v of M we have $0 \cdot u = 0 \cdot v$. The element $0 \cdot u$ is denoted by O (or simply by 0), and is called the *zero element* of the linear set in question. It can be shown that the familiar rules of algebra are still preserved. For example, if $ax = O$, $a \neq 0$, then $x = O$. If $ax = bx$ with $x \neq O$, then $a = b$, etc.

Let a linear set M be given and let, on M, a metric be introduced (thus the distance d between every two elements of M is defined satisfying (22.2.2)–(22.2.4)). If, moreover,

$$d(x + z, y + z) = d(x, y)$$

holds for every $x, y, z \in M$, then the metric is called *invariant.*

Definition 2. A liner set M equipped with an invariant metric d in such a way that $\lambda_n \to \lambda$ in R and $x_n \to x$ in $M \Rightarrow \lambda_n x_n \to \lambda x$ in M, is called a *linear metric space.*

REMARK 2. The above mentioned property of invariance of the metric is "almost obvious": It means that the distance between the elements u and v is the same as the distance between these elements "displaced paralelly along the element z". It can be shown that a metric, derived from the norm (see (iii) in the Definition 4 below) has both the properties required in Definition 2. Thus *every normed space is a linear metric space.*

Let us note, moreover, that as concerns the concepts of linear set, linear space and linear metric space, there is no uniformity in the literature.

Among linear metric spaces, linear normed spaces and Hilbert spaces are of particular importance. See Definitions 4 – 7 below.

Example 1. Spaces E_n, C, L_p (in particular, L_2) are linear metric spaces. In C and L_p, the sum of two functions and the product of a function and a constant are defined in the usual manner. We know, for example, that if $f \in L_p$, $g \in L_p$ $(1 \leqq p < \infty)$, then also $f + g \in L_p$ and $cf \in L_p$, where c is an arbitrary number which is, of course, real, or complex, according to whether a real, or complex space L_p is under consideration. In the space C, the zero element is the function which vanishes identically in $[a, b]$ (or in $\overline{\Omega}$), while in the space L_2, or L_p, it means every function equivalent to the zero function, i.e. every function vanishing everywhere in $[a, b]$ (or in $\overline{\Omega}$) except, possibly, for points which constitute a set of measure zero.

Definition 3. A linear subset L of elements of a linear metric space X, equipped with the metric of that space, is called a *linear subspace* (or briefly *subspace*) of the linear space X. (In particular, we may have $L = X$.) If, moreover, the subset L is closed (in the metric of the space X), then L is called a *closed* (linear) *subspace of* X.

Example 2. As can be shown, constant functions, i.e. functions of the form $f(x) = c$ (c being a constant), constitute a closed subspace L in $C([a, b])$.

A similar assertion on constant functions holds in the space $C(\overline{\Omega})$.

Definition 4. A set X is called the *linear normed metric space* (*normed space*, in brief), if

(i) elements of X constitute a linear set;

(ii) to each element $u \in X$ there is uniquely assigned a number $\|u\|$, called the *norm* of the element u, which satisfies the following so-called *axioms of the norm*:

$\|u\| \geqq 0$, while $\|u\| = 0$ if and only if $u = 0$ in X (i.e. if u is the zero element of the set X),

$\|au\| = |a| \, . \, \|u\|$ for every number a,

$\|u + v\| \leqq \|u\| + \|v\|$ (triangular inequality);

(iii) the distance $d(u, v)$ is defined by the formula

$$d(u, v) = \|u - v\| \, .$$

Let us note that if two normed spaces X, Y with generally different normes are considered, the norms are denoted by $\|u\|_X$, or $\|u\|_Y$, respectively, to avoid a misunderstanding.

Definition 5. A complete linear normed space is called a *Banach space* (*B-space*).

Example 3. Spaces E_n, C, L_p $(1 \leqq p < \infty$, in particular, $L_2)$ are Banach spaces provided the respective norms are defined by the relationships:

$$\text{in} \quad E_n: \qquad \|x\| = \sqrt{(x_1^2 + x_2^2 + \ldots + x_n^2)} \, ,$$

$$\text{in } C([a,b]): \qquad \|u\| = \max_{a\leqq x\leqq b}|u(x)|,$$

$$\text{in } C(\overline{\Omega}): \qquad \|u\| = \max_{x\in\overline{\Omega}}|u(x)|,$$

$$\text{in } L_p(a,b): \qquad \|u\| = \left(\int_a^b |u(x)|^p\,dx\right)^{1/p},$$

$$\text{in } L_p(\Omega): \qquad \|u\| = \left(\int_\Omega |u(x)|^p\,dx\right)^{1/p}.$$

(We use the brief notation

$$\int_\Omega |u(x)|^p\,dx \quad \text{instead of} \quad \underbrace{\iint\ldots\int_\Omega}_{n\text{-times}} |u(x_1, x_2, \ldots, x_n)|^p\,dx_1\,dx_2\ldots dx_n.)$$

REMARK 3. By introducing a metric, the concept of *convergence* is defined in X (see Definition 22.2.2); thus, in our case, the notation

$$\lim_{n\to\infty} u_n = u \quad \text{in the space } X \quad (\text{or briefly } u_n \to u \text{ in } X) \quad (u_n \in X,\ u \in X)$$

means that

$$\lim_{n\to\infty} \|u - u_n\| = 0.$$

(We speak also about the *convergence in norm*.) For example,

$$u_n \to u \quad \text{in} \quad L_p(a,b)$$

means that

$$\lim_{n\to\infty}\left(\int_a^b |u - u_n|^p\,dx\right)^{1/p} = 0,$$

or, what is the same, that

$$\lim_{n\to\infty}\int_a^b |u - u_n|^p\,dx = 0.$$

If a *series*

$$\sum_{n=1}^{\infty} u_n$$

($u_n \in X$) is given in the space X, then we construct the *sequence of partial sums* $\{s_k\}$,

$$s_k = \sum_{n=1}^{k} u_n, \quad k = 1, 2, \ldots,$$

and say that the given series is *convergent in the space* X and that its *sum* is s, if

$$\lim_{k\to\infty} s_k = s \quad \text{in } X.$$

We write

$$\sum_{n=1}^{\infty} u_n = s \quad \text{in } X.$$

If we consider, for example, the real space $L_2(a, b)$ and write

$$\sum_{n=1}^{\infty} u_n = s \quad \text{in } L_2(a, b),$$

it means that

$$\lim_{k\to\infty} \int_a^b (s - s_k)^2 \, dx = 0,$$

or, writing $\sum_{n=1}^{k} u_n$ for s_k, that

$$\lim_{k\to\infty} \int_a^b \Big(s - \sum_{n=1}^{k} u_n\Big)^2 dx = 0.$$

Theorem 1. *If* $u_n \to u$, *then* $\|u_n\| \to \|u\|$.

The converse is not true, in general.

Definition 6. A set P is called a *unitary* or a *pre-Hilbert* (*prehilbert*) *metric space* (briefly a *unitary* or a *pre-Hilbert* (*prehilbert*) *space*), if

(i) P is linear (see Definition 1);

(ii) to each pair of elements u, v from P there is uniquely assigned a number (u, v) (complex, in general), called the *scalar* (or *inner*) *product* of the elements u, v, such that

$$(u, v) = \overline{(v, u)}$$

(thus (v, u) is the complex conjugate of (u, v)),

$$(u_1 + u_2, v) = (u_1, v) + (u_2, v),$$

$$(au, v) = a(u, v)$$

(and, consequently, $(u, av) = \overline{a}(u, v)$),

$$(u, u) \geqq 0, \quad \text{while} \quad (u, u) = 0 \quad \text{if and only if} \quad u = 0;$$

(iii) the *norm* of an element u is defined by the formula

$$\|u\| = \sqrt{(u, u)}\,;$$

(iv) the *distance* is given by

$$d(u, v) = \|v - u\| .$$

Similarly as in the case of normed spaces, also here the notation $(u, v)_{P_1}$, or $(u, v)_{P_2}$ is used if P_1 and P_2 are two unitary spaces with different scalar products, respectively.

Definition 7. If a unitary space P is complete, it is called a *Hilbert space.*

In what follows, the notation H will most often be used for a Hilbert space.

REMARK 4. It follows from Definitions 5, 6 and 7, that a Hilbert space is also a Banach space.

In applications, very often *real Hilbert spaces* are encountered. Then the notation of complex conjugate, appearing in the first property of the scalar product, is superfluous and we have

$$(u, v) = (v, u).$$

Example 4. The complex space $L_2(a, b)$, or $L_2(\Omega)$ with the scalar product

$$(u, v) = \int_a^b u(x)\,\overline{v(x)}\,\mathrm{d}x,$$

or

$$(u, v) = \int_\Omega u(x)\,\overline{v(x)}\,\mathrm{d}x,$$

respectively, is a complex Hilbert space. The real space $L_2(a, b)$, or $L_2(\Omega)$ (of real functions) with the scalar product

$$(u, v) = \int_a^b u(x)\,v(x)\,\mathrm{d}x,$$

or

$$(u, v) = \int_\Omega u(x)\,v(x)\,\mathrm{d}x,$$

respectively, is a real Hilbert space.

The spaces C and L_p $(1 \leqq p < \infty,\ p \neq 2)$ are *not* Hilbert spaces.

Theorem 2. *For the scalar product and the norm the so-called Schwarz inequality is true:*

$$\left|(u, v)\right| \leqq \|u\| \,.\, \|v\| .$$

This inequality is often written in a rather more appropriate form

$$\left|(u, v)\right|^2 \leqq \|u\|^2 \,.\, \|v\|^2 .$$

Example 5. In the real space $L_2(a, b)$, or $L_2(\Omega)$, the last inequality becomes

$$\left(\int_a^b u(x)\, v(x)\, \mathrm{d}x\right)^2 \leqq \int_a^b u^2(x)\, \mathrm{d}x \,.\, \int_a^b v^2(x)\, \mathrm{d}x,$$

or

$$\left(\int_\Omega u(x)\, v(x)\, \mathrm{d}x\right)^2 \leqq \int_\Omega u^2(x)\, \mathrm{d}x \,.\, \int_\Omega v^2(x)\, \mathrm{d}x,$$

respectively.

Definition 8. The elements u, v are called *orthogonal*, if $(u, v) = 0$. Notation $u \perp v$. An element u is called *normed*, or *normalized*, if it has a unit norm, i.e. if $\|u\| = 1$.

Theorem 3. *If L is a closed subspace* (Definition 3) *of a Hilbert space H, then each element $u \in H$ can be uniquely represented in the form*

$$u = v + z, \tag{1}$$

where $v \in L$, $z \perp L$ (i.e. z is orthogonal to *all* elements of L).

The element v is called the *orthogonal projection* of the element u into the subspace L.

REMARK 5. It can be shown that all elements of H which are orthogonal to a given closed subspace L constitute another closed subspace (the so-called *orthogonal complement of L*), let us denote it by M. In the sense of equation (1) we then write

$$H = L + M \tag{2}$$

(the notation $H = L \oplus M$ is also often used) and we say that H is a *direct sum* of the subspaces L and M.

REMARK 6. Theorem 3 is a certain generalization of the well-known decomposition of a vector in the space V_3 of three-dimensional vectors: Construct a plane ρ passing through the origin and a straight line l perpendicular to it, passing as well through the origin. Vectors, lying in the plane ρ, or in the straight line l, form a two-dimensional, or one-dimensional subspace of the space V_3, respectively. Each vector $u \in V_3$ can be uniquely written in the form $u = u_\rho + u_l$, i.e. as the sum of two orthogonal vectors, the first of which lies in ρ, the second in l.

Theorem 4. *Let $u_n \in H$ $(n = 1, 2, \ldots)$ and let $u_i \perp u_k$ for $i \neq k$. Then the series*

$$\sum_{n=1}^{\infty} u_n$$

is convergent (in the sense of the metric of the Hilbert space H, see Remark 3) if and only if the series

$$\sum_{n=1}^{\infty} \|u_n\|^2$$

is convergent.

Definition 9. We say that the elements

$$u_1, u_2, \ldots, u_n, \ldots \in H$$

constitute an *orthogonal system* in H, if

$$(u_i, u_k) = 0 \quad \text{whenever} \quad i \neq k.$$

If, moreover, all these elements are normed, i.e. if

$$(u_k, u_k) = \|u_k\|^2 = 1$$

(Definition 8) holds for all $k = 1, 2, \ldots$, the given system is called *orthonormal.*

An orthonormal system is thus characterized by the relations

$$(u_i, u_k) = 0 \quad \text{if} \quad i \neq k, \qquad (u_i, u_k) = 1 \quad \text{if} \quad i = k.$$

Definition 10. Let

$$u_1, u_2, \ldots, u_n, \ldots \tag{3}$$

be an *orthonormal* system in H, let u be an arbitrary element of H. The numbers

$$a_n = (u, u_n), \quad n = 1, 2, \ldots, \tag{4}$$

are called the *Fourier coefficients of the element u with respect to the (orthonormal) system* (3). The series

$$\sum_{n=1}^{\infty} a_n u_n \tag{5}$$

is called the *Fourier series of the element u with respect to that system.*

Theorem 5. *A necessary and sufficient condition for a series*

$$\sum_{n=1}^{\infty} b_n u_n \tag{6}$$

(the u_n being elements of the *orthonormal* system (3)) *to be convergent (in H) is that the series*

$$\sum_{n=1}^{\infty} |b_n|^2 \tag{7}$$

converges (cf. Theorem 4). *If the series* (7) *converges, then the series* (6) *converges to a certain element* $v \in H$, *and* b_n *are Fourier coefficients of* v *with respect to the system* (3).

Theorem 6. *For the Fourier coefficients* (4) *of any element* $u \in H$ *we have*

$$\sum_{n=1}^{\infty} |a_n|^2 \leqq \|u\|^2 \tag{8}$$

(the so-called *Bessel inequality*).

REMARK 7. From Theorem 5 it then follows that, for any element $u \in H$, the corresponding Fourier series (5) is convergent in H (but not necessarily to the element u; see Theorem 7).

Theorem 7. *A necessary and sufficient condition for the series* (5) *to converge exactly to the element* u *is that*

$$\sum_{n=1}^{\infty} |a_n|^2 = \|u\|^2 . \tag{9}$$

Definition 11. If (9) is satisfied for *every* element $u \in H$ (i.e. if for every $u \in H$ the corresponding Fourier series (5) converges to u), the orthonormal system (3) is called *complete in* H. Equation (9) is called the *Parseval equality* (or *equation of completeness*).

Definition 12. The system (3) is called *closed* in H if H does not contain any non-zero element u orthogonal to each element of the system (3).

We note that as far as the terms "complete" and"closed" are concerned, terminology is not consistent in the literature (due, perhaps, to the following theorem):

Theorem 8. *The system* (3) *is complete in* H *if and only if it is closed.*

Theorem 9. *In a separable* (Definition 22.3.3) *Hilbert space there is at least one complete orthonormal system.*

REMARK 8. In $L_2(0, l)$, a typical example of a complete orthonormal system is furnished by the system of functions

$$\varphi_n(x) = \sqrt{\left(\frac{2}{l}\right)} \sin \frac{n\pi x}{l}, \quad n = 1, 2, \ldots . \tag{10}$$

For any function $u \in L_2(0, l)$ we thus have, in L_2,

$$u = \sum_{n=1}^{\infty} a_n \varphi_n \quad (\text{with} \quad a_n = \sqrt{\left(\frac{2}{l}\right)} \int_0^l u(x) \sin \frac{n\pi x}{l} \, dx),$$

i.e. (see Remark 3)

$$\left\| u - \sum_{n=1}^{k} a_n \varphi_n \right\| \to 0 \quad \text{for} \quad k \to \infty. \tag{11}$$

However, (11) does *not* imply pointwise convergence in the interval $[0, l]$, see § 16.3.

REMARK 9. If the elements

$$u_1, u_2, u_3, \ldots \tag{12}$$

are not orthogonal in H, the concept of a *complete sequence* may also be defined. We say that the elements (12) constitute a complete sequence (or a *base*) in H, if to each element $u \in H$ and to each $\varepsilon > 0$ a positive integer k and constants $b_1, b_2, \ldots, b_k$ can be found such that

$$\rho\left(u, \sum_{n=1}^{k} b_n u_n\right) = \left\| u - \sum_{n=1}^{k} b_n u_n \right\| < \varepsilon.$$

Using the familiar orthonormalization process (see § 16.2), a *complete orthonormal system* (= an *orthonormal base*) can be obtained from any complete sequence. For example, the orthonormalization of the sequence

$$1, x, x^2, x^3, \ldots$$

which is complete in $L_2(-1, 1)$ yields a complete orthonormal system of functions

$$\sqrt{\left(n + \tfrac{1}{2}\right)}\, P_n(x),$$

where $P_n(x)$ are Legendre polynomials.

REMARK 10 (*Sobolev spaces*). Simple examples of Hilbert spaces were given in Example 4. An other example of a Hilbert space is the so-called *Sobolev space* $W_2^{(k)}(\Omega)$ which plays an important role in modern methods of solving problems in partial differential equations. Here we present the definition of this space for the case of a *bounded* region Ω in E_n with the so-called *Lipschitz boundary* (*Lipschitz region*, in brief). Exact definition of these regions is rather complicated and can be found, e.g., in [389], Chap. 28. To that type of regions there belong, in E_2, bounded regions (multiply connected, in general) with boundaries constituted by a finite number of smooth, or piecewise smooth curves, without cuspidal points (a circle, a rectangle, a triangle, an annulus, etc.), in E_3 bounded regions with smooth or piecewise smooth boundaries, without corresponding singularities, i.e. edges of regression (§ 9.13), etc. (a sphere, an ellipsoid, a cube, a pyramide, a.s.o.).

Let $C^{(k)}(\Omega)$, or $C^{(k)}(\overline{\Omega})$ be the set of all functions which have continuous (partial) derivatives up to the order k inclusive in Ω, or $\overline{\Omega}$, respectively. (Instead of $C^{(0)}(\Omega)$, or $C^{(0)}(\overline{\Omega})$ (thus if continuity of the functions themselves is required only), the brief notation $C(\Omega)$, or $C(\overline{\Omega})$ is used.)

On the set $C^{(k)}(\overline{\Omega})$, let us define the scalar product

$$(u, v)_{W_2^{(k)}(\Omega)} \tag{13}$$

of two (real) functions $u(x) = u(x_1, \ldots, x_n)$, $v(x) = v(x_1, \ldots, x_n)$ as a sum of scalar products, in $L_2(\Omega)$, of the functions u, v and of their (partial) derivatives up to the order k inclusive. Thus, for $n = 2$, $k = 1$, for example, we have

$$(u, v)_{W_2^{(1)}(\Omega)} = (u, v)_{L_2(\Omega)} + \left(\frac{\partial u}{\partial x_1}, \frac{\partial v}{\partial x_1}\right)_{L_2(\Omega)} + \left(\frac{\partial u}{\partial x_2}, \frac{\partial v}{\partial x_2}\right)_{L_2(\Omega)} =$$

$$= \iint_\Omega u(x_1, x_2)\, v(x_1, x_2)\, dx_1\, dx_2 + \iint_\Omega \frac{\partial u}{\partial x_1}\frac{\partial v}{\partial x_1}\, dx_1\, dx_2 + \iint_\Omega \frac{\partial u}{\partial x_2}\frac{\partial v}{\partial x_2}\, dx_1\, dx_2.$$

Let us define the norm and the distance of these functions by

$$\|u\|_{W_2^{(k)}(\Omega)} = \sqrt{(u, u)_{W_2^{(k)}(\Omega)}}, \quad d(u, v)_{W_2^{(k)}(\Omega)} = \|v - u\|_{W_2^{(k)}(\Omega)}. \tag{14}$$

In this way, the set $C^{(k)}(\overline{\Omega})$ is converted into a metric space. Denote it by $S_2^{(k)}(\Omega)$. It can be shown that for $k \geqq 1$, this space is *not* complete.

Its completion (Remark 22.3.1) can be made with the help of functions from $L_2(\Omega)$ with the so-called *generalized derivatives*. Let us introduce this concept in brief: Let $i = (i_1, \ldots, i_N)$ be an n-dimensional vector the components of which are non-negative integers (the so-called *multiindex*). Denote $|i| = i_1 + \ldots + i_n$. Let us write briefly

$$D^i u \quad \text{for} \quad \frac{\partial^{|i|} u}{\partial x_1^{i_1} \ldots \partial x_n^{i_n}}.$$

For example, if $n = 2$, $i = (1, 2)$, we have

$$D^i u = \frac{\partial^3 u}{\partial x_1 \partial x_2^2}.$$

Further, let $C^{(\infty)}(\Omega)$, or $C^{(\infty)}(\overline{\Omega})$ be the set of functions infinitely times continuously differentiable in Ω, or in $\overline{\Omega}$, respectively. Denote by $C_0^{(\infty)}(\Omega)$ the set of such functions from $C^{(\infty)}(\Omega)$ which have the so-called *compact support* in Ω. Here, the compact support of a function u, denoted by

$$\operatorname{supp} u,$$

means the closure of the set of points $x \in \Omega$ at which $u(x) \neq 0$. The set $\operatorname{supp} u$ being closed and lying in Ω, by definition, each function from $C_0^{(\infty)}(\Omega)$ – and, consequently, also each of its derivatives – is equal to zero in a certain neighbourhood of the boundary of the region Ω (different for different functions from $C_0^{(\infty)}(\Omega)$, in general).

Let u be an arbitrary function from $C^{(k)}(\overline{\Omega})$, φ an arbitrary function from $C_0^{(\infty)}(\Omega)$. Applying $|i|$-times the Green theorem and using the just mentioned property of functions from $C_0^{(\infty)}(\Omega)$, one obtains

$$\int_\Omega u(x)\, D^i\varphi(x)\, \mathrm{d}x = (-1)^{|i|} \int_\Omega D^i u(x)\, \varphi(x)\, \mathrm{d}x \quad \text{for every} \quad |i| \leqq k. \tag{15}$$

(Here, the brief notation for the integral over Ω has been used similarly as in Remark 22.2.3.) The relations (15) have been derived for sufficiently smooth functions (we assumed $u \in C^{(k)}(\overline{\Omega})$). However, they can be satisfied as well for some functions from $L_2(\Omega)$ not belonging to $C^{(k)}(\overline{\Omega})$. (An example of such a function is, for $n = 1$ and $k = 1$, the function defined in the interval $[0, 2]$ by

$$u(x) = \begin{cases} x & \text{for } 0 \leqq x \leqq 1, \\ 2 - x & \text{for } 1 < x \leqq 2 \end{cases}$$

which does not belong to $C^{(1)}([0, 2])$, because it does not have the first derivative at the point $x = 1$.)

We define:

Let $u \in L_2(\Omega)$ and a multiindex $i = (i_1, \ldots, i_n)$ be given. Let the relation

$$\int_\Omega u(x)\, D^i\varphi(x)\, \mathrm{d}x = (-1)^{|i|} \int_\Omega v_i(x)\, \varphi(x)\, \mathrm{d}x \tag{16}$$

be satisfied for *every* $\varphi \in C_0^{(\infty)}(\Omega)$. Then we say that the function u has the *i-th generalized derivative* v_i (of order $|i|$).

It is usual to denote this derivative by the same symbol $D^i u$ as in the classical case. The function $v_i = D^i u$ is by the function u uniquely determined. If the function u has the classical derivative $D^i u$, then the generalized derivative $D^i u$ is equal to that classical derivative.

It can be shown that the completion (Remark 22.3.1) of the space $S_2^{(k)}(\Omega)$ can be done by adjoining, to that space, all functions from $L_2(\Omega)$ which have, in Ω, square integrable generalized derivatives up to the order k inclusive (and by extending the scalar product (13) and the norm and metric (14) to those functions). The so obtained complete space (thus a Hilbert space) is called the *Sobolev space* and is denoted by $W_2^{(k)}(\Omega)$, often also by $H_k(\Omega)$, or $H^k(\Omega)$.

Let us note that the original set of functions from the space $S_2^{(k)}(\Omega)$ (thus from $C^{(k)}(\overline{\Omega})$) is dense in this space. For details see e.g. [389], Chap. 29.

A generalization of Sobolev spaces are the so-called *weighted Sobolev spaces* (see e.g. [283]) which are applied to the solution of boundary value problems with various singularities (in coefficients of the given differential equation, etc.).

REMARK 11 (*Traces of functions from the space* $W_2^{(k)}(\Omega)$). As said above, the elements of $W_2^{(k)}(\Omega)$ consist of functions from $C^{(k)}(\overline{\Omega})$ and of such functions from $L_2(\Omega)$ which have square integrable generalized derivatives up to the order k in Ω. At the same time, the set $C^{(k)}(\overline{\Omega})$ is dense in $W_2^{(k)}(\Omega)$. As announced above, Sobolev spaces play a fundamental role in the theory of so-called weak solutions of problems in partial differential equations. Such a weak solution u is to be found among functions from $W_2^{(k)}(\Omega)$, satisfying certain boundary conditions (see § 18.9). However, it is to be made clear what does it mean that a function $u \in W_2^{(k)}(\Omega)$, or its derivatives, assume prescribed values on the boundary S of the region Ω:

For a function $u(x)$ continuous in $\overline{\Omega}$, and thus the more for every function $u \in$ $\in C^{(k)}(\overline{\Omega})$, its values $u(S)$ on the boundary are uniquely given. Let us call the function $u(S)$ the *trace* of the function $u(x)$ on the boundary. However, in general, the elements of $W_2^{(k)}(\Omega)$ are functions from $L_2(\Omega)$ only (although they have some other properties), and it is not clear, in this case, what is meant if we say that $u(S)$ is the trace of such a function.

Let us investigate the space $W_2^{(1)}(\Omega)$, at first. Let $u \in W_2^{(1)}(\Omega)$. This function need not belong to $C^{(1)}(\overline{\Omega})$, of course. However, the functions from $C^{(1)}(\overline{\Omega})$, being dense in $W_2^{(1)}(\Omega)$, the function $u(x)$ can be taken for the limit, in $W_2^{(1)}(\Omega)$, of an appropriate sequence of functions $u_n \in C^{(1)}(\overline{\Omega})$. It can be shown that, at the same time, the sequence of corresponding traces $u_n(S)$ converges in $L_2(S)$ to a certain function $v(S)$, independently of the choice of the sequence $\{u_n(x)\}$ converging to $u(x)$ in $W_2^{(1)}(\Omega)$. This function $v \in L_2(S)$ is called the *trace of the function* $u \in W_2^{(1)}(\Omega)$ *on* S.

Similarly, traces of the functions $\partial u/\partial x_j$ $(j = 1, \ldots, n)$ on S can be defined if $u \in W_2^{(2)}(\Omega)$, giving the possibility of defining, for example, $\partial u/\partial \nu$ on S (the outward normal derivative) in the sense of traces, etc. In this way, various boundary conditions for functions from the space $W_2^{(k)}(\Omega)$ can be formulated.

In particular, the subspace of such functions v from $W_2^{(k)}(\Omega)$ for which we have

$$D^i v = 0 \quad \text{on} \quad S \quad \text{for} \quad |i| \leqq k-1$$

is usually denoted by $\mathring{W}_2^k(\Omega)$ (or $H_0^k(\Omega)$). (On the other hand, $H_2^0(\Omega) = L_2(\Omega)$.)

For details see e.g. [389], Chap. 30. See also § 18.9.

REMARK 12 (*Distributions*). A further important concept is the concept of a distribution:

In the set $C_0^{(\infty)}(\Omega)$ (see Remark 10), let the convergence be introduced in the following way: We say that $\varphi_n \to \varphi$ in $C_0^{(\infty)}(\Omega)$, if a subregion Ω' of Ω ($\overline{\Omega'} \subset \Omega$) exists such that all the functions φ_n and the function φ have their supports (Remark 10) in Ω', i.e. if

$$\operatorname{supp} \varphi_n \subset \Omega', \quad n = 1, 2, \ldots, \quad \operatorname{supp} \varphi \subset \Omega'$$

and if, at the same time, the sequence of functions φ_n as well as sequences of all their derivatives $D^i\varphi_n$ converge in the metric of the space $C(\overline{\Omega'})$ (thus uniformly in $\overline{\Omega'}$) to the function φ and to the corresponding derivatives $D^i\varphi$, respectively. Let us denote that space (thus the set $C_0^{(\infty)}(\Omega)$ equipped with this definition of convergence) by $D(\Omega)$, briefly by D.

The dual space to $D(\Omega)$ (thus the space of all bounded linear functionals f on that space, see Remark 22.5.9), is called the *space of distributions.* It is denoted by D'.

Thus, a distribution f is a bounded linear functional on $D(\Omega)$. Let us denote the value of f at the point $\varphi \in D(\Omega)$ by

$$\langle \varphi, f \rangle.$$

The *derivative* $D^i f$ of this distribution is defined by

$$\langle \varphi, D^i f \rangle = (-1)^{|i|} \langle D^i \varphi, f \rangle \quad \text{for all} \quad \varphi \in D(\Omega).$$

Convergence in D' is defined as follows: We say that $f_n \to f$ in D', if

$$\lim_{n \to \infty} \langle \varphi, f_n \rangle = \langle \varphi, f \rangle \quad \text{for all} \quad \varphi \in D(\Omega).$$

The applicability of distributions in different fields of mathematics, in particular in the theory of boundary value problems of partial differential equations, lies in the fact that they enable to formulate many problems in a sufficiently general and, at the same time, rather natural way. Many simple functional spaces can be easily embedded (see Remark 13 below) into the space of distributions. For example, we can write

$$L_{1,\mathrm{loc}}(\Omega) \subset D'(\Omega),$$

putting

$$\langle \varphi, f \rangle = \int_\Omega \varphi f \, dx \quad \text{for all} \quad \varphi \in D(\Omega)$$

for every locally Lebesgue integrable function f. The *Dirac distribution* δ_{x_0} with the "singular point" $x_0 \in \Omega$ is defined by

$$\langle \varphi, \delta_{x_0} \rangle = \varphi(x_0) \quad \text{for all} \quad \varphi \in D(\Omega)$$

which makes it possible to express various singularities, e.g. isolated loads in the theory of elasticity, etc., in an appropriate way.

REMARK 13 (*The Spaces $W_p^{(k)}(\Omega)$, Embedding theorems*). More general than the space $W_2^{(k)}(\Omega)$ is the space $W_p^{(k)}(\Omega)$, with $1 \leqq p < \infty$, elements of which are functions from the space $L_p(\Omega)$, having generalized derivatives, up to the order k inclusive, in that space. For details see e.g. [284], [348].

The $W_p^{(k)}(\Omega)$ spaces are Banach spaces. For $p = 2$ we obtain the Hilbert space $W_2^{(k)}(\Omega)$ discussed in Remark 10.

It can be shown that a function from a space $W_p^{(k)}(\Omega)$ belongs simultaneously to a space $L_q(\Omega)$ with q "better" than p. In details:

Let X and Y be (linear) normed spaces and let every $u \in X$ belong simultaneously to the space Y. Then we write

$$X \subset Y \tag{17}$$

and say that the *space X is embedded*, or, in more details, *algebraically imbeded, into the space Y*. If, moreover, a constant $c > 0$, independent of $u \in X$, exists such that

$$\|u\|_Y \leqq c\|u\|_X \quad \text{for all} \quad u \in X, \tag{18}$$

we say that X *is continuously*, or *topologically embedded into* Y and write

$$X \subsetneq Y. \tag{19}$$

Let Ω be a bounded region in E_n with a Lipschitz boundary (Remark 10), k a positive integer, $1 \leqq p < \infty$. Then it can be shown that

$$W_p^{(k)}(\Omega) \subsetneq L_q(\Omega) \tag{20}$$

for every q satisfying

$$1 \leqq q \leqq \frac{pn}{n - kp}, \qquad \text{if} \quad kp < n, \tag{21}$$

$$1 \leqq q < \infty, \qquad \text{if} \quad kp = n. \tag{22}$$

If

$$kp > n, \tag{23}$$

we have even

$$W_p^{(k)}(\Omega) \subsetneq C(\overline{\Omega}). \tag{24}$$

Thus, if (23) holds, every function $u \in W_p^{(k)}(\Omega)$ is continuous in $\overline{\Omega}$ and a constant $c > 0$ (independent of $u \in W_p^{(k)}(\Omega)$) exists such that we have

$$\|u\|_{C(\overline{\Omega})} = \max_{x \in \overline{\Omega}} |u(x)| \leqq c\|u\|_{W_p^{(k)}(\Omega)} \quad \text{for all} \quad u \in W_p^{(k)}(\Omega). \tag{25}$$

(More exactly – because functions from $W_p^{(k)}(\Omega)$ are certain functions from $L_p(\Omega)$: if (23) is satisfied, then every function from $W_p^{(k)}(\Omega)$ can be converted into a function continuous in $\overline{\Omega}$ when changed, if necessary, on a set of Lebesgue measure zero; at the same time, (25) holds.)

Example 6. Let $n = 3$. By (21) every function $u \in W_2^{(1)}(\Omega)$ belongs to each space $L_q(\Omega)$ with $1 \leqq q \leqq 6$.

Example 7. Let $n = 1$ (thus, functions of one variable are considered). Then, by (23) and (24), every function from $W_2^{(1)}(a, b)$ is continuous in $[a, b]$, or can be made continuous there when changed in a proper way on a set of measure zero, and a constant $c > 0$ exists, independent of u, such that for every $u \in W_2^{(1)}(a, b)$ we have

$$\|u\|_{C([a,b])} = \max_{a \leqq x \leqq b} |u(x)| \leqq c\|u\|_{W_2^{(1)}(a,b)}. \tag{26}$$

22.5. Linear and Other Operators in Metric Spaces. Banach's Theorem on Contraction Mapping. Functionals. Adjoint Operators, Adjoint (Dual) Spaces. Completely Continuous Operators.

Definition 1. Let X and Y be two metric spaces. If to each $u \in X$ a uniquely determined $v \in Y$ is assigned, we write briefly

$$v = Au, \quad u \in X, \quad v \in Y, \tag{1}$$

and say that on X (or in X) an *operator* (or *mapping*) A is given which maps X into Y (or we speak of an operator A from X into Y, in brief).

REMARK 1. A similar definition can be given for more general sets X, Y than for metric spaces (cf. Definition 1.23.6). However, just the case considered in Definition 1 is most interesting from the point of view of applications of functional analysis.

REMARK 2. In (1), the element $u \in X$ is called the *original* and the element $v = Au \in Y$ the *image* of the original u. Two operators A, B are called *equal*, if $Au = Bu$ for all $u \in X$. The set of all $v \in Y$ which are obtained by (1) for all $u \in X$ is called the *range* of the operator A and is denoted by $R(A)$. If $R(A) = Y$ (i.e. if all elements of the space Y are obtained if u runs through the space X), we speak about a mapping from X *onto* Y (or we say that the operator A is *surjective*). If to any two different originals there correspond different images, i.e. if

$$u_1 \neq u_2 \quad \Rightarrow \quad Au_1 \neq Au_2,$$

the operator A is called *simple* (or *injective*). A simple surjective operator is called *one-to-one* (or *bijective*). In this case there exists the so-called *inverse operator* A^{-1} which assigns to every $v \in Y$ just that $u \in X$ for which $Au = v$. We write $u = A^{-1}v$. We have obviously

$$u = A^{-1}Au, \quad v = AA^{-1}v \quad \text{for all} \quad u \in X \quad \text{and} \quad v \in Y. \tag{2}$$

If $Y = X$, we say that the operator A *maps the space X into itself*, or we speak briefly about an *operator in the space X*.

An operator need not be defined on the entire space X, but only on a subset, or subspace $D(A) \subset X$, on the so-called *domain of definition* of the operator A (denoted also by D_A). This case is often encountered in applications. (If, for example, $X = L_2(a, b)$ and A is a differential operator, then it is not possible to take the whole space $L_2(a, b)$ for its domain of definition, but only a subspace of those (sufficiently smooth) functions from this space to which the operator A can be applied.) All the above-given definitions then remain valid, if X is replaced by the domain of definition $D(A)$ of the operator A. Two operators A, B are then called *equal* if they have the same domain of definition D and if

$$Au = Bu \quad \text{for every} \quad u \in D.$$

REMARK 3. If Y is the space of real, or complex numbers, the operator is called a *functional* (*real*, or *complex*, respectively). An example of a functional defined on the (entire) space $L_2(a, b)$ (or, more generally, on $L_p(a, b)$ with $p \geqq 1$) is the operator f given by

$$fu = \int_a^b u(x)\mathrm{d}x, \tag{3}$$

which, therefore, to every function $u \in L_2(a, b)$ (or $u \in L_p(a, b)$) assigns the number (3).

Theorem 1 (*The Banach Fixed-Point Theorem* or *Contraction Mapping Theorem*). *Let the operator A, defined on a complete metric space X, maps X into itself (i.e. if $u \in X$, then also $v = Au \in X$). Let a number α $(0 < \alpha < 1)$ exist such that*

$$d(Au, Av) \leqq \alpha d(u, v) \tag{4}$$

holds for any two elements $u, v \in X$ (d being the distance between u, v in X, see Definition 22.2.1; *thus the operator A "contracts distances"). Then the equation*

$$u = Au \tag{5}$$

has exactly one solution u_0 in X. The element u_0 (the so-called fixed-point of the operator A) can be obtained by successive approximations as a limit (in the metric of the space X) of the sequence

$$u_2 = Au_1, \quad u_3 = Au_2, \quad u_4 = Au_3, \quad \ldots, \tag{6}$$

where the initial element u_1 may be chosen arbitrarily.

The speed of convergence is given by the following estimate:

$$d(u_0, u_n) \leqq \frac{\alpha^{n-1}}{1-\alpha} d(u_1, Au_1) = \frac{\alpha^{n-1}}{1-\alpha} d(u_1, u_2).$$

REMARK 4. Using this theorem, existence (and uniqueness) of solutions of various problems can be proved, such as problems in the field of differential and integral equations, of finite and infinite systems of linear algebraic equations, etc. See the following example.

Example 1. Consider a nonlinear integral equation

$$u(x) = \lambda \int_a^b K\big(x, t, u(t)\big)\,\mathrm{d}t, \quad a \leqq x \leqq b, \tag{7}$$

where the function $K(x,t,z)$ is continuous and bounded in absolute value by a constant c in a parallelepiped $\overline{Q}$ ($a \leqq x \leqq b$, $a \leqq t \leqq b$, $|z| \leqq k$) and, in addition, satisfies the Lipschitz condition with respect to z in $\overline{Q}$, i.e. there exists a constant L such that the inequality

$$\big|K(x, t, z_2) - K(x, t, z_1)\big| \leqq L\,|z_2 - z_1| \tag{8}$$

holds for all (x, t, z_1), $(x, t, z_2) \in \overline{Q}$. The assertion is that for every λ sufficiently small, more exactly, for every λ satisfying the inequalities

$$|\lambda|\, c(b-a) \leqq k, \tag{9}$$

$$|\lambda|\, L(b-a) < 1, \tag{10}$$

there exists one and only one continuous function $u(x)$ satisfying equation (7).

For the space X let us choose the set of all functions continuous in $[a, b]$ such that $|u(x)| \leqq k$, with the metric of the space $C([a, b])$, i.e.

$$d(u, v) = \max_{a \leqq x \leqq b} \big|v(x) - u(x)\big|. \tag{11}$$

In view of (11) (uniform convergence), the space X is complete. If condition (9) is satisfied, the operator

$$Au = \lambda \int_a^b K\big(x, t, u(t)\big)\,\mathrm{d}t \tag{12}$$

(we write briefly Au instead of $(Au)(x)$) maps the space X into itself (because for every $x \in [a, b]$ we have

$$|v(x)| = |Au| = \left|\lambda \int_a^b K\big(x, t, u(t)\big)\mathrm{d}t\right| \leqq |\lambda| \int_a^b \Big|K\big(x, t, u(t)\big)\Big|\mathrm{d}t \leqq |\lambda|\,(b-a)c \leqq k$$

by (9); moreover, $v(x)$ is obviously a continuous function in $[a, b]$).

Furthermore, in view of (8) we have for every pair of functions u, v from X

$$d(Au, Av) = \max_{a \leqq x \leqq b} \left|\lambda \int_a^b K(x, t, v(t))\mathrm{d}t - \lambda \int_a^b K(x, t, u(t))\mathrm{d}t\right| \leqq$$

$$\leqq |\lambda|\, L \int_a^b |v(t) - u(t)|\mathrm{d}t \leqq |\lambda|\, L(b-a) \max_{a \leqq t \leqq b} |v(t) - u(t)| =$$

$$= |\lambda|\, L(b-a) d(u, v).$$

Thus, if (10) is true, condition (4) is satisfied and equation (7) possesses, by Theorem 1, a unique solution. This solution can be obtained as the limit of a uniformly convergent sequence of successive approximations (6).

Definition 2. An operator A is called *continuous at a point* $u_0 \in D(A)$ if for every sequence $\{u_n\}$ $(u_n \in D(A))$, $u_n \to u_0$ in the metric of the space X we have $Au_n \to Au_0$ in the metric of the space Y.

Definition 3 (*Linear Operators in Linear Spaces*). An operator A is called *linear*, if

(i) its domain of definition $D(A)$ is a linear set M (Definition 22.4.1),

(ii) for any elements $u_k \in M$ and any numbers c_k we have

$$A(c_1 u_1 + \ldots + c_m u_m) = c_1 Au_1 + \ldots + c_m Au_m. \tag{13}$$

Theorem 2. *A linear operator, continuous at some point* $u_1 \in D(A)$, *is continuous at every point* $u \in D(A)$.

Definition 4. Let X, Y be linear normed spaces (see Definition 22.4.4), A a linear operator from $D(A) \subset X$ into Y. The operator A is called *bounded* if a positive number C, independent of u, exists such that

$$\|Au\|_Y \leqq C\|u\|_X \tag{14}$$

holds for every $u \in D(A)$. (Here, naturally, $\|u\|_X$ denotes the norm of the element u in the space X, $\|Au\|_Y$ the norm of the element Au in the space Y.)

Example 2. In the complex space $L_2(a, b)$ with the norm

$$\|u\| = \sqrt{\left[\int_a^b |u(x)|^2 \mathrm{d}x\right]}, \tag{15}$$

the functional f given by

$$fu = \int_a^b u(x)\overline{v(x)}\mathrm{d}x, \tag{16}$$

where $v(x)$ is a fixed function from $L_2(a, b)$, is a bounded linear operator. Linearity is obvious. Further, by the Schwarz inequality (Theorem 22.4.2), we have

$$|fu| = \left|\int_a^b u(x)\overline{v(x)}\mathrm{d}x\right| = |(u, v)| \leqq \|u\| \cdot \|v\| .$$

Thus, in (14), it suffices to put

$$C = \|v\| .$$

On the other hand, consider the operator A, given by

$$Au = \frac{\mathrm{d}u}{\mathrm{d}x} \tag{17}$$

on the linear subspace $D(A) \subset L_2(a, b)$, consisting of those functions from $L_2(a, b)$ which are continuous, together with their first derivatives, in $[a, b]$. This operator, as an operator from $D(A)$ (with the metric of the space $L_2(a, b)$) into $L_2(a, b)$, is evidently linear, but *not* bounded. In fact, there exist functions from $D(A)$ with $\|u\| = 1$ for which $\|Au\| = \|\mathrm{d}u/\mathrm{d}x\|$ is as large as we please (we can, for instance, consider functions of the form

$$u(x) = \sqrt{\left(\frac{2}{b-a}\right)} \sin \frac{n\pi(x-a)}{b-a} \tag{18}$$

with n sufficiently large).

REMARK 5. Among the numbers C for which (14) holds, there exists a unique least number which is called the *norm of the operator* A and is denoted by $\|A\|$ or n_A. We have

$$\|A\| = n_A = \sup_{\|u\|_X = 1, u \in D(A)} \|Au\|_Y , \tag{19}$$

i.e. the norm of the operator A can be found as the least upper bound of the set of numbers $\|Au\|_Y$ with u ranging over all unit elements of $D(A)$ (or, in other words, with u ranging over the surface of the unit sphere in $D(A)$).

From the definition of the norm of an operator it follows that

$$\|Au\|_Y \leqq \|A\| \cdot \|u\|_X$$

for every $u \in D(A)$.

Theorem 3. *A linear operator A is continuous on $D(A)$ if and only if it is bounded.*

REMARK 6. *In the following text of this paragraph, B_1 and B_2 are Banach spaces* (i.e. complete normed linear spaces, see Definition 22.4.5) *and A is a continuous* (= *bounded*) *linear operator from B_1 (or from $D(A) \subset B_1$) into B_2.*

Theorem 4. *If $D(A)$ is a linear set dense in B_1, the operator A (linear and bounded on $D(A)$, see the foregoing remark) can be (uniquely) extended from $D(A)$ onto the entire space B_1 so that the norm of the operator A is preserved.*

REMARK 7. The operator extended in this manner (let us denote it by A') has the entire space B_1 as its domain of definition and satisfies the equality $A'u = Au$ for every element $u \in D(A)$. Moreover, we have $\|A'\| = \|A\|$. The process described in Theorem 4 is referred to as the *continuous extension of an operator.*

Theorem 5. *Let M be the set of bounded linear operators defined on the entire space B_1. For every element $u \in B_1$ let there exist a number $K(u)$ (thus depending on u, in general) such that $\|Au\|_{B_2} \leqq K(u) . \|u\|_{B_1}$ holds for every operator $A \in M$. Then the system M is uniformly bounded, i.e. a number K (independent of u and A) exists such that for all $u \in B_1$ and for every operator $A \in M$ we have*

$$\|Au\|_{B_2} \leqq K\|u\|_{B_1} \quad (\textit{briefly} \quad \|A\| \leqq K).$$

Theorem 6 (*The Banach Theorem on Inverse Operators*). *Let a bounded operator A map the entire space B_1 in one-to-one correspondence onto the (entire) space B_2. Then the inverse operator A^{-1}* (see Remark 2) *is also a bounded linear operator.*

REMARK 8. An important class of bounded linear operators are *bounded linear functionals*, i.e. bounded linear operators with Au being a real or complex number (see Remark 3). For functionals a more powerful theorem than Theorem 4 is true:

Theorem 7 (*The Hahn-Banach Theorem*). *A bounded linear functional f defined on a linear subspace L of a (linear) normed space B (L need not be dense in B) can be extended onto the entire space B with its norm preserved.*

REMARK 9 (*Dual Space*). Let us consider all bounded linear functionals f defined on a linear normed space B and define the sum of two functionals and the product of a functional and of a number (real or complex) as follows:

$$(f_1 + f_2)u = f_1u + f_2u; \quad (af)u = a(fu) \tag{20}$$

(for all $u \in B$). Furthermore, define the norm of a functional f by

$$\|f\| = n_f = \sup_{\|u\|_B = 1} |fu|$$

in accordance with Remark 5. Then all these functionals constitute a linear normed space which can be shown to be complete (and thus a Banach space). This space

is called the *adjoint space* or the *dual space* (briefly *dual*) *to the space* B (or *over the space* B) and is denoted by B^* (also the notation B' is often used).

Similarly, the dual space B^{**} to B^* can be defined.

If $B^* = B$ (or if B^* can be identified with B in a way described below), the space B is called *self-adjoint*. If $B^{**} = B$, it is called *reflexive*.

For example, by the Riesz theorem (Theorem 22.6.1), every bounded linear functional f on a Hilbert space H can be uniquely represented by an element $v \in H$, so that $fu = (u, v)$ holds for all $u \in H$. At the same time, $\|v\| = \|f\|$, the sum of two functionals f_1, f_2 is represented by the sum of the corresponding elements v_1, $v_2 \in H$, etc. Conversely, (u, v) is a bounded linear functional in H – with the norm $\|v\|$ – for every $v \in H$. Because the correspondence between the elements $f \in H^*$ and $v \in H$ is one-to-one (while $\|f\| = \|v\|$, etc.), it is possible to identify functionals $f \in X^*$ with corresponding elements $v \in H$ and write $H^* = H$. *In this sense, every Hilbert space is self-adjoint.* A similar consideration leads to the conclusion that *every Hilbert space is reflexive.*

Definition 5. A sequence $\{u_n\}$ of elements of B is said to be *weakly convergent* to an element $u_0 \in B$, if for every bounded linear functional f on B (thus for every $f \in B^*$) we have

$$fu_n \to fu_0.$$

Notation

$$u_n \rightharpoonup u_0, \quad \text{or} \quad u_n \xrightarrow{w} u_0. \tag{21}$$

Theorem 8. *If a sequence* $\{u_n\}$ *converges to* u_0, *then it is also weakly convergent to the same element* u_0.

The converse is not true, in general.

Theorem 9. *Let* B *be a reflexive Banach space. Then from every bounded sequence* $\{u_n\}$ *a subsequence* $\{u_{n_k}\}$ *can be chosen which is weakly convergent to an element* $u_0 \in B$.

REMARK 10. In particular, *from every bounded sequence in a Hilbert space a weakly convergent subsequence can be chosen.*

Definition 6 (*Adjoint Operator*). Let A be a bounded linear operator which maps B_1 into B_2, thus $v = Au$, $u \in B_1$, $v \in B_2$. Consider a bounded linear functional $f \in B_2^*$ (see Remark 9). Then obviously the functional given by

$$fv = fAu = gu \tag{22}$$

is a bounded linear functional on B_1. Thus, by means of the operator A, to every bounded linear functional $f \in B_2^*$ there is assigned a bounded linear functional $g \in B_1^*$ by the relation (21); let us write

$$g = A^* f. \tag{23}$$

The operator A^* is called the *adjoint operator* to the operator A.

REMARK 11. The definition of adjoint operators is simpler and very natural in a Hilbert space, see Remark 22.6.2. See also Example 22.6.2.

Definition 7. A continuous (= bounded) linear operator from B_1 into B_2 is called *completely* (*absolutely*) *continuous* (or *compact*) if it maps every bounded set $M \subset B_1$ onto a relatively compact (Definition 22.3.4) set $M' \subset B_2$.

The above given definition can be well generalized. For example, B_1, B_2 can be merely linear metric spaces. Let us note that there is no conformity in the literature as concerns the concept of a completely continuous, absolutely continuous, or compact operator. Under compact operators also nonlinear operators are often understood which map bounded sets into relatively compact sets.

Example 3. Let $K(x, t)$ be a function square integrable in a square $\overline{Q}$ ($a \leqq x \leqq b$, $a \leqq t \leqq b$), thus $K \in L_2(Q)$. It can be shown that the operator A given by

$$Au = \int_a^b K(x, t)\, u(t)\, \mathrm{d}t, \quad u \in L_2(a, b), \tag{24}$$

(we write again briefly Au instead of $(Au)(x)$) is completely continuous from $L_2(a, b)$ into $L_2(a, b)$. Its norm is

$$\|A\| = \sqrt{\left[\int_a^b \int_a^b |K(x, t)|^2 \mathrm{d}x\, \mathrm{d}t\right]}\,.$$

Definition 8. A sequence of operators A_n is said to be *uniformely convergent* to an operator A, if $\|A - A_n\| \to 0$ if $n \to \infty$.

For $\|A\|$ see Remark 5.

Theorem 10. *If a sequence of completely continuous operators A_n (from B_1 into B_2) is uniformly convergent to an operator A, then A is a completely continuous operator.*

REMARK 12. In the theory of integral equations, solvability of Fredholm equations of the form

$$u(x) - \mu \int_a^b K(x, t)\, u(t)\, \mathrm{d}t = v(x) \tag{25}$$

in dependence on the parameter μ is investigated. In a similar way, operator equations

$$Au - \lambda u = v \tag{26}$$

in various spaces can be discussed. If A is a completely continuous operator, similar results are obtained. See the next paragraph.

22.6. Operators and Operator Equations in Hilbert Spaces

REMARK 1. A Hilbert space is (see Remark 22.4.4) a special case of a Banach space. Consequently, everything stated above in § 22.5, concerning operators in a Banach space, is true also for operators in Hilbert space. However, in a Hilbert space much more about operators, or about corresponding operator equations, can be said.

In Part (a) *of this paragraph, we deal with linear bounded operators defined in the entire Hilbert space* H, a fact which will not be explicitly restated in the following text. (A bounded linear operator defined only on a linear subspace L, dense in H, can be taken for already extended onto the entire space by Theorem 22.5.4.)

In Part (b), *unbounded operators are considered*, typical representants of which are differential operators.

(a) Bounded operators

Theorem 1 (*The Riesz Theorem*). *Any bounded linear functional* f *in a Hilbert space* H *can be uniquely represented in the form of a scalar product, i.e. to every bounded linear functional* f *on* H *a unique* $v \in H$ *exists such that*

$$fu = (u, v) \quad \textit{holds for all } u \in H.$$

Moreover, we have

$$\|v\| = \|f\|.$$

Example 1. In particular, if H is the space $L_2(a, b)$ with the inner product

$$(u, v) = \int_a^b u(x)\,\overline{v(x)}\,\mathrm{d}x, \tag{1}$$

then to every bounded functional f there corresponds exactly one function $v \in L_2(a, b)$ such that

$$fu = \int_a^b u(x)\,\overline{v(x)}\,\mathrm{d}x \quad \text{holds for all} \quad u \in L_2(a, b). \tag{2}$$

(The converse is obvious: For every fixed $v \in L_2(a, b)$, (1) is a bounded linear functional in $L_2(a, b)$ (see example 22.5.2).) Moreover, by Theorem 1, $\|v\| = \|f\|$.

REMARK 2 (*Adjoint Operator*). If A is a bounded linear operator in H, then

$$(Au, v),$$

with v fixed, is a bounded linear functional in H. By Theorem 1, there is exactly one element $v^* \in H$ such that

$$(Au, v) = (u, v^*) \quad \text{for all} \quad u \in H. \tag{3}$$

Thus by (3) (with A fixed) to every $v \in H$ a unique $v^* \in H$ is assigned – let us write

$$v^* = A^* v \tag{4}$$

– such that

$$(Au, v) = (u, A^* v) \quad \text{holds for all} \quad u, v \in H. \tag{5}$$

The operator A^* is called the *adjoint operator* to A.

Example 2. The adjoint operator to the operator A in Example 22.5.3, given by

$$Au = \int_a^b K(x, t)\, u(t)\, \mathrm{d}t, \quad u \in L_2(a, b),$$

is

$$A^* v = \int_a^b \overline{K(t, x)}\, v(t)\, \mathrm{d}t, \quad v \in L_2(a, b).$$

Thus A^* is an integral operator whose kernel is obtained from the original kernel $K(x, t)$ by interchanging the variables and taking the complex conjugate value. To verify that (5) is then satisfied, it is sufficient to observe that

$$(Au, v) = \int_a^b \left(\int_a^b K(x, t)\, u(t)\, \mathrm{d}t \right) \overline{v(x)}\, \mathrm{d}x = \int_a^b \int_a^b K(x, t)\, u(t)\, \overline{v(x)}\, \mathrm{d}x\, \mathrm{d}t =$$

$$= \int_a^b \int_a^b K(t, x)\, u(x)\, \overline{v(t)}\, \mathrm{d}t\, \mathrm{d}x = \int_a^b u(x) \overline{\left(\int_a^b \overline{K(t, x)}\, v(t)\, \mathrm{d}t \right)}\, \mathrm{d}x = (u, A^* v).$$

Theorem 2. *The norm of the adjoint operator A^* is equal to the norm of the operator A, i.e.*

$$\|A^*\| = \|A\|.$$

(See also Example 2, where this fact is obvious.)

Definition 1. If $A = A^*$, then A is called a *self-adjoint operator.* Then for every $u, v \in H$ we have

$$(Au, v) = (u, Av).$$

Example 3. An integral operator with a real symmetric kernel, for which therefore $K(t, x) = K(x, t)$, is a self-adjoint operator (see Example 2).

Theorem 3. *A necessary and sufficient condition for a bounded linear operator A in a complex Hilbert space to be self-adjoint is that (Au, u) be real for every $u \in H$.*

Definition 2. A self-adjoint operator A is called *positive*, if for every $u \in H$ we have

$$(Au, u) \geqq 0$$

while

$$(Au, u) = 0 \quad \text{only if } u = 0 \text{ in } H.$$

It is called *positive definite* if such a constant $m > 0$, independent of u, exists that

$$(Au, u) \geqq m \|u\|^2$$

holds for all $u \in H$.

Evidently a positive definite operator is positive, but not vice versa, in general.

Theorem 4. *Let A be a positive definite operator in a real Hilbert space H, let $f \in H$. Then the equation*

$$Au = f \tag{6}$$

has exactly one solution $u_0 \in H$. This solution minimizes, in H, the quadratic functional (functional of energy)

$$Fv = (Av, v) - 2(f, v). \tag{7}$$

(Conversely, an element u, minimizing in H the functional (7), is the solution of equation (6).)

Let us note that for a *complex* Hilbert space the theorem still holds if functional (7) is replaced by the functional

$$Fv = (Av, v) - (f, v) - (v, f) = (Av, v) - 2\,\mathrm{Re}(f, v)$$

which (in consequence of positive definiteness of the operator A) assumes, on H, only real values.

REMARK 3. Let us emphasize once more that Theorem 4 has been stated for linear *bounded* operators defined in the *entire* space H. Thus from the practical point of view, this theorem can be applied to integral operators, matrices, etc., but not to differential operators which do *not* have the just mentioned properties in current functional spaces. For that case see Theorems 9 and 10 below.

REMARK 4. More generally, let us consider, in H, the equation (cf. Remark 22.5.12)

$$Au - \lambda u = f \quad (\text{or } (Au - \lambda I)u = f, \tag{8}$$

where I is the *identity operator*, i.e. $Iu = u$ holds for every $u \in H$) and also the corresponding homogeneous equation

$$Au - \lambda u = 0 \quad (\text{or } (A - \lambda I)u = 0). \tag{9}$$

A value λ such that the operator $A - \lambda I$ possesses, in H, a bounded inverse operator (usually denoted by R_λ), is called a *regular value* (*regular point*) of the operator A; the operator R_λ is called the *resolvent operator* (or *resolvent*). For every regular value λ, equation (8) has exactly one solution $u \in H$ for every $f \in H$. In particular, equation (9) possesses only the trivial solution $u = 0$.

Those values of λ which are not regular constitute the so-called *spectrum* of the operator A.

A value λ such that equation (9) has (except the trivial solution $u = 0$) a non-zero solution $u \in H$, is called an *eigenvalue* of the operator A. This non-zero solution u is called an *eigenelement* (an *eigenvector*) corresponding to that eigenvalue λ.

Let us note that the spectrum of an operator A need not consist only of eigenvalues of this operator, in general.

Theorem 5. *If A is a self-adjoint operator in H* (Definition 1), *then*

(i) *any complex number $\lambda = \alpha + \mathrm{i}\beta$ (α, β real, $\beta \neq 0$) is a regular value of the operator A. The operator A can thus have only real eigenvalues. If, in addition, A is positive* (Definition 2), *then A can have only positive eigenvalues.*

(ii) *eigenelements corresponding to different eigenvalues are orthogonal.*

About whether an operator A has eigenvalues at all, and about their structure, an information is given in the following theorem:

Theorem 6. *If A is a selfadjoint completely continuous* (Definition 22.5.7) *non-zero operator in a separable* (Definition 22.3.3) *Hilbert space H, then*

(i) *it has at least one non-zero eigenvalue λ (real,* by Theorem 5*);*

(ii) *outside every interval $[-\varepsilon, \varepsilon]$, where ε is an arbitrary positive number, there can lie only a finite number of eigenvalues, and to each of them there corresponds only a finite number of linearly independent eigenelements;*

(iii) *an orthonormal* (see the text preceding Theorem 7 below) *system of elements formed by all linearly independent eigenelements of the operator A including eigenelements corresponding to the eigenvalue $\lambda = 0$ (if $\lambda = 0$ is an eigenvalue) is complete in H.*

Example 4. An example of an operator discussed in Theorem 6 is furnished by an integral operator with a real symmetric kernel $K(x, t) \in L_2$ (see Examples 2 and 3 and Example 22.5.3).

Before stating Theorem 7, let us note that it is customary to order eigenvalues and corresponding linearly independent eigenelements in such a way that there is one-to-one correspondence between them. If, for example, the eigenvalues are arranged into a nonincreasing sequence and if to the first eigenvalue λ there correspond two linearly independent eigenelements v_1, v_2 (then λ is called a *double eigenvalue*, or *eigenvalue of order*, or *multiplicity* 2), then we require for that λ to appear twice in that sequence, writing

$$\lambda_1 = \lambda_2 > \lambda_3 \dots .$$

Moreover, it is usual, in theoretical considerations, to assume that all linearly independent eigenelements have been already orthonormalized by the orthonormalization process shown in Remark 16.2.15.

Theorem 7. *Let A be a completely continuous positive* (and thus self-adjoint) *operator in a separable Hilbert space of infinite dimension. Then this operator has a countable set of eigenvalues, all of them being positive, and to each of them there correspond a finite number of linearly independent eigenelements. The orthonormal system*

$$v_1, \ v_2, \ v_3, \ \dots$$

of eigenelements corresponding to the system

$$\lambda_1 \geqq \lambda_2 \geqq \lambda_3 \geqq \dots > 0$$

of eigenvalues in the sense of the preceding text is complete in H. Moreover,

$$\lim_{n\to\infty} \lambda_n = 0.$$

For the first eigenvalue λ_1 we have

$$\lambda_1 = \max_{v\in H, v\neq 0} \frac{(Av, v)}{(v, v)} = \frac{(Av_1, v_1)}{(v_1, v_1)},$$

while for the n-th ($n \geqq 2$) eigenvalue

$$\lambda_n = \max_{\substack{v\in H, v\neq 0, \\ (v,v_1)=0,\dots,(v,v_{n-1})=0}} \frac{(Av, v)}{(v, v)} = \frac{(Av_n, v_n)}{(v_n, v_n)},$$

holds.

REMARK 5. (i) Thus the so-called *Rayleigh quotient*

$$\frac{(Av, v)}{(v, v)} \tag{10}$$

($v \in H$, $v \neq 0$) assumes, on H, its maximal value, equal to the largest eigenvalue λ_1, just for the first eigenelement v_1. Then λ_2 is obtained as maximum of (10) if v runs through the subspace H_1 of such elements of H which are orthogonal to v_1, this maximum being attained for the second eigenelement v_2. Then the Rayleigh quotient is maximized in the subspace H_2 of such elements of H which are orthogonal both to v_1 and v_2, to obtain λ_3 just for $v = v_3$, etc.

As concerns assumptions of Theorem 7, a remark similar to Remark 3 can be added. Let us note, further, that Theorem 7, being formulated for positive operators, is the more valid for positive definite operators.

(ii) Theorem 7 with the following Theorem 8 imply that if $0 \neq \lambda \neq \lambda_n$ ($n = 1, 2, \ldots$), then the equation

$$Au - \lambda u = f$$

is uniquely solvable for every right-hand side $f \in H$.

(iii) As concerns application of the Rayleigh quotient (10) to obtain successively the eigenvalues $\lambda_1, \lambda_2, \ldots$, a similar process can be applied without the assumption of positiveness of the operator A; here self-adjointness and complete continuity are essential. However, care is to be taken here, because the eigenvalues can be positive, as well as negative, or equal to zero, in that case. To $\lambda = 0$ an infinite number of corresponding linearly independent eigenelements may exist. For details see e.g. [389], Chap. 38.

(iv) Let us note, finally, that Theorem 7 can be formulated as well for finite-dimensional spaces (for matrices in n-dimensional vector space, for example). Naturally, in that case there exist only a finite number of eigenvalues, etc.

REMARK 6. Now, let A be a completely continuous (not necessarily self-adjoint) operator (for example, an integral operator with a kernel $K \in L_2$ which need not satisfy the condition $K(t, x) = \overline{K(x, t)}$), and let A^* be the adjoint operator to A. Let us consider the equations

$$Au - \lambda u = f, \tag{11}$$

$$A^*u - \overline{\lambda}u = g, \tag{12}$$

$$Au - \lambda u = 0, \tag{13}$$

$$A^*u - \overline{\lambda}u = 0. \tag{14}$$

Theorem 8. *Let A be a completely continuous operator in H, $\lambda \neq 0$. Then equation* (11) *possesses a unique solution for any element $f \in H$ if and only if equation* (13) *has, in H, only a zero solution (i.e. if λ is not an eigenvalue of the operator A). The same statement is true for equations* (12) *and* (14). *If equation* (13) *has a nonzero solution (i.e. if λ is an eigenvalue of the operator A), then equation* (11) *is solvable (not uniquely, in this case) only for those $f \in H$ which*

are orthogonal to all solutions of equation (14). *A similar statement is true for equations* (14) *and* (12). *Further: if* λ *is an eigenvalue of equation* (13), *then* $\overline{\lambda}$ *is an eigenvalue of equation* (14). *Both equations have then the same number of linearly independent solutions.*

From the first part of Theorem 8, there follows the so-called *Fredholm alternative*: *Let* A *be a completely continuous operator,* $\lambda \neq 0$. *Then either equation* (11) *is uniquely solvable for every right-hand side* $f \in H$, *or equation* (13) *has a non-zero solution.*

A similar assertion holds for equations (12), (14).

(b) Unbounded operators

REMARK 7. In Example 22.5.2 an example of an unbounded operator has been given. In the following text of this paragraph, we shall consider linear *unbounded* operators defined on linear sets $D(A)$ *dense* in a Hilbert space H. This case is most often encountered in applications.

Example 5. An example of such an operator is the operator A given by

$$Au = -\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} \tag{15}$$

(the reason for the sign minus becomes clear in Example 6) on its domain of definition $D(A)$ consisting of all continuous functions with two continuous derivatives in $[a, b]$ (thus of all functions from $C^{(2)}([a, b])$), such that $u(a) = 0$, $u(b) = 0$. These functions constitute a linear set in the space $H = L_2(a, b)$ with the well-known scalar product (1). It can be shown (see, e.g., [389], Chap. 8) that this set is dense in H.

An other example is the operator

$$Au = -\Delta u = -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)$$

considered on a dense (in $L_2(\Omega)$) linear set $D(A)$ consisting of all continuous functions $u(x, y)$ with two continuous partial derivatives in a closed bounded region $\overline{\Omega}$ with a smooth, or piecewise smooth (or a lipschitzian, see Remark 22.4.10) boundary S, which satisfy the condition $u = 0$ on S.

REMARK 8 (*Adjoint Operator*). For some elements $v \in H$ there exists an element $v^* \in H$ such that

$$(Au, v) = (u, v^*) \tag{16}$$

holds for every $u \in D(A)$. (The elements $v = 0$, $v^* = 0$, for example, have the just mentioned property.) Since the linear set $D(A)$ is assumed to be dense in

H (Remark 7) it can be shown that the element v^* is by the element v uniquely determined. Thus, on the set (say M) of all these v, an operator A^* is given by

$$v^* = A^*v, \quad v \in M, \tag{17}$$

such that

$$(Au, v) = (u, A^*v) \tag{18}$$

holds for every $u \in D(A)$ and every $v \in M = D(A^*)$. The operator A^* (which is clearly linear and, in general, unbounded) is called the *adjoint operator to the operator* A. If $A^* = A$ (i.e. if $D(A^*) = D(A)$ and $A^*v = Av$ holds for every $v \in D(A)$), A is called a *self-adjoint operator.*

We recall the fact that $D(A)$ is assumed to be a set dense in H as stated in Remark 7.

Definition 3. A linear operator A is called *symmetric on* $D(A)$, if the relation

$$(Au, v) = (u, Av) \tag{19}$$

holds for every u and v from $D(A)$.

For a symmetric operator we have $A \subset A^*$, i.e. $D(A) \subset D(A^*)$ and $A^*u = Au$ for $u \in D(A)$. (This fact follows immediately from the definition of a symmetric and an adjoint operator.)

Definition 4. A *symmetric* operator A is called *positive on* $D(A)$, if for every $u \in D_A$ we have

$$(Au, u) \geqq 0,$$

while

$$(Au, u) = 0 \quad \text{only if} \quad u = 0.$$

It is called *positive definite on* $D(A)$, if such a number $m > 0$ exists independent of $u \in D_A$, that

$$(Au, u) \geqq m \|u\|^2$$

holds for every $u \in D_A$.

Example 6. Let A be the operator given by

$$Au = -\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}, \tag{20}$$

defined on the linear set $D(A)$ of all functions from $C^{(2)}([a, b])$ satisfying the boundary conditions

$$u(a) = 0, \quad u(b) = 0. \tag{21}$$

$D(A)$ is dense in $H = L_2(a, b)$ (Example 5). It is easy to show that A is a (symmetric) positive operator in $D(A)$:

Symmetry: Integrating by parts and using conditions (21), we have

$$(Au, v) = -\int_a^b u''\overline{v}\,\mathrm{d}x = -[u'\overline{v}]_a^b + \int_a^b u'\overline{v'}\,\mathrm{d}x = \int_a^b u'\overline{v'}\,\mathrm{d}x =$$
$$= [u\overline{v'}]_a^b - \int_a^b u\overline{v''}\,\mathrm{d}x = -\int_a^b u\overline{v''}\,\mathrm{d}x = (u, Av)$$

for every $u \in D(A)$, $v \in D(A)$.

Positiveness:

$$(Au, u) = -\int_a^b u''\overline{u}\,\mathrm{d}x = -[u'\overline{u}]_a^b + \int_a^b u'\overline{u'}\,\mathrm{d}x = \int_a^b u'\overline{u'}\,\mathrm{d}x = \int_a^b |u'|^2\,\mathrm{d}x \geqq 0 \tag{22}$$

for every $u \in D(A)$; further, if $(Au, u) = 0$, then (22) implies $u' \equiv 0$, hence $u \equiv c = \text{const.}$ in $[a, b]$. However, $u \in D(A)$, so that (21) holds, and thus $u \equiv 0$ in $[a, b]$.

REMARK 9. It can be shown that the considered operator is even positive definite (see e.g. [389], Chap. 8).

The same assertions hold for the second operator considered in Example 5.

Theorem 9 (*on Minimum of Functional of Energy*). *Let A be a positive operator on its domain of definition $D(A)$ dense in a real Hilbert space H, $f \in H$. Then:*

(i) *If the equation*

$$Au = f \tag{23}$$

has a solution $u_0 \in D(A)$, then u_0 minimizes, on $D(A)$, the quadratic functional (the so-called functional of energy, or energy functional)

$$Fu = (Au, u) - 2(f, u). \tag{24}$$

(ii) *Conversely, an element $u_0 \in D(A)$ which minimizes the functional* (24) *on $D(A)$, i.e. such that*

$$\min_{u \in D(A)} Fu = Fu_0,$$

is (the unique) solution of the equation (23).

Let us note (cf. Theorem 4) that for a *complex* Hilbert space Theorem 9 and the text following it still holds, if the functional (24) is replaced by the functional

$$Fu = (Au, u) - (f, u) - (u, f) = (Au, u) - 2\,\mathrm{Re}(f, u).$$

REMARK 10 (*The Space H_A, Generalized Solutions*). In contrast to Theorem 4 (stated for *bounded* positive definite operators definited on the *entire* space H), Theorem 9 does not represent an existence theorem. Here, neither a solution of (23) nor the minimum of (24) on $D(A)$ need exist. For example, if A is the operator from Example 6 and if $f \in L_2(a, b)$ is such a "sufficiently" discontinuous function in $[a, b]$ that it cannot be made continuous when changed on a set of measure zero (a piecewise constant function, for example), then the equation $Au = f$ cannot have a solution u_0 from $D(A)$, because $u_0 \in C^{(2)}([a, b])$ implies that Au_0 is continuous in $[a, b]$, while f has *not* this property.

If the operator A is positive definite on $D(A)$, then it is possible to ensure existence of a solution of equation (23) in a generalized sense. This can be done in the following way: For every pair of elements $u, v \in D(A)$ let us define a new scalar product $(u, v)_A$ (the so-called *energetic* (*energy*) *scalar product*) by

$$(u, v)_A = (Au, v), \tag{25}$$

and on its base the so-called *energetic* (*energy*) *norm* and *distance* by

$$\|u\|_A = \sqrt{(u, u)_A}\,, \quad \rho_A(u, v) = \|v - u\|_A\,, \tag{26}$$

respectively. In this way, $D(A)$ is converted into a metric space – let us denote it by S_A – with the metric (26). This space is not complete, in general. Let us complete it according to Remark 22.3.1. It can be shown that, A being positive definite, this completion can be done by joining certain elements from H to the set D_A (see Remark 12) and by extending the scalar product $(u, v)_A$, defined originally by the relation (25) for elements $u, v \in D_A$ only, to these new elements. The so obtained complete space is called the *energetic* (*energy*) *space* and is denoted by H_A. It is thus a Hilbert space with the scalar product

$$(u, v)_A, \quad u, v \in H_A. \tag{27}$$

The functional (24) is then extended onto the whole space H_A by

$$Fu = (u, u)_A - 2(f, u). \tag{28}$$

Now, it can be shown (completeness of H_A is essentially applied here) that this functional really assumes its minimum on H_A, for an element $u_0 \in H_A$ uniquely determined by the right-hand side f of equation (23). With regard to Theorem 9, more exactly to its second part, the element u_0 is called the *generalized solution of equation* (23).

Summarizing, we have:

Theorem 10. *Let H be a Hilbert space, A an operator positive definite on its domain of definition $D(A)$, dense in H, $f \in H$. Let us consider, on $D(A)$, the equation*

$$Au = f \tag{29}$$

and the functional (24). *Let H_A be the space introduced in Remark* 10. *Then the functional* (24) *can be extended by* (28) *onto the whole space H_A, and the so extended functional attains its minimum on H_A. The element u_0 for which this minimum is realized, is uniquely determined by the right-hand side f of equation* (29).

REMARK 11. As announced above, the element u_0 is called the *generalized solution* of equation (29). Thus, in this sense, every equation of the form $Au = f$ with a positive definite operator A (and with $f \in H$) is solvable. According to the definition, its generalized solution minimizes, in H_A, the functional (28). To the minimization of that functional variational methods can then be applied (see Chap. 24).

If the functional of energy assumes its minimum on $D(A)$, then the minimizing element u_0 is the solution of equation (29) in the "ordinary" (classical) sense.

REMARK 12. The whole theory discussed above has been motived, first of all, by the problematics of boundary value problems in differential (in particular, partial differential) equations. In that case, the space L_2 is mostly taken for H, and the completion of the space S_A (Remark 10) can be done by joining functions from L_2 with a sufficient number of generalized derivatives (Remark 22.4.10). In such a case, the embedding (Remark 22.4.13) of the space H_A into the space H can be shown to be compact, i.e. every set bounded in H_A is relatively compact in H. Moreover, in the just discussed problematics, concerning differential equations, H_A as well as H are separable spaces of infinite dimensions. This justifies assumptions of the following theorem on eigenvalues for (unbounded) positive definite operators. However, before stating that theorem, let us add still a small remark: As seen before (Remark 10, Theorem 10), even in the case when the operator A is positive definite (and thus special enough), the equation (29), i.e. the equation

$$Au = f,$$

need not have a solution in an "ordinary" sense (i.e. from $D(A)$). The same holds for an eigenvalue problem. Thus also here the concepts of an eigenvalue, or of an eigenelement, need a certain generalization: *We say that λ is an eigenvalue of the operator A (or of the equation*

$$Au - \lambda u = 0\,) \tag{30}$$

and that $u \in H_A$, $u \neq 0$, is the corresponding eigenelement, if

$$(u, v)_A - \lambda(u, v) = 0 \tag{31}$$

holds for all $v \in H_A$. This generalization which makes it possible to state a theorem on existence of eigenvalues, is very natural. In fact, if λ is an eigenvalue in the "ordinary" sense, i.e. if such an element $u \in D(A)$, $u \neq 0$, exists that (30) holds, then, multiplying (scalarly) this equation by an arbitrary element $v \in H$, we obtain

$$(Au, v) - \lambda(u, v) = 0.$$

The generalization consists in writing $(u, v)_A$ in (31) instead of (Au, v); here it is no more necessary to assume $u \in D(A)$. (In particular, in the case of differential equations, the "classical" smoothness of eigenfunctions need not be assumed – such eigenfunctions need not exist at all.)

Theorem 11. *Let us consider the equation*

$$Au - \lambda u = 0 \tag{32}$$

with an unbounded operator A *positive definite on a linear set* $D(A)$ *dense in a separable Hilbert space of infinite dimension. Let the embedding of the space* H_A *into the space* H *be compact* (Remark 12). *Then equation* (32) (*or, in other words, the operator* A) *has a countable set of eigenvalues, each of them being positive, and to every eigenvalue there corresponds a finite number of linearly independent eigenelements* (generalized, in general. i.e. from the space H_A, see the foregoing remark). *The system of all linearly independent eigenelements*

$$u_1, \ u_2, \ u_3, \ \ldots \tag{33}$$

(which can be assumed as already orthonormalized in H_A, or in H), *corresponding to the system of eigenvalues*

$$0 < \lambda_1 \leqq \lambda_2 \leqq \lambda_3 \leqq \ldots \tag{34}$$

in the sense of the text preceding Theorem 7, *is complete in* H_A *as well as in* H. *Moreover, we have*

$$\lim_{n \to \infty} \lambda_n = +\infty.$$

The so-called Rayleigh quotient

$$\frac{(u, u)_A}{(u, u)}, \quad u \in H_A, \ u \neq 0, \tag{35}$$

attains, on H_A, *its minimum, equal to* λ_1, *exactly for the first eigenelement* u_1, *i.e.*

$$\lambda_1 = \min_{u \in H_A, u \neq 0} \frac{(u, u)_A}{(u, u)} = \frac{(u_1, u_1)_A}{(u_1, u_1)};$$

λ_2 is obtained as minimum of (35) *on the subspace of H_A of such elements which are orthogonal, in H_A, to the element u_1, this minimum being attained for the second eigenelement u_2,*

$$\lambda_2 = \min_{\substack{u \in H_A, u \neq 0 \\ (u,u_1)_A = 0}} \frac{(u, u)_A}{(u, u)} = \frac{(u_2, u_2)_A}{(u_2, u_2)},$$

etc., in general

$$\lambda_n = \min_{\substack{u \in H_A, u \neq 0 \\ (u,u_1)_A = 0, \ldots, (u,u_{n-1})_A = 0}} \frac{(u, u)_A}{(u, u)} = \frac{(u_n, u_n)_A}{(u_n, u_n)}.$$

REMARK 13. For $\lambda \neq \lambda_n$ $(n = 1, 2, \ldots)$ the equation

$$Au = \lambda u = f$$

is uniquely solvable (in a generalized sense, if necessary, i.e. with a solution from H_A) for every $f \in H$.

REMARK 14. The reader may be rather surprised by the fact that in Theorem 11 we have, for the eigenvalues, $\lambda_n \to +\infty$, while in Theorem 7 we had $\lambda_n \to 0$. However, in Theorem 11 unbounded operators are considered, in Theorem 7 bounded operators, even completely continuous. In applications, typical representants of unbounded operators are differential operators, of bounded operators integral operators, which – very roughly speaking – have "inverse" properties.

REMARK 15. For details, generalizations, approximate methods in eigenvalue problems, etc., see, e.g. [389]. See also Chap. 24.

22.7. Abstract Functions. The Bochner Integral

Definition 1. Let J be a bounded closed interval,

$$J = [a, b], \tag{1}$$

X a linear normed space. We say that an *abstract function*

$$f: J \to X \tag{2}$$

is given on J if to every $t \in J$ a unique element $u \in X$ is assigned.

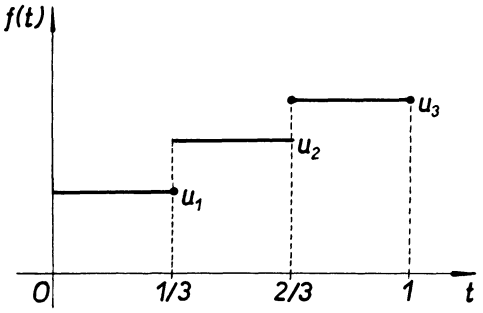


Fig. 22.3.

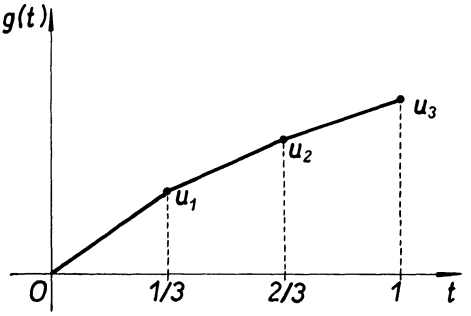


Fig. 22.4.

Let us note that an abstract function can be defined in a more general way (see e.g. [284]).

Example 1. Let $J = [0, 1]$, let u_1, u_2, u_3 be elements from the space X. The function $f(t)$, defined by

$$f(t) = \begin{cases} u_1 & \text{for } 0 \leqq t \leqq \frac{1}{3}, \\ u_2 & \text{for } \frac{1}{3} < t < \frac{2}{3}, \\ u_3 & \text{for } \frac{2}{3} \leqq t \leqq 1, \end{cases} \tag{3}$$

is an example of an abstract function defined on J. Its "symbolical graph" is sketched in Fig. 22.3. A "symbolical graph" of another abstract function $g(t)$ given with the help of these three elements by

$$g(t) = \begin{cases} 3tu_1 & \text{for } 0 \leqq t \leqq \frac{1}{3}, \\ u_1 + 3\left(t - \frac{1}{3}\right)(u_2 - u_1) & \text{for } \frac{1}{3} < t < \frac{2}{3}, \\ u_2 + 3\left(t - \frac{2}{3}\right)(u_3 - u_2) & \text{for } \frac{2}{3} \leqq t \leqq 1, \end{cases}$$

is given in Fig. 22.4. For $t = 1/6$, it assumes the "value" $u_1/2$.

An abstract function (2) is called *continuous at an inner point* t_0 of the interval J if for every $\varepsilon > 0$ such a $\delta > 0$ can be found that

$$|\Delta t| < \delta \quad \Rightarrow \quad \left\| f(t_0 + \Delta t) - f(t_0) \right\|_X < \varepsilon. \tag{4}$$

Here, $\|u\|_X$ is the norm of the element u in the space X (and $t_0 + \Delta t$ is assumed to belong to J, of course).

Analogously, *continuity from the right*, or *from the left* is defined, respectively, in a similar way as in "classical" analysis.

The function f is called *continuous in the interval* (a, b) if it is continuous at each point of that interval. It is called *continuous in the interval* $J = [a, b]$ if it is

continuous in (a, b) and continuous from the right, or from the left at the point a, or b, respectively.

Similarly as continuity, the *limit* $l \in X$ (or *limit from the right*, or *left*) at a point t_0 is defined. Inequalities (4) are then replaced by the inequalities

$$0 < |\Delta t| < \delta \quad \Rightarrow \quad \|f(t_0 + \Delta t) - l\|_X < \varepsilon.$$

An example of a continuous abstract function in the interval J is the function $g(t)$ "sketched" in Fig. 22.4. The function (3) has, for example, at the point $t = \frac{1}{3}$ the limit from the left equal to the element $u_1 \in X$, from the right to the element $u_2 \in X$. Thus it is discontinuous at that point.

The function (2) is said to have a *derivative* $f' = h \in X$ at an inner point t_0 of the interval J, if

$$\lim_{\Delta t \to 0} \left\| \frac{f(t_0 + \Delta t) - f(t_0)}{\Delta t} - h \right\|_X = 0. \tag{5}$$

Similarly the *derivative from the right*, or *from the left* is defined.

For an abstract function the so-called *Bochner integral* can be defined which represents a certain analogue of the Lebesgue integral of functions of a real variable. First, the so-called *simple function* is defined: The function (2) is called simple if it attains, in the interval J, only a finite number of values $g_1, \ldots, g_n \in X$, on Lebesgue measurable sets $S_1, \ldots, S_n$ of measures $\mu_1, \ldots, \mu_n$ $(\mu_1 + \ldots + \mu_n = b - a)$, respectively. The Bochner integral of such a function is then defined by

$$\int_J f(t)\,\mathrm{d}t = \sum_{i=1}^{n} g_i \mu_i. \tag{6}$$

An example of a simple function is the function (3). According to (6) we have, for that function,

$$\int_0^1 f(t)\,\mathrm{d}t = u_1 \cdot \tfrac{1}{3} + u_2 \cdot \tfrac{1}{3} + u_3 \cdot \tfrac{1}{3} = \tfrac{1}{3}(u_1 + u_2 + u_3).$$

It is clear that the integral in (6), defined by the sum on the right-hand side of (6), is an element of X. In our example it equals to one third of the sum of the elements u_1, u_2, u_3.

Measurable abstract functions in the Bochner sense are functions which can be approximated, with an arbitrary accuracy, by simple functions. More precisely: An abstract function (2) is called *strongly Bochner measurable*, if there exists a sequence of simple functions $f_n(t)$ such that

$$\lim_{n \to \infty} \|f(t) - f_n(t)\|_X = 0 \quad \text{for almost all} \quad t \in J. \tag{7}$$

If, moreover,

$$\lim_{n\to\infty}\int_J \|f(t)-f_n(t)\|_X \mathrm{d}t = 0,$$

we say that the function (2) is *Bochner integrable* and define

$$\int_J f(t)\,\mathrm{d}t = \lim_{n\to\infty}\int_J f_n(t)\,\mathrm{d}t. \tag{8}$$

In applications, the space $L_2(J, X)$ of so-called *square integrable* (*in the Bochner sense*) *abstract functions* is often encountered, i.e. of Bochner integrable functions for which

$$\int_J \|f(t)\|_X^2 \mathrm{d}t < +\infty.$$

In particular, if X is a Hilbert space H with the scalar product $(u_1, u_2)_H$, then $L_2(J, H)$ is as well a Hilbert space with the scalar product

$$(f, g)_{L_2(J,H)} = \int_J \big(f(t), g(t)\big)_H \mathrm{d}t.$$

For details and generalization see [284]. See also [390].

22.8. The Gâteaux Differential and Related Concepts

In nonlinear functional analysis as well as in nonlinear problems of partial differential equations, the concept of the so-called *Gâteaux differential* of a functional is often encountered.

Definition 1. Let X be a normed linear space (Definition 22.4.4). A functional F (nonlinear in general), defined on X, is called *continuous at a point* $u_0 \in X$, if

$$u \to u_0 \quad \text{in } X \quad \Rightarrow \quad Fu \to Fu_0, \tag{1}$$

or, what is the same, if

$$u \to u_0 \quad \text{in } X \quad \Rightarrow \quad |Fu_0 - Fu| \to 0. \tag{2}$$

It is called *continuous in the space* X if it is continuous at every point $u_0 \in X$.

Definition 2. We say that the functional F has the *Gâteaux differential at a point* $u_0 \in X$, if for every fixed $v \in X$ there exists the (finite) derivative

$$\frac{\mathrm{d}}{\mathrm{d}t}F(u_0 + tv)\bigg|_{t=0} = F'(u_0, v) \tag{3}$$

and if, moreover, the functional $F'(u_0, v)$ is linear in v.

The most usual notation

$$F'(u_0, v), \quad \mathrm{d}F(u_0, v), \quad \mathrm{D}F(u_0, v). \tag{4}$$

REMARK 1. Let us note that u_0 and v being fixed,

$$f_v(t) = F(u_0 + tv) \tag{5}$$

is an ordinary function of the variable t and that (3) is its ordinary derivative with respect to t at the point $t = 0$. Thus $F'(u_0, v)$ is a functional again. The fact that u_0 being fixed, the value of this functional depends on v, is pointed out in the notation (4). Cf. also Definition 7 below.

Let us note that the terminology is not uniform in the literature. See Remark 11.

Definition 3. The *second Gâteaux differential* is defined by

$$F''(u_0, v_0, w) = \frac{\mathrm{d}}{\mathrm{d}t} F'(u_0 + tw, v_0)\Big|_{t=0}. \tag{6}$$

Example 1. Let $X = W_2^{(1)}(0, 1)$. (For the Sobolev spaces $W_2^{(k)}$ see Remark 22.4.10.) Let us investigate the functional

$$Fu = \frac{1}{2}\int_0^1 u'^2 \mathrm{d}x + 5\int_0^1 u^4 \mathrm{d}x. \tag{7}$$

Let us show, first of all, that for every $u \in W_2^{(1)}(0, 1)$ both integrals in (7) have sense. For the first one, this fact is obvious, because in $W_2^{(1)}(0, 1)$ we have (as follows from (22.4.13), (22.4.14) in the quoted Remark 22.4.10)

$$\|u\|^2_{W_2^{(1)}(0,1)} = \int_0^1 u^2(x)\,\mathrm{d}x + \int_0^1 u'^2(x)\,\mathrm{d}x. \tag{8}$$

Thus the first of the integrals (7) is convergent for every $u \in W_2^{(1)}(0, 1)$.

To establish convergence of the second one, let us remind embedding results discussed in Remark 22.4.13 (see also Example 22.4.7): If $u \in W_2^{(1)}(0, 1)$, then u is continuous in $[0, 1]$ (or can be made continuous there if changed in a proper way on a set of measure zero) and a constant $c > 0$ exists, independent of u, such that for every $u \in W_2^{(1)}(0, 1)$ we have

$$\|u\|_{C([0,1])} = \max_{0 \leqq x \leqq 1} |u(x)| \leqq c\,\|u\|_{W_2^{(1)}(0,1)}. \tag{9}$$

So the function u being continuous in $[0, 1]$ (or equivalent to a continuous function), the second of the integrals (7) is finite. Thus the functional (7) is well defined on $W_2^{(1)}(0, 1)$.

The *Gâteaux differential*:

$$f_v(t) = F(u_0 + tv) = \frac{1}{2}\int_0^1 (u_0' + tv')^2 \mathrm{d}x + 5\int_0^1 (u_0 + tv)^4 \mathrm{d}x =$$
$$= \frac{1}{2}\int_0^1 u_0'^2\,\mathrm{d}x + t\int_0^1 u_0'v'\,\mathrm{d}x + \frac{t^2}{2}\int_0^1 v'^2\mathrm{d}x + 5\int_0^1 u_0^4\mathrm{d}x +$$
$$+ 20t\int_0^1 u_0^3 v\,\mathrm{d}x + 30t^2\int_0^1 u_0^2v^2\mathrm{d}x + 20t^3\int_0^1 u_0v^3\mathrm{d}x + 5t^4\int_0^1 v^4\mathrm{d}x. \quad (10)$$

Thus

$$f_v'(t) = \int_0^1 u_0'v'\,\mathrm{d}x + t\int_0^1 v'^2\mathrm{d}x + 20\int_0^1 u_0^3 v\,\mathrm{d}x +$$
$$+ 60t\int_0^1 u_0^2v^2\,\mathrm{d}x + 60t^2\int_0^1 u_0v^3\mathrm{d}x + 20t^3\int_0^1 v^4\mathrm{d}x \quad (11)$$

and

$$F'(u_0, v) = f_v'(0) = \int_0^1 u_0'v'\,\mathrm{d}x + 20\int_0^1 u_0^3 v\,\mathrm{d}x. \quad (12)$$

Obviously, for every fixed u_0, the functinal (12) is linear in v.

The *second differential*:

$$F'(u_0 + tw, v_0) = \int_0^1 (u_0' + tw')v_0'\,\mathrm{d}x + 20\int_0^1 (u_0 + tw)^3 v_0\,\mathrm{d}x =$$
$$= \int_0^1 u_0'v_0'\,\mathrm{d}x + t\int_0^1 w'v_0'\,\mathrm{d}x + 20\int_0^1 u_0^3 v_0\,\mathrm{d}x +$$
$$+ 60t\int_0^1 u_0^2v_0w\,\mathrm{d}x + 60t^2\int_0^1 u_0v_0w^2\,\mathrm{d}x + 20t^3\int_0^1 v_0w^3\,\mathrm{d}x. \quad (13)$$

Thus

$$\frac{\mathrm{d}}{\mathrm{d}t}F'(u_0 + tw, v_0) = \int_0^1 v_0'w'\,\mathrm{d}x + 60\int_0^1 u_0^2v_0w\,\mathrm{d}x +$$
$$+ 120t\int_0^1 u_0v_0w^2\mathrm{d}x + 60t^2\int_0^1 v_0w^3\mathrm{d}x$$

and

$$F''(u_0, v_0, w) = \frac{\mathrm{d}}{\mathrm{d}t}F'(u_0 + tw, v_0)\Big|_{t=0} = \int_0^1 v_0'w'\,\mathrm{d}x + 60\int_0^1 u_0^2v_0w\,\mathrm{d}x. \quad (14)$$

The following theorems are significant in the theory of the so-called weak solutions of nonlinear boundary value problems (see § 18.9):

Theorem 1. *Let the functional F attain its minimum on X at a point $u_0 \in X$ and let the Gâteaux differential $F'(u_0, v)$ exist. Then*

$$F'(u_0, v) = 0 \quad \text{for every} \quad v \in X. \tag{15}$$

REMARK 2. The same conclusion holds if that minimum is only local.

Definition 4. The functional F is called *convex*, or *strictly convex* in the space X, respectively, if

$$(1-t)Fu + tFv \geqq F\big((1-t)u + tv\big), \tag{16}$$

or

$$(1-t)Fu + tFv > F\big((1-t)u + tv\big) \tag{17}$$

holds for every pair of elements $u, v \in X$ $(u \neq v)$ and for all $t \in (0, 1)$.

REMARK 3. This definition represents a very natural analogue of the definition of convexity of "ordinary" functions.

Definition 5. The functional F is called *coercive* on X if

$$\lim_{\|u\|_X \to \infty} Fu = +\infty. \tag{18}$$

Theorem 2. *Let X be a reflexive Banach space* (Remark 22.5.9 and Definition 22.4.5). *Let F be a continuous convex and coercive functional on X. Then there exists a point $u_0 \in X$ in which the functional F attains its minimum on X.*

If, moreover, F is strictly convex, then such a point is unique.

REMARK 4. According to Theorem 1, (15) is then satisfied at that point.

REMARK 5. As said above, the function (5) is an ordinary function of one variable t, for which then ordinary theorems, as the mean-value theorem, etc., are valid under corresponding assumptions. This makes it possible to find relatively simple criteria establishing assumptions of Theorem 2:

Theorem 3. *Let X be a reflexive Banach space* (for example a Hilbert space). *Let the functional F fulfil the following assumptions:*

(i) *F is defined on the entire space X and has the first and second Gâteaux differentials everywhere in X.*

(ii) *$F'(u, v)$ is bounded in the following sense: let M be the set of all $u \in X$ such that $\|u\|_X \leqq r$* (thus a sphere in X with its centre at the point $u = 0$ and radius

r). *Then a constant* $K(r)$ (dependent on r, but independent of $u \in M$) *exists such that*

$$|F'(u, v)| \leqq K(r) \|v\|_X \tag{19}$$

holds for all $u \in M$ *and* $v \in X$.

(iii) *A constant* $k > 0$ *exists* (independent of u, v) *such that*

$$F''(u, v, v) \geqq k \|v\|_X^2 \tag{20}$$

holds for all $u, v \in X$.

Then there exists exactly one point $u_0 \in X$ *in which the functional* F *attains its minimum on* X.

At that point

$$F'(u_0, v) = 0$$

then holds for all $v \in X$.

How to apply this theorem is shown in Example 18.9.2.

REMARK 6. If it is known that a functional is the Gâteaux differential $F'(u, v)$ of a functional F (conditions for it can be found e.g. in [160]), then the "original" functional (with $F(0) = 0$, if required) can be found as follows:

$$Fu = \int_0^1 F'(tu, u)\, \mathrm{d}t. \tag{21}$$

For example, (12) being known, the functional (7) is easily obtained:

$$Fu = \int_0^1 \left(\int_0^1 tu'u'\, \mathrm{d}x + 20 \int_0^1 (tu)^3 u\, \mathrm{d}x \right) \mathrm{d}t = \frac{1}{2} \int_0^1 u'^2\, \mathrm{d}x + 5 \int_0^1 u^4 \mathrm{d}x.$$

REMARK 7 (*Monotone operators*). The above problematics, discussed in the "language of functionals", can be considered as well in the "language of operators":

Let A be an operator (nonlinear, in general) from a Banach space B into its dual space B^* (Remark 22.5.9). Denote by $\langle f, u \rangle$ the value of the functional $f \in B^*$ at the point $u \in B$.

Definition 6. The operator A is called

(i) *potential* on B, if a functional F exists such that for all $u, v \in B$ we have

$$F'(u, v) = \langle Au, v \rangle \tag{22}$$

(for $F'(u, v)$ see Definition 2);

(ii) *bounded* on B, if it maps every set bounded in B onto a set bounded in B^*;

(iii) *monotone*, or *strictly monotone* on B, if for every couple of elements u, $v \in B$ we have

$$\langle Au - Av, u - v\rangle \geqq 0, \tag{23}$$

or

$$\langle Au - Av, u - v\rangle > 0, \quad u \neq v, \tag{24}$$

respectively.

(iv) *coercive* on B, if

$$\lim_{\|u\|_B \to \infty} \frac{\langle Au, u\rangle}{\|u\|_B} = +\infty. \tag{25}$$

Theorem 4. *Let A be a* (nonlinear) *potential, bounded, monotone and coercive operator from a reflexive* (Remark 22.5.9) *Banach space B into its dual B^*. Then the equation*

$$Au = f$$

has a solution $u_0 \in B$ (i.e.

$$\langle Au_0, v\rangle = \langle f, v\rangle \quad \textit{holds for each} \quad v \in B) \tag{26}$$

for every $f \in B^*$.

If, moreover, A is strictly monotone,then this solution is unique.

REMARK 8. Theorem 4 can be generalized, in particular the requirement of potentiality can be removed. Solvability of the equation $Au = f$ in a reflexive Banach space can be proved e.g. for continuous monotone and coercive operators.

REMARK 9. Let us note that also the Gâteuax differential can be defined in the "language of operators".

Definition 7. Let the operator A (nonlinear, in general) map a linear normed space X into a normed space Y. Let u_0 be a fixed element from X. Let, for every $v \in X$, the limit (in the norm of the space Y)

$$\lim_{t \to 0} \frac{A(u_0 + tv) - Au_0}{t} = C_{u_0} v \tag{27}$$

exist and let the operator C_{u_0} be linear in v. Then we say that the operator A has the *Gâteaux* (or *weak*) *differential at the point* u_0. The operator C_{u_0} (depending on u_0, in general) from X into Y is called the *Gâteaux* (or *weak*) *derivative of the operator A at the point u_0* and $C_{u_0} v$ is the *Gâteaux* (or *weak*) *differential* (*at that point*) *in the direction of v.*

See also Remark 11.

REMARK 10. In Definition 2, the operator A was a functional and we could speak simply about the derivative (3) instead of the limit (27). This way is usually more convenient in applications.

Rather similar to the definition of the Gâteaux differential, there is the definition of the Fréchet differential:

Definition 8. We say that the operator A from a normed space X into a normed space Y has the *Fréchet* (or *strong*) *differential at the point* $u_0 \in X$, if such a linear operator D_{u_0} exists from X to Y that for every $v \in X$ we have

$$A(u_0 + v) - Au_0 = D_{u_0}v + R(u_0, v), \tag{28}$$

where

$$\lim \frac{\|R(u_0, v)\|_Y}{\|v\|_X} = 0 \quad \text{for} \quad \|v\|_X \to 0. \tag{29}$$

The operator D_{u_0}, depending on u_0, in general, is often denoted by $A'(x_0)$ and is called the *Fréchet derivative of the operator* A *at the point* u_0, while $C_{u_0}v = A'(x_0)v$ is then called the *Fréchet* (or *strong*) *differential of the operator* A *at the point* u_0 *in the direction of* v.

See also Remark 11.

If A is linear, then $A' = A$ for every u_0.

REMARK 11. As concerns Definitions 7 and 8, there is no uniformity in the literature. What we call *diferential* here is often called *derivative* (*in the direction of* v). Also other definitions are in use. In particular, not only linearity in v, but also continuity of the operators C_{u_0}, D_{u_0} is often required.

Theorem 5. *If the operator* A *has the Fréchet differential at the point* u_0, *then it has the Gâteaux diferential at that point as well, and both the differentials are equal.*

23. CALCULUS OF VARIATIONS

By FRANTIŠEK NOŽIČKA

References: [6], [47], [51], [65], [99], [103], [130], [149], [154], [161], [171], [214], [233], [260], [264], [304], [363], [464].

The calculus of variations is one of the classical branches of mathematical optimization. It has numerous applications in physics, especially in mechanics. Usually, the problem is to find such a function – from among functions possessing prescribed properties – for which the given integral (functional), whose value depends on these functions, assumes its extremum value. From the geometrical point of view, the problem can be stated as to find such a manifold (curve, surface, hypersurface) in a given class of smooth manifolds that gives the (at least local) minimum or maximum to the given functional, with respect to the class of the manifolds considered.

In accord with the type of the functional whose extremum is to be found, it is convenient to divide the problems of the calculus of variations into certain categories for which general theoretical procedures have been worked out.

The notation

$$K = \{(x, y) \in E_2 \mid y = y(x),\ x \in [a, b]\}$$

(i.e. K is the set of points (x, y) in E_2 for which $x \in [a, b]$, $y = y(x)$) is frequently used in this chapter to describe a curve K in the xy-plane.

A. PROBLEMS OF THE FIRST CATEGORY (ELEMENTARY PROBLEMS OF THE CALCULUS OF VARIATIONS)

23.1. Curves of the r–th Class, Distance of Order r between Two Curves, ε–Neighbourhood of Order r of a Curve

Definition 1. A curve K in the Euclidean plane E_2 (with the Cartesian coordinates x and y) described by the equation

$$y = y(x),\ x \in [a, b]\,,$$

i.e. the set of points

$$K = \{(x, y) \in E_2 \,|\, y = y(x),\ x \in [a, b]\}, \tag{1}$$

is called the *curve of the r–th class* (or *of the class T_r*) *in the interval* $[a, b]$ if the function y has a continuous derivative $y^{(r)}$ in an open interval $(\bar{a}, \bar{b})$ containing the closed interval $[a, b]$ (thus y has also continuous derivatives of all lower orders including zero).

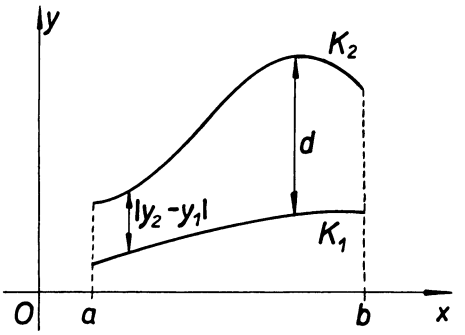


Fig. 23.1.

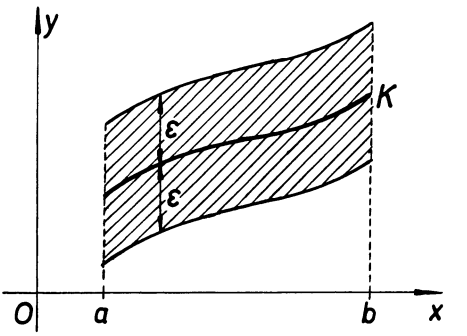


Fig. 23.2.

Definition 2. By the *distance between two curves*

$$\begin{aligned} K_1 &= \{(x, y) \in E_2 \,|\, y = y_1(x),\ x \in [a, b]\}, \\ K_2 &= \{(x, y) \in E_2 \,|\, y = y_2(x),\ x \in [a, b]\} \end{aligned} \tag{2}$$

(more exactly, by the *distance of order zero*) we mean the number

$$d(K_1, K_2) = \max_{x \in [a, b]} |y_2(x) - y_1(x)|,$$

i.e. the maximum of the differences (taken in absolute value) of their y–coordinates for all x in $[a, b]$ (Fig. 23.1).

Definition 3. If the curves K_1 and K_2 from (2) are of the r–th class, then the largest of the numbers

$$\max_{x \in [a, b]} |y_2(x) - y_1(x)|,\ \max_{x \in [a, b]} |y_2'(x) - y_1'(x)|, \ldots,\ \max_{x \in [a, b]} |y_2^{(r)}(x) - y_1^{(r)}(x)|,$$

i.e. the number

$$d_r(K_1, K_2) = \max_{k \in \{0, \ldots, r\}} \left\{ \max_{x \in [a, b]} |y_2^{(k)}(x) - y_1^{(k)}(x)| \right\},$$

where $y_i^{(0)}(x) = y_i(x)$, $i = 1, 2$, is called the *distance of order r between the curves considered.*

Apparently $d_0(K_1, K_2) = d(K_1, K_2)$.

REMARK 1. The family of all curves

$$\tilde{K} = \{(x, y) \in E_2 \,|\, y = \tilde{y}(x), x \in [a, b]\} \tag{3}$$

whose distance (of order zero) from the given curve K described by (1) is less than ε ($\varepsilon > 0$) is called the *ε–neighbourhood of the curve K.* The ε–neighbourhood of the curve K is thus the family of all curves described by (3) that lie in the zone indicated in Fig. 23.2. More generally, by the *ε–neighbourhood of order r of a curve K*, which is at least of class T_r, we mean all the curves $\tilde{K}$ of class T_r whose distance of order r from the curve K is less than ε, i.e. all the curves $\tilde{K}$ for which $d_r(K, \tilde{K}) < \varepsilon$. Apparently each curve that lies in an ε–neighbourhood of order r of the curve K lies also in its ε–neighbourhood of order zero.

23.2. Extrema of Functionals of the Form $\int_a^b F(x, y, y')\,dx$

Let $F(x, u, v)$ be a continuous function of variables x, u and v on an open subset Ω of the three–dimensional Euclidean space (with the Cartesian coordinates x, u and v). The function

$$F(x, y(x), y'(x))$$

is then continuous in $[a, b]$ for any curve K of class T_1 in $[a, b]$ described by (23.1.1) and such that $(x, y(x), y'(x)) \in \Omega$. Hence the integral

$$I(K) = \int_a^b F(x, y(x), y'(x))\,dx \tag{1}$$

represents a certain number (depending on the choice of the curve K from the family of curves considered) for each such curve K. Thus,

$$I(K) = \int_a^b F(x, y, y')\,dx \tag{2}$$

is a certain functional (see Remark 22.5.3); its *domain of definition* is the family L of all curves of class T_1 in $[a, b]$ described by (23.1.1) and possessing the property $(x, y(x), y'(x)) \in \Omega$ for all $x \in [a, b]$.

The extrema of the functional (2) are defined similarly to those of a function:

Definition 1. We say that the functional (2) assumes its *absolute minimum* or *absolute maximum* on its domain of definition L along the curve

$$K_0 = \{(x, y) \in E_2 \,|\, y = y_0(x),\ x \in [a, b]\} \in L \tag{3}$$

if

$$I(K) \geqq I(K_0), \quad \text{or} \quad I(K) \leqq I(K_0) \tag{4}$$

holds, respectively, for all curves K from the family L.

Definition 2. If there exists such an ε–neighbourhood of order zero of the curve K_0 that (4) holds for all curves $K \in L$ belonging to this neighbourhood, we say that the functional (2) assumes its *strong relative minimum* or *strong relative maximum* on L, respectively, along the curve K_0.

Definition 3. If there exists such an ε–neighbourhood of the first order of the curve K_0 that (4) holds for all curves $K \in L$ from this neighbourhood, we say that the functional (2) assumes its *weak relative minimum* or *weak relative maximum* on L, respectively, along the curve K_0.

Obviously the following theorem holds:

Theorem 1. *An absolute extremum is a strong and a weak relative extremum as well.*

(Obviously, the converse statement does not hold.)

A similar statement holds for the strong and the weak relative extrema.

REMARK 1. The previous facts implie the following corollary: *In order to find necessary conditions for an extremum of the functional* (2), *it is sufficient to find necessary conditions for its weak relative extremum.*

REMARK 2. Sufficient conditions for the existence of an extremum of a functional are, even in the considered simplest case, rather complicated and we are not going to state them here as well as in the following paragraphs (cf. references given at the beginning of this chapter). In mechanical and physical problems, that belong by their character to the calculus of variations, the existence of the extremum often follows from the formulation of the problem itself.

23.3. Variation of a Function and Variation of the Functional I

Let us consider the functional (23.2.2) on its domain of definition L. Suppose that the function F, regarded as a function of the variables x, u and v, has continuous partial derivatives up to the second order in Ω. (For the notation see the preceding paragraph.)

Definition 1. Let the curve K_0 given by the function $y_0(x)$ from (23.2.3) be a fixed curve of class T_1 in $[a, b]$ and K be an arbitrary curve of class T_1 in $[a, b]$ given by a function $y(x)$ from (23.1.1). The difference

$$\delta y_0(x) = y(x) - y_0(x) \quad (x \in [a, b]) \tag{1}$$

is called the *variation of the function* $y_0(x)$ *in* $[a, b]$.

The variation $\delta y_0(x)$ of a function is thus also a function that apparently depends on the choice of the curves K_0 and K.

Let us put

$$I(K) = \int_a^b F(x, y(x), y'(x))\,\mathrm{d}x\,, \tag{2}$$

$$I(K_0) = \int_a^b F(x, y_0(x), y_0'(x))\,\mathrm{d}x\,, \tag{2'}$$

where K and K_0 are assumed to be curves from the family L. From (1), (2) and (2') we can express their difference in the form

$$I(K) - I(K_0) = \int_a^b \{F(x, y_0 + \delta y_0, y_0' + \delta y_0') - F(x, y_0, y_0')\}\,\mathrm{d}x\,, \tag{3}$$

where $y_0 = y_0(x)$, $\delta y_0 = y(x) - y_0(x)$ and $\delta y_0' = y'(x) - y_0'(x)$. Confining ourselves to a sufficiently small ε–neighbourhood of the first order of the curve K_0, that belongs to the family L, and applying the mean value theorem to the function $F(x, y, y')$ (regarded as a function of three variables), we can rewrite the right–hand side of (3) as

$$\begin{aligned} &\int_a^b \{F(x, y_0 + \delta y_0, y_0' + \delta y_0') - F(x, y_0, y_0')\}\,\mathrm{d}x = \\ &= \int_a^b \left\{ \frac{\partial F}{\partial y}(x, y_0, y_0')\delta y_0 + \frac{\partial F}{\partial y'}(x, y_0, y_0')\delta y_0' \right\} \mathrm{d}x + d_1(K, K_0)\eta\,, \end{aligned} \tag{4}$$

where $d_1(K, K_0)$ is the distance of the first order between the curves K and K_0, and η is a function depending in general on the choice of the curves K and K_0 and

tending to zero as $d_1 \to 0$. Hence the expression

$$\delta I(K_0) = \int_a^b \left\{ \frac{\partial F}{\partial y}(x, y_0, y_0')\delta y_0 + \frac{\partial F}{\partial y'}(x, y_0, y_0')\delta y_0' \right\} \mathrm{d}x \tag{5}$$

is equal to the increment $I(K) - I(K_0)$ of the functional I if we neglect terms of order higher than $d_1(K, K_0)$.

Definition 2. The expression $\delta I(K_0)$ defined in (5) is called the *variation of the functional* (23.2.2) *along the curve* K_0.

Hence $\delta I(K_0)$ is also a functional depending on $y_0(x)$ and $\delta y_0(x)$. It is the principal part of the increment $I(K) - I(K_0)$ of the functional I, linear in δy_0 and $\delta y_0'$.

REMARK 1. Besides the given definition of the variation of the functional I, we can introduce another definition which, in the considered case of the functional (23.2.2), is equivalent. Let K_0 from (23.2.3) and K from (23.1.1) be fixed curves from the family L. Let δy_0 and $\delta y_0'$ preserve their meaning defined above and let I be the functional (23.2.2). Define a function Φ of one variable t as

$$\Phi(t) = \int_a^b F(x, y_0 + t\delta y_0, y_0' + t\delta y_0')\,\mathrm{d}x\,. \tag{6}$$

From (4) we can express the increment

$$\Phi(t) - \Phi(0) = t\int_a^b \left\{ \frac{\partial F}{\partial y}(x, y_0, y_0')\delta y_0 + \frac{\partial F}{\partial y}(x, y_0, y_0')\delta y_0' \right\} \mathrm{d}x + d_1(K_t, K_0)\eta_1\,,$$

where

$$K_t = \{(x, y) \in E_2 \mid y = y_0(x) + t\delta y_0(x),\ x \in [a, b]\}\,,$$

$d_1(K_t, K_0)/t$ is a bounded function, and $\eta_1 \to 0$ as $t \to 0$. This formula and the formulae (5), (6) imply

$$\lim_{t\to 0} \frac{\Phi(t) - \Phi(0)}{t} = \frac{\mathrm{d}\Phi}{\mathrm{d}t}(0) = \int_a^b \left\{ \frac{\partial F}{\partial y}(x, y_0, y_0')\delta y_0 + \frac{\partial F}{\partial y}(x, y_0, y_0')\delta y_0' \right\} \mathrm{d}x\,.$$

Hence, the variation $\delta I(K_0)$ of the functional I along the curve K_0 can also be defined as the derivative of the function $\Phi(t)$ with respect to t at the point $t = 0$.

REMARK 2. The reader, who is familiar with the elements of functional analysis knows that the variation defined in this way is the Gâteaux differential of the functional I and that a necessary condition for the extremum of this functional is that its Gâteaux differential vanishes. This condition is properly formulated for our aims in Theorem 1 of the following paragraph.

REMARK 3. The variation of the functional I from (23.2.2) along the curve $K = \{(x, y) \in E_2 \mid y = y(x), x \in [a, b]\}$, where $K \in L$, is usually written as

$$\delta I(K) = \int_a^b \{F'_y(x, y, y')\delta y + F'_{y'}(x, y, y')\delta y'\}\, dx \tag{7}$$

with the notation F'_y and $F'_{y'}$ used for $\partial F/\partial y$ and $\partial F/\partial y'$, respectively.

Among all the curves described by the equation $y = \overline{y}(x)$ $(x \in [a, b])$, different from the curve K chosen and belonging to the family L, let us now consider only

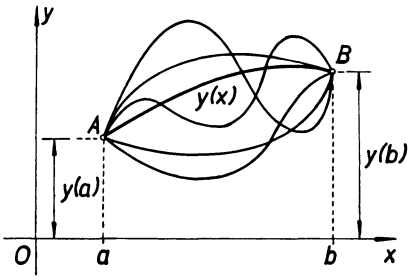


Fig. 23.3.

those curves for which $\overline{y}(a) = y(a)$ and $\overline{y}(b) = y(b)$ (Fig. 23.3). Then we have

$$\delta y(a) = \overline{y}(a) - y(a) = 0\,, \quad \delta y(b) = \overline{y}(b) - y(b) = 0\,. \tag{8}$$

For the curves from the domain L of definition of the functional I, that satisfy the boundary conditions (8), the variation of the functional can be rewritten (after integrating in (7) by parts) in the form

$$\delta I(K) = \int_a^b \left\{F'_y(x, y, y') - \frac{\mathrm{d}}{\mathrm{d}x}F'_{y'}(x, y, y')\right\} \delta y\, \mathrm{d}x\,, \tag{9}$$

which is called the *Lagrange form of the variation of the functional* (23.2.2), or in the form

$$\delta I(K) = \int_a^b \{F'_{y'}(x, y, y') - N(x)\}\, \delta y'\, \mathrm{d}x\,, \tag{9'}$$

where

$$N(x) = \int_a^x F'_y(x, y, y')\, \mathrm{d}x$$

which is called the *Du Bois–Reymond form of the variation* (7).

23.4. Necessary Condition for an Extremum of the Functional I

Let us again consider the functional

$$I = \int_a^b F(x, y, y')\,\mathrm{d}x$$

on its domain of definition L (§ 23.2).

Theorem 1. *Let the function $F(x, u, v)$ possess continuous partial derivatives up to the second order in Ω. If the functional I assumes its extremum* (absolute, strong relative, or weak relative) *on the set of all curves $K \in L$ with the property*

$$y(a) = y_0(a), \quad y(b) = y_0(b)$$

just along the curve $K_0 \in L$ with the description

$$K_0 = \{(x, y) \in E_2 \,|\, y = y_0(x),\ x \in [a, b]\}\ ,$$

then the function $y_0(x)$ satisfies the differential equation

$$F'_y - \frac{\mathrm{d}}{\mathrm{d}x} F'_{y'} = 0\,. \tag{1}$$

Further, the following assertions hold: At all the points of the curve K_0 for which $F''_{y'y'} \neq 0$, there exists a continuous second derivative $y''_0(x)$. If the curve K_0 gives a minimum to the functional I, then $F''_{y'y'} \geqq 0$ holds along this curve; if it gives a maximum, then we have $F''_{y'y'} \leqq 0$ along it.

Equation (1) is called the *Euler differential equation corresponding to the given variational problem* and every its solution $y(x)$ on the interval $[a, b]$ is called the *extremal* of the variational problem given; the Euler equation (1) is a second order differential equation and its solution contains two constants of integration to be determined from the given boundary conditions.

23.5. Special Cases of the Euler Equation. The Brachistochrone Problem

1. If the functional given has the form

$$I = \int_a^b F(x, y')\,\mathrm{d}x\,,$$

then $F'_y = 0$ and the Euler equation (23.4.1) reduces to

$$\frac{\mathrm{d}}{\mathrm{d}x} F'_{y'} = 0 .$$

This implies

$$F'_{y'} = C \quad (C = \text{const.}) ,$$

which is the first integral of the Euler equation.

2. If the functional given is of the form

$$I = \int_a^b F(y, y')\,\mathrm{d}x ,$$

if the curve described by the equation $y = y(x)$ $(x \in [a, b])$ is the extremal of the corresponding variational problem, and if $F''_{y'y'} \neq 0$ along this curve, then, by Theorem 23.4.1,

$$\frac{\mathrm{d}}{\mathrm{d}x}(F(y, y') - y'F'_{y'}(y, y')) =$$

$$= F'_y y' + F'_{y'} y'' - y'' F'_{y'} - y' \frac{\mathrm{d}}{\mathrm{d}x} F'_{y'} = y'(F'_y - \frac{\mathrm{d}}{\mathrm{d}x} F'_{y'}) = 0$$

along this curve and thus

$$F(y, y') - y'F'_{y'}(y, y') = C \quad (C = \text{ const.}) \tag{1}$$

is the first integral of the corresponding Euler equation.

Example 1 (The *Brachistochrone Problem*). From among all smooth curves joining the points $A(x_1, y_1)$ and $B(x_2, y_2)$, $y_1 > y_2$, $x_1 < x_2$ (Fig. 23.4), let us find the curve along which a particle moving from the rest at the point A to the point B in the earth gravity field (described by the vector $\boldsymbol{P}$ with components $P_x = 0$ and $P_y = -mg$, where g is the acceleration of gravity) reaches the point B in the shortest time. (Practically: We are to find the shape of such a groove joining the points A and B that a stone sliding from A along the groove under the influence of gravity (friction being neglected) reaches the point B in the shortest possible time.)

Since the force with which the constraint (groove) acts on the particle (stone) is perpendicular to the direction of motion, it does no work. The energy conservation principle thus implies

$$\tfrac{1}{2}mv^2 = mg(y_1 - y) ,$$

i.e.

$$v^2 = 2g(y_1 - y) . \tag{2}$$

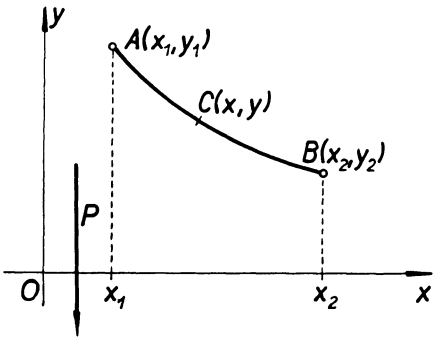


Fig. 23.4.

At the same time we have

$$v = \frac{\mathrm{d}s}{\mathrm{d}t} = \sqrt{(1 + y'^2)}\,\frac{\mathrm{d}x}{\mathrm{d}t} \tag{2'}$$

along the trajectory of the particle, where the functions $x = x(t)$, $y = y(x(t))$ represent the time course of the motion along a smooth curve described by the equation $y = y(x)$ and joining the points A and B. In (2'), s denotes the arc of this curve, $y' = \mathrm{d}y/\mathrm{d}x$, and t denotes time. According to (2) we have $y \leqq y_1$ for $x \in [x_1, x_2]$. If we had $y(x) = y_1$ at some point $x \in (x_1, x_2)$, then the scalar speed v would be zero at this point due to (2) and further motion would require a certain impulse, what is impossible because of the physical interpretation of the problem. We thus have $y(x) < y_1$ for $x \in (x_1, x_2]$ and, consequently, $v > 0$. This fact and the equations (2) and (2') imply

$$\mathrm{d}t = \frac{\mathrm{d}s}{v} = \sqrt{\left(\frac{1 + y'^2}{2g(y_1 - y)}\right)}\,\mathrm{d}x, \quad x \in (x_1, x_2].$$

From this, we obtain the formula

$$T = \int_{x_1}^{x_2} \sqrt{\left(\frac{1 + y'^2}{2g(y_1 - y)}\right)}\,\mathrm{d}x \tag{3}$$

for the total time T necessary to traverse the curve. The physical problem given is thus reduced to the following variational problem: Find, among all curves of class T_1 with end points A and B, the curve that gives the minimum value to the functional (3). The functional T in (3) is a functional of type 2 but the function

$$F(y, y') = \sqrt{\left(\frac{1 + y'^2}{2g(y_1 - y)}\right)}$$

does not satisfy the assumptions stated in the beginning of § 23.3 (it is even undefined for $y \geqq y_1$). However, the function F considered has continuous second order partial derivatives in the domain $x \in (x_1, \infty)$, $y < y_1$ and $y' \in (-\infty, \infty)$. Then a simple calculation yields $F''_{y'y'} > 0$ in this domain.

The physical nature of the problem indicates that among all the smooth curves joining the points A and B and described by the equation $y = y(x)$, where $y(x) < y_1$ for $x \in (x_1, x_2]$, there exists a curve possessing the property required. We will find a necessary condition for the existence of this extremal and obtain the desired result from it. Choose an arbitrary positive number ε such that it is less than $x_2 - x_1$. Then the arc of this extremal curve with end points $A'(x_1 + \varepsilon, y(x_1 + \varepsilon))$ and $B(x_2, y_2)$ apparently gives the minimum to the functional

$$\int_{x_1+\varepsilon}^{x_2} \sqrt{\left(\frac{1+y'^2}{2g(y_1-y)}\right)}\, \mathrm{d}x \tag{3'}$$

on the family of all curves of class T_1 described by the equation $y = y(x)$, with end points A' and B, and possessing the property $y(x) < y_1$ for $x \in [x_1 + \varepsilon, x_2]$. The case of the "modified" functional (3') thus can be included in the special case 2 discussed above. From (1) and (3'), we obtain the first integral of the Euler equation in the form

$$\sqrt{\left(\frac{1+y'^2}{2g(y_1-y)}\right)} - y' \frac{y'}{\sqrt{[2g(y_1-y)]}} \frac{1}{\sqrt{(1+y'^2)}} = C$$

or, after rearrangement,

$$1 = C\sqrt{[2g(y_1-y)]}\,\sqrt{(1+y'^2)}\ ,$$

i.e.

$$1 = 2C^2 g(y_1-y)(1+y'^2)\,, \quad C \neq 0\,.$$

Putting $K = 1/(2gC^2)$, we can write

$$\frac{K}{1+y'^2} = y_1 - y\,.$$

By the parameter method (completely analogous to the method of Remark 17.5.3 and Example 17.5.2) we get (putting $y' = \tan\varphi$, so that $y = y_1 - K\cos^2\varphi$) the parametric equations of the curve sought in the form

$$x = \frac{K}{2}(\sin 2\varphi + 2\varphi) + \tilde{C}\,, \quad y = y_1 - \frac{K}{2}(1 + \cos 2\varphi)\,.$$

These are parametric equations of the extremal of the "modified" variational problem and thus also the equations of the curve to be found. The constants K and $\tilde{C}$ are determined by the condition that the extremal curve passes through the points $A(x_1, y_1)$ and $B(x_2, y_2)$. An easy analysis leads to the conclusion that the curve found is an arc of a cycloid.

B. PROBLEMS OF THE SECOND CATEGORY (EXTREMA OF FUNCTIONALS OF THE FORM $\int_a^b F(x, y_1, \ldots, y_n, y_1', \ldots, y_n')\,\mathrm{d}x$)

23.6. Some Concepts and Definitions

Definition 1. A curve K in the Euclidean space E_{n+1} (with the Cartesian coordinates $x, y_1, \ldots, y_n$) with the description

$$K = \{(x, y_1, \ldots, y_n) \in E_{n+1} \mid y_i = y_i(x)\ (i = 1, \ldots, n),\ x \in [a, b]\} \qquad (1)$$

is called the *curve of the r-th class* (or of *the class T_r*) *in the interval* $[a, b]$ if the functions $y_i(x)$ $(i = 1, \ldots, n)$ have continuous derivatives of order r in an open interval $(\bar{a}, \bar{b})$ containing the closed interval $[a, b]$ (the functions $y_i(x)$ thus also have continuous derivatives of all lower orders, including zero).

Definition 2. Let K be a curve described by (1) and let

$$K_0 = \{(x, y_1, \ldots, y_n) \in E_{n+1} \mid y_i = y_{0i}(x)\ (i = 1, \ldots, n),\ x \in [a, b]\}\ . \qquad (2)$$

If both K and K_0 are curves of class T_r in E_{n+1}, then the largest of the values

$$|y_i(x) - y_{0i}(x)|,\ |y_i'(x) - y_{0i}'(x)|, \ldots, |y_i^{(k)}(x) - y_{0i}^{(k)}(x)|$$

for $i = 1, \ldots, n$ and $x \in [a, b]$ is called the *distance $d_k(K, K_0)$ of order k between these two curves* $(k \in \{0, \ldots, r\})$.

The concept of the *ε-neighbourhood* is introduced similarly to the case of a plane curve (§ 23.1). If the curve K_0 from (2) belongs to the class T_r, then the family $U_k(K_0; \varepsilon)$ (where $k \in \{0, \ldots, r\}$) of all curves K described by (1) and satisfying the inequality $d_k(K, K_0) < \varepsilon$ $(\varepsilon > 0)$ is called the *ε-neighbourhood of order k of the curve K_0*.

23.7. Formulation of the Variational Problem

Let a function $F(x, u_1, \ldots, u_n, v_1, \ldots, v_n)$ of $2n + 1$ variables (representing Cartesian coordinates of a point in the $(2n+1)$–dimensional Euclidean space E_{2n+1}) be given. Let F have continuous partial derivatives up to the second order in an open set Ω of the space E_{2n+1}. We denote by L the family of all curves (23.6.1) of class T_1 in $[a, b]$ that have common end points A and B (on each curve, the point A corresponds to the value $x = a$ and the point B to the value $x = b$) and for which

$$(x, y_1(x), \ldots, y_n(x), y_1'(x), \ldots, y_n'(x)) \in \Omega$$

for $x \in [a, b]$.

Our task is as follows:

Find such a curve K_0 from the family L which makes the functional

$$I = \int_a^b F(x, y_1, \ldots, y_n, y_1', \ldots, y_n')\,\mathrm{d}x \tag{1}$$

have an extremal value on the family L.

The family L of the curves considered is called the *domain of definition of the functional* (1).

REMARK 1. The concepts of *absolute extremum, strong relative extremum* and *weak relative extremum* of the functional (1) are defined in a manner similar to that in § 23.2. In what follows we will give only *necessary* conditions that are satisfied by any curve of the family L which gives extremum to the functional (1).

23.8. Necessary Conditions for an Extremum of the Functional I

Theorem 1. *Let K be a curve of class T_1, that is described by* (23.6.1), *has end points $A(a, y_1(a), \ldots, y_n(a))$ and $B(b, y_1(b), \ldots, y_n(b))$, and belongs to the domain of definition L of the functional* (23.7.1) *considered above. Further, let the function F satisfy the assumptions of* § 23.7. *If the functional* (23.7.1) *assumes its extremum* (absolute, strong relative, or weak relative) *on L along this curve, then the functions $y_i(x)$ satisfy the system of differential equations*

$$\frac{\partial F}{\partial y_i} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial F}{\partial y_i'} = 0 \quad (i = 1, \ldots, n). \tag{1}$$

REMARK 1. The system (1) is called the *system of Euler equations corresponding to the variational problem given*; each curve (23.6.1) which fulfills (1) is called the *extremal* of this problem. The general solution of the system of differential equations (1) (if it exists) depends on $2n$ constants of integration $c_1, \ldots, c_{2n}$ and, consequently, it has the form

$$y_i = y_i(x, c_1, \ldots, c_{2n}) \quad (i = 1, \ldots, n).$$

The constants $c_1, \ldots, c_{2n}$ are determined from the boundary conditions, i.e. from the requirement that the extremal passes through the points A and B.

Theorem 2. *Let the assumptions of Theorem* 1 *be satisfied. If a curve K belonging to the family L yields a minimum of the functional* (23.7.1) *on the family L of the curves considered, then the inequalities* (*the Legendre conditions*)

$$F''_{y'_1 y'_1} \geqq 0, \begin{vmatrix} F''_{y'_1 y'_1}, & F''_{y'_1 y'_2} \\ F''_{y'_2 y'_1}, & F''_{y'_2 y'_2} \end{vmatrix} \geqq 0, \ldots, \begin{vmatrix} F''_{y'_1 y'_1}, & F''_{y'_1 y'_2}, & \ldots, & F''_{y'_1 y'_n} \\ F''_{y'_2 y'_1}, & F''_{y'_2 y'_2}, & \ldots, & F''_{y'_2 y'_n} \\ \ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots & \ldots\ldots & \ldots\ldots\ldots\ldots \\ F''_{y'_n y'_1}, & F''_{y'_n y'_2}, & \ldots, & F''_{y'_n y'_n} \end{vmatrix} \geqq 0 \qquad (2)$$

hold at every point of the extremal K. In case of a maximum of the considered functional along the extremal K, the inequalities (2) *hold at every its point except that the inequality symbols are alternately* $\leqq, \geqq, \leqq, \geqq, \ldots$.

Example 1. Let us find the extremal of the functional

$$\int_0^{\pi/2} (y'^2 + z'^2 + 2yz)\, dx$$

with the boundary conditions

$$y(0) = 0, \ z(0) = 0, \ y\left(\frac{\pi}{2}\right) = 1, \ z\left(\frac{\pi}{2}\right) = -1.$$

In this particular case, the system of Euler equations (1) reduces to two differential equations

$$y'' - z = 0, \quad z'' - y = 0. \qquad (3)$$

Differentiating the first of these equations twice with respect to x and using the second one, we get the differential equation

$$y^{(4)} - y = 0$$

whose general solution is

$$y = c_1 e^x + c_2 e^{-x} + c_3 \cos x + c_4 \sin x.$$

This and the first of equations (3) yield

$$z = c_1 e^x + c_2 e^{-x} - c_3 \cos x - c_4 \sin x .$$

From the boundary conditions given we calculate that $c_1 = c_2 = c_3 = 0$, $c_4 = 1$. The extremal sought is thus the curve described by the equations

$$y = \sin x, \quad z = -\sin x \quad (x \in [0, \tfrac{1}{2}\pi]) .$$

C. PROBLEMS OF THE THIRD CATEGORY (EXTREMA OF FUNCTIONALS OF THE FORM $\int_a^b F(x, y, y', .., , y^{(n)})\,dx$)

23.9. Formulation of the Problem

Many problems of physics and technology lead to the mathematical problem of finding the extrema of functionals of the form

$$I = \int_a^b F(x, y, y', \ldots, y^{(n)})\,dx , \tag{1}$$

where the integrand depends not only on the derivative $y'(x)$ but also on the derivatives of higher orders of the function $y(x)$. The function $y(x)$ appears in the description of the curve

$$K = \{(x, y) \in E_2 \,|\, y = y(x),\ x \in [a, b]\} \tag{2}$$

from the domain of definition of the functional (1). To introduce the domain of definition L more precisely, we suppose that the given function $F(x, u_0, u_1, \ldots, u_n)$ of $n + 2$ independent variables is continuous on an open set Ω of the $(n + 2)$-dimensional Euclidean space (with Cartesian coordinates $x, u_0, u_1, \ldots, u_n$). Let

$$K_0 = \{(x, y) \in E_2 \,|\, y = y_0(x),\ x \in [a, b]\} \tag{2'}$$

be a curve of class T_n and have the property

$$(x, y_0(x), y_0'(x), \ldots, y_0^{(n)}(x)) \in \Omega \quad \text{for } x \in [a, b] .$$

The domain L of definition of the functional (1) is then introduced as the family of all curves of class T_n that are described by (2), for which

$$(x, y(x), y'(x), \dots, y^{(n)}(x)) \in \Omega \text{ for } x \in [a, b],$$

and that satisfy the boundary conditions

$$\begin{aligned} y(a) = y_0(a), y'(a) = y_0'(a), \dots, y^{(n-1)}(a) = y_0^{(n-1)}(a), \\ y(b) = y_0(b), y'(b) = y_0'(b), \dots, y^{(n-1)}(b) = y_0^{(n-1)}(b), \end{aligned} \tag{3}$$

where $y_0(a), \dots, y_0^{(n-1)}(b)$ are given numbers.

The corresponding variational problem is to find such a curve from L that yields an extremum of the functional (1).

23.10. A Necessary Condition for the Extremum of the Functional (23.9.1)

Under the assumption that the function $F(x, y, y', \dots, y^{(n)})$ in (23.9.1) has continuous partial derivatives up to order $n + 2$ with respect to its arguments $x, y, y', \dots, y^{(n)}$ in the above considered domain Ω, the following theorem holds:

Theorem 1. *If the functional* (23.9.1) *assumes its extremum on the family* L *along the curve* $K_0 \in L$ *described by* (23.9.2'), *then the equation*

$$F_y' - \frac{\mathrm{d}}{\mathrm{d}x} F_{y'}' + \frac{\mathrm{d}^2}{\mathrm{d}x^2} F_{y''}' - \dots + (-1)^n \frac{\mathrm{d}^n}{\mathrm{d}x^n} F_{y^{(n)}}' = 0 \tag{1}$$

is satisfied at all points of the curve K_0.

REMARK 1. The differential equation (1) is called the *Euler–Poisson equation*. In the general case (if $F''_{y^{(n)}y^{(n)}} \neq 0$), it is a differential equation of order $2n$ and its solution (if it exists) involves $2n$ constants of integration. The general solution of the equation (1) has thus the form

$$y = f(x, \alpha_1, \beta_1, \alpha_2, \beta_2, \dots \alpha_n, \beta_n).$$

The constants α_i and β_i $(i = 1, \dots, n)$ can be determined from the $2n$ conditions (23.9.3) at the end points of the integral curve of equation (1), which is called the *extremal* of the variational problem given.

Example 1. The following problem is studied in the theory of elasticity: A cylindrical solid beam, elastic and homogeneous, has its ends fixed at the same height above the ground (Fig.23.5). The profile of the deflected axis of the beam is to be determined.

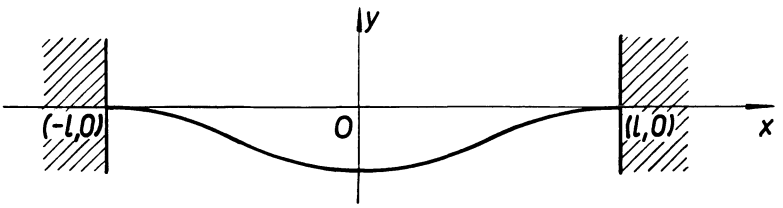


Fig. 23.5.

The formula

$$E = \int_{-l}^{l} \left[\frac{1}{2} \mu \frac{y''^2}{1 + y'^2} + \varrho y \sqrt{(1 + y'^2)} \right] \mathrm{d}x ,$$

where μ and ϱ are known positive constants, holds for the total potential energy E of the beam. If we assume the deflection of the beam to be very small as compared with its length, we can put $1 + y'^2 \approx 1$ and the formula for the potential energy becomes

$$E = \int_{-l}^{l} \left(\frac{1}{2} \mu y''^2 + \varrho y \right) \mathrm{d}x . \tag{2}$$

A well known physical principle now states: If a mechanical system is in stable equilibrium, then its total potential energy is minimal. Hence our physical problem leads to the variational problem of finding the minimum of the functional (2) on the family L of all curves of class T_2 with the description

$$K = \{(x, y) \in E_2 \,|\, y = y(x),\ x \in [-l, l]\}$$

and with the boundary conditions (characterizing fixed ends)

$$y(-l) = y(l) = 0 , \quad y'(-l) = y'(l) = 0 . \tag{3}$$

In case of the functional (2), the Euler–Poisson equation (1) reduces to the equation

$$\varrho + \frac{\mathrm{d}^2}{\mathrm{d}x^2} (\mu y'') = 0 ,$$

i.e.

$$\varrho + \mu y^{(4)} = 0 . \tag{4}$$

The general solution of the differential equation (4) depends on four constants of integration α, β, γ and δ, and has the form

$$y = \alpha x^3 + \beta x^2 + \gamma x + \delta - \frac{\varrho}{24\mu} x^4 .$$

The boundary conditions (3) then imply

$$y = \frac{\varrho}{24\mu}(-x^4 + 2l^2 x^2 - l^4) .$$

23.11. Generalization to the Case of an Arbitrary Finite Number of Functions Sought

A more general problem analogous to that formulated in § 23.9 consists in finding the extremum of the functional

$$I = \int_a^b F\left(x, y_1, y_1', \ldots, y_1^{(n_1)}, y_2, y_2', \ldots, y_2^{(n_2)}, \ldots, y_m, y_m', \ldots, y_m^{(n_m)}\right) \mathrm{d}x$$

whose domain of definition L is the family of curves possessing the following properties:

(i) Each curve $K \in L$ is described by

$$K = \{(x, y_1, \ldots, y_m) \in E_{m+1} \mid y_i = y_i(x) \quad (i = 1, \ldots, m),\ x \in [a, b]\} ,$$

where the functions $y_i(x)$ have continuous derivatives $y_i^{(n_i)}(x)$ $(i = 1, \ldots, m)$ in some open interval containing the interval $[a, b]$.

(ii) For all $x \in [a, b]$,

$$\left(x, y_1(x), y_1'(x), \ldots, y_1^{(n_1)}(x), y_2(x), y_2'(x), \ldots, y_2^{(n_2)}(x), \ldots \right.$$
$$\left. \ldots, y_m(x), y_m'(x), \ldots, y_m^{(n_m)}(x)\right) \in \Omega ,$$

where Ω is an open set in the space of the arguments of the function F. Moreover, we suppose that the function F has continuous partial derivatives of the required (sufficiently high) order in Ω.

In the considered more general case we arrive at the following *Euler–Poisson system of equations*:

$$\sum_{k=0}^{n_i} (-1)^k \frac{\mathrm{d}^k}{\mathrm{d}x^k} F'_{y_i^{(k)}} = 0 \quad (i = 1, \ldots, m) ,$$

which represents necessary conditions for the curve yielding the extremum of the given functional under the boundary conditions prescribed.

D. PROBLEMS OF THE FOURTH CATEGORY (FUNCTIONALS DEPENDING ON A FUNCTION OF n VARIABLES)

23.12. Some Concepts and Definitions

Let us consider a bounded region G with the boundary S in the n–dimensional Euclidean space E_n (with coordinates $x_1, x_2, \ldots, x_n$).

Definition 1. A function $\varphi(x_1, \ldots, x_n)$ is called the *function of the class* T_r *in* $\overline{G} = G \cup S$ if it has continuous partial derivatives up to the order r in a region $\widetilde{G}$ containing $\overline{G}$.

Definition 2. Let $\varphi(x_1, \ldots, x_n)$ be of class T_1 in $\overline{G}$. Let E_{n+1} be the $(n+1)$--dimensional Euclidean space with the Cartesian coordinates $x_1, \ldots, x_n, u$. The set

$$N = \{(x_1, \ldots, x_n, u) \in E_{n+1} \,|\, u = \varphi(x_1, \ldots, x_n),\ (x_1, \ldots, x_n) \in \overline{G}\} \qquad (1)$$

is then called an *explicitly described regular hypersurface in* E_{n+1} *on the considered closed region* $\overline{G}$.

Definition 3. Let N_1 and N_2 be two regular hypersurfaces in E_{n+1} with the descriptions

$$N_1 = \{(x_1, \ldots, x_n, u) \in E_{n+1} \,|\, u = \varphi(x_1, \ldots, x_n),\ (x_1, \ldots, x_n) \in \overline{G}\}\ ,$$
$$N_2 = \{(x_1, \ldots, x_n, u) \in E_{n+1} \,|\, u = \psi(x_1, \ldots, x_n),\ (x_1, \ldots, x_n) \in \overline{G}\}\ .$$

The number

$$d(N_1, N_2) = \max_{(x_1, \ldots, x_n) \in \overline{G}} |\varphi(x_1, \ldots, x_n) - \psi(x_1, \ldots, x_n)|\,,$$

is called the *distance* (*of order zero*) *of the hypersurfaces* N_1 *and* N_2, and the number

$$d_1(N_1, N_2) = \max\Big(\max_{(x_1, \ldots, x_n) \in \overline{G}} |\varphi - \psi|,\ \max_{(x_1, \ldots, x_n) \in \overline{G}} |\varphi_1 - \psi_1|, \ldots$$
$$\ldots,\ \max_{(x_1, \ldots, x_n) \in \overline{G}} |\varphi_n - \psi_n|\Big)\,,$$

where

$$\varphi_i = \frac{\partial \varphi}{\partial x_i}, \quad \psi_i = \frac{\partial \psi}{\partial x_i} \quad (i = 1, \ldots, n),$$

is called the *distance of the first order of the hypersurfaces* N_1 *and* N_2. The family of all regular hypersurfaces described by

$$\tilde{N} = \{(x_1, \ldots, x_n, u) \in E_{n+1} \mid u = \tilde{\varphi}(x_1, \ldots, x_n), \ (x_1, \ldots, x_n) \in \overline{G}\} \tag{2}$$

whose distance of order zero from the regular hypersurface (1) is less than ε ($\varepsilon > 0$) is called the *ε-neighbourhood of the hypersurface N in $\overline{G}$*. Similarly, the family of all regular hypersurfaces $\tilde{N}$ in $\overline{G}$, for which $d_1(\tilde{N}, N) < \varepsilon$, is called the *$\varepsilon$-neighbourhood of the first order of the regular hypersurface* (1) *in* $\overline{G}$.

Let the function $F(x_i, \varphi, \varphi_i) = F\left(x_1, \ldots, x_n, \varphi, \frac{\partial \varphi}{\partial x_1}, \ldots, \frac{\partial \varphi}{\partial x_n}\right)$ have continuous partial derivatives with respect to its $2n + 1$ arguments $x_1, \ldots, x_n, \varphi, \varphi_1, \ldots$ $\ldots, \varphi_n$ up to the third order on an open subset Ω of the $(2n + 1)$-dimensional space. Let us denote by L the family of all regular hypersurfaces that are described by (1) and for which

$$(x_1, \ldots, x_n, \varphi(x_1, \ldots, x_n), \varphi_1(x_1, \ldots, x_n), \ldots, \varphi_n(x_1, \ldots, x_n)) \in \Omega$$

for $(x_1, \ldots, x_n) \in \overline{G}$. Further let us put

$$I = \int_G \cdots \int F(x_i, \varphi, \varphi_i) \, \mathrm{d}x_1 \ldots \mathrm{d}x_n . \tag{3}$$

Definition 4. We say that the functional (3) assumes its *absolute minimum* or *absolute maximum on the family L along the regular hypersurface N* described by (1) and belonging to L if

$$I(\tilde{N}) \geqq I(N), \quad \text{or} \quad I(\tilde{N}) \leqq I(N) \tag{4}$$

holds, respectively, for any hypersurface $\tilde{N} \in L$.

If there exists such an ε-neighbourhood of order zero of the regular hypersurface N described by (1) and belonging to L that (4) holds for any regular hypersurface $\tilde{N}$ from this neighbourhood, we say that the functional (3) assumes its *strong relative minimum*, or *strong relative maximum on L*, respectively, along the hypersurface N. If (4) holds for an ε-neighbourhood of the first order of the hypersurface N, we speak about a *weak relative minimum*, or *weak relative maximum* of the functional (3), respectively.

Obviously, necessary conditions for a weak relative extremum are also necessary conditions both for a strong relative extremum and an absolute extremum.

23.13. Formulation of the Variational Problem and Necessary Conditions for an Extremum

Under the above stated assumptions, the variational problem is formulated in the following way: In the set of all regular hypersurfaces N that are described by (23.12.1), belong to L and satisfy the boundary condition

$$\varphi(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) \quad \text{for} \quad (x_1, \ldots, x_n) \in S, \tag{1}$$

where f is a given function continuous at the points of the boundary S of the region G, a hypersurface is to be found along which the functional (23.12.3) assumes its extremum.

Necessary conditions for an extremum of the functional (23.12.3) are given in the following theorem:

Theorem 1. *Let the regular hypersurface N, that is described by* (23.12.1), *belongs to L and satisfies the boundary condition* (1), *make the functional* (23.12.3) *have an extremum on the set of all regular hypersurfaces belonging to the family L and satisfying the given boundary condition. Then the partial differential equation*

$$F'_{\varphi} - \sum_{i=1}^{n} \frac{\partial}{\partial x_i} F'_{\varphi_i} = 0, \tag{2}$$

where

$$F'_{\varphi} = \frac{\partial F}{\partial \varphi}, \quad F'_{\varphi_i} = \frac{\partial F}{\partial \varphi_i} \quad (i = 1, \ldots, n),$$

holds at the points of the hypersurface N.

REMARK 1. The partial differential equation (2) is called the *Euler–Ostrogradski equation* and every regular hypersurface described by (23.12.1), with $\varphi(x_1, \ldots, x_n)$ solving equation (2), is called an *extremal n-dimensional manifold* (an *extremal hypersurface*) of the variational problem considered.

Example 1. The integral

$$I = \int_G \cdots \int \left(\sum_{i=1}^{n} \varphi_i^2 \right) \mathrm{d}x_1 \ldots \mathrm{d}x_n$$

is called the *n-dimensional Dirichlet integral.* Let a continuous function $f(x_1, \ldots, x_n)$ be given at the points of the boundary S of the region G. The Euler–Ostrogradski equation (2) holds for an extremum of this functional on the

set of all functions that are of class T_1 in $\overline{G}$ and satisfy the boundary condition $\varphi|_S = f|_S$. Equation (2) reads, in this case,

$$\Delta\varphi = \sum_{i=1}^{n} \frac{\partial^2 \varphi}{\partial x_i^2} = 0 ,$$

which is the Laplace equation.

E. PROBLEMS OF THE FIFTH CATEGORY (VARIATIONAL PROBLEMS WITH "MOVING (FREE) ENDS OF ADMISSIBLE CURVES")

23.14. Formulation of the Simplest Problem

In variational problems belonging to the first three categories, the domains of definition of functionals consisted of curves with common end points. These problems can be generalized in a certain sense. In what follows, we are going to present a certain generalization of a variational problem of the first category.

Let K_1 and K_2 be two curves with the descriptions

$$\begin{aligned} K_1 &= \{(x, y) \in E_2 \,|\, y = \varphi(x),\ x \in [a_1, b_1]\} , \\ K_2 &= \{(x, y) \in E_2 \,|\, y = \psi(x),\ x \in [a_2, b_2]\} \end{aligned} \tag{1}$$

which belong to the class T_1 in their domain of definition and have no common points (i.e. $K_1 \cap K_2 = \emptyset$). Let $F(x, y, y')$ be a function satisfying the assumptions stated in § 23.3 (i.e., F has continuous partial derivatives up to the second order on an open subset Ω of the three-dimensional space of its three arguments). The problem is: From among all the curves of class T_1 with the description

$$K = \{(x, y) \in E_2 \,|\, y = y(x),\ x \in [x_1, x_2]\} , \tag{1'}$$

which have one end point $A(x_1, y_1)$ on the curve K_1 and the other end point $B(x_2, y_2)$ on the curve K_2 (A and B are thus not fixed points, but A is some point on the curve K_1 and B is some point on the curve K_2), we have to find such a curve that makes the functional

$$I = \int_{x_1}^{x_2} F(x, y, y')\,\mathrm{d}x \tag{2}$$

have an extremum (Fig. 23.6).

23.15. Necessary Conditions for an Extremum

Let us denote by L the family of all the curves K of class T_1 with the description (23.14.1′) for which $(x, y(x), y'(x)) \in \Omega$ for $x \in [x_1, x_2]$ and which satisfy the boundary conditions

$$A(x_1, y(x_1)) \in K_1, \quad B(x_2, y(x_2)) \in K_2 .$$

Then the following theorem holds:

Theorem 1. *Let the curve $K \in L$ yield an extremum of the functional* (23.14.2) *on the family L. Then the Euler equation*

$$F'_y - \frac{\mathrm{d}}{\mathrm{d}x} F'_{y'} = 0 \tag{1}$$

holds at points of the curve K and the conditions

$$\{F + (\varphi' - y')F'_{y'}\}_A = 0, \quad \{F + (\psi' - y')F'_{y'}\}_B = 0 \tag{1'}$$

are satisfied at its end points A and B.

REMARK 1. The conditions (1′) are called the *transversality conditions* corresponding to the variational problem given. The symbols $\{\}_A$ and $\{\}_B$ mean that the coordinates x_1, $y_1 = y(x_1)$ of the points A and the coordinates x_2, $y_2 = y(x_2)$ of the points B, respectively, are to be substituted into the expression in braces.

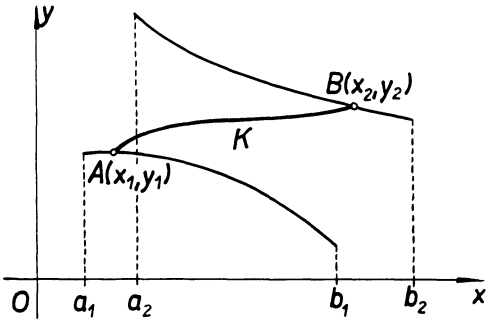


Fig. 23.6.

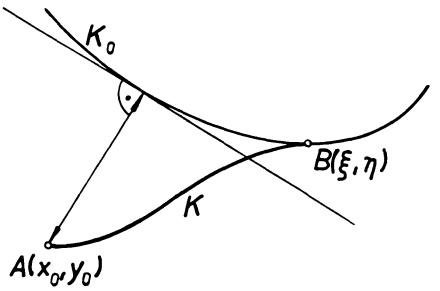


Fig. 23.7.

REMARK 2 (*A More General Form of the Transversality Conditions*). Let the curves K_1 and K_2 be implicitly described by the equations

$$\Phi(x, y) = 0 ,$$
$$\Psi(x, y) = 0 ,$$

where the functions Φ and Ψ have continuous partial derivatives of the first order in a two–dimensional domain containing the curves K_1 and K_2. Further let

$$\left(\frac{\partial \Phi}{\partial x}\right)^2 + \left(\frac{\partial \Phi}{\partial y}\right)^2 > 0, \quad \left(\frac{\partial \Psi}{\partial x}\right)^2 + \left(\frac{\partial \Psi}{\partial y}\right)^2 > 0$$

hold at each point (x, y) of the curves K_1 and K_2. Then the transversality conditions have the form

$$\begin{aligned} \left\{(F - y' F'_{y'})\frac{\partial \Phi}{\partial y}\right\}_A &= \left\{F'_{y'}\frac{\partial \Phi}{\partial x}\right\}_A, \\ \left\{(F - y' F'_{y'})\frac{\partial \Psi}{\partial y}\right\}_B &= \left\{F'_{y'}\frac{\partial \Psi}{\partial x}\right\}_B. \end{aligned} \tag{2}$$

In particular, if the curves K_1 and K_2 are straight lines parallel to the y–axis, then the conditions (2) reduce to

$$\{F'_{y'}\}_A = 0, \quad \{F'_{y'}\}_B = 0. \tag{3}$$

Notice that in case when the considered family of curves consists only of the curves that have one end point common (fixed) while the other end point lies on a given curve, one of the two transversality conditions (1′) or (2) no longer applies.

Example 1. Let K_0 be a plane curve of class T_1 described by the equation $y = \varphi(x)$, $x \in [a, b]$. Our task is to find the smallest distance of a given point $A(x_0, y_0)$ from this curve (assuming that $A \notin K_0$).

We consider all curves K of class T_1 described by the equation $y = y(x)$ that have one end point at the fixed point $A(x_0, y_0)$ and the other point $B(\xi, \eta)$ is a "free point" on the given curve K_0 (Fig. 23.7). The length of the arc of each such curve K is (assuming $\xi > x_0$)

$$I = I(K) = \int_{x_0}^{\xi} \sqrt{(1 + y'^2)}\, \mathrm{d}x. \tag{4}$$

Hence we are to find an extremum of the functional (4) on the family of all curves of class T_1 that are described by the equation $y = y(x)$, have one end point at the fixed point A and the other end point on the given curve K_0. In this case, the Euler equation reduces to the equation

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{y'}{\sqrt{(1 + y'^2)}} = 0,$$

which implies that $y' = c$ ($c =$ const.) and $y = cx + d$ ($d =$ const.). Consequently, the extremal is a straight line. This straight line passes through the point $A(x_0, y_0)$

and the unknown constants c and d thus fulfil the relation

$$y_0 = cx_0 + d. \tag{5}$$

The transversality condition (1′) gives in this case

$$\sqrt{(1+y'^2)} + (\varphi' - y')\frac{y'}{\sqrt{(1+y'^2)}} = 0,$$

i.e.

$$1 + \varphi' y' = 0 \tag{6}$$

at the point B of the extremal lying on the given curve K_0. It is apparent from (6) that the transversality condition is the orthogonality condition in this particular case. At the point B, we have $y' = c$, $\varphi'|_B = \varphi'(\xi)$. Under the assumption $\varphi'(\xi) \neq 0$, (6) implies

$$c = -\frac{1}{\varphi'(\xi)}.$$

The constant d in (5) can be determined from the condition $\varphi(\xi) = c\xi + d$, i.e.

$$d = \varphi(\xi) - c\xi = \varphi(\xi) + \frac{\xi}{\varphi'(\xi)}.$$

Rewriting the condition (5) in the form

$$y_0 = -\frac{1}{\varphi'(\xi)}x_0 + \varphi(\xi) + \frac{\xi}{\varphi'(\xi)} = \frac{1}{\varphi'(\xi)}(\xi - x_0) + \varphi(\xi),$$

we calculate the first coordinate ξ of the point B. Its second coordinate is calculated from the relation $\eta = \varphi(\xi)$.

In a similar way we can determine the distance between two curves in E_2.

REMARK 3. Variational problems with "moving (free) ends of admissible curves" can be easily generalized to the case of functionals of the form (23.7.1), i.e. functionals defined for curves in an Euclidean space of arbitrary dimension. Similar problems with "moving (free) ends of admissible curves" also arise in the case of functionals depending on derivatives of higher orders. The transversality conditions in these two cases are naturally more complicated (see, e.g., [51], [65], [130]).

F. PROBLEMS OF THE SIXTH CATEGORY (THE ISOPERIMETRIC PROBLEM IN THE SIMPLEST CASE)

23.16. Formulation of the Problem

Let $F(x, y, y')$ and $\Phi(x, y, y')$ be two given functions that have continuous partial derivatives up to the second order in an open subset Ω of the three–dimensional space of their three arguments. Let us denote by L the family of all curves K of class T_1 with the description

$$K = \{(x, y) \in E_2 \,|\, y = y(x),\ x \in [a, b]\} ,$$

for which $(x, y(x), y'(x)) \in \Omega$ for $x \in [a, b]$. The problem is: From among all the curves K from the family L, for which the functional

$$G = \int_a^b \Phi(x, y, y')\,\mathrm{d}x \tag{1}$$

assumes the same prescribed value (i.e. $G =$ const.), such a curve is to be found that makes the functional

$$I = \int_a^b F(x, y, y')\,\mathrm{d}x \tag{2}$$

have an extremum on the family of the curves considered.

This problem can be solved by the procedure shown in the following text, assuming that the curve sought is *not* an extremal of the functional G (on the family L). The variational problem formulated in this way and completed with the boundary conditions prescribed on the ends of admissible curves is the basic type of the so–called category of *isoperimetric problems.*

23.17. A Necessary Condition for an Extremum

Theorem 1. *Let the curve $K_0 \in L$ described by the equation $y = y_0(x)$ $(x \in [a, b])$ yield an extremum of the functional*

$$I = \int_a^b F(x, y, y')\,\mathrm{d}x$$

on the family of all curves from L that are described by the equation $y = y(x)$ $(x \in [a, b])$ and satisfy the conditions

$$y(a) = y_0(a), \quad y(b) = y_0(b),$$
$$G(K) = \int_a^b \Phi(x, y(x), y'(x))\,dx = C \tag{1}$$

(where C, $y_0(a)$, $y_0(b)$ are given numbers). If the curve K_0 is not an extremal of the functional

$$G = \int_a^b \Phi(x, y, y')\,dx$$

on the family of all curves $K \in L$ having common end points $A(a, y_0(a))$ and $B(b, y_0(b))$, then there exists a constant λ such that the curve K_0 is an extremal of the functional

$$H = \int_a^b [F(x, y, y') + \lambda\Phi(x, y, y')]\,dx \tag{1'}$$

on the family of all curves $K \in L$ possessing the properties (1).

REMARK 1. Theorem 1 gives a method for solving the considered isoperimetric problem. The problem reduces to a problem of the first category, i.e. to the problem of finding an extremum of the functional (1') under the conditions given. Hence, the extremals for the corresponding variational problem satisfy the Euler differential equation of the variational problem with the functional (1'), i.e. the equation

$$F'_y - \frac{d}{dx}F'_{y'} + \lambda\left(\Phi'_y - \frac{d}{dx}\Phi'_{y'}\right) = 0. \tag{2}$$

Example 1. From among all the curves K of class T_1 in $[a, b]$ that are described by the equation $y = y(x)$ $(x \in [a, b])$, have the end points $A(a, 0)$ and $B(b, 0)$, are of the same length l, greater than $b - a$, and lie in the half-plane $y \geqq 0$ (Fig. 23.8), such a curve is to be found that the area enclosed by it and the segment AB is maximal.

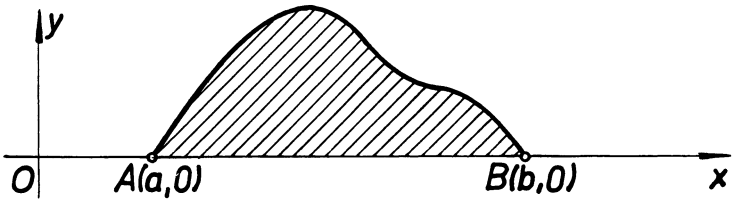


Fig. 23.8.

The area enclosed by the segment AB and a curve K of the considered family of curves is given by the integral

$$I(K) = \int_a^b y \, dx, \tag{3}$$

the length of the curve by the integral

$$G(K) = \int_a^b \sqrt{(1 + y'^2)} \, dx.$$

According to our assumptions, we have

$$G(K) = \int_a^b \sqrt{(1 + y'^2)} \, dx = l. \tag{4}$$

The boundary conditions for the family of admissible curves are

$$y(a) = y(b) = 0. \tag{5}$$

Let us consider, first, a variational problem of the first category with the functional G whose domain of definition is the family of the curves of class T_1 described by the equation $y = y(x)$ $(x \in [a, b])$ and possessing only the property (5). According to case 1 of § 23.5, we have

$$\frac{y'}{\sqrt{(1 + y'^2)}} = \text{const.}$$

at the points of the extremal of this variational problem. Therefore, y' is constant at the points of each such extremal. From this and from the boundary conditions (5), we conclude that $y = 0$ $(x \in [a, b])$ holds for the extremal sought. However, the length of the segment with the end points A and B is $b - a$ and is thus less than l. Hence, such an extremal of the discussed variational problem with the functional G cannot be, at the same time, an extremal of the given isoperimetric problem. We can thus apply Theorem 1.

In this particular case, the Euler differential equation reduces to (see § 23.5)

$$y + \lambda \sqrt{(1 + y'^2)} - y' \lambda \frac{y'}{\sqrt{(1 + y'^2)}} = \alpha \quad (\alpha = \text{const.})$$

and, after rearrangement, to

$$y \sqrt{(1 + y'^2)} + \lambda = \alpha \sqrt{(1 + y'^2)}.$$

This further implies

$$y = \alpha - \frac{\lambda}{\sqrt{(1 + y'^2)}} . \tag{6}$$

Put $y' = \tan\varphi$ (where φ is taken from the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$) so that

$$y = \alpha - \lambda \cos\varphi . \tag{7}$$

Differentiating with respect to x, we obtain

$$y' = \lambda \sin\varphi \frac{\mathrm{d}\varphi}{\mathrm{d}x}$$

and, substituting $y' = \tan\varphi$, we further have

$$\mathrm{d}x = \lambda \cos\varphi \, \mathrm{d}\varphi ,$$

i.e.

$$x = \lambda \sin\varphi + \beta \quad (\beta = \text{ const. }) . \tag{8}$$

The relations (7) and (8) finally imply

$$(x - \beta)^2 + (y - \alpha)^2 = \lambda^2 . \tag{9}$$

The values of λ, α, β are determined by the value of l and by the conditions (5). The condition $l > b - a$ yields $\lambda \neq 0$. In fact, let $\lambda = 0$. Then (6) gives $y = \alpha = \text{const.}$, and the boundary conditions (5) imply that the extremal of the problem is the segment with end points $A(a, 0)$ and $B(b, 0)$ and the length $l = b - a$, which contradicts the assumption $l > b - a$. Therefore $\lambda \neq 0$ and we obtain from (9) that the extremal is the arc of a circle with end points A and B. The radius $|\lambda|$ and the centre $S(\beta, \alpha)$ of the circle can be determined from the conditions imposed. A detailed analysis gives the following results:

If $l < \frac{\pi}{2}(b - a)$, then the sought curve of required properties is the above mentioned arc of a circle.

In the case $l = \frac{\pi}{2}(b - a)$, we obtain the arc of the circle with radius $|\lambda| = \frac{b-a}{2}$ and the centre $S\left(\frac{a+b}{2}, 0\right)$, i.e. the semicircle described by the equation $y = \sqrt{[(b + a)x - ab - x^2]}$ $(x \in [a, b])$, and the derivative y' is improper at the points A and B. In this case, the assumptions of § 23.5 are not satisfied (the curve is not of class T_1 in $[a, b]$, but the curve solves the problem considered, anyway.

In the case $l > \frac{\pi}{2}(b-a)$, there is no curve with the description $y = y(x)$ $(x \in [a, b])$ that would solve the given problem.

REMARK 2. Isoperimetric problems can be generalized to isoperimetric problems with "moving ends of admissible curves", and also to the case of functionals depending on curves in a space of arbitrary dimension, as well as to functionals of the form (23.9.1).

G. PROBLEMS OF THE SEVENTH CATEGORY (PARAMETRIC VARIATIONAL PROBLEMS)

23.18. Formulation of the Problem

In variational problems of the preceding categories, we considered functionals whose domains of definition were curves in the space E_{n+1} with an explicit description of the form (23.6.1). Naturally, we can also take into account functionals defined on a family L of curves given *parametrically* in the space E_n, i.e. curves with the description

$$K = \{(y_1, \dots, y_n) \in E_n \mid y_i = y_i(t)\ (i = 1, \dots, n),\ t \in [t_1, t_2]\}\ . \tag{1}$$

Since any curve with this description – considered as a set of points in E_n – is independent of an arbitrary regular transformation of the parameter t, only such functionals

$$I = \int_{t_1}^{t_2} F(y_1, \dots, y_n, y_1', \dots, y_n')\, dt \tag{2}$$

are admissible which are invariant with respect to a regular transformation of the parameter. A special property of the function F can guarantee that this functional would be invariant: Namely, the function F has to be a *positive homogeneous function of the first degree with respect to the variables* y_i' $(i = 1, \dots, n)$, i.e.,

$$F(y_1, \dots, y_n, ky_1', \dots, ky_n') = kF(y_1, \dots, y_n, y_1', \dots, y_n')$$

has to hold in its domain of definition, for an arbitrary $k > 0$. Since

$$F = \sum_{i=1}^{n} F'_{y_i'} y_i' \tag{3}$$

holds for such homogeneous functions according to the Euler theorem (12.6.1), we arrive at the equations (assuming corresponding differentiability of the function F)

$$\sum_{i=1}^{n} F''_{y_i' y_j'} y_i' = 0 \quad (j = 1, \dots, n)\,. \tag{3'}$$

The variational problem with a functional of the form (2) is called the *parametric* (or *homogeneous*) *variational problem.*

23.19. Necessary Conditions for an Extremum of the Functional I

Let the function $F(y_1, \dots, y_n, y_1', \dots, y_n')$ possess continuous partial derivatives up to the second order in an open set Ω of the $2n$-dimensional space of the arguments $y_1, \dots, y_n, y_1', \dots, y_n'$ and let it be a positive homogeneous function of the first degree with respect to the variables y_i' $(i = 1, \dots, n)$ in Ω. Let us denote by L the family of all curves K with the parametric description (23.18.1), where all the functions $y_i(t)$ $(i = 1, \dots, n)$ have continuous derivatives of the first order in an open interval $(\bar{t}_1, \bar{t}_2)$ containing the interval $[t_1, t_2]$ and where $(y_1(t), \dots, y_n(t), y_1'(t), \dots, y_n'(t)) \in \Omega$ for $t \in [t_1, t_2]$. Since (23.18.3) implies

$$F'_{y_k} = \sum_{i=1}^{n} F''_{y_i' y_k} y_i',$$

we get, taking (23.18.3′) into account, the relation

$$\sum_{k=1}^{n} y_k' \left(F'_{y_k} - \frac{\mathrm{d}}{\mathrm{d}x} F'_{y_k'} \right) = 0 \tag{1}$$

to be satisfied at all points of each curve $K \in L$.

The following theorem holds:

Theorem 1. *If the functional* (23.18.2) *assumes its extremum on the family* L *along a curve* $K \in L$, *then the system of Euler equations*

$$F'_{y_i} - \frac{\mathrm{d}}{\mathrm{d}x} F'_{y_i'} = 0 \quad (i = 1, \dots, n)$$

holds at the points of this curve K.

Note that one of these equations is redundant, in consequence of (1).

H. PROBLEMS OF THE EIGHT CATEGORY (VARIATIONAL PROBLEMS WITH CONSTRAINTS)

23.20. Formulation of the Variational Problem and Necessary Conditions for an Extremum

In applications to geometry and mechanics, more general variational problems are considered. In these problems, the family of admissible curves consists of smooth curves in E_{n+1} (with an explicit description) that lie on a given smooth d–dimensional manifold (surface, hypersurface) $S_d \subset E_{n+1}$, $(2 \leqq d \leqq n)$. The corresponding functional I to be minimized or maximized has the form (23.7.1). Assume that the manifold S_d has the implicit description

$$S_d = \{(x, y_1, \ldots, y_n) \in G \,|\, \varphi_j(x, y_1, \ldots, y_n) = 0 \quad (j = 1, \ldots, n - d + 1)\} ,$$

where G is such a domain in E_{n+1} in which the functions φ_j have continuous partial derivatives of the first order, and the matrix consisting of $n - d + 1$ rows

$$\frac{\partial \varphi_j}{\partial x}, \frac{\partial \varphi_j}{\partial y_1}, \ldots, \frac{\partial \varphi_j}{\partial y_n} \quad (j = 1, \ldots, n - d + 1)$$

has the maximal possible rank, i.e. $n - d + 1$, at every point of the set S_d.

In what follows, we are going to study the variational problem with the functional

$$I = \int_a^b F(x, y_1, \ldots, y_n, y_1', \ldots, y_n') \, \mathrm{d}x \tag{1}$$

where the function F is assumed to have continuous partial derivatives up to the second order in an open set Ω of the $(2n + 1)$–dimensional space of its arguments. We will take the family of all curves possessing the following properties for the domain of the functional (1):

(i) The curves have the description

$$K = \{(x, y_1, \ldots, y_n) \in E_{n+1} \,|\, y_i = y_i(x) \ (i = 1, \ldots, n), \ x \in [a, b]\} \tag{2}$$

and are of class T_1 in $[a, b]$.

(ii) The curves have common end points

$$A(a, y_1(a), \ldots, y_n(a)), \quad B(b, y_1(b), \ldots, y_n(b)) .$$

(iii) For every curve K with the properties (i) and (ii)

$$(x, y_1(x), \dots, y_n(x), y_1'(x), \dots, y_n'(x)) \in \Omega \quad \text{for all} \quad x \in [a, b].$$

(iv) $K \subset S_d$ for every curve K with the properties (i), (ii) and (iii).

The problem of finding an extremum of the functional (1) on the family L of all curves possessing the properties (i) to (iv) is called the *variational problem with constraints* or the *Lagrange variational problem.*

Theorem 1. *If the functional* (1) *assumes its extremum on the family L along a curve $K \in L$ then there exist functions $\lambda_j(x)$ $(j = 1, \dots, n - d + 1)$, defined on $[a, b]$ and such that the relations*

$$\begin{gathered} F_{y_i}' - \frac{\mathrm{d}}{\mathrm{d}x} F_{y_i'}' + \sum_{j=1}^{n-d+1} \lambda_j \frac{\partial \varphi_j}{\partial y_i} = 0 \quad (i = 1, \dots, n), \\ \varphi_j(x, y_1, \dots, y_n) = 0 \quad (j = 1, \dots, n - d + 1) \end{gathered} \tag{3}$$

hold at the points of the curve K.

23.21. Variational Problems with Generalized Constraints

Variational problems with constraints formulated in § 23.20 can be generalized in such a way that the constraints are considered in the form

$$\tilde{\varphi}_j(x, y_1, \dots, y_n, y_1', \dots, y_n') = 0 \quad (j = 1, \dots, n - d + 1). \tag{1}$$

Under some assumptions (not presented here) on the functions $\tilde{\varphi}_j$, the domain L of the definition of the functional (23.20.1) (in which the function F satisfies the assumptions of § 23.20) is introduced as the family of curves possessing the properties (i), (ii) and (iii) of § 23.20 and satisfying the conditions (1). Necessary conditions for the existence of an extremum of the functional (23.20.1) on this family L are similar to the conditions (23.20.3) and have the form

$$\begin{gathered} F_{y_i}' - \frac{\mathrm{d}}{\mathrm{d}x} F_{y_i'}' + \sum_{j=1}^{n-d+1} \lambda_j \left(\frac{\partial \varphi_j}{\partial y_i} - \frac{\mathrm{d}}{\mathrm{d}x} \frac{\partial \varphi_j}{\partial y_i'} \right) = 0 \quad (i = 1, \dots, n), \\ \tilde{\varphi}_j(x, y_1, \dots, y_n, y_i', \dots, y_n') = 0 \quad (j = 1, \dots, n - d + 1). \end{gathered}$$

23.22. Canonical Form of the Euler Equations. Hamiltonian Equations

If the function $F(x, y_1, \dots, y_n, y_1', \dots, y_n')$ in the definition of the functional (23.7.1) satisfies the assumptions of § 23.7, then the Euler equations (23.8.1) can be rewritten as

$$F_{y_i}' - \left(\sum_{j=1}^{n} F_{y_i' y_j}'' y_j' + \sum_{j=1}^{n} F_{y_i' y_j'}'' y_j'' \right) = 0 \quad (i = 1, \dots, n), \tag{1}$$

which are differential equations of the second order for the functions $y_i(x)$ $(i = 1, \dots, n)$ to be found. It is known from the theory of ordinary differential equations (§ 17.18) that every system of second order differential equations can be transformed into an equivalent system of first order differential equations. In our particular case, this can be carried out with the help of the so-called *Legendre transformation*, given by

$$p_i = F_{y_i'}'(x, y_1, \dots, y_n, y_1', \dots, y_n') \quad (i = 1, \dots, n) \tag{2}$$

provided that the quantities y_i' can be expressed from (2) as functions of the variables $x, y_1, \dots, y_n, p_1, \dots, p_n$, i.e.

$$y_i' = y_i'(x, y_1, \dots, y_n, p_1, \dots, p_n) \quad (i = 1, \dots, n). \tag{2'}$$

If we introduce the function

$$H(x, y_1, \dots, y_n, p_1, \dots, p_n) = \\ = \sum_{i=1}^{n} F_{y_i'}'(x, y_1, \dots, y_n, y_1', \dots, y_n') y_i' - F(x, y_1, \dots, y_n, y_1', \dots, y_n'), \tag{3}$$

where the quantities y_i' on the right-hand side are those of (2'), then the Euler differential equations (1) are equivalent to the system of $2n$ differential equations

$$\begin{aligned} \frac{\mathrm{d}y_i}{\mathrm{d}x} &= \frac{\partial H}{\partial p_i}(x, y_1, \dots, y_n, p_1, \dots, p_n), \\ \frac{\mathrm{d}p_i}{\mathrm{d}x} &= -\frac{\partial H}{\partial y_i}(x, y_1, \dots, y_n, p_1, \dots, p_n) \end{aligned} \tag{4}$$

of the first order for the $2n$ sought functions $y_i(x)$ and $p_i(x)$ $(i = 1, \dots, n)$.

The function H defined by (3) is called the *Hamilton* (or *Hamiltonian*) *function* and the system (4) is called the *system of Hamilton differential equations*.

CONCLUDING REMARK. Categories I to VIII in no case cover the wide field of problems of the calculus of variations. We could introduce a lot of further categories, as for instance the problems with constrained extrema of functionals defined on a family of more general curves than those belonging to the class T_1, or the problems with moving ends which are subsets of given smooth manifolds, or the problems obtained as certain combinations of problems of the above presented categories, and a number of other variational problems (see references introduced at the beginning of this chapter).

24. VARIATIONAL METHODS FOR NUMERICAL SOLUTION OF BOUNDARY-VALUE PROBLEMS FOR DIFFERENTIAL EQUATIONS. FINITE ELEMENT METHOD. BOUNDARY ELEMENT METHOD

By MILAN PRÁGER

References: [19], [20], [23], [25], [30], [34], [58], [83], [84], [88], [89], [103], [109], [139], [140], [142], [153], [165], [167], [176], [177], [210], [216], [226], [229], [242], [264], [279], [280], [314], [315], [324], [334], [348], [350], [351], [354], [389], [390], [419], [449], [460], [461], [463], [488], [492], [493], [511], [512], [513].

24.1. Introduction. Theoretical Background. Table of Boundary-Value Problems

Among numerical methods of solution of linear boundary-value problems for ordinary differential equations and elliptic partial differential equations, variational methods play an important role. Theoretical results for these methods have been obtained mostly by means of functional analysis. Thus the reader is recommended to have a look at Definitions 22.4.6 and 22.4.7 at first, concerning the concept of the Hilbert space, and at Remark 22.4.10 on generalized derivatives and the Sobolev space. This space, currently used in variational methods, is mostly denoted by $H^k(\Omega)$ or $W_2^{(k)}(\Omega)$. In this chapter, the former symbol is used. If no misunderstanding can occur we use only the symbol H^k instead of $H^k(\Omega)$. Further, the reader is recommended to notice Theorem 22.6.9 on the minimum of the so-called functional of energy, Remark 22.6.10 on the energy space H_A and the generalized solution to operator equations of the form $Au = f$, and §§ 18.8 and 18.9 about generalized and weak solutions of differential equations. These concepts play a fundamental role in variational methods.

Let us consider a boundary-value problem for a differential equation in the form

$$Au = f\,,$$

where f is a given element of a real Hilbert space H, A is a symmetric positive definite operator (Definitions 22.6.3, 22.6.4) defined on the domain $D(A)$ dense in H and u the unknown solution.

In this formulation, the operator A corresponds to the given differential operator (the left-hand side of the given equation) and its domain of definition $D(A)$ corresponds to the order of the differential operator and to the boundary conditions (again to their left-hand sides). Since $D(A)$ is a linear set, this formulation describes only homogeneous boundary conditions. Nonhomogeneous boundary conditions can be included into the right-hand side f in the following way: We choose a sufficiently smooth function w satisfying the boundary conditions. The solution is supposed to be of the form $u = z + w$. We then have $A(z+w) = f$ or $Az = f - Aw$ with $z \in D(A)$. The domain of definition $D(A)$ is then a linear set of those sufficiently smooth functions that satisfy the corresponding homogeneous boundary condition.

Variational methods are based on Theorem 22.6.9 which affirms that, roughly speaking, the element (function) $u_0 \in D(A)$ is a solution of the equation $Au = f$ if and only if u_0 minimizes, on $D(A)$, the functional of energy

$$Fu = (Au, u) - 2(f, u).$$

Here, $(.,.)$ stands for the scalar product in H, mostly in $L_2(\Omega)$. However, even in simple cases it can happen that neither the solution of the given equation, nor the minimum of the corresponding functional F need exist. Then, following Remark 22.6.10 (or § 18.8), we introduce a new Hilbert space H_A as the completion (§ 22.2) of $D(A)$ with a new scalar product $(u, v)_A = (Au, v)$. Then the above functional of energy, defined originally only for $u \in D(A)$, is extended onto the whole space H_A by

$$Fu = (u, u)_A - 2(f, u), \qquad u \in H_A.$$

This functional attains really its minimum on H_A for an element $u_0 \in H_A$, uniquely determined by the operator A and the right-hand side f of the given equation (cf. the quoted paragraphs). This element is called the *generalized solution* of the equation $Au = f$. The existence and uniqueness of a (generalized) solution of this equation is thus ensured provided A is a symmetric positive definite operator on $D(A)$. If, moreover, u_0 belongs to $D(A)$, then it is the solution of equation $Au = f$ in the usual sense. In the case of differential equations, this happens if the given data (coefficients of the operator A, the function f, etc.) are sufficiently smooth. Such a solution is then called *classical*.

Now, by § 18.9, the functional F attains its minimum at such a point u_0 at which its first differential

$$F'(u_0, v) = 2(u_0, v)_A - 2(f, v)$$

vanishes. Instead of the problem to find the minimum of the functional F on H_A we thus can require the fulfilment of the "variational condition"

$$(u_0, v)_A = (f, v) \qquad \text{for all } v \in H_A . \tag{1}$$

The generalized and the *weak solution* (see § 18.9) of the problem are identical here.

REMARK 1. If we integrate equation (1) by parts (i.e. use the Green theorem, cf. § 18.9) and eliminate the "variation" v we obtain what we started from, i.e. a differential equation. But this is just what we are going to avoid because the application of the differential operator to the function u requires more smoothness of u than its use in the "variational condition". The possibility to use less smooth functions as a solution of boundary-value problems is an important feature of variational methods.

On the other hand, it can be shown that the space H_A contains (if A is a differential operator) the same functions as a subspace V of a properly chosen Sobolev space. The norm in the space H_A is then equivalent with the norm of that Sobolev space. The scalar product $(u, v)_A$ is a bilinear form (i.e. linear both in u and in v) on V satisfying the hypotheses of Theorem 18.9.1. And from it the existence of the weak solution to the given boundary-value problem follows once again. Moreover, Theorem 18.9.1 does not require the symmetry of the bilinear form.

Consequently, we will consider as a primary problem the problem of the representation of the linear functional (f, v) by a bilinear form. More exactly: Let a Sobolev space H^k and its subspace H_0^k (cf. Remark 22.4.11) be given. Let us choose a subspace V so that $H_0^k \subset V \subset H^k$. We denote by $\|u\|_V$ the norm of the element u of the space V. On the space V, let a bilinear form $a(u, v)$ be given (this form is denoted by $((u, v))$ in § 18.9; here we adopt the notation usual in the finite element context). Let $a(u, v)$ be continuous, i.e.

$$|a(u, v)| \leqq K \|u\|_V \|v\|_V \qquad \text{for all } u, v \in V , \tag{2}$$

and V-elliptic, i.e.

$$a(u, u) \geqq \alpha \|u\|_V^2 \qquad \text{for all } u \in V , \quad \alpha > 0 . \tag{3}$$

Problem 1. *We are to find an element $u_0 \in V$ such that*

$$a(u_0, v) = (f, v) \quad \textit{holds for all functions } v \in V . \tag{4}$$

Existence and uniqueness of such an element is guaranteed by Theorem 18.9.1 (see § 18.9 for details). Theorem 18.9.1 holds even if a general linear functional on the space V stands on the right-hand side. For usual boundary-value problems,

the scalar product (f, v) in $L_2(\Omega)$ is always present on the right-hand side of (4). But for some types of boundary conditions the right-hand side functional has to be modified by adding other terms and the generalized version of Problem 1 is then used. Our task in this chapter is to describe effective methods for numerical computation of the element u_0.

We recall that also here V is a subspace and that the boundary-value problem in the formulation (4) includes, for some types of boundary conditions, only their homogeneous versions. For nonhomogeneous conditions it is necessary to find, in advance, a function $w \in H^k$ that satisfies these nonhomogeneous conditions. The solution of the given problem is then sought in the form $u_0 = z + w$ and we require

$$a(z, v) = (f, v) - a(w, v) \quad \text{for all } v \in V \tag{4'}$$

instead of (4). Such a function w will be called *representing function.*

The most important boundary-value problems are summarized in Tab. 24.1, given at the end of this paragraph. To each problem, the corresponding bilinear form, the right-hand side functional (f, v) or its generalization, the subspace V and the representing function w are shown.

REMARK 2. If $a(u, v)$ is symmetric, the functional $a(u, u) - 2(f, u)$ attains its minimum over V just at the element u_0.

REMARK 3. In the case of a symmetric form, $a(u, v)$ may be considered as a new scalar product on V; the conditions (2) and (3) represent then the equivalence of the norm given by the scalar product (we recall that $a(u, v) = (u, v)_A$) and the norm on V. This new scalar product is often called the *energy scalar product* and the corresponding norm is called the *energy norm.*

REMARK 4. *Nonsymmetric* bilinear forms are obtained in a natural way if the differential operator contains lower derivatives of odd order, e.g. if a second-order operator contains first derivatives:

Example 1. Let us have the equation

$$-\Delta u + \frac{\partial u}{\partial x} = f$$

in a domain $\Omega \subset E_2$ with a Lipschitz boundary S (Remark 22.4.10) and with the boundary condition $u = 0$ on S.

Multiplying the equation by a function $v \in H_0^1$ and integrating over Ω, we obtain

$$-\int_\Omega \Delta u\, v \, \mathrm{d}x\mathrm{d}y + \int_\Omega \frac{\partial u}{\partial x} v \, \mathrm{d}x\mathrm{d}y = \int_\Omega f v \, \mathrm{d}x\mathrm{d}y \,.$$

Integrating by parts and using the given boundary condition we have (for more details see e.g. [389], Chaps. 8 and 32; cf. also Example 18.9.1)

$$\int_\Omega \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} + \frac{\partial u}{\partial x} v \right) \mathrm{d}x\mathrm{d}y = \int_\Omega f v \, \mathrm{d}x\mathrm{d}y \,.$$

Thus, we set

$$a(u, v) = \int_\Omega \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} + \frac{\partial u}{\partial x} v \right) \mathrm{d}x\mathrm{d}y \,.$$

Changing the role of u and v, we obtain

$$a(v, u) = \int_\Omega \left(\frac{\partial v}{\partial x}\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} u \right) \mathrm{d}x\mathrm{d}y$$

and finally, integrating the last term by parts with the use of the boundary condition,

$$a(v, u) = \int_\Omega \left(\frac{\partial v}{\partial x}\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\frac{\partial u}{\partial y} - \frac{\partial u}{\partial x} v \right) \mathrm{d}x\mathrm{d}y \,.$$

We see that, in this case, $a(u, v) \neq a(v, u)$ and thus a is not symmetric.

TABLE 24.1.

Most Current Types of Boundary-Value Problems

Notation:

DE ... differential equation
BC ... boundary conditions
CS ... condition of solvability
BF ... bilinear form
RS ... right-hand side functional (f, v)
FS ... fundamental space V with elements denoted by v
RF ... function w a priori satisfying some kinds of boundary conditions

Use:

For given DE and BC, it is necessary to verify CS. Then, we find BF, RS, FS, and eventually (in the case of nonhomogeneous boundary conditions of some kind) RF in the table. In this way, we obtain information necessary for the construction of (4) or (4').

(A) *Ordinary differential equations* (one-dimensional case)

(1) DE: $-(pu')' + ru = f$ on a bounded interval $I = (a, b)$, p, r bounded,
$p(x) \geqq p_0 > 0$, $r(x) \geqq 0$, $f \in L_2(I)$

BC: $u(a) = A, \quad u(b) = B$

CS: none

BF: $\displaystyle\int_a^b (pu'v' + ruv)\,\mathrm{d}x$

RS: $\displaystyle\int_a^b fv\,\mathrm{d}x$

FS: $V = H_0^1$

RF: $w \in H^1, \quad w(a) = A, \quad w(b) = B$

(2) DE: the same as in (1)

BC: $u(a) = A, \quad p(b)u'(b) + \beta u(b) = B, \quad \beta \geqq 0$

CS: none

BF: $\displaystyle\int_a^b (pu'v' + ruv)\,\mathrm{d}x + \beta u(b)v(b)$

RS: $\displaystyle\int_a^b fv\,\mathrm{d}x + Bv(b)$

FS: $V \subset H^1, \quad v(a) = 0$

RF: $w \in H^1, \quad w(a) = A$

(3) DE: the same as in (1)

BC: $p(a)u'(a) - \alpha u(a) = A, \quad p(b)u'(b) + \beta u(b) = B, \quad \alpha \geqq 0, \quad \beta \geqq 0,$
$\displaystyle\int_a^b r\,\mathrm{d}x + \alpha + \beta > 0$

CS: none

BF: $\displaystyle\int_a^b (pu'v' + ruv)\,\mathrm{d}x + \alpha u(a)v(a) + \beta u(b)v(b)$

RS: $\displaystyle\int_a^b fv\,\mathrm{d}x + Av(a) + Bv(b)$

FS: $V = H^1$

RF: none

(4) DE: $-(pu')' = f$ on a bounded interval $I = (a, b)$, p bounded, $p(x) \geqq p_0 > 0$, $f \in L_2(I)$

BC: $u'(a) = A, \quad u'(b) = B$

CS: $\displaystyle\int_a^b f\,\mathrm{d}x - Ap(a) + Bp(b) = 0$

BF: $\displaystyle\int_a^b pu'v'\,dx$

RS: $\displaystyle\int_a^b fv\,dx - Ap(a)v(a) + Bp(b)v(b)$

FS: $V = H^1, \quad \displaystyle\int_a^b v\,dx = 0$

RF: none

(5) DE: $(pu'')''-(qu')'+ru = f$ on a bounded interval $I = (a, b)$, p, q, r bounded, $p(x) \geqq p_0 > 0$, $q(x) \geqq 0$, $r(x) \geqq 0$, $f \in L_2(I)$

BC: $u(a) = A$, $u'(a) = B$, $u(b) = C$, $u'(b) = D$

CS: none

BF: $\displaystyle\int_a^b (pu''v'' + qu'v' + ruv)\,dx$

RS: $\displaystyle\int_a^b fv\,dx$

FS: $V = H_0^2$

RF: $w \in H^2$, $w(a) = A$, $w'(a) = B$, $w(b) = C$, $w'(b) = D$

(6) DE: the same as in (5)

BC: $u(a) = A$, $u''(a) = B$, $u(b) = C$, $u''(b) = D$

CS: none

BF: the same as in (5)

RS: $\displaystyle\int_a^b fv\,dx - Bp(a)v'(a) + Dp(b)v'(b)$

FS: $V \subset H^2$, $v(a) = 0$, $v(b) = 0$

RF: $w \in H^2$, $w(a) = A$, $w(b) = C$

(7) DE: the same as in (5)

BC: $u(a) = A$, $u'(a) = B$, $u(b) = C$, $p(b)u''(b) + \alpha u'(b) = D$, $\alpha \geqq 0$

CS: none

BF: $\displaystyle\int_a^b (pu''v'' + qu'v' + ruv)\,dx + \alpha u'(b)v'(b)$

RS: $\displaystyle\int_a^b fv\,dx + Dv'(b)$

FS: $V \subset H^2$, $v(a) = 0$, $v'(a) = 0$, $v(b) = 0$

RF: $w \in H^2, \quad w(a) = A, \quad w'(a) = B, \quad w(b) = C$

(8) DE: the same as in (5)

BC: $u(a) = A, \quad u'(a) = B, \quad p(b)u''(b) = C, \quad [-(pu'')' + qu']_{x=b} = D$

CS: none

BF: the same as in (5)

RS: $\displaystyle\int_a^b fv\,\mathrm{d}x + Dv(b) + Cv'(b)$

FS: $V \subset H^2, \quad v(a) = 0, \quad v'(a) = 0$

RF: $w \in H^2, \quad w(a) = A, \quad w'(a) = B$

(9) DE: $\displaystyle\sum_{i=0}^{k}(-1)^i(p_iu^{(i)})^{(i)} = f$ on a bounded interval $I = (a, b)$, p_i, $i = 0, \ldots k$, bounded, $p_k(x) \geqq p > 0, \quad p_i(x) \geqq 0, \quad i = 0, \ldots, k-1, \quad f \in L_2(I)$

BC: $u(a) = A_0, \quad u'(a) = A_1, \quad \ldots, \quad u^{(k-1)}(a) = A_{k-1}$
$u(b) = B_0, \quad u'(b) = B_1, \quad \ldots, \quad u^{(k-1)}(b) = B_{k-1}$

CS: none

BF: $\displaystyle\int_a^b \sum_{i=0}^{k} p_iu^{(i)}v^{(i)}\,\mathrm{d}x$

RS: $\displaystyle\int_a^b fv\,\mathrm{d}x$

FS: $V = H_0^k$

RF: $w \in H^k, \quad w(a) = A_0, \quad w'(a) = A_1, \quad \ldots, \quad w^{(k-1)}(a) = A_{k-1}$
$w(b) = B_0, \quad w'(b) = B_1, \quad \ldots, \quad w^{(k-1)}(b) = B_{k-1}$

(B) *Partial differential equations of elliptic type*

(10) DE: $\displaystyle -\sum_{i,j=1}^{m}\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + cu = f$ on a bounded domain $\Omega \subset E_m$ with the boundary S, a_{ij}, $i, j = 1, \ldots, m$, c bounded,
$\displaystyle\sum_{i,j=1}^{m} a_{ij}(x)\xi_i\xi_j \geqq \alpha\sum_{i=1}^{m}\xi_i^2$, $\alpha > 0$, for all real vectors $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)$ and almost all points $x \in \Omega$; $c \geqq 0$, $f \in L_2(\Omega)$,
especially $-\Delta u + cu = f$

BC: $u = g$ on S

CS: none

BF: $\displaystyle\int_\Omega \left(\sum_{i,j=1}^{m} a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + cuv \right) \mathrm{d}x ,$

especially $\displaystyle\int_\Omega \left(\sum_{i=1}^{m} i \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} + cuv \right) \mathrm{d}x$

RS: $\displaystyle\int_\Omega fv \, \mathrm{d}x$

FS: $V = H_0^1$

RF: $w \in H^1, \quad w = g \text{ on } S$

(11) DE: the same as in (10)

BC: $u = g \text{ on } S_1 , \quad \operatorname{mes} S_1 \neq 0 ,$

$\displaystyle\sum_{i,j=1}^{m} a_{ij} \frac{\partial u}{\partial x_j} \nu_i + \sigma u = h \text{ on } S_2 , \quad S_1 \cup S_2 = S , \quad S_1 \cap S_2 = \emptyset ,$

$\nu_i = \cos(\nu, x_i)$, ν is the outward unit normal, σ bounded, $\sigma(s) \geqq 0$,

$h \in L_2(S_2)$,

especially $u = g \text{ on } S_1 , \quad \dfrac{\partial u}{\partial \nu} + \sigma u = h \text{ on } S_2$

CS: none

BF: $\displaystyle\int_\Omega \left(\sum_{i,j=1}^{m} a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + cuv \right) \mathrm{d}x + \int_{S_2} \sigma uv \, \mathrm{d}S ,$

especially $\displaystyle\int_\Omega \left(\sum_{i=1}^{m} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} + cuv \right) \mathrm{d}x + \int_{S_2} \sigma uv \, \mathrm{d}S$

RS: $\displaystyle\int_\Omega fv \, \mathrm{d}x + \int_{S_2} hv \, \mathrm{d}S$

FS: $V \subset H^1, \quad v = 0 \text{ on } S_1$

RF: $w \in H^1, \quad w = g \text{ on } S_1$

(12) DE: the same as in (10)

BC: $\displaystyle\sum_{i,j=1}^{m} a_{ij} \frac{\partial u}{\partial x_j} \nu_i + \sigma u = h \text{ on } S$, σ bounded, $\sigma(s) \geqq 0$, $\displaystyle\int_\Omega c \, \mathrm{d}x + \int_S \sigma \, \mathrm{d}S > 0$,

$h \in L_2(S_2)$,

$\nu_i = \cos(\nu, x_i)$, ν is the outward unit normal

CS: none

BF: $\displaystyle\int_\Omega \left(\sum_{i,j=1}^{m} a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + cuv \right) \mathrm{d}x + \int_S \sigma uv \, \mathrm{d}S ,$

especially $\displaystyle\int_\Omega \left(\sum_{i=1}^{m} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} + cuv \right) \mathrm{d}x + \int_S \sigma uv \, \mathrm{d}S$

RS: $\displaystyle\int_\Omega fv \, \mathrm{d}x + \int_S hv \, \mathrm{d}S$

FS: $V = H^1$

RF: none

(13) DE: $\displaystyle -\sum_{i,j=1}^{m} \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial u}{\partial x_j} \right) = f$ on a bounded domain $\Omega \subset E_m$ with the boundary S, a_{ij}, $i, j = 1, \ldots, m$, bounded,

$\displaystyle\sum_{i,j=1}^{m} a_{ij}(x) \xi_i \xi_j \geqq \alpha \sum_{i=1}^{m} \xi_i^2$, $\alpha > 0$, for all real vectors $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)$ and almost all points $x \in \Omega$, $f \in L_2(\Omega)$,

especially $-\Delta u = f$

BC: $\displaystyle\sum_{i,j=1}^{m} a_{ij} \frac{\partial u}{\partial x_i} \nu_j = h$ on S, $h \in L_2(S)$, $\nu_i = \cos(\nu, x_i)$, ν is the outward unit normal,

especially $\displaystyle\frac{\partial u}{\partial \nu} = h$ on S

CS: $\displaystyle\int_\Omega f \, \mathrm{d}x + \int_S h \, \mathrm{d}S = 0$

BF: $\displaystyle\int_\Omega \sum_{i,j=1}^{m} a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} \, \mathrm{d}x ,$

especially $\displaystyle\int_\Omega \sum_{i=1}^{m} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, \mathrm{d}x$

RS: $\displaystyle\int_\Omega fv \, \mathrm{d}x + \int_S hv \, \mathrm{d}S$

FS: $V = H^1$, $\displaystyle\int_\Omega v \, \mathrm{d}x = 0$

RF: none

(14) DE: $\Delta^2 u = f$ on a bounded domain $\Omega \subset E_2$ with the boundary S, $f \in L_2(\Omega)$

BC: $u = g$ on S, $\displaystyle\frac{\partial u}{\partial \nu} = h$ on S, ν is the outward unit normal

CS: none

BF: $\displaystyle\int_\Omega \left(\frac{\partial^2 u}{\partial x^2}\frac{\partial^2 v}{\partial x^2} + 2\frac{\partial^2 u}{\partial x \partial y}\frac{\partial^2 v}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2}\frac{\partial^2 v}{\partial y^2}\right) \mathrm{d}x\mathrm{d}y$

RS: $\displaystyle\int_\Omega fv \,\mathrm{d}x\mathrm{d}y$

FS: $V = H_0^2$

RF: $w \in H^2$, $\quad w = g$ on S, $\quad \dfrac{\partial w}{\partial \nu} = h$ on S

(15) DE: the same as in (14)

BC: $u = g$ on S, $\quad Mu = h$ on S,
where $Mu = \sigma \Delta u + (1-\sigma)\dfrac{\partial^2 u}{\partial \nu^2}$, σ is a number, $0 \leqq \sigma < 1$, $h \in L_2(S)$,
ν is the outward unit normal

CS: none

BF: $\displaystyle\int_\Omega \left[\left(\frac{\partial^2 u}{\partial x^2} + \sigma\frac{\partial^2 u}{\partial y^2}\right)\frac{\partial^2 v}{\partial x^2} + 2(1-\sigma)\frac{\partial^2 u}{\partial x \partial y}\frac{\partial^2 v}{\partial x \partial y} + \right.$
$\left. + \left(\frac{\partial^2 u}{\partial y^2} + \sigma\frac{\partial^2 u}{\partial x^2}\right)\frac{\partial^2 v}{\partial y^2}\right] \mathrm{d}x\mathrm{d}y$

RS: $\displaystyle\int_\Omega fv \,\mathrm{d}x\mathrm{d}y + \int_S h\frac{\partial v}{\partial \nu}\,\mathrm{d}s$

FS: $V \subset H^2$, $\quad v = 0$ on S

RF: $w \in H^2$, $\quad w = g$ on S

(16) DE: the same as in (14)

BC: $Mu = g$ on S, $\quad Mu$ the same as in (15),

$Nu = h$ on S, where $Nu = -\dfrac{\partial}{\partial \nu}\Delta u - (1-\sigma)\dfrac{\partial^3 u}{\partial \nu \partial \tau^2}$,
$0 \leqq \sigma < 1$, g, $h \in L_2(S)$, τ is the unit tangent,
ν is the outward unit normal

CS: $\displaystyle\int_\Omega f \,\mathrm{d}x\mathrm{d}y + \int_S h \,\mathrm{d}s = 0$

$\displaystyle\int_\Omega xf \,\mathrm{d}x\mathrm{d}y + \int_S xh \,\mathrm{d}s + \int_S g\nu_1 \,\mathrm{d}s = 0$

$\displaystyle\int_\Omega yf \,\mathrm{d}x\mathrm{d}y + \int_S yh \,\mathrm{d}s + \int_S g\nu_2 \,\mathrm{d}s = 0$

BF: the same as in (15)

RS: $\displaystyle\int_\Omega fv \,\mathrm{d}x\mathrm{d}y + \int_S \left(vh + \frac{\partial v}{\partial \nu}g\right)\mathrm{d}s$

FS: $V = H^2$, $\int_\Omega v\,\mathrm{d}x\mathrm{d}y = 0$, $\int_\Omega xv\,\mathrm{d}x\mathrm{d}y = 0$, $\int_\Omega yv\,\mathrm{d}x\mathrm{d}y = 0$

RF: none

(17) DE: $-\left(\lambda+\frac{\mu}{2}\right)\left(\frac{\partial^2 u_1}{\partial x^2}+\frac{\partial^2 u_2}{\partial x\partial y}\right)-\frac{\mu}{2}\left(\frac{\partial^2 u_1}{\partial x^2}+\frac{\partial^2 u_1}{\partial y^2}\right)=f_1$,

$-\left(\lambda+\frac{\mu}{2}\right)\left(\frac{\partial^2 u_1}{\partial x\partial y}+\frac{\partial^2 u_2}{\partial y^2}\right)-\frac{\mu}{2}\left(\frac{\partial^2 u_2}{\partial x^2}+\frac{\partial^2 u_2}{\partial y^2}\right)=f_2$

on a bounded domain $\Omega \subset E_2$ with the boundary S, $\lambda \geqq 0$, $\mu > 0$ are constants, f_1, $f_2 \in L_2(\Omega)$; *system of equations of plane elasticity.*

If we put

$\boldsymbol{u} = (u_1, u_2)$, $x = x_1$, $y = x_2$,

$\varepsilon_{ij}(\boldsymbol{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)$, $i, j = 1, 2$

and

$\sigma_{ij} = \lambda \operatorname{div} \boldsymbol{u}.\delta_{ij} + \mu\varepsilon_{ij}(\boldsymbol{u})$, where $\delta_{ij} = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases}$

the system can be written as

$-\frac{\partial \sigma_{11}}{\partial x_1}-\frac{\partial \sigma_{12}}{\partial x_2} = f_1$,

$-\frac{\partial \sigma_{21}}{\partial x_1}-\frac{\partial \sigma_{22}}{\partial x_2} = f_2$

BC: $u_1 = g_1$, $u_2 = g_2$ on S

CS: none

BF: $\int_\Omega \left\{\lambda\left(\frac{\partial u_1}{\partial x}+\frac{\partial u_2}{\partial y}\right)\left(\frac{\partial v_1}{\partial x}+\frac{\partial v_2}{\partial y}\right)+\right.$

$\left.+\mu\left[\frac{\partial u_1}{\partial x}\frac{\partial v_1}{\partial x}+\frac{1}{2}\left(\frac{\partial u_1}{\partial y}+\frac{\partial u_2}{\partial x}\right)\left(\frac{\partial v_1}{\partial y}+\frac{\partial v_2}{\partial x}\right)+\frac{\partial u_2}{\partial y}\frac{\partial v_2}{\partial y}\right]\right\}\mathrm{d}x\mathrm{d}y$

or

$\int_\Omega \{\lambda \operatorname{div} \boldsymbol{u} \operatorname{div} \boldsymbol{v}+$

$+\mu\left[\varepsilon_{11}(\boldsymbol{u})\varepsilon_{11}(\boldsymbol{v})+2\varepsilon_{12}(\boldsymbol{u})\varepsilon_{12}(\boldsymbol{v})+\varepsilon_{22}(\boldsymbol{u})\varepsilon_{22}(\boldsymbol{v})\right]\}\,\mathrm{d}x_1\mathrm{d}x_2$

RS: $\int_\Omega (f_1 v_1 + f_2 v_2)\,\mathrm{d}x\mathrm{d}y$

FS: $V = H_0^1 \times H_0^1$ (the space of pairs (v_1, v_2), $v_1 \in H_0^1$, $v_2 \in H_0^1$)

RF: $w_1, w_2 \in H^1$, $w_1 = g_1$, $w_2 = g_2$ on S

(18) DE: the same as in (17)

BC: $u_1 = g_1$, $u_2 = g_2$ on S_1,

$\left[\lambda\left(\frac{\partial u_1}{\partial x}+\frac{\partial u_2}{\partial y}\right)+\mu\frac{\partial u_1}{\partial x}\right]\nu_1+\frac{\mu}{2}\left(\frac{\partial u_1}{\partial y}+\frac{\partial u_2}{\partial x}\right)\nu_2 = h_1$,

$$\frac{\mu}{2}\left(\frac{\partial u_1}{\partial y}+\frac{\partial u_2}{\partial x}\right)\nu_1+\left[\lambda\left(\frac{\partial u_1}{\partial x}+\frac{\partial u_2}{\partial y}\right)+\mu\frac{\partial u_2}{\partial y}\right]\nu_2=h_2 \text{ on } S_2\,,$$

$\operatorname{mes} S_1 \neq 0$, $S_1 \cup S_2 = S$, $S_1 \cap S_2 = \emptyset$, $h_1,\ h_2 \in L_2(S_2)$, $\nu_i = \cos(x_i,\ \nu)$, ν is the outward unit normal,

or in the notation from (17)

$$\sigma_{11}\nu_1+\sigma_{12}\nu_2=h_1\,,$$
$$\sigma_{21}\nu_1+\sigma_{22}\nu_2=h_2 \text{ on } S_2$$

CS: none

BF: the same as in (17)

RS: $$\int_\Omega (f_1 v_1+f_2 v_2)\,\mathrm{d}x\mathrm{d}y+\int_{S_2}(h_1 v_1+h_2 v_2)\,\mathrm{d}s$$

FS: $V \subset H^1 \times H^1\,, \quad v_1=0\,,\ v_2=0 \text{ on } S_1$

RF: $w_1,\ w_2 \in H^1\,, \quad w_1=g_1\,,\ w_2=g_2 \text{ on } S_1$

(19) DE: the same as in (17)

BC: $$\left[\lambda\left(\frac{\partial u_1}{\partial x}+\frac{\partial u_2}{\partial y}\right)+\mu\frac{\partial u_1}{\partial x}\right]\nu_1+\frac{\mu}{2}\left(\frac{\partial u_1}{\partial y}+\frac{\partial u_2}{\partial x}\right)\nu_2=h_1\,,$$
$$\frac{\mu}{2}\left(\frac{\partial u_1}{\partial y}+\frac{\partial u_2}{\partial x}\right)\nu_1+\left[\lambda\left(\frac{\partial u_1}{\partial x}+\frac{\partial u_2}{\partial y}\right)+\mu\frac{\partial u_2}{\partial y}\right]\nu_2=h_2 \text{ on } S$$

$h_1,\ h_2 \in L_2(S)\,, \quad \nu_i=\cos(x_i,\ \nu)$, ν is the outward unit normal,

or in the notation from (17)

$$\sigma_{11}\nu_1+\sigma_{12}\nu_2=h_1\,,$$
$$\sigma_{21}\nu_1+\sigma_{22}\nu_2=h_2 \quad \text{on } S$$

CS: $$\int_\Omega f_1\,\mathrm{d}x\mathrm{d}y+\int_S h_1\,\mathrm{d}s=0\,, \quad \int_\Omega f_2\,\mathrm{d}x\mathrm{d}y+\int_S h_2\,\mathrm{d}s=0\,,$$
$$\int_\Omega (xf_2-yf_1)\,\mathrm{d}x\mathrm{d}y+\int_S (xh_2-yh_1)\,\mathrm{d}s=0$$

BF: the same as in (17)

RS: $$\int_\Omega (f_1 v_1+f_2 v_2)\,\mathrm{d}x\mathrm{d}y+\int_S (h_1 v_1+h_2 v_2)\,\mathrm{d}s$$

FS: $V \subset H^1 \times H^1\,, \ \int_\Omega v_1\,\mathrm{d}x\mathrm{d}y=0\,, \ \int_\Omega v_2\,\mathrm{d}x\mathrm{d}y=0\,, \ \int_\Omega (xv_2-yv_1)\,\mathrm{d}x\mathrm{d}y=0$

RF: none

24.2. Fundamental Approximation Methods

(a) The Ritz Method

This method is a fundamental method for finding an approximate solution of Problem 24.1.1 in the case of a *symmetric* bilinear form $a(u, v)$. The problem can then be equivalently formulated as a problem to find the minimum of the functional

$$F(u) = a(u, u) - 2(f, u) \tag{1}$$

on a space V that is a suitably chosen subspace of a Sobolev space H^k. The choice of V is determined by the type of the boundary conditions of the original differential problem.

The basic idea of the *Ritz method* consists in restricting the space V to a finite-dimensional subspace V_N and taking the element y_N that minimizes the functional (1) on the subspace V_N for the approximation of the weak solution to the given problem (the weak solution itself minimizes the functional (1) on V).

Let us choose a basis $\varphi_1, \ldots, \varphi_N$ in the subspace V_N. The element y_N will be given by

$$y_N = \sum_{k=1}^{N} c_k \varphi_k \tag{2}$$

and the problem to find y_N is the problem to find its coefficients $c_1, \ldots, c_N$.

Substituting (2) into (1), we obtain

$$a\left(\sum_{k=1}^{N} c_k \varphi_k, \sum_{j=1}^{N} c_j \varphi_j\right) - 2\left(f, \sum_{k=1}^{N} c_k \varphi_k\right)$$

or

$$\sum_{k=1}^{N}\sum_{j=1}^{N} a(\varphi_k, \varphi_j) c_k c_j - 2\sum_{k=1}^{N} c_k (f, \varphi_k).$$

This is a quadratic form in the variables $c_1, \ldots, c_N$. We apply standard conditions for its extremum, i.e., we put the derivatives with respect to $c_1, \ldots, c_N$ equal to zero; we thus obtain a system of linear algebraic equations

$$\sum_{k=1}^{N} a(\varphi_k, \varphi_j) c_k = (f, \varphi_j), \quad j = 1, \ldots, N. \tag{3}$$

The matrix $\boldsymbol{A}$ with the entries $a(\varphi_k, \varphi_j)$ is usually called the *Gram matrix* of the elements $\varphi_1, \ldots, \varphi_N$ in the energy scalar product (Remark 24.1.3). This matrix is symmetric (because of $a(\varphi_j, \varphi_k) = a(\varphi_k, \varphi_j)$). Moreover, it is positive definite

(because of (24.1.3)) and, consequently, nonsingular, so the system (3) has a unique solution for an arbitrary right-hand side. In this way, the coefficients c_k of the linear combination (2) and the approximate solution y_N are uniquely determined. In the context of the finite element method (cf. § 24.3), the matrix $\boldsymbol{A}$ is often called the *stiffness matrix* and the right-hand side vector is called the *load vector.*

The key question for the quality of the approximation obtained by the Ritz method is the choice of the subspace V_N and its basis functions. If we choose V_N so that it contains the true solution we have an excellent approximation, the true solution itself. It is, of course, difficult to make such a choice and, moreover, a change of the right-hand side produces another solution for which the subspace V_N need not be suitable. We thus require a good approximation of the *entire* space V by the subspace V_N. This can be achieved in two ways:

1. We choose an infinite sequence of linearly independent elements $\varphi_i \in V$, $i = 1, 2, \ldots$, and we require that, for an arbitrary element $v \in V$ and for an arbitrary $\varepsilon > 0$, such a finite linear combination of the elements φ_i can be found that

$$\left\| v - \sum_{i=1}^{N} k_i \varphi_i \right\| < \varepsilon$$

holds.

We introduce a whole sequence of subspaces V_N defined as linear hulls of elements $\varphi_1, \ldots, \varphi_N$, and a question arises how many elements of the basis to use. Further, we must take into account that the computation of the approximate solution (or at least of its values at some points) consists in the computation of the sum $\sum_{i=1}^{N} c_i \varphi_i$.

2. We use the approach of the finite element method (see § 24.3 for further details).

We have to realize that not only the subspace V_N itself, but also the choice of its basis is important for the numerical solution. Even though the stiffness matrix is positive definite it may be ill-conditioned (cf. § 30.3) with an inappropriate choice of the basis and the solution of the system (3) may be influenced or eventually entirely destroyed by round-off errors. Therefore, choosing the basis according to 1, our aim is to make the functions φ_i orthogonal, if possible. The system (3) can be solved very well by contemporary computers. With the approach of the finite elements method, the matrix has many zero entries (it is sparse) and systems with several tens or even hundreds of thousands of unknowns can be solved.

The *error of the approximate solution* $y_N - u$ fulfils

$$a(y_N - u, y_N - u) = F(y_N) - F(u). \tag{4}$$

The exact solution of the given problem, i.e. the function that minimizes the functional (1) on the space V, is denoted by u. From (4), we have immediately the following important assertion:

Theorem 1. *The sequence of Ritz approximations converges to the exact solution in the subspace V if and only if it is minimizing, i.e. if and only if*

$$\lim_{N\to\infty} F(y_N) = F(u)\,.$$

The next assertion gives an estimate for the norm of the Ritz approximation.

Theorem 2. *For the Ritz approximation y_N we have*

$$\|y_N - u\|_V \leqq K \inf_{v\in V_N} \|u - v\|_V\,,$$

where the constant K is independent of the subspace V_N.

This simple but important fact is called *Céa's lemma.* We can, with a certain inaccuracy, express it by saying that the *Ritz approximation of the exact solution is the best one among all the elements of the subspace V_N.* Because the form $a(u, v)$ is symmetric we have even $a(y_N - u, y_N - u) = \inf_{v\in V_N} a(v - u, v - u)$. Theorem 2 applies, however, for the non-symmetric case, too, cf. § 24.3.

Example 1. On the interval $I = (0, 1)$ let us consider the boundary-value problem

$$-u'' + u = 1\,,$$
$$u(0) = u(1) = 0\,.$$

From Tab. 24.1, we find out that the weak solution belongs to the space $V = H_0^1(I)$ and fulfils the equality

$$\int_0^1 (u'v' + uv)\,\mathrm{d}x = \int_0^1 v\,\mathrm{d}x$$

for all functions $v \in V$. We choose the subspace V_N in the first of the two mentioned ways and put

$$\varphi_k(x) = \sin k\pi x\,.$$

This choice leads to the system (see (3))

$$\sum_{k=1}^{N}\left(kj\pi^2 \int_0^1 \cos k\pi x \cos j\pi x\,\mathrm{d}x + \int_0^1 \sin k\pi x \sin j\pi x\,\mathrm{d}x\right)c_k = \int_0^1 \sin j\pi x\,\mathrm{d}x\,,$$

$j = 1, \ldots, N$, for the coefficients c_k. The orthogonality of trigonometric functions yields

$$\frac{1}{2}[(j\pi)^2 + 1]c_j = \frac{1}{j\pi}[1 - (-1)^j].$$

The approximate solution is then given by

$$y_N(x) = \sum_{j=1}^{N} c_j \sin j\pi x,$$

where

$$c_j = \begin{cases} \dfrac{4}{j\pi[(j\pi)^2 + 1]} & \text{for } j \text{ odd}, \\ 0 & \text{for } j \text{ even}. \end{cases}$$

From here we can obtain the exact solution as the sum of an infinite series

$$u = \sum_{i=1}^{\infty} \frac{4}{(2i-1)\pi[(2i-1)^2\pi^2 + 1]} \sin(2i-1)\pi x.$$

In this case, the use of the Ritz method was extremely easy. We have even obtained the exact solution as the sum of a trigonometric series. Let us note that here we solved a one-dimensional problem with constant coefficients and that the exact solution could be easily found immediately:

$$u = 1 - \frac{\cosh\left(x - \frac{1}{2}\right)}{\cosh\frac{1}{2}}.$$

So it was not necessary to look for an approximate solution here. We have made it only to illustrate the method. However, in the case of the equation $-\Delta u + u = 1$ on a rectangle Ω, with the boundary condition $u = 0$ on S, thus in the case of a similar problem in two dimensions, a simple formula for the exact solution is not known, whereas the Ritz method enables us to find the solution as the sum of a double trigonometric series in a completely analogous way as before.

The situation changes substantially in the case of equations with nonconstant coefficients:

Example 2. On the interval $I = (0, 1)$, let us consider the boundary value problem

$$-u'' + xu = 1,$$
$$u(0) = u(1) = 0.$$

We find in Tab. 24.1 that the weak solution belongs to $V = H_0^1(I)$ and fulfils the equality

$$\int_0^1 (u'v' + xuv)\,\mathrm{d}x = \int_0^1 v\,\mathrm{d}x$$

for all functions $v \in V$. We choose again

$$\varphi_k(x) = \sin k\pi x$$

for basis functions and obtain

$$a_{kj} = \int_0^1 kj\pi^2 \cos k\pi x \cos j\pi x\,\mathrm{d}x + \int_0^1 x \sin k\pi x \sin j\pi x\,\mathrm{d}x \tag{5}$$

for the elements a_{kj} of the stiffness matrix.

Due to the second integral, the orthogonality in the energy scalar product does not hold any more. We obtain

$$a_{kk} = \tfrac{1}{2}k^2\pi^2 + \tfrac{1}{4}$$

and, for $j \neq k$,

$$a_{kj} = 0 \qquad \text{for } k+j \text{ even},$$
$$a_{kj} = \frac{-4jk}{\pi^2(j^2-k^2)^2} \qquad \text{for } k+j \text{ odd}.$$

Because the right-hand side is determined in the same manner as in Example 1, we have the system

$$\begin{aligned}
\left(\frac{1}{2}\pi^2 + \frac{1}{4}\right)c_1 \quad & -\frac{8}{9}\frac{1}{\pi^2}c_2 \quad && -\frac{16}{225}\frac{1}{\pi^2}c_4 \ \ldots = \frac{2}{\pi}, \\
-\frac{8}{9}\frac{1}{\pi^2}c_1 + \left(2\pi^2 + \frac{1}{4}\right)c_2 \quad & -\frac{24}{25}\frac{1}{\pi^2}c_3 \quad && \ldots = 0, \\
-\frac{24}{25}\frac{1}{\pi^2}c_2 + \left(\frac{9\pi^2}{2} + \frac{1}{4}\right)c_3 \quad & && \ldots = \frac{2}{3\pi}, \\
& && \ldots\ldots\ldots\ldots
\end{aligned}$$

for the coefficients c_k, and this system is to be solved. It is obvious, that for this purpose an appropriate numerical method of Chap. 30 has to be used.

Putting, for example, $N = 4$, we obtain the values

$$c_1 = 0{\cdot}12280\,, \quad c_2 = 0{\cdot}00058\,, \quad c_3 = 0{\cdot}00475\,, \quad c_4 = 0{\cdot}00002\,.$$

REMARK 1. If the coefficients were more complicated functions, some of the methods of § 13.13 for the computation of the integrals in (5) would have to be used. It is clear that such an application of the Ritz method is possible only with the use of computers.

(b) The Galerkin Method

If the form $a(u, v)$ is not symmetric, Problem 24.1.1 is not equivalent with the problem of finding the minimum of the functional (1) and the Ritz method cannot be used. Therefore, we make the condition (24.1.4) a basis for another approximate method, the so-called *Galerkin method.*

We choose a finite-dimensional subspace V_N of the space V, look for an approximate solution y_N in V_N, and require that

$$a(y_N, v_N) = (f, v_N) \tag{6}$$

be fulfilled for all functions $v_N \in V_N$. If we choose a basis $\varphi_1, \ldots, \varphi_N$ in V_N and assume y_N in the form (2), i.e.

$$y_N = \sum_{k=1}^{N} c_k \varphi_k ,$$

where c_k are unknown coefficients, we use (6) in such a way that we successively take all basis functions for v_N. We again obtain the system (3) (with $a(\varphi_k, \varphi_j) \neq \neq a(\varphi_j, \varphi_k)$, in general). Assuming (24.1.2) and (24.1.3), this system has a unique solution again. Theorem 2 (Céa's lemma) holds for the corresponding approximate solution y_N and we can make conclusions about the quality of the approximation, or about the convergence of the Galerkin method as before.

If $a(u, v)$ is a symmetric bilinear form, the Galerkin method turns into the Ritz method. In such a case we often speak about the *Ritz-Galerkin method.* The Galerkin method can, however, be applied to a much wider class of problems, especially to problems of the form

$$Au = f ,$$

where A is an arbitrary operator defined on $D(A)$. This operator need not even be linear. The approximate solution y_N is supposed to belong to a finite-dimensional subspace V_N of $D(A)$ and one requires that

$$(Ay_N - f, v_N) = 0$$

holds for all elements $v_N \in V_N$. The idea is that the discrepancy between the left and right-hand side of the equation, i.e. the residual $Ay_N - f$, will be small for

a sufficiently "rich" subspace V_N because it will be orthogonal to all elements of this rich subspace V_N.

If we apply this idea, employed before in the linear case, to nonlinear problems, we get the formulation (6) where the form $a(u, v)$ is nonlinear with respect to its first argument. Partial answers to questions on convergence in the important case of second-order differential equations in two dimensions are contained in [139], [140], [160], [165].

24.3. The Finite Element Method

The *finite element method* (FEM) is, as we just mentioned, the Ritz or Galerkin method for Problem 24.1.1 with a special choice of the finite-dimensional space V_N and its basis. The choice of the basis does not influence the approximate solution y_N but it does influence the stiffness matrix $\boldsymbol{A}$. FEM tries to obtain the stiffness matrix with properties favourable from the viewpoint of the numerical solution of the corresponding system. We attain this goal making many elements of the matrix $\boldsymbol{A}$ vanish.

In the FEM, the construction of finite-dimensional subspaces V_h (we will now use this notation instead of V_N) of the space $V \subset H^k$ has the following features:

1) the given domain is partitioned into many subdomains,
2) the elements v_h of the subspace V_h are polynomials (or other functions of simple form) on each subdomain of the partition constructed,
3) in the space V_h, such a basis may be chosen that its elements are functions different from zero only on small domains (that are unions of a few subdomains of the partition).

We can say that the FEM suitably joins properties of variational and finite-difference methods and has the following advantages:

1) it enables the construction of irregular nets (i.e. the partition into subdomains),
2) it enables, in an easy way, the construction of methods of order higher than second,
3) it enables the construction of methods for the solution of equations of order higher than second.

(a) Decompositions and Finite Elements

The *decomposition* of the given domain Ω into a finite number of subdomains $K_1, \ldots, K_N$ is carried out in such a way that single subdomains of the decomposition are simple geometric figures so that polynomials can be defined on them in

a simple way. The subdomains of the decomposition are called *elements*. A decomposition in the FEM has to possess these properties:

1) each element is a closed subdomain with a nonempty interior,
2) the union of all elements is equal to the closure of the given domain, i.e.

$$\bigcup_{i=1}^{N} K_i = \overline{\Omega},$$

3) two different elements K_1, K_2 of the decomposition have disjoint interiors K_1^0, K_2^0, i.e.

$$K_1 \neq K_2 \quad \Longrightarrow \quad K_1^0 \cap K_2^0 = \emptyset,$$

4) the boundary of each element is a Lipschitz boundary (cf. Remark 22.4.10).

An important characteristic of the decomposition is the *discretization parameter* h that is defined as the maximum diameter of all elements of the decomposition. The decomposition is therefore denoted by $\mathcal{T}_h$ even though for the same h different decompositions may be given. This is generally adopted notation and we will use it, too. Let us further suppose that

5) in the one-dimensional case, the elements are closed intervals, in several dimensions, we will suppose that elements are polygons or polyhedra. Some sides or faces may be curved.

Further, we require that

6) every side (face) of the element K is either a part of the boundary S or a side (face) of another element K_1. In Fig. 24.1, the a) admissible and b) inadmissible coupling of two elements are shown,

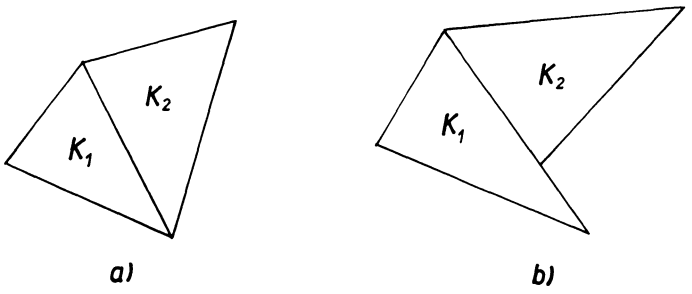


Fig. 24.1 a, b.

7) the interior of every side (face) of an arbitrary element K of the decomposition has no common points with the set Γ_0 of the boundary points from S in which the type of the boundary condition changes.

REMARK 1. If we choose an element of a common type described further, it is possible that the property 2) cannot be satisfied. We mention in § 24.4 how to treat such cases and how to estimate the corresponding error.

One-dimensional elements are, as we have already said, closed intervals. In the multidimensional case the choice may vary significantly. We will use triangles and quadrangles for the elements of decomposition in the plane. In three dimensions, we will use tetrahedra, hexahedra and prismatic elements.

Because the two-dimensional problems with triangular elements are most frequent the term *triangulation* is used instead of the term decomposition also for other types of decompositions. The terms *partition* or *net* or *grid* are utilized in the same sense, too.

For our further use, polynomials must be defined on the individual elements of decomposition. Therefore, we introduce the *finite element* as a triplet $\{K, P, \Sigma\}$, where

1) K is an element of the decomposition,
2) P is a space of polynomials (or other simple functions) defined on K,
3) Σ is a finite set of function values or values of derivatives of the polynomial from the space P at specific points of K.

The points, where the function values or the values of derivatives are specified, are called *nodes* and the function values or the values of derivatives at those points are called *nodal parameters.* The set Σ is therefore the set of nodal parameters. The number of the elements of the set Σ is also called the *number of degrees of freedom* of the given finite element. The nodal parameter itself is often called a *degree of freedom.* This applies particularly if the nodal parameter is another functional on the space P, e.g. $\int_K p(x)\,\mathrm{d}x$ for $p \in P$.

REMARK 2. The basis functions for the FEM are usually chosen in such a way that exactly one of their nodal parameters equals unity and all others are zero. Because the approximate solution is a linear combination of the basis functions, its function values or its values of derivatives are exactly the coefficients of this linear combination. The coefficients c_k computed from (24.2.3) thus have their own meaning as the values or values of the derivatives of the approximate solution.

It is very important to ensure that the set of the nodal parameters has the property of *unisolvency.* That means that to an arbitrary choice of values of nodal parameters, there exists one and only one polynomial from the space P whose nodal parameters attain the given values.

Now, we will give a survey of the most frequently used finite elements, i.e. the triplets consisting of the element of the decomposition of the domain, of the space of polynomials defined on it and of the set of nodal parameters. The survey will be

given separately for one, two and three dimensions. The space of all polynomials defined on K of degree not greater than k will be denoted by $P_k = P_k(K)$. We will not consider finite elements for higher dimensions.

(α) One-dimensional Finite Elements

The element, i.e. the element of the decomposition, is here an interval $[a, b]$. We accompany individual elements with illustrating figures where nodes, the supports of nodal parameters, are displayed. If the function value at the node is specified, we mark the node with a bold dot. A bold dot with a circle around it means that the function value and the value of the first order derivative are specified, and, finally, a bold dot with k circles around represents the function value and the values of derivatives from the first to the k-th order.

Linear element

P is the space P_1.

Nodal parameters: $p(a)$, $p(b)$ – values at boundary points (Fig. 24.2).

Quadratic element

P is the space P_2.

Nodal parameters: $p(a)$, $p\left(\frac{a+b}{2}\right)$, $p(b)$ (Fig. 24.3).

General Lagrange element

P is the space P_n.

Nodal parameters: $p(a)$, $p(a+\Delta)$, $p(a+2\Delta)$, …, $p(b)$, where $\Delta = (b-a)/n$ (Fig. 24.4 for $n = 4$).

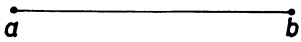


Fig. 24.2.

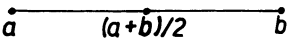


Fig. 24.3.

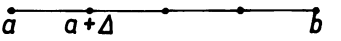


Fig. 24.4.

Cubic Hermite element

P is the space P_3.

Nodal parameters: $p(a)$, $p'(a)$, $p(b)$, $p'(b)$ (Fig. 24.5).

General Hermite element

P is the space P_{2n-1}.

Nodal parameters: $p(a)$, $p'(a)$, …, $p^{(n-1)}(a)$, $p(b)$, $p'(b)$, …, $p^{(n-1)}(b)$ (Fig. 24.6 for $n = 3$).

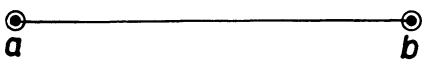


Fig. 24.5.

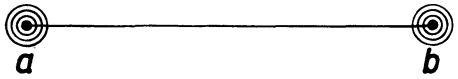


Fig. 24.6.

It follows from the properties of interpolation that all sets of nodal parameters for all finite elements shown are unisolvent.

For the reason of further manipulation, it is convenient to define finite elements on the *reference interval*, the interval [0, 1]. There is a one-to-one correspondence between the general interval and the reference interval given by

$$x = (b - a)\xi + a = a(1 - \xi) + b\xi\,,$$

where $x \in [a, b]$, $\xi \in [0, 1]$, or by the inverse transformation

$$\xi = \frac{x - a}{b - a}\,.$$

Because the linear transformation conserves the polynomial character of a function and its degree as well, it is sufficient to define finite elements only on the reference interval. In this notation, we have on the reference interval:

Linear element

Basis functions: $p_1(\xi) = 1 - \xi$, $p_2(\xi) = \xi$ (Fig. 24.7a).

Quadratic element

Basis functions: $p_1(\xi) = 2\xi^2 - 3\xi + 1$, $p_2(\xi) = 4\xi(1 - \xi)$, $p_3(\xi) = 2\xi^2 - \xi$ (Fig. 24.7b).

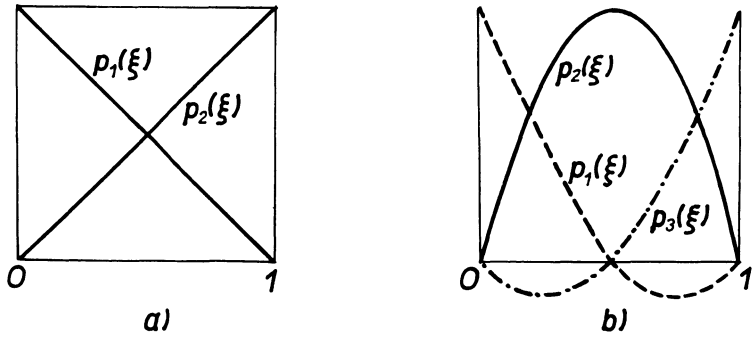


Fig. 24.7 a, b.

General Lagrange element

Basis functions: $p_k(\xi) = l_k^{(n)}(\xi)$, $k = 1, 2, \ldots, n+1$, where $l_k^{(n)}(\xi)$ is the elementary Lagrange interpolation polynomial (cf. § 32.6) for equidistant partition with n intervals.

Cubic Hermite element

Basis functions: $p_{00}(\xi) = 2\xi^3 - 3\xi^2 + 1$, $\quad p_{01}(\xi) = \xi^3 - 2\xi^2 + \xi$,
$p_{10}(\xi) = -2\xi^3 + 3\xi^2$, $\quad p_{11}(\xi) = \xi^3 - \xi^2$.

The value of the j-th derivative of the polynomial p_{ij} at the point i is equal to one, all other nodal parameters are zero.

For a general Hermite element the basis functions are not shown.

(β) Two-dimensional Finite Elements

(A) Triangular elements

We shall systematically use here the reference triangle T with the vertices $V_1 = (1, 0)$, $V_2 = (0, 1)$, $V_3 = (0, 0)$. The transformation of this triangle onto an arbitrary triangle with vertices $A_1 = (x_1, y_1)$, $A_2 = (x_2, y_2)$, $A_3 = (x_3, y_3)$ is given by

$$\begin{aligned} x &= (x_1 - x_3)\xi + (x_2 - x_3)\eta + x_3\,, \\ y &= (y_1 - y_3)\xi + (y_2 - y_3)\eta + y_3 \quad (\xi, \eta) \in T\,. \end{aligned} \tag{1}$$

The inverse transformation is given by

$$\begin{aligned} \xi = p_1(x, y) &= \frac{1}{J}[(y_2 - y_3)x - (x_2 - x_3)y + x_2y_3 - x_3y_2]\,, \\ \eta = p_2(x, y) &= \frac{1}{J}[(y_3 - y_1)x - (x_3 - x_1)y + x_3y_1 - x_1y_3]\,, \end{aligned} \tag{2}$$

where

$$J = \begin{vmatrix} 1, & x_1, & y_1 \\ 1, & x_2, & y_2 \\ 1, & x_3, & y_3 \end{vmatrix} = x_2y_3 - x_3y_2 - x_1y_3 + x_3y_1 + x_1y_2 - x_2y_1\,.$$

The magnitude of J equals double the area of the triangle $A_1A_2A_3$. We introduce one more polynomial,

$$p_3(x, y) = 1 - p_1(x, y) - p_2(x, y)\,,$$

i.e.

$$p_3(x, y) = \frac{1}{J}[(y_1 - y_2)x - (x_1 - x_2)y + (x_1y_2 - x_2y_1)]\,.$$

It is easily verified that

$$p_i(x_j, y_j) = \delta_{ij}, \quad i, j = 1, 2, 3, \text{ where } \delta_{ij} = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases}$$

The finite elements on a triangle, i.e. the space of polynomials, nodes, nodal parameters, and the corresponding basis functions will be introduced on the reference triangle in the variables ξ, η. To simplify the formulae, we use the notation $\vartheta = 1 - \xi - \eta$. All quantities are transformed onto a general triangle by (2). To indicate nodes and nodal parameters in figures we use the above introduced symbols. The circles now mean that all derivatives with respect to all variables up to the corresponding order are prescribed. If a value of the normal derivative on the boundary of the element is prescribed at a node we use an arrow at the node in the direction of the normal. The node itself is not marked if the function value is not prescribed at it.

Linear element

P is the space P_1.

Nodal parameters: the values at the vertices of the triangle (Fig. 24.8).

Basis functions: $p_1 = \xi$, $p_2 = \eta$, $p_3 = \vartheta$.

Quadratic element

P is the space P_2.

Nodal parameters: the values at the vertices and at the midpoints of the sides of the triangle (Fig. 24.9).

Basis functions: $p_1 = \xi(2\xi - 1)$, $p_2 = \eta(2\eta - 1)$, $p_3 = \vartheta(2\vartheta - 1)$,
$p_4 = 4\xi\eta$, $p_5 = 4\eta\vartheta$, $p_6 = 4\xi\vartheta$.

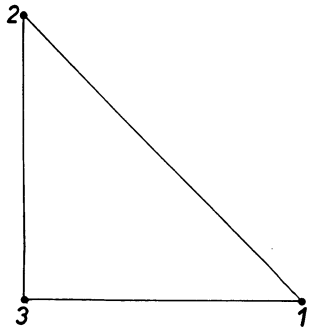


Fig. 24.8.

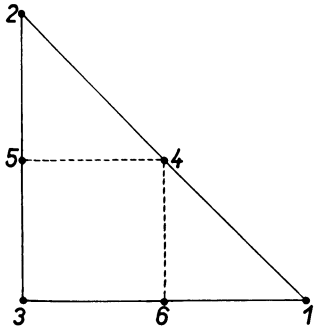


Fig. 24.9.

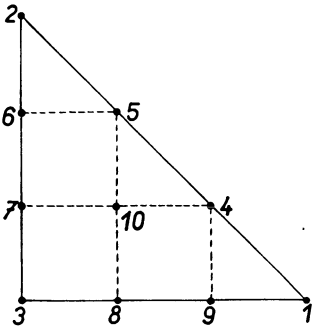


Fig. 24.10.

Cubic Lagrange element

P is the space P_3.

Nodal parameters: the values at the vertices, at the points of trisection of the sides and at the centre of gravity of the triangle (Fig. 24.10).

Basis functions:
$$p_1 = \tfrac{1}{2}\xi(3\xi-1)(3\xi-2), \quad p_2 = \tfrac{1}{2}\eta(3\eta-1)(3\eta-2),$$
$$p_3 = \tfrac{1}{2}\vartheta(3\vartheta-1)(3\vartheta-2), \quad p_4 = \tfrac{9}{2}\xi\eta(3\xi-1),$$
$$p_5 = \tfrac{9}{2}\xi\eta(3\eta-1), \quad p_6 = \tfrac{9}{2}\eta\vartheta(3\eta-1),$$
$$p_7 = \tfrac{9}{2}\eta\vartheta(3\vartheta-1), \quad p_8 = \tfrac{9}{2}\xi\vartheta(3\vartheta-1),$$
$$p_9 = \tfrac{9}{2}\xi\vartheta(3\xi-1), \quad p_{10} = 27\,\xi\eta\vartheta.$$

If we replace the nodal parameter u_{10}, i.e. the value at the centre of gravity, by the value

$$u_{10} = -\frac{1}{6}\sum_{i=1}^{3} u_i + \frac{1}{4}\sum_{i=4}^{9} u_i$$

and if we use the above basis functions we obtain a space that does not contain all polynomials of degree 3. The same result is obtained by omitting the centre of gravity as a node and utilizing the basis functions $r_i = p_i - \frac{1}{6}p_{10}$, $i = 1, 2, 3$, and $r_i = p_i + \frac{1}{4}p_{10}$, $i = 4, 5, \ldots, 9$. By this procedure, called the *elimination of an interior parameter*, we obtain a new element with nine nodal parameters and an incomplete space of polynomials of degree 3. The order of approximation by such elements will be theoretically equal to the order of approximation by elements from the space of all polynomials of degree only 2 but, practically, its error does not differ very much from the error attainable with the full space of polynomials of degree 3. (Cf. section (c) of this paragraph.)

General Lagrange element

P is the space P_m.

Nodal parameters: the values at the vertices of all triangles obtained by a partition of each side in m equal parts and by joining these division points by straight lines parallel to the sides of the triangle. We have $(m+1)(m+2)/2$ nodal parameters (Fig. 24.11 for $m = 4$).

Basis functions: the basis function for the vertex (1, 0) is
$$p_1 = \xi(m\xi-1)\Big(\frac{m}{2}\xi-1\Big)\ldots\Big(\frac{m\xi}{(m-1)}-1\Big).$$
We do not show the other basis functions.

Cubic Hermite element

P is the space P_3.

Nodal parameters: the function values and the values of both first derivatives at the vertices and the function value at the centre of gravity (Fig. 24.12).

Basis functions: $p_1 = \xi(3\xi - 2\xi^2 - 7\eta\vartheta)$, $p_2 = \eta(3\eta - 2\eta^2 - 7\xi\vartheta)$,
$p_3 = \vartheta(3\vartheta - 2\vartheta^2 - 7\xi\eta)$, $p_4 = 27\,\xi\eta\vartheta$.
(It is $p_i(A_i) = 1$, $i = 1, \ldots, 4$.)
$r_1 = \xi[\eta(\vartheta - \xi) + \vartheta(\eta - \xi)]$, $r_2 = \xi\eta(\eta - \vartheta)$, $r_3 = \xi\vartheta(\vartheta - \eta)$.
(It is $\dfrac{\partial r_i(A_i)}{\partial x} = 1$, $i = 1, 2, 3$.)
$s_1 = \xi\eta(\xi - \vartheta)$, $s_2 = \eta[\vartheta(\xi - \eta) + \xi(\vartheta - \eta)]$, $s_3 = \eta\vartheta(\vartheta - \xi)$.
(It is $\dfrac{\partial s_i(A_i)}{\partial y} = 1$, $i = 1, 2, 3$.)

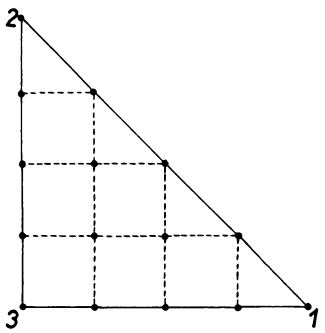


Fig. 24.11.

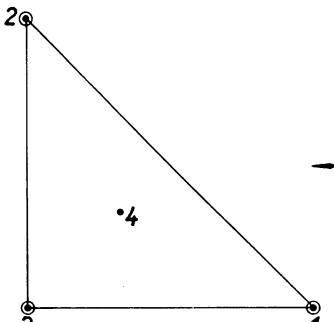


Fig. 24.12.

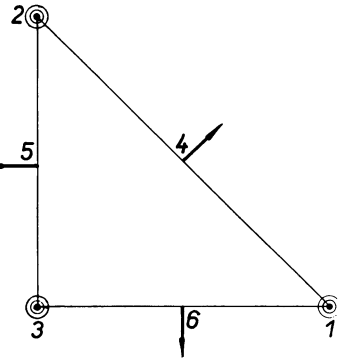


Fig. 24.13.

Quintic Hermite element

P is the space P_5.

Nodal parameters: the function values and the values of all derivatives of order 1 and 2 at the vertices and the values of the normal derivatives at the midpoints of the sides of the triangle; together 21 nodal parameters (Fig. 24.13).

Basis functions: We denote by $p_i^{(j,k)}$ the basis function for which $\dfrac{\partial^{j+k} p_i^{(j,k)}}{\partial\xi^j \partial\eta^k} = 1$ at the point V_i and all its other nodal parameters are zero. We have $V_4 = \left(\frac{1}{2}, \frac{1}{2}\right)$, $V_5 = \left(0, \frac{1}{2}\right)$, $V_6 = \left(\frac{1}{2}, 0\right)$:

$p_1^{(0,0)} = 6\xi^5 - 15\xi^4 + 10\xi^3 + 15\xi^2\eta^2\vartheta$,

$p_2^{(0,0)} = 6\eta^5 - 15\eta^4 + 10\eta^3 + 15\xi^2\eta^2\vartheta$,

$p_3^{(0,0)} = \vartheta^2[6\vartheta^3 - 15\vartheta^2 + 10\vartheta + 30\xi\eta(\xi + \eta)]$;

$$p_1^{(1,0)} = -3\xi^5 + 7\xi^4 - 4\xi^3 - \tfrac{7}{2}\xi^2\eta^2\vartheta,$$
$$p_2^{(1,0)} = \xi\eta^2(-8\eta^2 + \tfrac{37}{2}\xi\vartheta + 5\xi^2 + 14\eta - 5),$$
$$p_3^{(1,0)} = \xi\vartheta^2(-3\xi^2 + 6\xi\eta - 8\eta^2 - 2\vartheta + 3);$$
$$p_1^{(0,1)} = \xi^2\eta(-8\xi^2 + \tfrac{37}{2}\eta\vartheta + 5\eta^2 + 14\xi - 5),$$
$$p_2^{(0,1)} = -3\eta^5 + 7\eta^4 - 4\eta^3 - \tfrac{7}{2}\xi^2\eta^2\vartheta,$$
$$p_3^{(0,1)} = \eta\vartheta^2(-3\eta^2 + 6\xi\eta - 8\xi^2 - 2\vartheta + 3);$$
$$p_1^{(2,0)} = \tfrac{1}{2}\xi^3(1-\xi)^2 + \tfrac{1}{4}\xi^2\eta^2\vartheta,$$
$$p_2^{(2,0)} = \xi^2\eta^2(\tfrac{1}{2}\eta + \tfrac{5}{4}\vartheta),$$
$$p_3^{(2,0)} = \tfrac{1}{2}\xi^2\vartheta^2(1 - \xi + 2\eta);$$
$$p_1^{(1,1)} = \xi^2\eta(4\xi^2 + 7\xi\eta + 5\eta^2 - 6\xi - 7\eta + 2),$$
$$p_2^{(1,1)} = \xi\eta^2(5\xi^2 + 7\xi\eta + 4\eta^2 - 7\xi - 6\eta + 2),$$
$$p_3^{(1,1)} = \xi\eta(2\vartheta^2 - \vartheta^2);$$
$$p_1^{(0,2)} = \xi^2\eta^2(\tfrac{1}{2}\xi + \tfrac{5}{4}\vartheta),$$
$$p_2^{(0,2)} = \tfrac{1}{2}\eta^3(1-\eta)^2 + \tfrac{1}{4}\xi^2\eta^2\vartheta,$$
$$p_3^{(0,2)} = \tfrac{1}{2}\eta^2\vartheta^2(1 - \eta + 2\xi);$$
$$p_4 = -4\,.\,\sqrt{2}\,.\,\xi^2\eta^2\vartheta \text{ (here, we have } \frac{\partial p_4}{\partial \xi}\,.\,\frac{1}{\sqrt{2}} + \frac{\partial p_4}{\partial \eta}\,.\,\frac{1}{\sqrt{2}} = 1$$
at the point V_4),
$$p_5^{(1,0)} = 16\xi\eta^2\vartheta^2,$$
$$p_6^{(0,1)} = 16\xi^2\eta\vartheta^2.$$

This element with 21 nodal parameters is convenient for the solution of fourth-order problems. No triangular element shown above can be utilized for the solution of fourth-order problems in the sense of § 24.2. This element may be simplified similarly as the cubic Hermite element by omitting the three nodal parameters at the midpoints of sides. This simplification is, however, not substantial and one must take into account that the corresponding stiffness matrix will be of large order with many nonzero entries.

(B) Rectangular elements

We use the square with the vertices (0, 0), (0, 1), (1, 1), (1, 0) as the reference element. The transformation of this square in the ξ, η-plane onto the rectangle $ABCD$ in the x, y-plane with the sides parallel with the coordinate axes is simple,

$$x = h_1\xi + x_0\,, \quad y = h_2\eta + y_0 \tag{3}$$

(where x_0, y_0, h_1, h_2 have an obvious meaning). From the rectangle onto the square we transform by

$$\xi = \frac{x - x_0}{h_1}\,, \quad \eta = \frac{y - y_0}{h_2}\,. \tag{4}$$

The space of the polynomials used on rectangular elements are, as a rule, constructed as products of polynomials in single variables.

Bilinear Lagrange element

P is the space P_{bil}, the subspace of P_2 consisting of polynomials of the form

$$a + b\xi + c\eta + d\xi\eta\,.$$

Nodal parameters: the function values at the vertices (Fig. 24.14).

Basis functions: $p_1 = (1-\xi)(1-\eta)$, $p_2 = \xi(1-\eta)$, $p_3 = \xi\eta$, $p_4 = (1-\xi)\eta$.

Biquadratic Lagrange element

P is the space P_{biq}, the subspace of P_4 consisting of polynomials of the form $p_2(\xi)q_2(\eta)$, where p_2 and q_2 are arbitrary quadratic polynomials in one variable.

Nodal parameters: the function values at the vertices, at the midpoints of sides and at the centre of gravity (Fig. 24.15).

Basis functions:

$$\begin{aligned}
p_1 &= (\xi-1)(2\xi-1)(\eta-1)(2\eta-1),\\
p_2 &= \xi(2\xi-1)(\eta-1)(2\eta-1),\\
p_3 &= \xi(2\xi-1)\eta(2\eta-1),\\
p_4 &= (\xi-1)(2\xi-1)\eta(2\eta-1),\\
p_5 &= 4\xi(1-\xi)(\eta-1)(2\eta-1),\\
p_6 &= 4\xi(2\xi-1)\eta(1-\eta),\\
p_7 &= 4\xi(1-\xi)\eta(2\eta-1),\\
p_8 &= 4(\xi-1)(2\xi-1)\eta(1-\eta),\\
p_9 &= 16\xi(1-\xi)\eta(1-\eta).
\end{aligned}$$

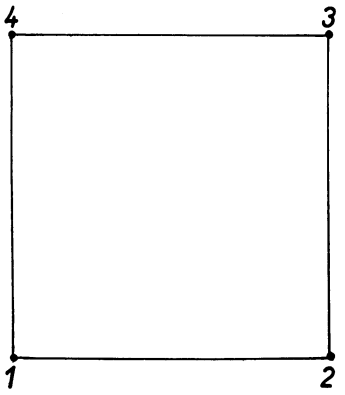


Fig. 24.14.

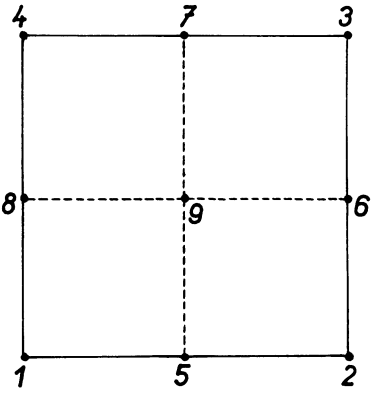


Fig. 24.15.

Rectangular Hermite elements

are constructed in a similar way. They are relatively complicated and we do not show them, see [83].

(C) Isoparametric elements

These elements are introduced with the aim to approximate a nonpolygonal domain better than only by a polygon. A more exact approximation of the domain implies a more exact approximate solution, cf. § 24.4. The construction of these elements uses a linear combination of basis functions on the reference element corresponding to any above introduced element. We then take the coordinates of the nodes of the resulting element for the coefficients of this linear combination and in this manner we obtain the transformation of the reference element onto the resulting element which may be a curvilinear triangle or a curvilinear quadrangle.

Isoparametric triangular quadratic element

We start from the quadratic Lagrange element and take the coordinates of the vertices A_1, A_2, A_3 of the curvilinear triangle (i.e. arbitrary three points not lying on the straight line) and those of the midpoints A_5 and A_6 of the sides A_1A_3 and A_2A_3, respectively, for the coefficients of basis functions corresponding to the values at the vertices and at the midpoints, of the sides $(\frac{1}{2}, 0)$ and $(0, \frac{1}{2})$, respectively. These two sides will be straight. We take the coordinates of an arbitrary point A_4 lying in the angle $A_1A_3A_2$ for the coefficients of the basis function corresponding to the midpoint $(\frac{1}{2}, \frac{1}{2})$. We thus get the transformation:

$$x = x_1\xi(2\xi - 1) + x_2\eta(2\eta - 1) + x_3\vartheta(2\vartheta - 1) + x_4 4\xi\vartheta + \\ + \frac{x_1 + x_3}{2} 4\xi\vartheta + \frac{x_2 + x_3}{2} 4\eta\vartheta = x_1\xi + x_2\eta + x_3\vartheta + \left(x_4 - \frac{x_1 + x_2}{2}\right) 4\xi\eta,$$

$$y = y_1\xi(2\xi - 1) + y_2\eta(2\eta - 1) + y_3\vartheta(2\vartheta - 1) + y_4 4\xi\eta + \\ + \frac{y_1 + y_3}{2} 4\xi\vartheta + \frac{y_2 + y_3}{2} 4\eta\vartheta = y_1\xi + y_2\eta + y_3\vartheta + \left(y_4 - \frac{y_1 + y_2}{2}\right) 4\xi\eta.$$

The points $(1, 0)$, $(0, 1)$, $(0, 0)$, $(\frac{1}{2}, \frac{1}{2})$ in the (ξ, η)-plane are transformed onto the points $A_1 = (x_1, y_1)$, $A_2 = (x_2, y_2)$, $A_3 = (x_3, y_3)$ and $A_4 = (x_4, y_4)$, respectively. This transformation will be denoted by F; we thus have $K = F(T)$. The reference triangle is denoted by T and the resulting curvilinear triangle by K. The transformation F is bilinear and, therefore, it transforms straight lines parallel to the axes ξ and η into straight lines (Fig. 24.16).

A curvilinear element is defined as a triplet $\{K, P_K, \Sigma\}$, where $K = F(T)$, P_K is the set of functions (not of polynomials!) of the form

$$p(x, y) = \hat{p}(F^{-1}(x, y)), \quad (x, y) \in K, \quad \hat{p} \in P_2.$$

This corresponds to the use of a quadratic polynomial $\hat{p}$ on the reference triangle. The set Σ of nodal parameters is the set of function values at the points A_i, $i = 1, \ldots, 6$. We have to assume that the inverse transformation F^{-1} exists at every

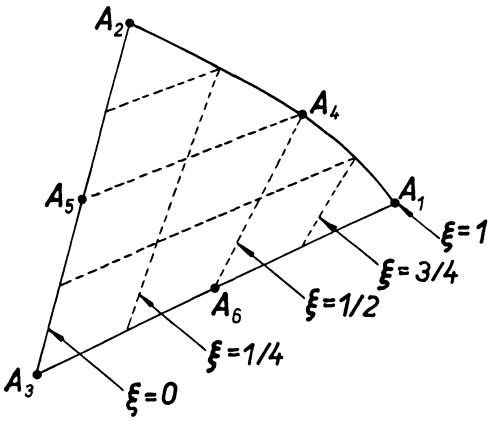


Fig. 24.16.

point $x \in K$. This fact depends on the position of A_4. If the point A_4 is the midpoint of A_1A_2, the inverse transformation exists (the triangle is not curved). It can be shown (see [242]) that the inverse transformation exists unless the point A_4 is too close to A_3A_1 or A_3A_2.

The curved part of boundary of the element K is a part of a parabola (since the points A_1, A_2, A_4 do not lie on a straight-line). From the form of the transformation F, it is seen that the curved part is determined only by A_1, A_2, A_4 (and not by A_3).

Example 1. Let $A_1 = (3, 2)$, $A_2 = (2, 3)$, $A_3 = (1, 1)$, $A_4 = (3, 3)$. The transformation F is

$$x = 3\xi + 2\eta + \vartheta + 2\xi\eta\,,$$
$$y = 2\xi + 3\eta + \vartheta + 2\xi\eta\,, \quad (\xi, \eta) \in T\,.$$

The inverse transformation exists and is given by

$$\xi = \tfrac{1}{4}\{2x - 2y - 3 + [4(x - y)^2 + 4(x + y) + 1]^{1/2}\}\,,$$
$$\eta = \tfrac{1}{4}\{-2x + 2y - 3 + [4(x - y)^2 + 4(x + y) + 1]^{1/2}\}\,.$$

The triangle $K = F(T)$ is shown in Fig. 24.17.

Isoparametric quadrangular bilinear element

The transformation of the reference square onto the quadrangle $A_1A_2A_3A_4$, $A_1 = (x_1, y_1)$, $A_2 = (x_2, y_2)$, $A_3 = (x_3, y_3)$, $A_4 = (x_4, y_4)$ is

$$x = x_1(1 - \xi)(1 - \eta) + x_2\xi(1 - \eta) + x_3\xi\eta + x_4(1 - \xi)\eta\,,$$
$$y = y_1(1 - \xi)(1 - \eta) + y_2\xi(1 - \eta) + y_3\xi\eta + y_4(1 - \xi)\eta\,.$$

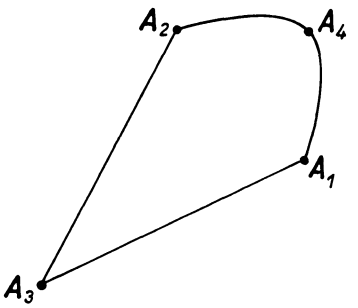


Fig. 24.17.

An arbitrary quadrangle can result. The space P is the space of functions $p(x, y) = \hat{p}(F^{-1}(x, y))$, where $\hat{p}(\xi, \eta) \in P_{\text{bil}}$.

Nodal parameters: function values at the vertices.

The inverse transformation exists if and only if the resulting quadrangle is convex. These elements do not have curved sides. They are introduced for the approximation of those straight parts of the boundary that are not parallel to the coordinate axes.

Isoparametric quadrangular biquadratic element

We use the basis functions p_j of the biquadratic Lagrange element for the transformation of the reference square to a curved quadrangle. We put

$$x = \sum_{j=1}^{9} p_j(\xi, \eta) x_j ,$$

$$y = \sum_{j=1}^{9} p_j(\xi, \eta) y_j .$$

The points (x_j, y_j) are the nodes of the resulting quadrangle. The space P is the space of functions $p(x, y) = \hat{p}(F^{-1}(x, y))$, where $\hat{p}(\xi, \eta) \in P_{\text{biq}}$.

It is necessary also here to guarantee the existence of the inverse transformation by fixing the points $A_5, \ldots, A_8$ not very far from the midpoints of the sides of the quadrangle and the point A_9 from its centre of gravity. The resulting curved element is in Fig. 24.18.

(γ) Three-dimensional Finite Elements

We will show here only a few simple elements of a low degree. The way how to construct higher elements is obvious from the preceding text but we will not present them because one obtains complicated formulae here. The transformations

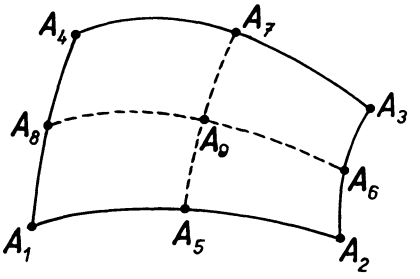


Fig. 24.18.

from reference figures onto general elements of given type are constructed similarly as in the two-dimensional case (cf. (1), (2) and (3), (4)).

(A) Tetrahedral elements

Linear tetrahedral element

The reference element is the tetrahedron with vertices (1, 0, 0), (0, 1, 0), (0, 0, 1) and (0, 0, 0).

P is the space P_1.

Nodal parameters: function values at the vertices (Fig. 24.19).

Basis functions: $p_1 = \xi$, $p_2 = \eta$, $p_3 = \zeta$, $p_4 = 1 - \xi - \eta - \zeta$.

(B) Hexahedral elements

Trilinear hexahedral element

The reference element is the cube with vertices (0, 0, 0), (1, 0, 0), (1, 1, 0), (0, 1, 0), (0, 0, 1), (1, 0, 1), (1, 1, 1) and (0, 1, 1).

P is the space P_{tril}, the subspace of P_3 consisting of all polynomials of the form

$$(a_1\xi + b_1)(a_2\eta + b_2)(a_3\zeta + b_3).$$

Nodal parameters: function values at the vertices (Fig. 24.20).

Basis functions:

$$\begin{aligned} p_1 &= (1-\xi)(1-\eta)(1-\zeta), & p_2 &= \xi(1-\eta)(1-\zeta), \\ p_3 &= \xi\eta(1-\zeta), & p_4 &= (1-\xi)\eta(1-\zeta), \\ p_5 &= (1-\xi)(1-\eta)\zeta, & p_6 &= \xi(1-\eta)\zeta, \\ p_7 &= \xi\eta\zeta, & p_8 &= (1-\xi)\eta\zeta. \end{aligned}$$

Onto an arbitrary hexahedron with quadrangular faces, the reference cube is usually transformed isoparametrically by

$$x = \sum_{j=1}^{8} p_j(\xi, \eta, \zeta)x_j\,,$$

$$y = \sum_{j=1}^{8} p_j(\xi, \eta, \zeta) y_j ,$$

$$z = \sum_{j=1}^{8} p_j(\xi, \eta, \zeta) z_j .$$

If the resulting hexahedron is convex, this transformation has an inverse and it is possible to transform the basis functions, too.

(C) Prismatic elements

Prismatic pentahedral element

The reference element is the triangular prism with the vertices (0, 0, 0), (1, 0, 0), (0, 1, 0), (0, 0, 1), (1, 0, 1) and (0, 1, 1).

P is the space of all polynomials of the form $(a_1\xi + b_1\eta + c_1)(a_2\zeta + b_2)$.

Nodal parameters: function values at the vertices (Fig. 24.21).

Basis functions: $p_1 = (1 - \xi - \eta)(1 - \zeta)$, $p_2 = \xi(1 - \zeta)$, $p_3 = (1 - \zeta)$,
$p_4 = (1 - \xi - \eta)\zeta$, $p_5 = \xi\zeta$, $p_6 = \eta\zeta$.

It is possible to obtain more general pentahedra by an isoparametric transformation of the reference prism.

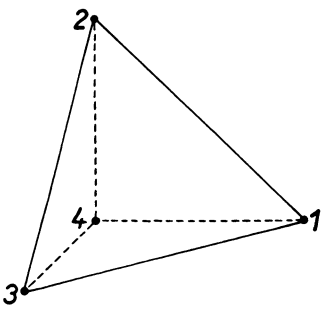


Fig. 24.19.

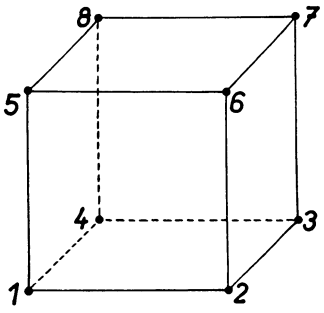


Fig. 24.20.

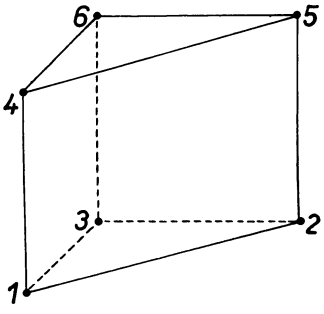


Fig. 24.21.

(b) The Finite Element Spaces

We will now show how to construct finite-dimensional spaces V_h of functions defined on the entire domain with the help of finite elements, i.e. of the triplets: element – space of polynomials – set of nodal parameters. The closure of the domain is the union of the elements of the decomposition and we recall that not every domain given is of this type but it can be approximated by such domains then.

At the construction we utilize the basic principle of the FEM, the one-to-one correspondence of the set of values of nodal parameters to the polynomial of given

type on a single element. Choosing the values of the nodal parameters for all elements of a given decomposition of the domain, the corresponding polynomials on each element are given. If some node is shared by several finite elements, the same values of the nodal parameters at this node for all the corresponding finite elements are to be taken. The function v_h defined on the entire domain is set up from the polynomials on the individual elements by taking

$$v_h|_K(A) = \sum_{j=1}^{n} u_j p_j(A), \tag{5}$$

where p_j are basis functions on the element K, u_j are the values of the nodal parameters and A an arbitrary point of K.

It is obvious that the set of all the functions of the form (5) (i.e. for all possible choices of the values of the nodal parameters u_j) is a linear set of finite dimension. The question arises, if this linear set is a finite-dimensional subspace V_h of the space V that appears in the definition of the weak solution of the basic Problem 24.1.1.

The answer is given by

Theorem 1. *Let $\mathcal{T}_h$ be a decomposition of Ω into convex elements. Let V_h be the subspace of such functions of $L_2(\Omega)$ that $v_h|_K$ is a polynomial for any $K \in \mathcal{T}_h$. Then $V_h \subset H^1(\Omega)$ if and only if $V_h \subset C(\overline{\Omega})$.*

In other words, a piecewise polynomial function belongs to $H^1(\Omega)$ if and only if it is continuous on $\overline{\Omega}$. (See [83], [280].)

REMARK 3. The above theorem can be generalized to the case of decompositions with nonconvex elements and with functions that are not polynomials. Such a generalization is employed when isoparametric elements are used.

REMARK 4. If a piecewise polynomial function v_h is to belong to the space $H^2(\Omega)$ then all its first derivatives have to belong to $H^1(\Omega)$ and, according to Theorem 1, they therefore have to be continuous.

Theorem 1 and Remark 4 thus enable us to reduce the investigation whether a piecewise polynomial function belongs to a Sobolev space, or not, to the investigation of continuity. For second order problems, we take $v_h \in V \subset H^1(\Omega)$, and for fourth order problems $v_h \in V \subset H^2(\Omega)$.

A substantial step in the investigation of continuity is to examine the continuity across the boundary of neighbouring elements. If we use elements only of one type from those shown above to decompose the domain, the corresponding function v_h is continuous.

If we want to combine elements of different types conserving the continuity, we have to fulfil certain conditions. We will show it for Lagrange elements, i.e. for the elements where only function values are used as nodal parameters. Only one nodal parameter corresponds to each node.

If we have two neighbouring elements K_1 and K_2 (let us recall that such elements have an entire side (face) in common) we require that the set of nodes of K_1 belonging to K_2 coincides with the set of nodes of K_2 lying in K_1.

It is thus impossible to couple the linear and the quadratic triangular elements because the midpoint of the side of the quadratic element is not a node for the linear element. The nodes of the linear element are marked in Fig. 24.22 by a cross, the nodes of the quadratic element by a bold dot. The node U violates the above condition.

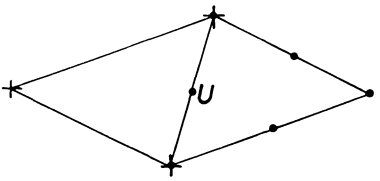


Fig. 24.22.

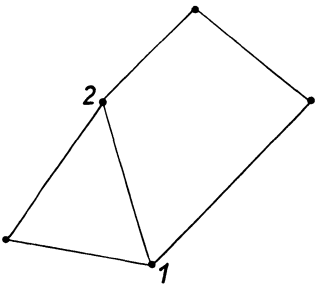


Fig. 24.23.

Let us put $S = K_1 \cap K_2$ (S is the common boundary of the elements K_1 and K_2) and let N_S be the set of common nodes on S. We further require that the space of the values on S of the polynomial defined on K_1 coincide with the space of the values on S of the polynomial defined on K_2. We denote this space of boundary values by P_S. Finally, we require that the boundary value $p \in P_S$ vanishing at all the nodes of N_S, i.e. satisfying

$$p(A_i) = 0 \quad \text{for } A_i \in N_S\,,$$

be zero identically.

The last requirement guarantees that the boundary values are uniquely determined only by the values at the boundary nodes.

If all the above requirements are fulfilled, the function v_h from (5) is continuous on the common boundary S of the elements K_1 and K_2.

Example 2. Let the domain be decomposed into a linear triangular element and an isoparametric quadrangular bilinear element, see Fig. 24.23. Let us investigate

the continuity of the function v_h in this domain. The common boundary S of both the elements contains the nodes common for both of them. These nodes are the boundary points of the common boundary and form the set N_S. We will show that for both the elements the boundary value of the corresponding space P depends only on the function values at these two points.

We start from the reference triangle and suppose that the original S_T of the common boundary has the equation $\xi + \eta = 1$. The basis function corresponding to the nodal parameter at the node (0, 0) lying outside S_T is $p_3 = 1 - \xi - \eta$ and vanishes on S_T. For $\xi + \eta = 1$ we thus have

$$v_h(\xi, 1 - \xi) = u_1\xi + u_2(1 - \xi)\,,$$

where u_1, u_2 are the values of the nodal parameters at the nodes of N_S. This fact holds as well after a linear transformation on an arbitrary triangle.

We start the investigation of the isoparametric quadrangular bilinear element on the reference square assuming that the original S_C of the common boundary is $\eta = 0$. The basis functions corresponding to the nodal parameters outside S_C are $p_3 = \xi\eta$ and $p_4 = (1 - \xi)\eta$ and vanish for $\eta = 0$. We thus have

$$v_h(\xi, 0) = u_2(1 - \xi) + u_1\xi\,.$$

The nodal parameter u_1 now corresponds to the right-hand boundary point of S_C, the parameter u_2 to the left-hand one.

The fact that the boundary value of a function from the space P (such a function need not be a polynomial on the entire element) is a linear polynomial uniquely determined by the nodal parameters u_1 and u_2 is preserved by the isoparametric transformation.

If both the parameters are zero, the boundary value is zero, too. The function v_h is, therefore, continuous on the union of both the elements.

The only one of the two- and three-dimensional elements shown above, that yields the continuity of both first derivatives of the function v_h (in addition to the continuity of v_h itself), is the quintic Hermite triangular element. (In one dimension, the cubic Hermite element is sufficient for the continuity of the derivative.) The boundary value for the quintic element on one side of the triangle is determined by the values at the boundary points of this side and by the first and second order derivatives along the boundary (tangent derivatives) at these points. The basis functions corresponding to other nodes (including the midpoint of the side) vanish at the boundary points of the chosen side as well as their first and second order derivatives. The boundary value is a polynomial of degree five and is uniquely determined by the nodal parameters that are equal for both the neighbouring elements. The function v_h is thus continuous.

The continuity of the derivatives can be investigated more easily if we change the coordinate system and take the derivatives along the boundary (tangent derivatives) or perpendicular to the boundary (normal derivatives). Both these derivatives are polynomials of degree four (in one variable following the boundary). The tangent derivative is obviously continuous. The normal derivative is determined by its values at the boundary points of the side, by its tangent derivative at the boundary points (we know the values of all derivatives of the second order there) and the value at the midpoint. These are five parameters at the nodes on the chosen side and they just suffice to determine the fourth degree polynomial uniquely. The normal derivative is thus continuous, too.

(c) Convergence of the Finite Element Method

We now state theorems on the convergence and on the error estimates for the FEM. These theorems make use of Céa's Lemma, Theorem 24.2.2. For an easy application of this theorem, we assume:

a) the domain is a polygon (polyhedron),
b) if it is necessary to use the representing function w (in the case of non-homogeneous boundary conditions), the computation of the bilinear form $a(w, v)$ is performed exactly,
c) the system (24.2.3) is assembled and solved exactly, i.e.
   c_1) the integrals involved in the computation of the coefficients $a(\varphi_k, \varphi_j)$ and the right-hand side (f, φ_j) are computed exactly,
   c_2) the system is solved exactly; we do not consider errors caused by numerical solution of the system of linear algebraic equations.

The problem of the error estimate of the approximate solution is thus reduced to the question how to find the distance of the exact solution from the space V_h.

If we are to investigate the error behaviour for $h \to 0$, i.e. for the case when the discretization parameter h (the maximal diameter of the elements from the decomposition) converges to zero, it is necessary to introduce some definitions.

Definition 1. The set of decompositions $\mathcal{F} = \{\mathcal{T}_h\}$ is called the *system of decompositions* if, for every $\varepsilon > 0$, such a decomposition $\mathcal{T}_h \in \mathcal{F}$ exists that $h < \varepsilon$.

Usually we have to impose a more strict requirement:

Definition 2. The system of decompositions $\mathcal{F} = \{\mathcal{T}_h\}$ is called *regular*, if there exists such a constant $M > 0$ that for an arbitrary decomposition $\mathcal{T}_h \in \mathcal{F}$ and for an arbitrary element $K \in \mathcal{T}_h$, such a ball (or a circle) B_K of radius ϱ_K exists that $B_K \subset K$ and

$$Mh_K \leqq \varrho_K ,$$

where h_K is the diameter of K.

Example 3. Let K be a triangle and B_K a circle inscribed to K. This is the maximal circle such that $B_K \subset K$. We then have $\operatorname{tg} \frac{\alpha}{2} > \varrho_K/h_K \geqq \frac{1}{2} \operatorname{tg} \frac{\alpha}{2}$, where h_K is the maximal side and α the minimal angle of K. The system of triangular decompositions (triangulations) is thus regular if and only if the angles of all triangles in all decompositions are bounded from below. This is the *condition of the minimal angle.*

Regular systems of triangulations can be obtained e.g. in such a way that we divide each triangle of the original decomposition by three lines joining the midpoints of sides. We obtain four congruent triangles similar to the original one, see Fig. 24.24. Another evident way, to refine the triangle by the medians, is not suitable because it reduces the angles, see Fig. 24.25.

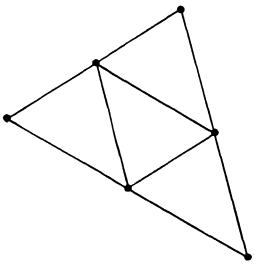


Fig. 24.24.

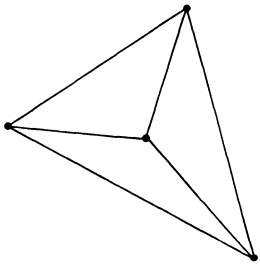


Fig. 24.25.

However, it is not possible to transfer this idea to three dimensions. We cannot decompose a tetrahedron into eight congruent tetrahedra. Nevertheless, it can be shown that regular systems of decomposition exist, see [279].

We can now state the convergence theorem for boundary-value problems for second-order differential equations.

Theorem 2. *Let u be the solution of Problem* 24.1.1 *and let the space $C^\infty(\overline{\Omega}) \cap V$ be dense in V (in the H^1 norm). Let $\mathcal{T} = \{\mathcal{T}_h\}$ be a regular system of decompositions and let the approximate solution u_h belong to the space $V_h = \{v \in V, H_0^1 \subset V \subset H^1; v|_K \in P_1(K)$ for all $K \in \mathcal{T}_h\}$. (The functions from V_h are thus continuous and piecewise linear.) Then*

$$\lim_{h \to 0} \|u - u_h\|_{1,\Omega} = 0\,. \tag{6}$$

If the solution is smoother, i.e. if $u \in H^p$ for $p > 1$ and if we use higher elements, we can find a quantitative error estimate.

Theorem 3. *Let the solution u of Problem* 24.1.1 *belong to $H^{s+1}(\Omega)$, where s is a positive integer. Let $\mathcal{T} = \{\mathcal{T}_h\}$ be a regular system of decompositions into simplexes (i.e. into triangles or tetrahedra) and let the approximate solution u_h belong*

to the space $V_h = \{v \in V,\ H_0^1 \subset V \subset H^1; v|_K = P_s(K)$ *for all* $K \in \mathcal{T}_h\}$. *Then there exist such constants* h_0 *and* $C > 0$ *that*

$$\|u - u_h\|_{1,\Omega} \leqq Ch^s\, |u|_{s+1,\Omega} \tag{7}$$

holds for an arbitrary decomposition $\mathcal{T}_h \in \mathcal{F}$, *where* $0 < h < h_0$.

The symbol $|u|_{k,\Omega}$ denotes the *seminorm* in the space $H^k(\Omega)$ defined by

$$|v|_{k,\Omega} = \left(\sum_{|i|=k} \left\|D^i v\right\|_{0,\Omega}^2\right)^{\frac{1}{2}}, \quad v \in H^k(\Omega).$$

The symbol D^i is an abbreviated notation for derivative, see Remark 22.4.10.

REMARK 5. Theorem 3 holds also for $u \in H^1$ but the estimate (7) is useless. Theorem 2 could be formulated for $u \in H^s$, $s > 1$, too, but the assertion of Theorem 3 is stronger because the convergence of the sequence of the approximate solutions (6) is a consequence of (7).

If we are interested only in the error of function values we use the norm of error in the space $H^0 = L_2$ instead of in H^1 and we have the following estimate:

Theorem 4. *Let* $\mathcal{F} = \{\mathcal{T}_h\}$ *be a regular system of decompositions into simplexes. Let* Ω *be convex and let the solution of Problem* 24.1.1 *belong to* $H^2(\Omega)$. *Let the approximate solution* u_h *belong to the space* $V_h = \{v \in V,\ H_0^1 \subset V \subset H^1;$ $v|_K \in P_1(K)$ *for all* $K \in \mathcal{T}_h\}$. *Then there exist such constants* h_0 *and* $C > 0$ *that*

$$\|u - u_h\|_{0,\Omega} \leqq Ch^2\, |u|_{2,\Omega}$$

for an arbitrary decomposition $\mathcal{T}_h \in \mathcal{F}$, *where* $0 < h < h_0$.

REMARK 6. If the solution in Theorem 4 is smoother and if we use higher elements (like in Theorem 3) we obtain the power h^{s+1} in the estimate.

REMARK 7. Theorems 2, 3 and 4 can be generalized to some other types of decompositions, e.g. rectangular or hexahedral.

The error of the approximate solution of a fourth order problem obtained with the use of the quintic Hermite element is estimated as follows:

Theorem 5. *Let* $\mathcal{F} = \{\mathcal{T}_h\}$ *be a regular system of triangulations. Let the solution of Problem* 24.1.1 *belong to* $H^6(\Omega)$. *Let the approximate solution* u_h *belong to the space* $V_h = \{v, \in V,\ H_0^2 \subset V \subset H^2; v|_K \in P_5(K)$ *for all* $K \in \mathcal{T}_h\}$. *(The functions*

from V_h are thus continuous along with their first derivatives.) Then there exist such constants h_0 and $C > 0$ that

$$\|u - u_h\|_{2,\Omega} \leqq Ch^4 |u|_{6,\Omega}$$

for an arbitrary decomposition $\mathcal{T}_h \in \mathcal{F}$, where $0 < h < h_0$.

24.4. Computational Aspects of the Finite Element Method

We made several assumptions in order to prove the convergence and the error estimates of the approximate solution with the mere use of Theorem 24.2.2. Now, we add several remarks to these assumptions.

If the domain is not polygonal, i.e. if it is not the union of the elements of the decomposition, even when isoparametric elements are used, it is necessary to approximate it. This can be carried out in different ways, the domain of decomposition can be inscribed to the given domain, or partially inscribed and partially circumscribed, etc. If the boundary of the given domain is smooth it is possible to describe the corresponding error. We refer the reader to [139], [140].

If it is necessary to use the function w representing the nonhomogeneous boundary conditions in the variational formulation (cf. Tab. 24.1), it is natural in the context of the FEM to choose this function so that $w \in V_h$, and this leads to an error. We choose, in fact, different functions for different triangulations. If we use linear elements in two dimensions we have a piecewise linear approximation of the boundary condition. The corresponding error is analyzed again in [139], [140].

The assembly and the solution of the system (24.2.3) or (24.2.6) is an important practical problem. The theory guarantees existence and uniqueness of the solution of the system but, because there can be thousands of unknown nodal parameters, the assembly of the matrix of the system is a certain organizational task that is to be solved during the coding of the problem. The numbering of the nodes and the nodal parameters plays an important role here and can influence the properties of the matrix of the system significantly.

It is necessary, too, to realize that the entries of the stiffness matrix and of the load vector are usually integrals that we are often unable to compute exactly. We therefore have to propose suitable methods for an approximate computation of these integrals, especially in two or three dimensions, and to find out their influence on the error of the approximate solution of the given boundary-value problem.

The Gauss quadrature formulae are mostly used in one dimension, see § 13.13. The basis for the construction of quadrature formulae in several dimensions is the following theorem:

Theorem 1. *Let K be an arbitrary element of the decomposition of a d-dimensional domain. Let $A_1, \ldots, A_p$ be points of K and let $c_1, \ldots, c_p$ be real numbers. Let*

$$\int_K f \, \mathrm{d}x = \sum_{i=1}^{p} c_i f(A_i)$$

hold for all polynomials f of degree k or less. Let $k+1 > d$ and $u \in W_1^{(k+1)}(K)$ (the Sobolev space of functions with all generalized derivatives of order $k+1$ integrable in K). We then have

$$\left| \int_K u \, \mathrm{d}x - \sum_{i=1}^{p} c_i u(A_i) \right| \leqq C h_K^{k+1} |u|_{k+1,1,K} ,$$

where C is independent of K and of u, h_K is the diameter of K and $|u|_{k+1,1,K}$ denotes the seminorm of the function u in the space $W_1^{(k+1)}(K)$ (see [83], p.11).

We show now some simplest quadrature formulae used in two or three dimensions. We remark that the corresponding accuracy estimates are not always a direct consequence of Theorem 1.

The following quadrature formulae are used for triangles:

$$\int_T u \, \mathrm{d}x \approx \frac{1}{3} \operatorname{mes} T [u(A_1) + u(A_2) + u(A_3)]$$

where A_i, $i = 1, 2, 3$, are the vertices and $\operatorname{mes} T$ the area of the triangle. The formula is exact for $u \in P_1(T)$ and has the order of accuracy h^2; Fig. 24.26.

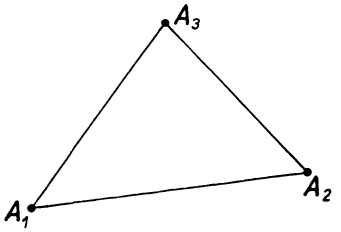


Fig. 24.26.

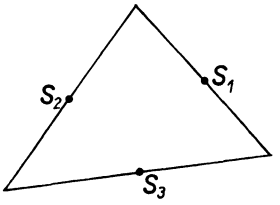


Fig. 24.27.

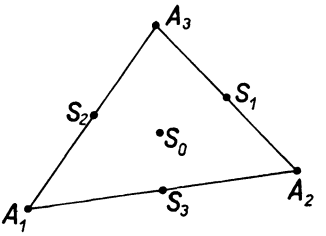


Fig. 24.28.

$$\int_T u \, \mathrm{d}x \approx \frac{1}{3} \operatorname{mes} T [u(S_1) + u(S_2) + u(S_3)]$$

where S_i, $i = 1, 2, 3$, are the midpoints of the sides of the triangle. The formula is exact for $u \in P_2(T)$ and has the order of accuracy h^3; Fig. 24.27.

$$\int_T u \, \mathrm{d}x \approx \operatorname{mes} T \{c_1 [u(A_1) + u(A_2) + u(A_3)] + c_2 [u(S_1) + u(S_2) + u(S_3)] + c_3 u(S_0)\}$$

where A_i and S_i are as above, S_0 is the centre of gravity of the triangle, $c_1 = \frac{3}{60}$, $c_2 = \frac{8}{60}$, $c_3 = \frac{27}{60}$. The formula is exact for $u \in P_3(T)$ and has the order of accuracy h^4; Fig. 24.28.

The formula

$$\operatorname{mes} K \,.\, u(G)$$

where G is the centre of gravity, is used for the simplex (i.e. the triangle in two and the tetrahedron in three dimensions). The formula is exact for $u \in P_1(K)$ and has the order of accuracy h^2.

The influence of numerical quadrature on the accuracy of the approximate solution of a two-dimensional problem is described, e.g., by the following theorem:

Theorem 2. *Let Ω be a polygonal domain and let a regular system of triangulations be given on Ω. Let u and u_h be the exact and the approximate solution, respectively, of a second order elliptic boundary-value problem. Let*

$$\|u - u_h\|_{1,\Omega} = O(h^k) \,.$$

Further, let $a^*(\varphi_i, \varphi_j)$ be the entries of the stiffness matrix $\boldsymbol{A}^*$ computed by numerical quadrature and, similarly, (f^*, φ_j) be the components of the numerically computed load vector $\boldsymbol{f}^*$ (see (24.2.3)). *Let $u_h^* = \sum_{k=1}^{N} c_k^* \varphi_k$ be the approximate solution with $\boldsymbol{c}^* = (c_1^*, \ldots, c_N^*)^{\mathrm{T}}$ satisfying*

$$\boldsymbol{A}^* \boldsymbol{c}^* = \boldsymbol{f}^* \,.$$

Let the quadrature formula used be exact for polynomials of degree $2k-2$ or less and let the coefficients of the equation, its right-hand side and its solution be sufficiently smooth. Then we have

$$\|u - u_h^*\|_{1,\Omega} = O(h^k) \,.$$

We conclude this paragraph with an example of the use of the FEM.

Example 1. We look for the solution of the differential equation

$$-\Delta u = 1$$

on the square Ω with the vertices (0, 0), (1, 0), (1, 1), (0, 1) and with the boundary condition $u = 0$ on its boundary.

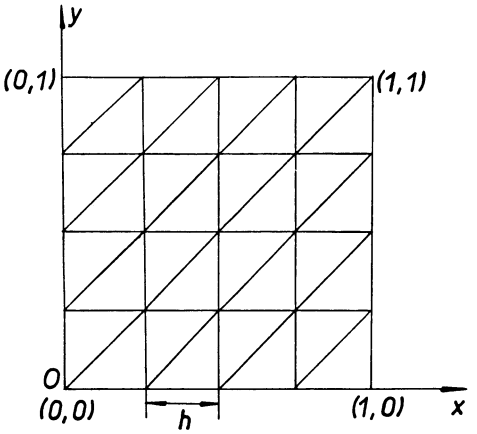


Fig. 24.29.

We use a triangular net obtained from a square net with the mesh-size $h = 1/n$ by dividing each square by the diagonal parallel to the bisector of the first quadrant (see Fig. 24.29 with $n = 4$) and linear triangular elements.

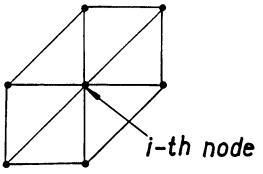


Fig. 24.30.

We first compute the matrix $\boldsymbol{A}_N$ and the right-hand side $\boldsymbol{b}_N$ according to (24.2.3). Each basis function corresponds to one nodal parameter, in our case to the function value at the node in the interior of the domain Ω. The basis function φ_i is thus different from zero only on the subdomain R_i, the union of the six triangles of decomposition (see Fig. 24.30). Consequently, the values $a(\varphi_i, \varphi_j)$ can be nonzero only if the subscripts i and j correspond to the same node ($i = j$) or to the neighbouring nodes (i.e. to such ones that are directly connected by a mesh line). The domain of integration for equal subscripts is the domain R_i, for different subscripts the domain Q_{ij} composed of two triangles of the decomposition (some of possible forms of Q_{ij} are shown in Fig. 24.31).

We have, using Tab. 24.1,

$$a(\varphi_i, \varphi_i) = \int_{R_i} \left[\left(\frac{\partial \varphi_i}{\partial x} \right)^2 + \left(\frac{\partial \varphi_i}{\partial y} \right)^2 \right] \mathrm{d}x\mathrm{d}y = \sum_{j=1}^{6} \int_{T_j} \left[\left(\frac{\partial \varphi_i}{\partial x} \right)^2 + \left(\frac{\partial \varphi_i}{\partial y} \right)^2 \right] \mathrm{d}x\mathrm{d}y .$$

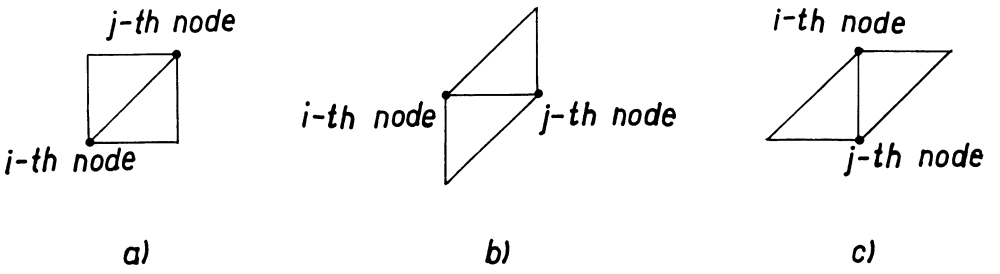


Fig. 24.31 a, b, c.

Triangles constituting R_i are denoted by T_j, $j = 1, \ldots, 6$. We carry out the computation separately on the individual triangles T_j, and the function φ_i is the basis function of a linear finite element on each of them.

The triangle T_k with the vertices (x_i, y_i), $(x_i + h, y_i)$ and $(x_i + h, y_i + h)$ is transformed onto the reference triangle T_{ref} by

$$\begin{aligned} \xi &= p_1(x, y) = \frac{x - x_i}{h} - \frac{y - y_i}{h}, \\ \eta &= p_2(x, y) = \frac{y - y_i}{h}. \end{aligned} \tag{1}$$

The i-th node (x_i, y_i) is transformed into the origin. We have

$$\varphi_i(x, y)\,|_{T_k} = p_3(\xi, \eta) = p_3(p_1(x, y), p_2(x, y)),$$

where $p_3(\xi, \eta) = 1 - \xi - \eta$, and, consequently,

$$\frac{\partial \varphi_i}{\partial x}\bigg|_{T_k} = \frac{\partial p_3}{\partial \xi}\frac{\partial p_1}{\partial x} + \frac{\partial p_3}{\partial \eta}\frac{\partial p_2}{\partial x} = -\frac{1}{h}.$$

Similarly, we find that $\dfrac{\partial \varphi_i}{\partial y}\bigg|_{T_k} = 0$. Finally,

$$\int_{T_k}\left[\left(\frac{\partial \varphi_i}{\partial x}\right)^2 + \left(\frac{\partial \varphi_i}{\partial y}\right)^2\right] \mathrm{d}x\mathrm{d}y = \int_{T_k} \frac{1}{h^2}\,\mathrm{d}x\mathrm{d}y = \frac{1}{h^2}\int_{T_{\text{ref}}} J\,\mathrm{d}\xi\mathrm{d}\eta = \frac{1}{h^2}.h^2.\frac{1}{2} = \frac{1}{2},$$

where J is the Jacobian of the transformation inverse to (1).

We proceed on all the triangles from R_i in a similar way. We find the value 1 for two of them and the value $\frac{1}{2}$ for the others.

Similarly,

$$\int_{T_k} \left(\frac{\partial \varphi_i}{\partial x} \frac{\partial \varphi_j}{\partial x} + \frac{\partial \varphi_i}{\partial y} \frac{\partial \varphi_j}{\partial y} \right) \mathrm{d}x\mathrm{d}y = 0 \quad \text{for the position of nodes in Fig. 24.31 a),}$$

$$= -\tfrac{1}{2} \quad \text{for the position of nodes in Fig. 24.31 b) or c).}$$

Together, we have

$$a(\varphi_i, \varphi_i) = 4\,, \quad a(\varphi_i, \varphi_j) = -1 \quad \text{for neighbours along axes,}$$

$$= 0 \quad \text{for diagonal neighbours.}$$

The computation of the components of the right-hand side vector gives

$$\int_{T_k} 1 \,.\, \varphi_i \,\mathrm{d}x\mathrm{d}y = h^2/6$$

for all the triangles constituting R_i. Together, we have

$$b_{N,i} = \int_{R_i} 1 \,.\, \varphi_i \,\mathrm{d}x\mathrm{d}y = h^2\,.$$

We assemble the matrix $\boldsymbol{A}_N$ employing the numbering of nodes from the left to the right and the rows from top to bottom. We introduce the square tridiagonal matrix of order $n-1$

$$\boldsymbol{M}_n = \begin{bmatrix} 4, & -1, & 0, & 0, & \dots \\ -1, & 4, & -1, & 0, & \dots \\ 0, & -1, & 4, & -1, & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots, & & & -1, & 4 \end{bmatrix}$$

and denote the identity matrix of the same order by $\boldsymbol{I}$. The matrix $\boldsymbol{A}_N$ of the system is block tridiagonal of block order $n-1$ and we have

$$\boldsymbol{A}_N = \begin{bmatrix} \boldsymbol{M}_n, & -\boldsymbol{I}, & \boldsymbol{0}, & \dots \\ -\boldsymbol{I}, & \boldsymbol{M}_n, & -\boldsymbol{I}, & \dots \\ \dots & \dots & \dots & \dots \\ \dots, & & -\boldsymbol{I}, & \boldsymbol{M}_n \end{bmatrix}.$$

The order of $\boldsymbol{A}_N$ is $N = (n-1)^2$. The right-hand side vector $\boldsymbol{b}_N$ has $(n-1)^2$ equal components of magnitude h^2.

The solution of the system

$$\boldsymbol{A}_N \boldsymbol{c} = \boldsymbol{b}_N\,,$$

where $\boldsymbol{c} = (c_1, \ldots, c_N)^{\mathrm{T}}$, yields the coefficients of the linear combination $\sum_{i=1}^{n} c_i \varphi_i$, i.e. the approximate solution of our problem. The coefficients c_i are the values of the approximate solution at the nodes owing to the properties of the basis functions. The system was solved for different n by the Gaussian elimination method (see § 30.1). The approximate values at the points $(\frac{1}{4}, \frac{1}{2})$ and $(\frac{1}{2}, \frac{1}{2})$ are as follows:

n	4	8	16	20
$u_h(\frac{1}{4}, \frac{1}{2})$	0·05469	0·05664	0·05716	0·05722
$u_h(\frac{1}{2}, \frac{1}{2})$	0·07031	0·07278	0·07345	0·07353

Finishing this paragraph, we recall that the FEM was generalized in many directions. It is used to solving nonlinear problems, in particular problems given by variational inequalities, it is used with spaces of finite elements not satisfying the condition $V_N \subset V$ (*nonconforming elements*), etc. The reader can find these applications in the bibliography shown at the beginning of this Chapter. A very comprehensive software for solving problems by the FEM has also been developed. These programs are very often products of professionals and thus of very good quality. We recommend the reader to use this software for solving his practical problems.

24.5. Computation of Eigenvalues and Eigenfunctions by the Finite Element Method

Variational methods, particularly the FEM, can be successfully applied to the numerical computation of eigenvalues and eigenfunctions of boundary-value problems for differential equations.

Let A be a linear (unbounded) symmetric positive definite operator with the domain of definition $D(A)$ in the space H and let us look for such a number λ and such an element $u \neq 0$ that

$$Au = \lambda u .$$

The existence of λ and u is proved, similarly as in the case of boundary-value problems, starting from the concept of the weak solution of an eigenvalue problem that we are now going to introduce:

Let $a(u, v)$ be a symmetric bilinear form defined on a space V, $H_0^k \subset V \subset H^k$, where H^k is a suitably chosen Sobolev space (§ 24.1). We suppose that the form $a(u, v)$ fulfils the conditions (24.1.2) and (24.1.3) and that

$$a(u, v) = (Au, v)$$

for all elements $u, v \in D(A)$.

The space V is the space H_A from Remark 22.6.10 with an equivalent norm. We find in Tab. 24.1 which bilinear form and which space V correspond to the differential operator and the given boundary conditions.

Definition 1. We say that λ is an *eigenvalue of the operator* A and that $u \in V$, $u \neq 0$, is the corresponding *eigenelement* (*eigenvector*) if

$$a(u, v) = \lambda(u, v) \tag{1}$$

holds for all $v \in V$.

REMARK 1. Because the space V is a subspace of a Sobolev space we speak in what follows about *eigenfunctions* instead of eigenelements, even though some of the further results hold for a general Hilbert space.

Problem 1. *We are to find such a number* λ *and such a function* $u \in V$, $u \neq 0$, *that* (1) *holds.*

Theorem 22.6.11 says that this eigenproblem has a countable set of positive eigenvalues and that a finite number of linearly independent eigenfunctions correspond to each of them. The system of all linearly independent eigenfunctions is complete in the space V. The eigenvalues and eigenfunctions are solutions of a sequence of minimization problems for the Rayleigh quotient

$$R(u) = \frac{a(u, u)}{(u, u)}, \quad u \neq 0,$$

on suitable subspaces of V.

REMARK 2. The assumption of symmetry of the bilinear form is important. It guarantees, among others, that the eigenvalues are *real*. The nonsymmetric case is more difficult and we will not deal with it.

If we apply the idea of the Ritz method to Problem 1 we obtain a method called the *Rayleigh-Ritz method*. We choose a finite-dimensional subspace V_N of V and look for pairs (λ_N, u_N) fulfilling

$$a(u_N, v_N) = \lambda_N(u_N, v_N) \quad \text{for all functions } v_N \in V_N. \tag{2}$$

We choose a basis $\varphi_1, \ldots, \varphi_N$ in the space V_N. We assume the eigenfunction u_N in the form $u_N = \sum_{i=1}^{N} c_i \varphi_i$ and take only basis functions for the functions v_N in (2)

(we know that it is sufficient). In this way, we obtain a generalized algebraic eigenproblem

$$\boldsymbol{A}_N \boldsymbol{c} = \lambda_N \boldsymbol{B}_N \boldsymbol{c}\,, \tag{3}$$

for the (nonzero) coefficient vector $\boldsymbol{c} = (c_1, \ldots, c_N)^{\mathrm{T}}$. The entries of the symmetric positive definite matrices $\boldsymbol{A}_N$ and $\boldsymbol{B}_N$ are given by

$$\begin{aligned} a_{N,i,j} &= a(\varphi_i, \varphi_j)\,, \\ b_{N,i,j} &= (\varphi_i, \varphi_j)\,. \end{aligned} \tag{4}$$

Solving the algebraic eigenproblem (3), we obtain N approximations of eigenvalues

$$0 < \lambda_{N,1} \leqq \lambda_{N,2} \leqq \cdots \leqq \lambda_{N,N}$$

and N eigenvectors, i.e. coefficient vectors $\boldsymbol{c}^{(k)} = (c_1^{(k)}, \ldots, c_N^{(k)})^{\mathrm{T}}$, $k = 1, \ldots, N$, and, from them, N approximations of the eigenfunctions

$$u_{N,k} = \sum_{i=1}^{N} c_i^{(k)} \varphi_i\,.$$

We have an error estimate for approximate eigenvalue:

Theorem 1. *Let the system of exact eigenfunctions as well as the system of approximate ones be orthonormal. Then for an arbitrary positive integer l and $N \geqq l$, there exists such a constant $C(l) > 0$ that*

$$\lambda_k \leqq \lambda_{N,k} \leqq \lambda_k + C(l) \sum_{r=1}^{k} \inf_{v_N \in V_N} \|u_r - v_N\|_V^2 \quad \text{for } k = 1, \ldots, l\,.$$

If we have a sequence of spaces V_N, $V_N \subset V_{N+1}$, $N = 1, 2, \ldots$, with the interpolation property, i.e.

$$\lim_{N \to \infty} \inf_{v_N \in V_N} \|u_k - v_N\|_V = 0 \quad \text{for } k = 1, \ldots, l\,,$$

then

$$\lim_{N \to \infty} \lambda_{N,k} = \lambda_k\,, \quad \lambda_{N,k} \geqq \lambda_{N+1,k}\,, \quad N = 1, 2, \ldots\,,$$

holds.

In a similar way it is possible to show that the approximation error of the eigenfunctions

$$\|u_{N,k} - u_k\|_V$$

is of the order of the interpolation error.

The most natural way how to choose the spaces V_N and their basis functions is the use of the FEM. For the simplest elements, we have the results:

Theorem 2. *Let Ω be a plane polygonal domain, let a regular system of triangulations* (Definition 24.3.2) *be given on it and let the linear triangular elements are chosen for the approximation. Then we have*

$$\lambda_k \leqq \lambda_{k,h} \leqq \lambda_k + C\lambda_k^2 h^2$$

for $k = 1, \ldots, \dim V_h$ and $h \in (0, C\lambda_k^{-1/2})$, where $C > 0$ is independent of h and k.

Theorem 3. *Let the assumptions of Theorem 2 be fulfilled and let λ_k be a simple eigenvalue. Let the system of exact eigenfunctions as well as the system of approximate ones be orthonormal. For a sufficiently small h, we then have*

$$\|u_k - u_{k,h}\|_0 \leqq C\lambda_k^2 h^2,$$
$$\|u_k - u_{k,h}\|_1 \leqq C\lambda_k h,$$

where $C > 0$ is independent of h and k.

See [449] for more details.

Example 1. Consider an equilateral triangle with its side equal to 1. Let an eigenvalue problem

$$\Delta u = \lambda u,$$
$$u = 0 \quad \text{on the boundary},$$

be given on the triangle and let us compute the minimum eigenvalue.

We use a uniform triangulation with the partition of a side into n parts (see Fig. 24.32, where $n = 4$) and linear triangular elements.

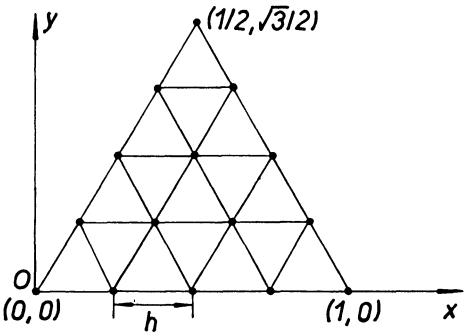


Fig. 24.32.

We have to compute the entries of the matrices $\boldsymbol{A}_N$ and $\boldsymbol{B}_N$ according to (4) before solving the algebraic problem (3). The order of the matrices is equal to the number of basis functions used and we have $N = (n-1)(n-2)/2$.

Each basis function corresponds to one nodal parameter and, because nodal parameters are function values at interior nodes of the grid, to one interior node. The basis function φ_i is thus nonzero only on a hexagonal domain S_i, the union of six triangles of the decomposition (see Fig. 24.33). It is obvious that the values $a(\varphi_i, \varphi_j)$ as well as (φ_i, φ_j) can be nonzero only if the subscripts i and j correspond to the same node ($i = j$) or to the adjacent nodes. The domain of integration reduces to the hexagon S_i for equal subscripts and to the intersection of the corresponding hexagons for different subscripts (referring, of course, to adjacent nodes). The intersection is a rhomb R_{ij} consisting of two triangles of the decomposition (Fig. 24.34).

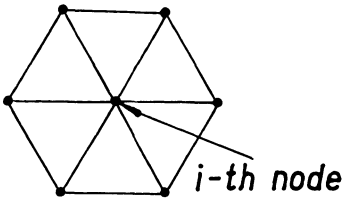


Fig. 24.33

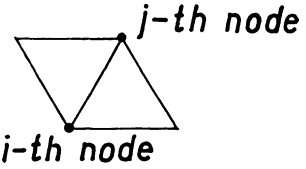


Fig. 24.34.

We thus have

$$a(\varphi_i, \varphi_i) = \int_{S_i} \left[\left(\frac{\partial \varphi_i}{\partial x} \right)^2 + \left(\frac{\partial \varphi_i}{\partial y} \right)^2 \right] \mathrm{d}x\mathrm{d}y = \\ = \sum_{j=1}^{6} \int_{T_j} \left[\left(\frac{\partial \varphi_i}{\partial x} \right)^2 + \left(\frac{\partial \varphi_i}{\partial y} \right)^2 \right] \mathrm{d}x\mathrm{d}y .$$

The triangles of the decomposition building up the hexagon S_i are denoted by T_j, $j = 1, \ldots, 6$. The computation is carried out over individual triangles of the decomposition where φ_i reduces to the corresponding basis function of the finite element.

The triangle T_k with the vertices (x_i, y_i), $(x_i + h, y_i)$ and $\left(x_i + \frac{1}{2}h, y_i + \frac{\sqrt{3}}{2}h\right)$ is transformed onto the reference triangle by

$$\xi = p_1(x, y) = \frac{x - x_i}{h} - \frac{y - y_i}{h\sqrt{3}}, \\ \eta = p_2(x, y) = \frac{2(y - y_i)}{h\sqrt{3}} . \tag{5}$$

The i-th node (x_i, y_i) is thus transformed into the origin. We have

$$\varphi_i(x, y)\,|_{T_k} = p_3(\xi, \eta) = p_3(p_1(x, y), p_2(x, y))\,,$$

where $p_3(\xi, \eta) = 1 - \xi - \eta$, and, consequently,

$$\frac{\partial \varphi_i}{\partial x}\,\Big|_{T_k} = \frac{\partial p_3}{\partial \xi}\frac{\partial p_1}{\partial x} + \frac{\partial p_3}{\partial \eta}\frac{\partial p_2}{\partial x} = -\frac{\partial p_1}{\partial x} - \frac{\partial p_2}{\partial x}\,.$$

We obtain $\dfrac{\partial \varphi_i}{\partial x}\,\Big|_{T_k} = -\dfrac{1}{h}$ and similarly $\dfrac{\partial \varphi_j}{\partial y}\,\Big|_{T_j} = -\dfrac{1}{h\sqrt{3}}$.

We have, therefore,

$$\int_{T_k}\left[\left(\frac{\partial \varphi_i}{\partial x}\right)^2 + \left(\frac{\partial \varphi_i}{\partial y}\right)^2\right] \mathrm{d}x\mathrm{d}y = \int_{T_k}\left(\frac{1}{h^2} + \frac{1}{3h^2}\right) \mathrm{d}x\mathrm{d}y =$$

$$= \frac{4}{3h^2}\int_{T_{\mathrm{ref}}} J\,\mathrm{d}\xi\mathrm{d}\eta = \frac{4}{3h^2}\cdot\frac{h^2\sqrt{3}}{2}\cdot\frac{1}{2} = \frac{\sqrt{3}}{3}\,.$$

The symbol J denotes the Jacobian of the transformation inverse to (5).

By a similar procedure, we obtain the same value if the triangle T_j is in the "reversed" position (with its "peak" pointing below).

We further find

$$\int_{T_k}\left(\frac{\partial \varphi_i}{\partial x}\frac{\partial \varphi_j}{\partial x} + \frac{\partial \varphi_i}{\partial y}\frac{\partial \varphi_j}{\partial y}\right) \mathrm{d}x\mathrm{d}y = -\frac{\sqrt{3}}{6}\,,$$

$$\int_{T_k} \varphi_i^2\,\mathrm{d}x\mathrm{d}y = \frac{h^2\sqrt{3}}{24}\,, \qquad \int_{T_k} \varphi_i\varphi_j\,\mathrm{d}x\mathrm{d}y = \frac{h^2\sqrt{3}}{48}$$

for i and j corresponding to adjacent nodes.

The contributions from all the triangles are equal because the grid is regular and we have

$$a(\varphi_i, \varphi_i) = 2\sqrt{3}\,, \qquad a(\varphi_i, \varphi_j) = -\frac{\sqrt{3}}{3}\,,$$

$$(\varphi_i, \varphi_i) = \frac{h^2\sqrt{3}}{4}\,, \qquad (\varphi_i, \varphi_j) = \frac{h^2\sqrt{3}}{24}\,.$$

Multiplying all the coefficients by $\sqrt{3}$ and numbering the nodes rowwise from top to bottom, we obtain

$$3^{\frac{1}{2}}\boldsymbol{A}_N = \begin{bmatrix}
6 & -1 & -1 & & & & & & & & & \\
-1 & 6 & -1 & -1 & -1 & & & & & & & \\
-1 & -1 & 6 & & -1 & -1 & & & & & & \\
 & -1 & & 6 & -1 & & -1 & -1 & & & & \\
 & -1 & -1 & -1 & 6 & -1 & & -1 & -1 & & & \\
 & & -1 & & -1 & 6 & & & -1 & -1 & & \\
 & & & \cdots & & & \cdots & & & & &
\end{bmatrix},$$

$$3^{\frac{1}{2}}\boldsymbol{B}_N = \frac{h^2}{8}\begin{bmatrix}
6 & 1 & 1 & & & & & & & & & \\
1 & 6 & 1 & 1 & 1 & & & & & & & \\
1 & 1 & 6 & & 1 & 1 & & & & & & \\
 & 1 & & 6 & 1 & & 1 & 1 & & & & \\
 & 1 & 1 & 1 & 6 & 1 & & 1 & 1 & & & \\
 & & 1 & & 1 & 6 & & & 1 & 1 & & \\
 & & & \cdots & & & \cdots & & & & &
\end{bmatrix}.$$

The partition into blocks corresponding to the rows of nodes is marked in both the matrices. The omitted entries are zeros. For the particular case of $n = 4$ with three interior points and $h = \frac{1}{4}$ (see Fig. 24.32), the matrices $\boldsymbol{A}_N$ and $\boldsymbol{B}_N$ are

$$3^{\frac{1}{2}}\boldsymbol{A}_3 = \begin{bmatrix} 6, & -1, & -1 \\ -1, & 6, & -1 \\ -1, & -1, & 6 \end{bmatrix}, \qquad 3^{\frac{1}{2}}\boldsymbol{B}_3 = \frac{1}{128}\begin{bmatrix} 6, & 1, & 1 \\ 1, & 6, & 1 \\ 1, & 1, & 6 \end{bmatrix}.$$

If the problem (3) is solved e.g. by a modification of the power method of Chap. 30 for different values of n, we obtain

n	4	8	12	20
approximate value of λ_1	64	55·395	53·851	53·072

The exact value is $\lambda_1 = \dfrac{16\pi^2}{3} \doteq 52{\cdot}638.$

24.6. Variational Methods for Numerical Solution of Parabolic Equations

We show here a further possible application of variational methods (see also § 18.10).

Let a parabolic equation

$$\frac{\partial u}{\partial t} + Au = f \tag{1}$$

be given, where A is an elliptic operator which is assumed to be symmetric and time independent, for simplicity. Equation (1) is completed by corresponding initial and boundary conditions.

The function u to be found thus depends, in addition to the space variables, also on the time variable t.

A typical parabolic problem is to find the solution of the differential equation

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(x, t) \quad \text{in } (a, b) \times I, \tag{2}$$

fulfilling the conditions

$$u(x, 0) = u_0(x) \quad \text{in } [a, b], \tag{3}$$

$$u(a, t) = 0, \quad u(b, t) = 0 \quad \text{in } \overline{I}, \tag{4}$$

where $I = (0, T)$.

The domain of definition of a parabolic problem is, in general, the Cartesian product of the domain Ω with the time interval I. The condition (3) is the *initial condition* of the problem, i.e. the prescribed value of the solution for $t = 0$, the conditions (4) are *boundary conditions*, given on $S \times \overline{I}$, where S is the boundary of Ω. In our case, the domain is the interval (a, b) and its boundary are its two endpoints. As a consequence of the conditions (3) and (4), we have $u_0(a) = u(a, 0) = 0$ and $u_0(b) = u(b, 0) = 0$. These requirements on initial and boundary conditions are called *consistency conditions*. The boundary condition can be nonhomogeneous, too, i.e., it can have a nonzero right-hand side.

Equations of the type (1) describe *nonstationary heat conduction problems, diffusion problems*, etc. First of all, it is necessary to give the definition of a weak solution of parabolic problems in order that we can use variational methods. The basic idea is that we move the time derivative $\frac{\partial u}{\partial t}$ to the right-hand side in (1) and, then, we proceed as with an elliptic problem according to § 24.1. We, therefore, introduce the bilinear form $a(u, v)$ corresponding to the operator A and the space V, $H_0^k \subset V \subset H^k$ (§ 24.1), where H^k is a suitable Sobolev space; the information necessary for most current problems can be found in Tab. 24.1. The bilinear

form $a(u, v)$ is symmetric for a symmetric operator. Let us assume that it satisfies (24.1.2) and (24.1.3).

Definition 1. We say that a function u is a *weak solution of the given parabolic problem* if $u(t) \in V$ for almost all $t \in (0, T)$ (i.e. for all $t \in (0, T)$ with a possible exception of a set of measure zero) and, moreover,

$$\left(\frac{\partial u}{\partial t}, v\right) + a(u, v) = (f, v) \quad \text{holds for all } v \in V \text{ and almost all } t \in (0, T) \tag{5}$$

and if, finally,

$$u(0) = u_0\,, \tag{6}$$

where $u_0 \in V$ is a given function, the initial condition.

REMARK 1. The foregoing definition says that the "variational condition" (5) is to be fulfilled for almost all t from the given time interval, namely for each value of t separately. It is necessary to guarantee that the derivative $\dfrac{\partial u}{\partial t}$ exists and that it is a linear continuous functional on the space V for each fixed t in order to justify the equality (5). The derivative of an abstract function is defined as the derivative of a mapping from the interval $(0, T)$ into the space V, according to § 22.8. Moreover, it is usually required that this derivative be an element of the space $L_2(\Omega)$ for almost all $t \in (0, T)$. Under these assumptions, the weak solution is well-defined. The details can be found in [390], [463].

REMARK 2. In the above formulation, the bilinear form $a(u, v)$ does not depend on the time variable t, i.e. its coefficients do not. In a similar way, the weak solution for a time dependent bilinear form $a(u, v)$ can be introduced, but it is then necessary to impose further requirements on it in order to guarantee the existence of the solution. Therefore, we have used the simpler formulation of the problem. We consider, also for simplicity, only symmetric bilinear forms $a(u, v)$.

The numerical solution of a parabolic equation requires two kinds of discretization, in time and in space. The procedures where only the time discretization or only the space discretization is carried out are called *semidiscrete methods.* Even though it is necessary to perform both the discretizations for solving the problem on a computer, the semidiscrete methods have their own importance because they make the theoretical investigations easier and, moreover, they enable us to combine analytic methods with the discretization.

We thus have two semidiscrete methods. The *time semidiscretization* or the *Rothe method*, is discussed in § 18.10; see [390] for more details. The *space semidiscretization* or *Galerkin semidiscrete method* starts directly from the methods discussed in the preceding paragraphs of this chapter and consists in the following:

We choose a finite-dimensional subspace V_N of the space V. Let the functions $\varphi_1, \ldots, \varphi_N$ form a basis in V_N. The approximate solution is assumed in the form

$$u_N = \sum_{k=1}^{N} c_k(t)\varphi_k .$$

The coefficients of the linear combination are now functions of time, in contrast to the elliptic case. We require the validity of (5) for the approximate solution u_N, taking successively φ_j, $j = 1, \ldots, N$, for v,

$$\left(\sum_{k=1}^{N} \frac{\mathrm{d}c_k(t)}{\mathrm{d}t}\varphi_k, \varphi_j\right) + \sum_{k=1}^{N} c_k(t)a(\varphi_k, \varphi_j) = (f, \varphi_j) \tag{7}$$

and we require that the initial values of c_k fulfil the equality

$$(u_0, \varphi_j) = \sum_{k=1}^{N} c_k(0)(\varphi_k, \varphi_j), \quad j = 1, \ldots, N. \tag{8}$$

Let us denote, as usual, the stiffness matrix by $\boldsymbol{A}$, its entries by $a_{kj} = a(\varphi_k, \varphi_j)$ and the matrix with the entries $b_{kj} = (\varphi_k, \varphi_j)$ by $\boldsymbol{B}$. The matrix $\boldsymbol{B}$ is often called the *mass matrix*. Further let $c(t)$ and $F(t)$ be the vectors

$$\boldsymbol{c}(t) = \begin{bmatrix} c_1(t) \\ c_2(t) \\ \vdots \\ c_N(t) \end{bmatrix} = (c_1(t), c_2(t), \ldots, c_N(t))^{\mathrm{T}},$$

$$\boldsymbol{F}(t) = \begin{bmatrix} (f, \varphi_1) \\ (f, \varphi_2) \\ \vdots \\ (f, \varphi_N) \end{bmatrix} = ((f, \varphi_1), (f, \varphi_2), \ldots, (f, \varphi_N))^{\mathrm{T}}.$$

Then (7) yields a system of ordinary differential equations

$$\boldsymbol{B}\boldsymbol{c}'(t) + \boldsymbol{A}\boldsymbol{c}(t) = \boldsymbol{F}(t) \tag{9}$$

for unknown functions $c_k(t)$ with the initial condition

$$\boldsymbol{B}\boldsymbol{c}(0) = \boldsymbol{U}^0, \tag{10}$$

where $\boldsymbol{U}^0 = (U_1^0, \ldots, U_N^0)^{\mathrm{T}}$ and $U_j^0 = (u_0, \varphi_j)$.

As a consequence of the symmetry of $a(u, v)$, both the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric and positive definite. Let $\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}$ be the Choleski factorization (cf. § 30.1) of the matrix $\boldsymbol{B}$. Let $\boldsymbol{d} = \boldsymbol{E}\boldsymbol{c}$ be the new unknown vector. The equation (9) yields

$$\boldsymbol{d}'(t) + \overline{\boldsymbol{A}}\boldsymbol{d}(t) = \boldsymbol{g}(t) \tag{11}$$

where $\overline{\boldsymbol{A}} = \boldsymbol{E}^{-\mathrm{T}}\boldsymbol{A}\boldsymbol{E}^{-1}$, $\boldsymbol{g} = \boldsymbol{E}^{-\mathrm{T}}\boldsymbol{F}$ and the initial condition (10) gives

$$\boldsymbol{d}(0) = \boldsymbol{d}^0\,, \tag{12}$$

where $\boldsymbol{d} = \boldsymbol{E}^{-\mathrm{T}}\boldsymbol{U}^0$.

The system (11) is a system of linear differential equations with constant coefficients. This is a consequence of the fact that the bilinear form is time-independent. We can write the solution of such a system as linear combinations of exponentials if we make use of the roots of the characteristic equation, see § 17.18. This can be useful for theoretical considerations. If we, however, wish to obtain numerical results we have to evaluate these combinations of exponentials. The usual way is, therefore, the numerical treatement of the problem (11), (12) by a method of Chap. 25. The system (11) is very stiff, see Remark 25.5.2. This is caused by the fact that the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ originate from a parabolic problem. This fact must be considered when a method for the numerical solution of the initial-value problem (11), (12) is looked for.

The discretization in space is most frequently carried out by the FEM. For the simplest elements, we have the following error estimate:

Theorem 1. *Let Ω be a convex polygonal domain, let a regular system of triangulations be given on Ω* (Definition 24.3.2) *and let $u_h(t)$ be the approximation obtained with the use of linear elements. Let us assume that the exact solution $u(t)$ of the problem* (5), (6) *belongs for almost all $t \in (0, T)$ to the space $H^2(\Omega)$. Then*

$$\max_{t\in(0,T)} \|u(t) - u_h(t)\|_{0,\Omega} \leqq C\left(1 + \left|\log\frac{T}{h^2}\right|\right) h^2 \max_{t\in(0,T)} \|u(t)\|_{2,\Omega}\,.$$

If we want to obtain full discretization from the semidiscrete problem (7), (8), we partition the interval $[0, T]$ into M parts by partition points t_k and put $u_{h,k} = u_h(t_k)$, $\tau_k = t_k - t_{k-1}$, $f_k = f(t_k)$ for each $k = 0, 1, \ldots, M$.

The approximate solution is obtained by applying, to the solution of the time dependent problem, either the *implicit Euler method*

$$\left(\frac{u_{h,k} - u_{h,k-1}}{\tau_k}, v\right) + a(u_{h,k}, v) = (f_k, v)\,, \quad v \in V_h\,, \quad k = 1, \ldots, M\,,$$
$$(u_{h,0}, v) = (u_0, v)\,, \quad v \in V_h\,,$$

or the *Crank-Nicolson method*

$$\left(\frac{u_{h,k}-u_{h,k-1}}{\tau_k}, v\right)+a\left(\frac{u_{h,k}+u_{h,k-1}}{2}, v\right)=\frac{1}{2}\left[(f_k, v)+(f_{k-1}, v)\right], \quad v\in V_h, \quad k=1, \ldots, M,$$

$$(u_{h,0}, v)=(u_0, v), \quad v\in V_h.$$

The unknown solution has the form

$$u_{h,k}=\sum_{i=1}^{N} c_{ik}\varphi_i$$

on the k-th time level and we have to solve a system of linear algebraic equations for the unknown coefficients c_{ik} if we proceed from the $(k-1)$-st time level to the k-th one.

REMARK 3. The use of the implicit Euler method for the solution of a semidiscrete problem (i.e. discretized in space by the FEM) is equivalent with the use of the Rothe method of § 18.10 followed by the FEM discretization. The time and space discretization may be interchanged. The Crank-Nicolson method cannot be derived from the Rothe method. Both methods used here for the time discretization correspond to simple A-stable methods for the solution of initial-value problems for ordinary differential equations, see Remark 25.5.2.

Further methods convenient for the adaptive choice of the time-step are in [242]. The error estimates can be found also in [242], and, in addition, in [463].

REMARK 4. Numerical solution of parabolic equations by the finite-difference method is considered in Chap. 27.

REMARK 5. Methods, similar to those considered in this paragraph, can be used in the case of a hyperbolic equation like e.g. the equation of the vibrating string or first-order hyperbolic systems. In these problems, the solutions need not be smooth. The problems with nonsmooth solutions are frequent (e.g. the shock waves) and we are interested just in them. The mere transfer of the ideas stated in this paragraph to hyperbolic problems will not be successful for nonsmooth solutions. It is necessary to take care of the choice of a suitable approximation for nonsmooth solutions. The reader is therefore referred to [242].

Example 1. Let the problem (2), (3), (4) be solved by the Galerkin semidiscrete method. We take $a=0$, $b=1$, $f(x, t)=t$ in (2). The initial condition will be $u_0(x)=\frac{1}{2}-\left|\frac{1}{2}-x\right|$. The interval $[0, 1]$ will be divided into 4 equal parts and linear elements will be used. We thus have three nodes P_1, P_2, P_3 with the coordinates

$x_1 = \frac{1}{4}$, $x_2 = \frac{1}{2}$, $x_3 = \frac{3}{4}$, respectively, and three function values at these points as the nodal parameters. The basis function φ_i at the node P_i is

$$\varphi_i(x) = 1 - 4|x_i - x| \quad \text{for } |x_i - x| \leqq \tfrac{1}{4},$$
$$\varphi_i(x) = 0 \quad \text{elsewhere.}$$

We will construct the system (9). The matrix $\boldsymbol{A}$ has the entries $a(\varphi_i, \varphi_j) = \int_0^1 \varphi_i' \varphi_j' \, dx$. For $i = j$ we have

$$a(\varphi_i, \varphi_i) = \int_{x_i - \frac{1}{4}}^{x_i} 4^2 \, dx + \int_{x_i}^{x_i + \frac{1}{4}} 4^2 \, dx = 8.$$

We further have

$$a(\varphi_i, \varphi_j) = \int_{x_i}^{x_j} 4.(-4) \, dx = -4$$

for $|i - j| = 1$ and $a(\varphi_i, \varphi_j) = 0$ for $|i - j| > 1$.

Similarly, the entries of $\boldsymbol{B}$ are $(\varphi_i, \varphi_j) = \int_0^1 \varphi_i \varphi_j \, dx$.

We have

$$(\varphi_i, \varphi_i) = \int_{x_i - \frac{1}{4}}^{x_i + \frac{1}{4}} (1 - 4|x_i - x|)^2 \, dx =$$
$$= \int_{x_i - \frac{1}{4}}^{x_i} [1 - 4(x_i - x)]^2 \, dx + \int_{x_i}^{x_i + \frac{1}{4}} [1 - 4(x - x_i)]^2 \, dx = \frac{1}{6}$$

for $i = j$ and

$$(\varphi_i, \varphi_j) = \int_{x_i}^{x_j} [1 - 4(x - x_i)][1 - 4(x_j - x)] \, dx = \frac{1}{24}$$

for $|i - j| = 1$ and $(\varphi_i, \varphi_j) = 0$ for $|i - j| > 1$.

Moreover, we have

$$(f, \varphi_j) = \int_{x_j - \frac{1}{4}}^{x_j + \frac{1}{4}} t \varphi_j \, dx = \frac{1}{4} t.$$

The right-hand side vector is $\boldsymbol{U}^0 = \left(\frac{1}{16}, \frac{5}{48}, \frac{1}{16}\right)$.

We have the system of differential equations

$$\begin{aligned}
\tfrac{1}{6}c_1' + \tfrac{1}{24}c_2' \qquad\quad + 8c_1 - 4c_2 \qquad\quad &= \tfrac{1}{4}t, \\
\tfrac{1}{24}c_1' + \tfrac{1}{6}c_2' + \tfrac{1}{24}c_3' - 4c_1 + 8c_2 - 4c_3 &= \tfrac{1}{4}t, \\
\tfrac{1}{24}c_2' + \tfrac{1}{6}c_3' \qquad\quad - 4c_2 + 8c_3 &= \tfrac{1}{4}t
\end{aligned}$$

with the initial conditions

$$\tfrac{1}{6}c_1(0)+\tfrac{1}{24}c_2(0)=\tfrac{1}{16}\,,\quad \tfrac{1}{24}c_1(0)+\tfrac{1}{6}c_2(0)+\tfrac{1}{24}c_3(0)=\tfrac{5}{48}\,,\quad \tfrac{1}{24}c_2(0)+\tfrac{1}{6}c_3(0)=\tfrac{1}{16}\,,$$

from which we determine the coefficient functions $c_i(t)$, $i = 1, 2, 3$.

If we want to solve this system analytically we proceed according to § 17.18. As concerns numerical solution, the system can be treated either in the present form or transformed into the form (11) (solved explicitly with respect to the derivatives). We find the Choleski decomposition of the matrix $\boldsymbol{B}$ in a numeric way and obtain $\boldsymbol{B} = \boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}$, where

$$\boldsymbol{E} = \begin{bmatrix} 0{\cdot}4082, & 0{\cdot}1021, & 0 \\ 0, & 0{\cdot}3953, & 0{\cdot}1054 \\ 0, & 0, & 0{\cdot}3944 \end{bmatrix}$$

The new unknown vector $\boldsymbol{d} = \boldsymbol{E}\boldsymbol{c}$ fulfils the system

$$\begin{aligned} d_1'(t) + 47{\cdot}98d_1(t) - 37{\cdot}176d_2(t) + 9{\cdot}934d_3(t) &= 0{\cdot}6123t\,, \\ d_2'(t) - 37{\cdot}176d_1(t) + 67{\cdot}209d_2(t) - 43{\cdot}614d_3(t) &= 0{\cdot}4744t\,, \\ d_3'(t) + 9{\cdot}934d_1(t) - 43{\cdot}614d_2(t) + 69{\cdot}943d_3(t) &= 0{\cdot}5070t \end{aligned}$$

with the initial conditions

$$d_1(0) = 0{\cdot}1531\,,\quad d_2(0) = 0{\cdot}2240\,,\quad d_3(0) = 0{\cdot}0986\,.$$

This system can be solved directly according to § 17.18 or numerically by some of the methods of Chap. 25.

24.7. The Boundary Element Method

Some of the boundary-value problems for the Laplace equation can be solved with the use of integral equations (cf. § 18.4). The method consists in the solution of an integral equation for the density of potential of a single layer or of a double layer which are then used to express the solution. This can also be a way for the numerical solution of the problem, because the solution of integral equations can be formulated as the variational problem 24.1.1. The application of the FEM to the solution of these integral equations is called the *boundary element method* or the *boundary integral equation method.*

This method can sometimes replace or complete the FEM. Its advantage is well seen on numerical solution of the *exterior* Dirichlet problem (Definition 18.4.6). Let Ω be a bounded, simply connected (see Remark 22.1.9) domain with a smooth boundary. Let us denote the complement of its closure $\overline{\Omega}$ by Ω'.

We are given the following problem in three dimensions:

$$\Delta u = 0 \quad \text{in } \Omega', \tag{1}$$

$$u = u_0 \quad \text{on } S, \tag{2}$$

$$u(x) \to 0 \quad \text{for } \|x\| \to \infty. \tag{3}$$

This is the exterior Dirichlet problem, the condition (3) is in fact "the boundary condition in infinity" and is necessary to guarantee uniqueness of the solution of the problem.

Because the domain Ω' is unbounded it cannot be decomposed into a finite number of triangles. The standard procedure is to substitute a domain $\Omega'_b = \Omega' \cap \cap K(O, b)$ for a sufficiently large b for the domain Ω'. The symbol $K(O, b)$ denotes the ball with centre at a point $O \in \Omega$ and radius b. Sometimes, it is necessary to choose the radius b very large in order to obtain the required accuracy of the approximate solution, and the demands of the FEM for the computer time and memory are enormous.

If we use a variational method to solve integral equations from § 18.4 quoted, we obtain the so-called *indirect method of boundary elements.* In order to describe the *direct method,* we recall some formulae for harmonic functions. We start from the formula (14.8.21) or from the analogous formula for the plane problem. We write this formula, independently of the dimension, as

$$\int_\Omega (u\Delta v - v\Delta u)\,\mathrm{d}X = \int_S \left(u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n}\right)\,\mathrm{d}S \tag{4}$$

where Ω is a bounded domain, S its boundary, n the unit outward normal to the boundary, $\mathrm{d}X$ the element of area or volume and $\mathrm{d}S$ the element of a curve or surface, according to the dimension of the problem.

Definition 1. The fundamental solution of the Laplace equation is the function

$$\psi_3(P, Q) = \frac{1}{4\pi r},$$

where r is the function of two points $P(x_1, x_2, x_3)$ and $Q(\xi_1, \xi_2, \xi_3)$,

$$r = [(x_1 - \xi_1)^2 + (x_2 - \xi_2)^2 + (x_3 - \xi_3)^2]^{\frac{1}{2}},$$

in the three-dimensional space and the function

$$\psi_2(P, Q) = \frac{1}{2\pi} \ln \frac{1}{r}$$

of two points $P(x_1, x_2)$ and $Q(\xi_1, \xi_2)$, where

$$r = [(x_1 - \xi_1)^2 + (x_2 - \xi_2)^2]^{\frac{1}{2}},$$

in the plane. (Cf. Definition 18.4.8.)

The fundamental solution is, for a fixed Q, the solution of the Laplace equation in the whole space (or plane) with the exception of the point $P = Q$.

We further have

$$\int_{\Omega} \Delta\psi(P, Q)\varphi(Q)\,\mathrm{d}Q = -\varphi(P) \tag{5}$$

for an arbitrary function $\varphi(P)$ that is sufficiently smooth and vanishes in the neighbourhood of the boundary S. Because of the singularity of $\psi(P, Q)$ at $P = Q$, the integral in (5) is to be understood as the limit of the integrals over the domain $\Omega - K(P, \varepsilon)$ for ε tending to zero.

The symbol $K(P, \varepsilon)$ denotes a circle or ball with the centre P and radius ε.

Substituting in (4) an arbitrary harmonic function for u and the fundamental solution for v we obtain

$$u(P) = \int_S \psi(P, Q)\frac{\partial u(Q)}{\partial n_Q}\,\mathrm{d}S_Q - \int_S u(Q)\frac{\partial \psi(P, Q)}{\partial n_Q}\,\mathrm{d}S_Q\,. \tag{6}$$

The subscript in the notation for the normal and for the element of the boundary denotes the variable with respect to which the differentiation or integration is carried out.

The value of a harmonic function at an arbitrary interior point of the domain Ω is thus expressed as the difference of two integrals over S containing its boundary values and values of its normal derivative. These integrals are the potentials of single layer (with the kernel ψ) and of double layer (with the kernel $\frac{\partial \psi}{\partial n}$), cf. Definition 18.4.10.

The properties of these potentials imply that the right-hand side of (6) makes sense even if the point P lies outside Ω. We arrive at

$$\int_S \psi(P, Q)\frac{\partial u(Q)}{\partial n_Q}\,\mathrm{d}S_Q - \int_S u(Q)\frac{\partial \psi(P, Q)}{\partial n_Q}\,\mathrm{d}S_Q = \begin{cases} u(P) & \text{for } P \in \Omega\,, \\ \frac{1}{2}u(P) & \text{for } P \in S\,, \\ 0 & \text{for } P \in \Omega'\,. \end{cases} \tag{7}$$

Similar formulae are obtained for a function harmonic in Ω' and fulfilling (3).

We underline that formula (7) is valid only if S is smooth. It can be modified for some kinds of non-smooth boundaries, e.g. for boundaries with corners in E_2.

The formula (7) is a basis for derivation of integral equations for the solution of boundary-value problems by the direct method of boundary elements. For the

solution of the interior Dirichlet problem, we start from (7) with $P \in S$ and obtain an integral equation of the first kind for the unknown function $\frac{\partial u}{\partial n}$ on S,

$$\int_S \psi(P, Q) \frac{\partial u(Q)}{\partial n_Q} \, \mathrm{d}S_Q = f(P) \tag{8}$$

where the terms containing $u(Q)$ on S (and being thus known) are included in the right-hand side. The solution of the interior Neumann problem leads to a Fredholm integral equation of the second kind

$$\frac{1}{2} u(P) + \int_S u(Q) \frac{\partial \psi(P, Q)}{\partial n_Q} \, \mathrm{d}S_Q = f(P), \tag{9}$$

where $f(P)$ is a known term containing $\frac{\partial u(Q)}{\partial n_Q}$ on S. In a similar way, exterior problems or problems with other types of boundary conditions are transformed into integral equations.

REMARK 1. The method using the formula (7) is called the *direct method* because the unknown function is the solution itself or its normal derivative on the boundary. We are often interested just in these quantities. This method is, therefore, sometimes preferred to the procedure following from § 18.4, where the unknown function in the integral equation is the potential density that does not yield directly the values of the solution or of its normal derivative on the boundary.

The value at an arbitrary interior point of Ω is obtained again by (7). It is necessary to compute both the boundary integrals in order to obtain one value. In the indirect method of the boundary elements it is necessary to compute only one integral, cf. the paragraph quoted.

REMARK 2. If we want to obtain the equation of the second kind from (7) also for the Dirichlet problem we can differentiate formula (7) with respect to the normal at the point P. As the right-hand side term, we obtain an integral whose kernel is the derivative of the double layer potential with respect to the normal and is called the *hypersingular integral*. Its evaluation represents some difficulty.

The method of boundary elements consists in application of the idea of the FEM to the numerical solution of (8) and (9) and to similar equations of § 18.4. The kernels of these equations have only a weak singularity and are suitable for this method.

It is obvious that we have to apply quadrature formulae to the integrals in (7) to compute the values of the approximate solution.

The Galerkin method is used for both the kinds of the equations but its convergence is analysed for each kind separately. We decompose the boundary S into

elements in both cases. The elements of the decomposition are finite arcs for plane domains or spatially curved triangles or quadrangles for three-dimensional domains. Let h be the decomposition parameter, i.e. the maximal element diameter. We use the space of piecewise constant functions as the space of finite elements. We denote this space by W_h. It is $W_h \subset L_2(S)$ (for $L_2(S)$ see e.g. [389], Chap. 30) and we have $v|_{K_k} = v_k$, where K_k is the element of decomposition and v_k a constant, for $v \in W_h$. The basis of this space is given by the functions φ_j with $\varphi_j = 1$ on K_j, $\varphi_j = 0$ elsewhere. The spaces of linear and quadratic elements can be used as well, see [25], [58]. Here, however, we confine ourselves to this simplest case for which the theorems on the error estimate for the approximate solution are given below.

The corresponding bilinear form for the equation (8) is

$$b(u, v) = \int_S \int_S \psi(P, Q) u(Q) v(P) \, \mathrm{d}S_P \mathrm{d}S_Q ,$$

and it is symmetric and positive definite on the space of boundary values. The right-hand side functional is

$$F(v) = \int_S f(P) v(P) \, \mathrm{d}S_P .$$

It is possible to apply the results of § 24.2 and § 24.3 or their analogies to this problem.

The approximate solution u_h is assumed in the form $\sum_{i=1}^{N} c_i \varphi_i$. We assemble the matrix $\boldsymbol{B}$ with the entries

$$b_{ij} = \int_S \int_S \psi(P, Q) \varphi_i(P) \varphi_j(Q) \, \mathrm{d}S_P \mathrm{d}S_Q$$

and the right-hand side vector $\boldsymbol{G}$ with the components

$$G_j = \int_S f(P) \varphi_j(P) \, \mathrm{d}P .$$

The coefficient vector $\boldsymbol{c} = (c_1, \ldots, c_N)^{\mathrm{T}}$ is obtained as the solution of the system

$$\boldsymbol{B}\boldsymbol{c} = \boldsymbol{G} . \tag{10}$$

Theorem 1. *Let the right-hand side of the equation* (8) *be a function from* $H^1(S)$ *and let* d_h *be the approximation of the normal derivative obtained by the boundary element method with piecewise constant elements. Then we have*

$$\left\| \frac{\partial u}{\partial n} - d_h \right\|_W \leqq Ch \, \|f\|_{H^1(S)} ,$$

where C is a constant independent of the solution and of the decomposition.

REMARK 3. The spaces $H^1(S)$ and $W = H^{-\frac{1}{2}}(S)$, appearing in Theorem 1, are Sobolev spaces of functions (or functionals) defined on the boundary S. Such spaces are not defined in Remark 22.4.10; we refer, therefore, to [389] Chap. 30 and [348], p. 81.

REMARK 4. The condition number of the corresponding matrix of the system is of order $O(h^{-1})$ for regular systems of decomposition. The system can thus be solved without a significant influence of round-off errors for a reasonable choice of the size of h.

Equations of the second kind, written in the form

$$(I - K)u = g\,, \tag{11}$$

where K is the corresponding integral operator, are solved in a similar way. The approximate solution u_h is assumed in the form $\sum_{i=1}^{N} c_i\varphi_i$. Applying the Galerkin method we obtain

$$\left((I - K)\sum_{i=1}^{N} c_i\varphi_i,\ \varphi_j\right) = (g,\ \varphi_j)\,, \quad j = 1, \ldots, N\,. \tag{12}$$

Let $\boldsymbol{D}$ be a diagonal matrix with the entries $(\varphi_i,\ \varphi_i)$ and $\boldsymbol{B}$ a matrix with the entries

$$b_{ij} = (K\varphi_i,\ \varphi_j)\,.$$

We rewrite (12) in the form

$$(\boldsymbol{D} - \boldsymbol{B})\boldsymbol{c} = \boldsymbol{G} \tag{13}$$

where $\boldsymbol{c} = (c_1, \ldots, c_N)^{\mathrm{T}}$ is the unknown coefficient vector and $\boldsymbol{G} = ((g,\ \varphi_1), \ldots$ $\ldots, (g,\ \varphi_N))^{\mathrm{T}}$ the right-hand side vector.

We have also an error estimate of the approximate solution here:

Theorem 2. *Let $u \in H^1(S)$ hold for the solution of the equation* (11). *Let u_h be the approximation of the solution obtained by the boundary element method with piecewise constatnt elements. Then there are such numbers $C > 0$ and $h_0 > 0$ that*

$$\|u - u_h\|_W \leqq Ch$$

for $h < h_0$ and the constant C depends on u but does not depend on the decomposition of the boundary of the domain. The space W is again the Sobolev space $H^{-\frac{1}{2}}(S)$ (see [242]).

REMARK 5. The matrix $\boldsymbol{B}$ is not symmetric here and the condition number of the symmetrized matrix $(\boldsymbol{D} - \boldsymbol{B})^{\mathrm{T}}(\boldsymbol{D} - \boldsymbol{B})$ is bounded independently of h. This guarantees good solvability of the system (13) for an arbitrarily fine partition.

Finishing this paragraph, we can state that an advantageous feature of the boundary element method is the dimension reduction of the problem, a two-dimensional problem is reduced to a problem on a one-dimensional boundary, a three--dimensional problem to a problem on a two-dimensional boundary. The method is suitable for domains whose boundary is geometrically complicated. The method can be, or even has to be combined with the FEM (e.g. for the Poisson equation, i.e. for a nonhomogeneous equation).

A drawback of the method is that we need to know the fundamental solution, and, through this, we are practically limited to equations with constant coefficients and that for every value at an interior point we must evaluate formula (7).

Further, less substantial drawbacks are that the computation of the coefficients b_{ij} is laborious and that the matrix $\boldsymbol{B}$ is full. This restricts our choice of methods for the numerical solution of the corresponding algebraic systems.

The method for approximate solution of the above integral equations that is based on collocation is also sometimes called the boundary element method, see [58], [25].

In the conclusion, let us show a simple illustrative example.

Example 1. We solve the Dirichlet problem

$$\begin{aligned} \Delta u &= 0 \quad \text{in } K(0, R), \\ u &= R\cos s, \quad s \in [0, 2\pi] \end{aligned}$$

on the circle K with centre at the origin and radius R.

The exact solution is $u = x$ or, in polar coordinates, $u = \varrho\cos\varphi$.

We use the above boundary element method with constant elements and with the partition of the boundary into N parts. We put $h = 2\pi/N$. The value of the normal derivative to the boundary is assumed in the form $\dfrac{\partial u}{\partial n} = \sum_{i=1}^{N} d_i\varphi_i$. For the entries b_{ij} of the matrix $\boldsymbol{B}$, we obtain

$$b_{ij} = \int_S \int_S \frac{1}{2\pi} \ln\frac{1}{r}\varphi_i\varphi_j \,\mathrm{d}S_P\mathrm{d}S_Q\,,$$

or with the parametrization of the boundary

$$b_{ij} = \frac{R^2}{2\pi}\int_{(j-1)h}^{jh}\int_{(i-1)h}^{ih} \ln\frac{1}{r}\,\mathrm{d}s_P\mathrm{d}s_Q\,.$$

For the distance r we have (see Fig. 24.35)

$$r = 2R \sin \frac{|s_P - s_Q|}{2}$$

and, consequently,

$$b_{ij} = \frac{R^2}{2\pi} \int_{(j-1)h}^{jh} \int_{(i-1)h}^{ih} -\ln\left(2R \sin \frac{|s_P - s_Q|}{2}\right) \mathrm{d}s_P \mathrm{d}s_Q = -\frac{R^2}{2\pi} h^2 \ln 2R - \frac{R^2}{2\pi} I_{ij},$$

where

$$I_{ij} = \int_{(j-1)h}^{jh} \int_{(i-1)h}^{ih} \ln \sin \frac{|s_P - s_Q|}{2} \mathrm{d}s_P \mathrm{d}s_Q .$$

The integrals I_{ij} depend only on the difference $|i - j|$ and do not depend on R; it is, however, necessary to compute them numerically. The computation must be carried out with some care because for some combination of the subscripts, e.g. for $i = j$, the integrals are singular.

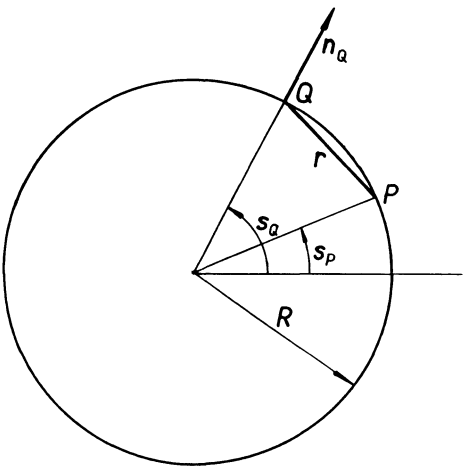


Fig. 24.35.

We have

$$G_j = \frac{1}{2} \int_S R \cos s_P \, \varphi_j \, \mathrm{d}s_P + \int_S \int_S R \cos s_Q \, \frac{\partial}{\partial n_Q} \left(\frac{1}{2\pi} \ln \frac{1}{r}\right) \mathrm{d}s_Q \varphi_j \mathrm{d}s_P$$

for the components of the right-hand side vector. It is

$$\frac{\partial}{\partial n_Q} \left(\frac{1}{2\pi} \ln \frac{1}{r}\right) = \frac{1}{2\pi} \frac{\cos(n_Q, \overrightarrow{QP})}{r} = -\frac{1}{4\pi R} \quad \text{for } P, Q \in S .$$

The last equality is again obvious from Fig. 24.35.

The second integral in the expression for G_j is thus zero and

$$G_j = \frac{1}{2}R^2 \int_{(j-1)h}^{jh} \cos s_P \, \mathrm{d}s_P = R^2 \sin \frac{h}{2} \cos \frac{2j-1}{2} h .$$

We choose $N = 8$. The values of the integrals I_{ij} computed numerically with the use of the trapezoidal rule with 32 parts are

$\lvert i-j \rvert$	0	1	2	3	4	5	6	7
I_{ij}	−1·4859	−0·6651	−0·2303	−0·0583	−0·0080	−0·0583	−0·2303	−0·6651

Further we put $R = 1$ and obtain the matrix $\boldsymbol{B}$ and the right-hand side vector $\boldsymbol{G}$:

$$\boldsymbol{B} = \begin{bmatrix} 0{\cdot}1684, & 0{\cdot}0378, & -0{\cdot}0314, & -0{\cdot}0588, & -0{\cdot}0668, & -0{\cdot}0588, & -0{\cdot}0314, & 0{\cdot}0378 \\ 0{\cdot}0378, & 0{\cdot}1684, & 0{\cdot}0378, & -0{\cdot}0314, & -0{\cdot}0588, & -0{\cdot}0668, & -0{\cdot}0588, & -0{\cdot}0314 \\ -0{\cdot}0314, & 0{\cdot}0378, & 0{\cdot}1684, & 0{\cdot}0378, & -0{\cdot}0314, & -0{\cdot}0588, & -0{\cdot}0668, & -0{\cdot}0588 \\ -0{\cdot}0588, & -0{\cdot}0314, & 0{\cdot}0378, & 0{\cdot}1684, & 0{\cdot}0378, & -0{\cdot}0314, & -0{\cdot}0588, & -0{\cdot}0668 \\ -0{\cdot}0668, & -0{\cdot}0588, & -0{\cdot}0314, & 0{\cdot}0378, & 0{\cdot}1684, & 0{\cdot}0378, & -0{\cdot}0314, & -0{\cdot}0588 \\ -0{\cdot}0588, & -0{\cdot}0668, & -0{\cdot}0588, & -0{\cdot}0314, & 0{\cdot}0378, & 0{\cdot}1684, & 0{\cdot}0378, & -0{\cdot}0314 \\ -0{\cdot}0314, & -0{\cdot}0588, & -0{\cdot}0668, & -0{\cdot}0588, & -0{\cdot}0314, & 0{\cdot}0378, & 0{\cdot}1684, & 0{\cdot}0378 \\ 0{\cdot}0378, & -0{\cdot}0314, & -0{\cdot}0588, & -0{\cdot}0668, & -0{\cdot}0588, & -0{\cdot}0314, & 0{\cdot}0378, & 0{\cdot}1684 \end{bmatrix},$$

$$\boldsymbol{G} = (\,0.3536,\ 0.1464, -0.1464, -0.3536, -0.3536, -0.1464,\ 0.1464,\ 0.3536\,)^{\mathrm{T}} .$$

The solution of (10) yields piecewise constant approximate values of the normal derivative of the solution. They are

$$\begin{aligned} d_1 &= \ \ 0{\cdot}9509, & d_2 &= \ \ 0{\cdot}3939, & d_3 &= -0{\cdot}3939, & d_4 &= -0{\cdot}9509, \\ d_5 &= -0{\cdot}9509, & d_6 &= -0{\cdot}3939, & d_7 &= \ \ 0{\cdot}3939, & d_8 &= \ \ 0{\cdot}9509. \end{aligned}$$

The relative error (the same for all values) with respect to the exact value at the midpoint of the corresponding arc is 2·9 %. The approximate solution at an arbitrary point of $K(0, 1)$ is then obtained from the discretized version of (7).

25. APPROXIMATE SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS

By EMIL VITÁSEK

References: [21], [24], [49], [88], [104], [107], [145], [150], [155], [170], [193], [212], [234], [235], [291], [293], [322], [332], [358], [378], [389], [395], [420], [425], [445], [446], [456], [460], [461], [475.].

In this chapter we describe basic methods for numerical solution of initial value problems and boundary value problems for ordinary differential equations. We also describe some methods for the approximation of eigenvalues of differential operators. Since these problems involve many rather different items we begin with an introductory paragraph in which we specify the individual problems that will be dealt with in the further text.

25.1. Introduction

By the *initial value problem* for the system of m differential equations of the first order

$$\begin{aligned} {}^1y' &= {}^1f(x, {}^1y, \ldots, {}^my), \\ &\vdots \\ {}^my' &= {}^mf(x, {}^1y, \ldots, {}^my) \end{aligned} \tag{1}$$

we mean the problem of finding m functions ${}^1y, \ldots, {}^my$ of x which are defined, continuous and continuously differentiable in an interval $[a, b]$, which satisfy (1) for any $x \in [a, b]$ and for which

$${}^1y(\xi) = \eta_1, \ \ldots, \ {}^my(\xi) = \eta_m \tag{2}$$

holds where $\boldsymbol{\eta} = (\eta_1, \ \ldots, \ \eta_m)^{\mathrm{T}}$ is a given m-dimensional vector and ξ is a point from $[a, b]$. This ξ is usually equal to a and we restrict ourselves to this special case. The conditions (2) imposed on the solution of (1) are called the *initial conditions.*

Using the vector notation (cf. § 17.2), the system (1) and the initial conditions (2) can be written in the form

$$\begin{aligned} \boldsymbol{y}' &= \boldsymbol{f}(x, \boldsymbol{y}), \\ \boldsymbol{y}(\xi) &= \boldsymbol{\eta}, \end{aligned}$$

where

$$\boldsymbol{y} = \begin{bmatrix} {}^1y \\ \vdots \\ {}^my \end{bmatrix} = \left({}^1y, \ \ldots, \ {}^my\right)^{\mathrm{T}},$$

$$\boldsymbol{f}(x, \boldsymbol{y}) = \begin{bmatrix} {}^1f(x, {}^1y, \ \ldots, \ {}^my) \\ \vdots \\ {}^mf(x, {}^1y, \ \ldots, \ {}^my) \end{bmatrix}$$

and

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_m \end{bmatrix}.$$

Since any m-th order differential equation

$$y^{(m)} = f(x, y, y', \ \ldots, y^{(m-1)}) \tag{3}$$

can be written as a system of m first-order equations

$$\begin{aligned} {}^1y' &= {}^2y, \\ {}^2y' &= {}^3y, \\ &\ldots \\ {}^{m-1}y' &= {}^my, \\ {}^my' &= f(x, {}^1y, \ \ldots, \ {}^my) \end{aligned}$$

by means of introducing new unknown functions by

$$\begin{aligned} {}^1y &= y, \\ {}^2y &= y', \\ &\ldots \\ {}^my &= y^{(m-1)}, \end{aligned} \tag{4}$$

it is clear what is understood by the initial value problem for the equation (3).

By the *boundary value problem* or, in more detail, by the *two-point boundary value problem for the system* (1), we mean the problem of finding such a solution of (1) for which we have

$$\boldsymbol{r}(\boldsymbol{y}(a), \boldsymbol{y}(b)) = \boldsymbol{o}\,. \tag{5}$$

Here $\boldsymbol{r}$ is a given m-dimensional vector function of $2m$ variables, $\boldsymbol{y}(x)$ is an m-dimensional vector with components ${}^1y(x), \ldots, {}^my(x)$ and a and b are two different points from the interval in which the solution is sought. These points are usually the end points of this interval.

If it is possible to write the conditions (5), called the *boundary conditions*, in the form

$$\boldsymbol{r}_a(\boldsymbol{y}(a)) = \boldsymbol{o}\,, \quad \boldsymbol{r}_b(\boldsymbol{y}(b)) = \boldsymbol{o}\,,$$

where $\boldsymbol{r}_a$ and $\boldsymbol{r}_b$ are m_a- and m_b-dimensional vector functions of m variables ($m = m_a + m_b$), respectively, we speak about the *separated boundary conditions.* The conditions of the form

$$\boldsymbol{U}\boldsymbol{y}(a) + \boldsymbol{V}\boldsymbol{y}(b) = \boldsymbol{c}\,,$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are $m \times m$ matrices and $\boldsymbol{c}$ is an m-dimensional vector, are called *linear boundary conditions* and the conditions

$$\boldsymbol{V}_a\boldsymbol{y}(a) = \boldsymbol{v}_a\,, \quad \boldsymbol{V}_b\boldsymbol{y}(b) = \boldsymbol{v}_b\,, \tag{6}$$

where $\boldsymbol{V}_a$ and $\boldsymbol{V}_b$ are $m_a \times m$ and $m_b \times m$ matrices and $\boldsymbol{v}_a$ and $\boldsymbol{v}_b$ are m_a- and m_b-dimensional vectors, respectively, are *linear separated boundary conditions.*

Taking into account the transformation (4) we again know what is a boundary value problem for an equation of order m.

Note that the same number of boundary conditions as the number of unknown functions is not necessary in the definition of a boundary value problem and it is also possible to study the conditions of the type (5) combining the values of the solution at more than two points of the given interval.

If we replace $\boldsymbol{V}_a$ in (6) by the identity matrix and omit the second equation at all we see that the initial value problem is a very special case of the boundary value problem. This fact has rather serious consequences: Whilst the existence theorems are known for relatively large classes of nonlinear initial value problems (cf., e.g., § 17.2), the solution of a boundary value problem may not exist or may not be determined uniquely even in the case of a very simple linear equation. This fact can be most easily comprehended from the following almost trivial example.

Example 1. Let us solve the differential equation

$$y'' + y = 0\,. \tag{7}$$

Any solution of this equation can be written in the form

$$y(x) = C_1 \sin x + C_2 \cos x,$$

where C_1 and C_2 are arbitrary constants. But from here we immediately see that any function of the form $C \sin x$, where C is any number, is the solution of (7) with the boundary conditions

$$y(0) = 0, \quad y(\pi) = 0. \tag{8}$$

Thus, the boundary value problem (7), (8) has infinitely many solutions. If we replace the conditions (8) by

$$y(0) = 0, \quad y(\pi) = 1,$$

the solution does not exist.

On the basis of this example, one can conclude that theoretical questions (concerning existence and uniqueness of solution, etc.) in boundary value problems are substantially more complicated than in the case of initial value problems. For that reason we will be able to describe numerical methods for solving initial value problems directly for the general nonlinear problem (1), (2). In the case of boundary value problems, we restrict ourselves, on the contrary, very often to linear problems only. The main reason for this restriction is that nonlinear boundary value problems are often solved in such a way that one constructs a sequence of linear boundary value problems, the solutions of which converge to the solution of the original problem. The realization of such a procedure then depends on our ability to solve linear problems efficiently.

The *eigenvalue problem* in the case of differential equations is the problem of finding such values of a parameter λ for which the system of differential equations

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}, \lambda) \tag{9}$$

($\boldsymbol{y}$ is an m-dimensional vector and $\boldsymbol{f}$ is an m-dimensional vector function of $m+2$ variables for which $\boldsymbol{f}(x, \boldsymbol{o}, \lambda) = \boldsymbol{o}$), the right-hand term of which depends, besides on x and $\boldsymbol{y}$, also on the parameter λ, with boundary conditions

$$\boldsymbol{r}(\boldsymbol{y}(a), \boldsymbol{y}(b), \lambda) = \boldsymbol{o} \tag{10}$$

($\boldsymbol{r}(\boldsymbol{o}, \boldsymbol{o}, \lambda) = \boldsymbol{o}$) has a nontrivial (i.e. not identically equal to zero) solution. The value of the parameter λ satisfying these conditions is called the *eigenvalue* and the corresponding (nontrivial) solution *eigenfunction of the problem* (9), (10).

REMARK 1. The assumptions $\boldsymbol{f}(x, \boldsymbol{o}, \lambda) = \boldsymbol{o}$ and $\boldsymbol{r}(\boldsymbol{o}, \boldsymbol{o}, \lambda) = \boldsymbol{o}$ imply that $\boldsymbol{y} = \boldsymbol{o}$ is a solution of (9), (10) for *any* λ. Thus, the eigenvalue problem consists in finding such values of λ for which the problem (9), (10) has, moreover, a nontrivial solution (see also § 17.17).

REMARK 2. If we introduce another unknown function ${}^{m+1}y = \lambda$ and add a further equation ${}^{m+1}y' = 0$ in (9), the eigenvalue problem (9), (10) may be looked at as a general boundary value problem (1), (5). The eigenvalue problem (9), (10) is thus equivalent with the problem of finding the vector $\tilde{\boldsymbol{y}} = (\boldsymbol{y}^{\mathrm{T}}, {}^{m+1}y)^{\mathrm{T}}$ which solves the equation

$$\tilde{\boldsymbol{y}}' = \tilde{\boldsymbol{f}}(x, \tilde{\boldsymbol{y}})$$

with the boundary conditions

$$\tilde{\boldsymbol{r}}(\tilde{\boldsymbol{y}}(a), \tilde{\boldsymbol{y}}(b)) = \boldsymbol{o},$$

where

$$\tilde{\boldsymbol{f}}(x, \tilde{\boldsymbol{y}}) = \begin{bmatrix} \boldsymbol{f}(x, \boldsymbol{y}, {}^{m+1}y) \\ \boldsymbol{o} \end{bmatrix}$$

and

$$\tilde{\boldsymbol{r}}(u_1, \ldots, u_m, u_{m+1}, v_1, \ldots, v_m, v_{m+1}) = \\ = \boldsymbol{r}(u_1, \ldots, u_m, v_1, \ldots, v_m, v_{m+1}).$$

This remark may be useful rather from the practical point of view since it allows, at least formally, to transfer the ideas of studying boundary value problems to eigenvalue problems.

From the concrete methods for the numerical solution of initial value problems we describe in detail Runge-Kutta methods and linear k-step (multistep) methods, including the predictor-corrector methods. We also describe the Gragg method which forms a starting point for effective using of idea of the Richardson extrapolation and which is rather popular at present.

For the numerical solution of boundary value problems and eigenvalue problems we introduce some methods of transforming such problems into initial value problems and the finite difference method. As far as the variational methods are concerned, we refer to Chap. 24 which is especially devoted to these methods.

A. INITIAL VALUE PROBLEMS

25.2. Euler's Method. Error Estimates

Euler's method is the simplest of all the methods for approximate solution of initial value problems. The application of Euler's method in practical numerical problems is not recommended because its efficiency is very limited. Nevertheless, we will study this method in a particular paragraph in considerable detail since it very clearly exhibits certain features which are peculiar also for more complicated methods.

Since the existence theorems for initial value problems for systems of equations are formally identical with the corresponding ones for one equation, we restrict ourselves to a single differential equation

$$y' = f(x, y) \tag{1}$$

with the initial condition

$$y(a) = \eta\,. \tag{2}$$

In Euler's method the values y_n approximating the values of the exact solution at the points $x_n = a + nh$, $n = 0, \ldots, N$ ($h = (b-a)/N$, N positive integer, is the so-called *integration step* or *steplength*), are calculated recursively according to the following formulae:

$$\begin{aligned} y_0 &= \eta\,, \\ y_{n+1} &= y_n + hf(x_n, y_n)\,, \quad n = 0, \ldots, N-1\,. \end{aligned} \tag{3}$$

These formulae make possible an obvious geometric interpretation: We consider the differential equation $y' = f(x, y)$ as an equation defining a field of directions (cf. § 17.2) in the strip $a \leqq x \leqq b$ of the (x, y)-plane. The problem of solving the differential equation is then geometrically equivalent to the problem of determining a curve $y = y(x)$ passing through the given initial point (a, η) and having its slope at each point coincident with the slope prescribed by the direction field. The points (x_n, y_n) defined by (3) can be considered as vertices of a polygonal graph which passes through the given initial point and possesses the property that each its link has the direction prescribed by the direction field at its left-hand end point.

From this interpretation we see that an exact solution, the graph of which is a straight-line, is computed by Euler's method *exactly*. In all the other cases, the quantity $e_n = y_n - y(x_n)$, called the *total* or *accumulated discretization* or *truncation error*, is generally not equal to zero. Thus, in order to be acceptable, it

is necessary for the studied method to be capable of making the total discretization error as small as necessary. Since the only parameter in Euler's method is the integration step, the discretization error can be controlled only by changes of this integration step.

Thus, to have some opinion on the magnitude of the total discretization error we must have some information on the behaviour of the discretization error, as a function of h, at our disposal. There are several possible levels of achievement in the study of the error.

If we simply know that

$$\lim_{\substack{h \to 0 \\ x_n = x}} e_n = 0 \tag{4}$$

for any $x \in [a, b]$, we speak about the *convergence of the method.* Underline that the approach to the limit in (4) is understood for a fixed $x = x_n$. The index n has, consequently, to grow when h is diminishing. The study of the behaviour of y_n as $h \to 0$ with n *fixed* has no practical sense, since one has, obviously, $x_n = a + nh \to a$ and $y_n \to y_0 = \eta$ as $h \to 0$.

If we find a function $\varphi(h)$ for which $\varphi(h) \to 0$ as $h \to 0$ and for which $e_n = O(\varphi(h))$ as $h \to 0$, then this function φ indicates the *rate of convergence* of the given method. Namely, the symbol $O(\varphi(h))$ (cf. § 11.4) means that there exists a constant M such that $|e_n| \leqq M\varphi(h)$ for small h's; thus, the discretization error converges to zero at least as fast as the known function φ (in practical situations, the function φ is almost exclusively of the form h^p).

An information on the behaviour of the discretization error stronger that the rate of convergence is the *error bound* $|e_n| \leqq \psi(h)$ where ψ is a known function.

Finally, if we find a function η such that $e_n/\chi(h) \to 1$ for $h \to 0$ we call the function χ the *asymptotic estimate of the error.* We call it also the *main* (or *principal*) *part of the total discretization error.*

At the first glance, the most useful information about the behaviour of the discretization error is the error bound since it allows to choose the magnitude of the integration step a priori in such a way that the total discretization error is smaller than a tolerance prescribed in advance. Though this consideration is theoretically correct it is practically meaningless. The main reason is that the a priori error bound is usually extremely pessimistic so that it gives values many times (very often even several million times) greater than the reality. The consequence of this feature is that algorithms constructed on the basis of a priori error bounds are practically always very inefficient and they may be even infeasible. For that reason, the rate of convergence or the asymptotic estimate of the error are often of the same desirability as a priori error bounds.

On the other hand, we must be content with the fact that the method under investigation is convergent in situations when we are not able to say more.

In the following text we introduce all the mentioned characteristics of convergence of Euler's method. First we introduce a further quantity which also gives some information about the total discretization error. It is the so-called *local discretization error* which is committed by performing one step of the method on the assumption that all values necessary for performing it are accurate. Thus, the local discretization error for Euler's method is given by

$$l(y(x);\, h) = y(x+h) - y(x) - hf(x,\, y(x))\,, \tag{5}$$

where y is the exact solution of the given differential equation.

It is clear that the necessary condition for the convergence is that the local discretization error is small since the total discretization error is the result of the accumulation of local errors.

Theorem 1. *Let the right-hand term of the differential equation* (1)

(i) *be defined and continuous as a function of two variables in the strip* $R = \{(x,\, y);\, a \leqq x \leqq b,\ -\infty < y < \infty\}$;

(ii) *satisfy in* R *the Lipschitz condition with respect to* y *with a constant independent of* x, *i.e., let there exist a constant* $L > 0$ *such that*

$$|f(x,\, y) - f(x,\, z)| \leqq L\,|y - z| \tag{6}$$

for any $x \in [a,\, b]$ *and for any* y *and* z.

Further , let y_n *be the approximate solution of the problem* (1), (2) *computed by Euler's method* (3) *and let* y *be the exact solution. Then the total discretization error* $e_n = y_n - y(x_n)$ *satisfies*

$$|e_n| \leqq \omega(h)\frac{\mathrm{e}^{L(x_n - a)} - 1}{L}\,, \quad n = 0,\, \ldots,\, N\,, \tag{7}$$

where ω *is the modulus of continuity of the function* y' , *i.e.,*

$$\omega(h) = \sup_{\substack{x,\, x^* \in [a,\, b] \\ |x - x^*| \leqq h}} |y'(x) - y'(x^*)|\,. \tag{8}$$

REMARK 1. The global assumptions (i) and (ii) from Theorem 1 guarantee the existence and uniqueness of the solution of (1), (2) in the whole interval $[a,\, b]$ whilst the usual local assumptions (cf., for example, Theorems 17.2.1 and 17.2.2) guarantee the existence of the solution only in some neighbourhood of the point $(a,\, \eta)$. This is, from the practical point of view, very agreeable. On the other hand, one must realize that especially the assumption (ii) is rather restrictive and that

it may happen very often that it is not satisfied (with the important exception of a linear equation). But in the practical situation, we usually solve such equations for which some a priori information about the position of the solution is available. Thus, we are mostly able to achieve the fulfilment of the assumption (ii) properly changing the definition of f in that part of the strip R in which the solution surely does not lie.

It follows from Remark 1 that the function y' is continuous in $[a, b]$ so that $\omega(h) \to 0$ as $h \to 0$. Since the function $\{\exp[L(x_n - a)] - 1\}/L$ is bounded on $[a, b]$ the formula (7) proves the convergence of Euler's method. At the same time, the function $\omega(h)$ indicates its rate of convergence.

Theorem 2. *Let the assumptions* (i) *and* (ii) *from Theorem* 1 *be satisfied and let the exact solution of the differential equation* (1) *with the initial condition* (2) *have two continuous derivatives in* $[a, b]$. *Let*

$$M(x) = \frac{1}{2} \max_{t \in [a, x]} |y''(t)| \tag{9}$$

and let y_n *be the approximate solution computed by Euler's method. Then*

$$|e_n| \leqq hM(x_n) \frac{e^{L(x_n - a)} - 1}{L}, \quad n = 0, \ldots, N. \tag{10}$$

Theorem 2 answers two questions concerning the convergence of Euler's method. If we know the function $M(x)$ defined by (9) or if we are able to bound it, the formula (10) represents the error bound. If we know only that M exists without knowing its concrete form, the formula (10) expresses the fact that the rate of convergence of Euler's method is h in this case.

REMARK 2. It is not possible to increase the rate of convergence of Euler's method by further increasing the regularity of the exact solution.

REMARK 3. The local discretization error of Euler's method behaves as $h\omega(h)$ on the assumptions of Theorem 1 and as h^2 on the assumptions of Theorem 2. Thus, the accumulation of local errors results in loosing one power of h.

REMARK 4. If f has continuous partial derivatives with respect to both variables in the set

$$R_\eta = \{(x, y);\ a \leqq x \leqq b,\ |y| \leqq Y\}, \tag{11}$$

where

$$Y = |\eta| e^{L(b-a)} + \max_{x \in [a, b]} |f(x, 0)| \frac{e^{L(b-a)} - 1}{L}, \tag{12}$$

then we have

$$M(x) \leqq \frac{1}{2} \max_{(x,y)\in R_\eta} |f_x(x, y) + f_y(x, y)f(x, y)| , \tag{13}$$

where f_x and f_y mean the partial derivatives of the function f with respect to x and y, respectively.

Theorem 3. *Let the assumptions* (i) *and* (ii) *from Theorem* 1 *be satisfied and let, moreover, the right-hand term f of the given differential equation have continuous first and second partial derivatives with respect to both variables in* R_η *defined by* (11). *Then the total discretization error of Euler's method can be written in the form*

$$e_n = e(x_n)h + O(h^2) , \tag{14}$$

where e is the solution of the differential equation

$$e' = f_y(x, y(x))e - \tfrac{1}{2}y''(x) \tag{15}$$

with the initial condition $e(a) = 0$.

Thus, the expression $e(x_n)h$ in the right-hand term of (14) represents the assymptotic estimate (or the principal part) of the total discretization error. Numerical experiments show that this estimate usually gives a very good idea of the behaviour of the real discretization error. It is clear at the first glance, that it is very difficult to compute e directly since it is the solution of a further differential equation the right-hand term of which depends, moreover, on the solution sought. But the fact that such a function exists, without knowing its particular form, allows to construct the error estimate of the form

$$y(x; h) - y(x) = 2\,[y(x; h) - y(x; h/2)] + O(h^2) , \tag{16}$$

where $y(x; h)$ denotes the approximate solution at x computed by Euler's method with an integration step h.

Ignoring the higher-order terms, the estimate of the total discretization error can be gained by computing the approximate solution twice with two different integration steps (h and $h/2$). This approach to estimating the error is often called the *deferred approach to limit.*

Example 1. Let us solve the differential equation $y' = -y$ with the initial condition $y\,(0) = 1$.

The exact solution is $\exp(-x)$, the Lipschitz constant is equal to unity and $M(x) = \frac{1}{2}$. Thus, the expression $h\,[\exp(x_n) - 1]\,/2$ represents the error bound in the sense of Theorem 2. The differential equation (15) is in this case, obviously,

$e' = -e - \frac{1}{2}\exp(x)$ so that $e(x) = -\frac{1}{2}x\exp(-x)$. The results for $h = 2^{-6}$ can be found in Tab. 25.1. It illustrates our theoretical considerations very well. Really, the error bound given by (10) is very pessimistic. For example, at the point $x = 5$ it gives approximately 5000 times larger value than in reality and this occurs even in the trivial case when we know the function $M(x)$ exactly and are thus not obliged to estimate it. On the other hand, the estimates obtained according to Theorem 3 or formula (16) agree with reality very well.

TABLE 25.1

x_n	1	2	3	4	5
y_n	0·364987	0·133235	0·048622	0·017746	0·006477
e_n	−0·002892	−0·002120	−0·001165	−0·000570	−0·000261
error bound according to (10)	0·013424	0·049914	0·149106	0·418735	1·15666
$e(x_n)h$	−0·002874	−0·002114	−0·001167	−0·000572	−0·000263
error estimate according to (16)	−0·002902	−0·002123	−0·001165	−0·000560	−0·000260

REMARK 5. All which was stated for Euler's method is valid not only for one differential equation but also for a system of m equations of the first order. The only difference is that the scalar quantities y and f must be interpreted as m-dimensional vectors.

REMARK 6. The solution of the problem (1), (2) may be also approximated by the sequence $\{y_n(x)\}$ constructed by means of the recurrent formula

$$y_{n+1}(x) = \eta + \int_a^x f(t, y_n(t))\,\mathrm{d}t, \quad n = 0, 1, \ldots . \tag{17}$$

These approximations, named *Picard's approximations*, are used exclusively for theoretical purposes since the algorithms based on them cannot be compared with the discrete methods as far as their efficiency is concerned.

REMARK 7. The restriction to the case of equidistant points x_n is not substantional and the properties of Euler's method in which this assumption is violated are practically the same as the properties of Euler's method with the fixed integration step.

25.3. General One-Step Method

In Euler's method, the approximate solution y_{n+1} at the point x_{n+1} was computed exclusively on the basis of knowing the approximate solution y_n at the point x_n. Thus, Euler's method can be considered as a special case of the general one-step methods when the *general one-step method* means any algorithm for solving the problem (25.2.1), (25.2.2) in which the approximation y_{n+1} at x_{n+1} is computed only from the values x_n, y_n and h. It is useful to write the dependence of the approximate solution y_{n+1} on x_n, y_n and h in the form

$$y_{n+1} = y_n + h\Phi_f(x_n, y_n, h), \tag{1}$$

where Φ_f is a function of three variables which depends on the right-hand term of the given differential equation (for example, $\Phi_f(x, y, h) = f(x, y)$ in the case of Euler's method).

Before introducing important special cases, we formulate some general properties of Φ which allow to formulate general convergence theorems.

The general one-step method is said to be *regular* if the corresponding function Φ_f:

(i) is defined and continuous in the domain $D = \{(x, y, h);\ a \leqq x \leqq b,\ -\infty < y < \infty,\ 0 \leqq h \leqq h_0\}$ (h_0 is a positive constant),

(ii) satisfies the Lipschitz condition with respect to the second variable with a constant independent of the remaining variables, i.e., there exists a constant L such that

$$|\Phi_f(x, y, h) - \Phi_f(x, z, h)| \leqq L\,|y - z| \tag{2}$$

for any two points (x, y, h) and (x, z, h) from D.

The general one-step method is said to be *consistent with the given differential equation* (*consistent*, for short) if it is regular and if

$$\Phi_f(x, y, 0) = f(x, y). \tag{3}$$

The *local error* $l(y(x);\ h)$ of the general one-step method is defined by the formula

$$l(y(x);\ h) = y(x + h) - y(x) - \Phi_f(x, y(x), h), \tag{4}$$

where y is the exact solution of the problem (25.2.1), (25.2.2).

The largest positive integer p for which

$$l(y(x);\ h) = O(h^{p+1}) \tag{5}$$

is called the *order* of the general one-step method.

The order of the method depends on the smoothness of the solution of the given differential equation and is one of important characteristics of the method.

The basic characteristic properties of a general one-step method are introduced in the following theorems:

Theorem 1. *Let y_n be the approximate solution computed by a regular general one-step method whose local discretization error can be estimated according to the formula*

$$|l(y(x_n); h)| \leqq \psi(h), \quad n = 0, \ldots, N, \tag{6}$$

and let y be the corresponding exact solution. Then the total discretization error $e_n = y_n - y(x_n)$ is bounded as follows:

$$|e_n| \leqq \frac{e^{L(x_n - a)} - 1}{L} \frac{1}{h} \psi(h). \tag{7}$$

It is seen from Theorem 1 that the behaviour of the total discretization error is fully controlled by the local discretization error. Thus, the local discretization error really represents an important characteristic of the given method.

Theorem 2. *The local discretization error of a general one-step method which is consistent with the given differential equation satisfies the inequality*

$$|l(y(x); h)| \leqq h[\omega(h) + \varphi(h)], \tag{8}$$

where ω is the modulus of continuity of the first derivative of the exact solution $y(x)$ (i.e., it is defined by (25.2.8)) *and $\varphi(h)$ is defined by*

$$\varphi(h) = \max_{x \in [a, b]} |\Phi_f(x, y(x), h) - \Phi_f(x, y(x), 0)| .$$

Since $\omega(h)$ and $\varphi(h)$ tend to zero as $h \to 0$, Theorems 1 and 2 imply the convergence of a consistent general one-step method. Theorem 1 and (5) imply that the total discretization error of a general one-step method of order p behaves as h^p. If we know not only that (5) holds, i.e., that there exists a constant M such that

$$|l(y(x); h)| \leqq M h^{p+1}, \tag{9}$$

but if we are able, moreover, to estimate M, then the inequality (7) represents an error bound. This bound exhibits the same negative properties as that one in the case of Euler's method, namely, it is extremaly pessimistic. For that reason we present, in addition, a theorem characterizing the asymptotic behaviour of the total discretization error.

Theorem 3. *Consider a regular general one-step method of order p whose local error can be written in the form*

$$l(y(x);\, h) = \varphi(x,\, y(x))h^{p+1} + O(h^{p+2})\,. \tag{10}$$

Then we have

$$e_n = e(x_n)h^p + O(h^{p+1})\,, \tag{11}$$

where e is the solution of the differential equation

$$e' = f_y(x,\, y(x))e - \varphi(x,\, y(x)) \tag{12}$$

with the initial condition $e\,(a) = 0\,.$

Note that the assumption (10) is satisfied if the solution of (25.2.1), (25.2.2) (or, which is, in essence, the same, the right-hand term of (25.2.1)) and the function Φ_f are sufficiently smooth.

The quantity $e(x_n)h^p$ is the main (principal) part of the total discretization error and gives a very good idea of the behaviour of the real error. Like in the case of Euler's method, it is very difficult to compute it directly. The existence of such a function can be used indirectly for obtaining the formula

$$y(x;\, h) - y(x) = \frac{2^p}{2^p - 1}[y(x;\, h) - y(x;\, h/2)] + O(h^{p+1}) \tag{13}$$

giving the same information about the error as the asymptotic formula (11). The use of (13) to an error estimate is called the *deferred approach to the limit.*

In the following text we introduce some special one-step methods. The main idea of constructing them is to achieve the highest possible order.

(a) Taylor's Expansion Methods

In this special one-step method we put

$$\Phi_f(x,\, y,\, h) = f(x,\, y) + \frac{1}{2}hf^{(1)}(x,\, y) + \cdots + \frac{1}{p!}h^{p-1}f^{(p-1)}(x,\, y)\,, \tag{14}$$

where the functions $f^{(s)}(x,\, y)$ are defined recurrently by

$$\begin{aligned} f^{(0)}(x,\, y) &= f(x,\, y)\,, \\ f^{(s)}(x,\, y) &= \frac{\partial f^{(s-1)}(x,\, y)}{\partial x} + \frac{\partial f^{(s-1)}(x,\, y)}{\partial y} f(x,\, y)\,. \end{aligned} \tag{15}$$

The function $f^{(s)}(x, y(x))$ equals to the $(s+1)$-st derivative of the function $y(x)$ which is the solution of the differential equation (25.2.1). Consequently, the local error of the method (14) is $y^{(p+1)}(\xi)h^{p+1}/(p+1)!$ and the method has the order p. Thus, we have found a way of constructing a general one-step method, the order of which is arbitrarily high. Nevertheless, the Taylor expansion method is used very rarely in practice. The main reason is that the evaluation of functions $f^{(s)}$ which are necessary for the computation of the values of the function Φ_f may lead and usually also really leads to very complicated expressions.

(b) Runge-Kutta Methods

These methods are the most practicable one-step methods and the main disadvantage of the Taylor expansion method is avoided here in such a way that the function Φ_f is taken in the form

$$\Phi_f(x, y, h) = w_1 k_1 + \cdots + w_s k_s\,, \tag{16}$$

where

$$\begin{aligned} k_1 &= f(x, y)\,, \\ k_i &= f\Big(x + b_i h,\ y + h\sum_{j=1}^{i-1} c_{ij} k_j\Big)\,, \quad i = 2, \ldots, s\,. \end{aligned} \tag{17}$$

The numbers w_i, b_i, c_{ij} in (16) and (17) are chosen in such a way that the function Φ_f given by (16) differs from the right-hand term of (14) starting from the term of order h^p. Thus, the parameters w_i, b_i and c_{ij} are taken so that the method is of order p. In contrast to the Taylor expansion method, the necessity of the computation of derivatives of the right-hand term of the given differential equation is replaced by computing the values of f at more than one point.

Denote by $p(s)$ the maximal attainable order of the Runge-Kutta method which uses s values of f (such a method is called the *s-stage Runge-Kutta method*). This function has the property that $p(s) \to \infty$ as $s \to \infty$; moreover, it is $p(s) = s$ for $s \leqq 4$, $p(5) = 4$, $p(6) = 5$, $p(7) = 5$, $p(8) = 6, \ldots$, and there exist infinitely many s-stage methods of order $p(s)$. Consequently, in the class of Runge-Kutta methods we have methods of arbitrarily high orders at our disposal.

In the following lines we introduce some important particular Runge-Kutta methods.

The *second-order method*

$$y_{n+1} = y_n + hf(x_n + \tfrac{1}{2}h,\ y_n + \tfrac{1}{2}hf(x_n, y_n)) \tag{18}$$

and the method

$$y_{n+1} = y_n + \tfrac{1}{2}h\,[f(x_n, y_n) + f(x_n + h,\ y_n + hf(x_n, y_n))] \tag{19}$$

whose order is also two, are usually connected with the name *modified Euler's method.* It is just the method (18) which was suggested by Runge in the last century, giving thus a start of the development of any other method of this type.

Heun's method

$$\begin{aligned} y_{n+1} &= y_n + \tfrac{1}{4}h(k_1 + 3k_3)\,, \\ k_1 &= f(x_n,\, y_n)\,, \\ k_2 &= f(x_n + \tfrac{1}{3}h,\, y_n + \tfrac{1}{3}hk_1)\,, \\ k_3 &= f(x_n + \tfrac{2}{3}h,\, y_n + \tfrac{2}{3}hk_2) \end{aligned} \tag{20}$$

has the order 3.

Two well-known fourth-order methods are

$$\begin{aligned} y_{n+1} &= y_n + \tfrac{1}{6}h(k_1 + 2k_2 + 2k_3 + k_4)\,, \\ k_1 &= f(x_n,\, y_n)\,, \\ k_2 &= f(x_n + \tfrac{1}{2}h,\, y_n + \tfrac{1}{2}hk_1)\,, \\ k_3 &= f(x_n + \tfrac{1}{2}h,\, y_n + \tfrac{1}{2}hk_2)\,, \\ k_4 &= f(x_n + h,\, y_n + hk_3) \end{aligned} \tag{21}$$

and

$$\begin{aligned} y_{n+1} &= y_n + \tfrac{1}{8}h(k_1 + 3k_2 + 3k_3 + k_4)\,, \\ k_1 &= f(x_n,\, y_n)\,, \\ k_2 &= f(x_n + \tfrac{1}{3}h,\, y_n + \tfrac{1}{3}hk_1)\,, \\ k_3 &= f(x_n + \tfrac{2}{3}h,\, y_n - \tfrac{1}{3}hk_1 + hk_2)\,, \\ k_4 &= f(x_n + h,\, y_n + hk_1 - hk_2 + hk_3)\,. \end{aligned} \tag{22}$$

By far the most used Runge-Kutta method is the method (21); it is even just this particular method which is very often understood when speaking about the Runge--Kutta method. For that reason, this method will be referred to as the *standard Runge-Kutta method.*

The methods of order higher than four are used seldom. Thus, we introduce, from these methods, only *Fehlberg's method*

$$\begin{aligned} y_{n+1} &= y_n + h(\tfrac{16}{135}k_1 + \tfrac{6656}{12825}k_3 + \tfrac{28561}{56430}k_4 - \tfrac{9}{50}k_5 + \tfrac{2}{55}k_6)\,, \\ k_1 &= f(x_n,\, y_n)\,, \\ k_2 &= f(x_n + \tfrac{1}{4}h,\, y_n + \tfrac{1}{4}hk_1)\,, \\ k_3 &= f(x_n + \tfrac{3}{8}h,\, y_n + \tfrac{1}{32}h(3k_1 + 9k_2))\,, \\ k_4 &= f(x_n + \tfrac{12}{13}h,\, y_n + \tfrac{1}{2197}h(1932k_1 - 7200k_2 + 7296k_3))\,, \\ k_5 &= f(x_n + h,\, y_n + h(\tfrac{439}{216}k_1 - 8k_2 + \tfrac{3680}{513}k_3 - \tfrac{845}{4104}k_4))\,, \\ k_6 &= f(x_n + \tfrac{1}{2}h,\, y_n + h(-\tfrac{8}{27}k_1 + 2k_2 - \tfrac{3544}{2565}k_3 + \tfrac{1859}{4104}k_4 - \tfrac{11}{40}k_5))\,. \end{aligned} \tag{23}$$

This method has the order five and it possesses the property that, combining its k's in another way, we obtain another approximation

$$y^*_{n+1} = y_n + h(\tfrac{25}{216}k_1 + \tfrac{1408}{2565}k_3 + \tfrac{2197}{4104}k_4 - \tfrac{1}{5}k_5)$$

the order of which is four. The number $y_{n+1} - y^*_{n+1}$ can be then considered as the error estimate of y^*_{n+1}.

REMARK 1. If the assumptions (i) and (ii) from Theorem 25.2.1 are satisfied, then any Runge-Kutta method is regular; if, moreover, the method under consideration is of order at least one, i.e., if $w_1 + \cdots + w_s = 1$, it is also consistent with the given differential equation. Then, any Runge-Kutta method is at least convergent on these general assumptions. This fact contrasts with the behaviour of general Taylor's expansion method which need not be feasible for a differential equation satisfying only (i) and (ii) since the corresponding derivatives of f need not exist.

Example 1. Let us solve the differential equation $y' = y$ with the initial condition $y(0) = 1$ on the interval $[0, 5]$ by the standard Runge-Kutta method.

The results can be seen for $h = 2^{-3}$ in Tab. 25.2. The constant M from (9) which is needed for using Theorem 1 has been bounded by the formula

$$M \leqq 6KM_0(1 + K + K^2 + K^3 + K^4), \tag{24}$$

where M_0 and K are chosen in such a way that

$$|f(x, y)| \leqq M_0, \quad \left|\frac{\partial^{i+k} f(x, y)}{\partial x^i \partial y^k}\right| \leqq \frac{K}{M_0^{i+k-1}}, \quad i + k \leqq 4. \tag{25}$$

TABLE 25.2

x_n	1	2	3	4	5
y_n	2·718277	7·389029	20·08543	54·59775	148·4118
e_n	−0·000005	−0·000027	−0·00011	−0·00040	−0·0014
error bound according to Theorem 1	0·05	0·40	2·95	21·8	161·33
$e(x_n)h^4$	−0·000006	−0·000030	−0·00012	−0·00044	−0·0015
error estimate according to (13)	−0·000005	−0·000027	−0·00011	−0·00039	−0·0013

It is again seen, like in Euler's method, that the a priori error bound gained on the basis of Theorem 1 is even more drastically pessimistic whilst the asymptotic estimate as well as the estimate according to (13) give fully sufficient information about the behaviour of the actual error.

The bound given by (24) is the well-known *Bieberbach estimate* of the local discretization error of the standard Runge-Kutta method. Similar estimates can be constructed also for all other Runge-Kutta methods. Example 1, however, shows that such estimates have only limited practical sense.

REMARK 2. Like in the case of Euler's method, all what was stated about the general one-step method is also valid for the system of m equations of the first order. It is again sufficient to suppose that the quantities y and Φ are m-dimensional vectors. But it is necessary to note that the order of some Runge-Kutta methods may be smaller if the method is used for solving systems of equations than if it is used for solving one equation only. Such situation, however, does not occur in the particular methods introduced above.

25.4. Linear k-Step Method

In § 25.3 we have seen that the computation of the approximate solution of the given differential equation by a one-step method proceeds in such a way that we solve, in fact, a new initial value problem with the initial condition specified at the point at which the approximate solution has been found in the preceding step. This is advantageous since the logical structure of the computer code is simple, the change of the integration step is easy, etc. On the other hand, such methods have some significant disadvantages. One of them, namely the fact that the estimation of the local discretization error is (at least in the case of Runge-Kutta methods) very complicated, was met already in § 25.3 (cf. (25.3.24)). Also the fact that we forget the preceding values of the approximate solution at the moment when the solution at the present point has been computed does not seem to be prescient. On the contrary, it appears to be natural that if one uses not only the value at the point $x = x_n$ but also, for example, the value at the point $x = x_{n-1}$ to compute the approximate solution at the point $x = x_{n+1}$ one must obtain a more accurate result. Thus, it seems that the methods which make use of the information from some preceding points for the construction of the approximate solution and which, consequently, are multistep in this sense are more efficient (i.e., less laborious and more accurate) than one-step methods.

In the present paragraph we describe some particular methods of this type. We first start with investigating their common properties.

By a *multistep* or, in more detail, by a *k-step method* for the solution of (25.2.1), (25.2.2), we mean the method given by the formula

$$\sum_{i=0}^{k} \alpha_i y_{n+i} = h \sum_{i=0}^{k} \beta_i f(x_{n+i}, y_{n+i}), \quad n = 0, 1, \ldots, \tag{1}$$

where $\alpha_0, \ldots, \alpha_k, \beta_0, \ldots, \beta_k$ are constants. We will assume in what follows that $\alpha_k = 1$ and that α_0 and β_0 are not simultaneously equal to zero.

The equation (1) must be regarded as the equation for determining y_{n+k} if we suppose that $y_n, y_{n+1}, \ldots, y_{n+k-1}$ are already known. This procedure is then repeated for $n = 0, 1, \ldots$ till we achieve the point at which we are interested in the solution. The form of (1) shows that we take into account only such multistep methods in which the relation among y_{n+i} and $f(x_{n+i}, y_{n+i})$ (i.e., among the approximations of the solution and of its derivatives) is *linear*. One can also imagine methods in which this relation is nonlinear. Such methods have so few advantages in comparison with the linear methods that they are practically out of use.

It is clear that the approximation y_{n+k} is uniquely determined by (1) if $\beta_k = 0$. This is the case of the so-called *explicit method*. In a general case, i.e., if $\beta_k \neq 0$ (such a method is called *implicit*), (1) represents a generally nonlinear equation for the unknown y_{n+k}. If the right-hand term of the given differential equation satisfies the assumption (ii) from Theorem 25.2.1 and if we limit ourselves to integration step satisfying

$$h < \frac{1}{L\,|\beta_k|}, \tag{2}$$

then (1) has one and only one solution and, moreover, it can be determined by successive approximations

$$\begin{gathered} y_{n+k}^{(s+1)} = h\beta_k f(x_{n+k}, y_{n+k}^{(s)}) - \sum_{i=0}^{k-1} \alpha_i y_{n+i} + h \sum_{i=0}^{k-1} \beta_i f(x_{n+i}, y_{n+i}), \\ s = 0, 1, \ldots, \quad y_{n+k}^{(0)} \text{ arbitrary.} \end{gathered} \tag{3}$$

Taking into account the necessity to solve a nonlinear equation when using an implicit method, we see that implicit methods are more laborious than explicit ones. Consequently, it seems that there is no sense in using implicit methods at all. However, we will show, in the following text, such convenient properties of implicit methods which fully balance this disadvantage.

To be able to start the computation it is seen from (1), used for $n = 0$, that we must know the approximate solutions $y_0, \ldots, y_{k-1}$ at the points $x_0, \ldots, x_{k-1}$. This feature is typical for k-step methods and the problem of obtaining the values $y_0, \ldots, y_{k-1}$ must be always solved before the application of (1). Runge-Kutta methods are very often used for this purpose.

By *convergence of a k-step method* we mean, like in the case of Euler's method or a general one-step method, that

$$\lim_{\substack{h\to 0\\ x_n=x}} y_n = y(x) \tag{4}$$

is satisfied for any $x \in [a, b]$ and for any solution of the initial value problem (25.2.1), (25.2.2). Since the approximate solution depends not only on the integration step h and on y_0, but moreover on $y_1, \ldots, y_{k-1}$, it is necessary to require the fulfilment of (4) for *any* approximate solution with the initial conditions $y_s = y_s(h)$, $s = 0, \ldots, k-1$, satisfying

$$\lim_{h\to 0} y_s(h) = \eta, \quad s = 0, \ldots, k-1. \tag{5}$$

Thus, the method is convergent if the exact solution is well approximated for small h's by any approximate solution which is determined by this h and by the initial conditions which well approximate the accurate initial condition.

The *local discretization error* is defined by

$$l(y(x); h) = \sum_{i=0}^{k} \alpha_i y(x+ih) - h\sum_{i=0}^{k} \beta_i y'(x+ih) \tag{6}$$

(cf. (25.3.4)), where y is the exact solution of the initial value problem (25.2.1), (25.2.2). If this solution is sufficiently smooth, then

$$l(y(x); h) = C_0 y(x) + C_1 y'(x)h + \cdots + C_q y^{(q)}(x)h^q + \ldots,$$

where

$$\begin{gathered} C_0 = \alpha_0 + \cdots + \alpha_k, \\ C_1 = \alpha_1 + 2\alpha_2 + \cdots + k\alpha_k - (\beta_0 + \cdots + \beta_k), \\ C_q = \frac{1}{q!}(\alpha_1 + 2^q\alpha_2 + \cdots + k^q\alpha_k) - \frac{1}{(q-1)!}\left(\beta_1 + 2^{q-1}\beta_2 + \cdots + k^{q-1}\beta_k\right), \\ q = 2, 3, \ldots. \end{gathered} \tag{7}$$

We say that the *linear k-step method has order p* if $C_0 = C_1 = \cdots = C_p = 0$, $C_{p+1} \neq 0$. The constant C_{p+1} is, in this situation, called the *error constant.*

If the method is of order p and if the exact solution is sufficiently smooth, then we have

$$l(y(x); h) = C_{p+1} y^{(p+1)}(x)h^{p+1} + O(h^{p+2}). \tag{8}$$

The method whose order is at least one is said to be *consistent.*

If we introduce the first and second characteristic polynomial of the method (1) by

$$\varrho(\xi) = \sum_{i=0}^{k} \alpha_i \xi^i, \quad \sigma(\xi) = \sum_{i=0}^{k} \beta_i \xi^i, \tag{9}$$

the consistency can be expressed as

$$\varrho(1) = 0, \quad \varrho'(1) = \sigma(1). \tag{10}$$

In § 25.3 we stated that a general one-step method which is consistent with the given differential equation is convergent and that the higher the order of such a method is, the higher is the rate of its convergence. In order to obtain analogous statements also for a linear k-step method we must confine ourselves to D-stable methods.

We say that a linear k-step method is *D-stable* if the modulus of no root of its first characteristic polynomial ϱ exceeds 1 and if the roots of modulus 1 are simple.

Theorem 1. *Let y be the solution of the initial value problem* (25.2.1), (25.2.2) *and let the assumptions* (i) *and* (ii) *from Theorem* 25.2.1 *be satisfied. Further, let y_n be the approximate solution computed by a D-stable linear k-step method* (1) *of order $p \geqq 1$ with the initial conditions y_s, $s = 0, \ldots, k-1$, for which we have $|y_s - y(x_s)| \leqq \delta$ for $s = 0, \ldots, k-1$. Finally, let*

$$l(y(x_n); h) \leqq \psi(h), \quad n = 0, 1, \ldots. \tag{11}$$

Then, for any h satisfying (2),

$$|y_n - y(x_n)| \leqq \Gamma^* \left[Ak\delta + (x_n - a)\frac{1}{h}\psi(h) \right] \mathrm{e}^{\Gamma^* LB(x_n - a)}, \tag{12}$$

where L is the Lipschitz constant of the right-hand term of the given differential equation,

$$A = \sum_{i=0}^{k} |\alpha_i|, \quad B = \sum_{i=0}^{k} |\beta_i|,$$
$$\Gamma^* = \Gamma/(1 - h|\beta_k| L), \quad \Gamma = \sup_{s=0,1,\ldots} |\gamma_s|$$

and the numbers γ_i are defined by

$$1/\left(\alpha_k + \alpha_{k-1}\xi + \cdots + \alpha_0 \xi^k\right) = \gamma_0 + \gamma_1 \xi + \gamma_2 \xi^2 + \ldots.$$

We see from (12) that the total discretization error consists of two parts: The first part, represented by the term involving the estimate of the local truncation

error, corresponds exactly to the discretization error of a one-step method with a similar estimate of local error. The second part of the error, represented by the term involving the estimate of the error in initial conditions, owes its existence to the fact that the approximate solution is determined by k starting values. The user of a linear k-step method must have this structure of the total discretization error always in mind since improper initial values may totally destroy the approximate solution.

As far as the local discretization error is concerned we have the following theorem:

Theorem 2. *Let the right-hand term of the given differential equation satisfy the assumptions* (i) *and* (ii) *from Theorem* 25.2.1. *Then the local discretization error of a linear k-step method of order* $p \geqq 1$ *can be estimated as*

$$|l(y(x_n); h)| \leqq Kh\omega(h)\,, \tag{13}$$

where K *is a constant depending only on* α_i*'s and* β_i*'s but independent of* h, *and* ω *is the modulus of continuity of* y'. *If, moreover, the exact solution has* $p+1$ *continuous derivatives in* $[a, b]$, *then*

$$|l(y(x_n); h)| \leqq GYh^{p+1}\,, \tag{14}$$

where G *is a constant independent of* h *and*

$$Y = \max_{x \in [a,b]} |y^{(p+1)}(x)|\,.$$

Combining Theorems 1 and 2, we obtain the assertions concerning the convergence, the rate of convergence and the error estimate of the total discretization error of a linear k-step method.

Example 1. Let us solve the differential equation $y' = -y$ with the initial condition $y(0) = 1$ on the interval $[0, 1]$ by the two-step method

$$y_{n+2} + 4y_{n+1} - 5y_n = 2h\,[2f(x_{n+1}, y_{n+1}) + f(x_n, y_n)]$$

the order of which is 3. (This formula was constructed in such a way that it has maximal possible order and is, at the same time, explicit.)

The results are presented for $h = 0{\cdot}1$ and $h = 0{\cdot}05$ in Tab. 25.3. Both necessary additional initial conditions were taken from the exact solution. It is seen at the first glance that the results have no reasonable sense. In fact, the method used is not D-stable (the roots of ϱ are 1 and -5). Thus, this example drastically shows that the D-stability is really a necessary condition for convergence.

TABLE 25.3

x	$h = 0{\cdot}1$		$h = 0{\cdot}05$	
	approximate solution	error	approximate solution	error
0·0	1·000000	0	1·000000	0
0·1	0·904837	0	0·904836	−0·000001
0·2	0·818715	−0·000015	0·818711	−0·000019
0·3	0·740872	0·000054	0·740315	−0·000503
0·4	0·669997	−0·000323	0·656994	−0·013326
0·5	0·608200	0·001669	0·252935	−0·353596
0·6	0·539907	−0·008905	−8·833877	−9·382689
0·7	0·543768	0·047183	−248·4746	−248·9711
0·8	0·198971	−0·250358	−6606·041	−6606·490
0·9	1·734617	1·328048	−175303·9	−175304·3
1·0	−6·677257	−7·045136	−4651727	−4651728

As far as the assymptotic behaviour of the total discretization error is concerned, the corresponding assertion is substantially more complicated than in the case of a one-step method. In order to be able to formulate it, it is first necessary to introduce some concepts and notations. Moreover, we will deal only with D-stable linear k-step methods of order at least 1 and such that the corresponding polynomials ϱ and σ have no common factors. Any individual linear k-step method which will be introduced in the following text satisfies these assumptions. As a consequence of the D-stability, moduli of all roots of ϱ are not greater than 1 and the roots of modulus 1 are simple. Further, since the method under consideration is of order at least 1, the number 1 must be a simple root of ϱ. Put $\xi_1 = 1$ and denote by $\xi_2, \ldots, \xi_m$ the remaining roots of ϱ with modulus 1 (naturally, if such roots exist). The roots $\xi_1, \ldots, \xi_m$ are all simple (the method is D-stable) and we will call them *essential roots*. The number

$$\lambda_s = \frac{\sigma(\xi_s)}{\xi_s \varrho'(\xi_s)}, \quad s = 1, \ldots, m, \tag{15}$$

is called the *growth parameter* corresponding to the essential root ξ_s. Put further

$$\xi_s = \exp(\mathrm{i}\varphi_s), \quad s = 1, \ldots, m, \quad \mathrm{i} = \sqrt{(-1)}\ . \tag{16}$$

Moreover, let the order of accuracy of the initial conditions which determine the approximate solution be h^q and, in addition, let

$$y_s - y(x_s) = \gamma_s h^q + O(h^{q+1}), \quad s = 0, \ldots, k-1, \tag{17}$$

where γ_s are constants (independent of h). Finally, put

$$D_j = \sum_{s=0}^{k-1} \alpha_{js}\gamma_s\,, \quad j = 1, \ldots, m\,, \tag{18}$$

where α_{js}'s are the coefficients of the polynomial $\varrho(\xi)/(\xi - \xi_j)$, i.e.,

$$\frac{\varrho(\xi)}{\xi - \xi_j} = \alpha_{j0} + \alpha_{j1}\xi + \cdots + \alpha_{j,k-1}\xi^{k-1}\,. \tag{19}$$

Theorem 3. *Let the solution y of the differential equation* (25.2.1), *the right-hand term of which satisfies the assumptions* (i) *and* (ii) *from Theorem* 25.2.1, *have $p+2$ continuous derivatives in $[a, b]$. Further, let y_n be the approximate solution computed by means of a D-stable k-step method of order $p \geqq 1$ with polynomials ϱ and σ having no common factors. Finally, let the initial conditions satisfy* (17). *Then*

$$\begin{aligned} e_n = y_n - y(x_n) = e(x_n)h^p + O(h^{p+1}) + \\ + h^q \sum_{j=1}^{m} \frac{D_j}{\varrho'(\xi_j)} \exp(\mathrm{i}n\varphi_j) e_j(x_n) + O(h^{q+1})\,, \end{aligned} \tag{20}$$

where e is the solution of the differential equation

$$e' = \frac{\partial f}{\partial y}(x, y(x))e - C_{p+1}y^{(p+1)}(x) \tag{21}$$

with the initial condition $e(a) = 0$, e_j, $j = 1, \ldots, m$, is the solution of the differential equation

$$e_j' = \lambda_j \frac{\partial f}{\partial y}(x, y(x))e_j$$

with the initial condition $e_j(a) = 1$, λ_j is given by (15) *and D_j by* (18).

On the basis of Theorem 1, we were already able to divide the total discretization error into two parts: the genuine discretization error arising from the replacement of an infinite-dimensional problem by a finite-dimensional one and the error arising from an inaccurate fulfilment of initial conditions. The genuine discretization error is represented by the term involving e in Theorem 3. This term corresponds exactly to the principal part of the total discretization error of a one-step method and it is a smooth function of x. The errors due to inaccuracies in initial conditions are represented by the terms involving e_j's. If $m = 1$, the sum in (20) reduces to its first term. The starting error is then proportional to the exact solution (note that $\lambda_1 = 1$) and the result is like when using a one-step method with slightly wrong initial values. If $m > 1$, additional terms appear in the sum in (20). These terms can be characterized as oscillations (owing to the factor $\exp(\mathrm{i}n\varphi_j)$). If the initial conditions

are relatively inexact (i.e., if $q \leqq p$), this circumstance may negatively influence the possibility of using a formula similar to (25.3.13) for the error estimate. Moreover, if we approximate a differential equation with an exponentially decreasing solution and if, for some j, the corresponding growth parameter is negative, then the function e_j oscillates with an exponentially growing amplitude. These oscillations may cause difficulties even in the case when $q > p$, i.e., when the initial conditions are relatively exact. Such linear k-step methods are referred to as the *weakly stable methods* and the reasons indicated above show that these methods are suspicious. We finish the comment to Theorem 3 with the recommendation of highest care when using weakly stable methods.

In the conclusion of this paragraph we introduce some often used linear k-step methods:

(a) Methods of Numerical Integration. Adams-Bashforth Method. Adams-Moulton Method

These methods are based on the extrapolation or interpolation of the right-hand term of the given differential equation followed by integration.

The *Adams-Bashforth method* is given by

$$y_{n+k} - y_{n+k-1} = h \sum_{j=0}^{k-1} \beta_j^{(k)} f(x_{n+j}, y_{n+j}), \tag{22}$$

where

$$\beta_j^{(k)} = (-1)^{k-1-j} \sum_{s=k-1-j}^{k-1} \binom{s}{k-1-j} \gamma_s, \quad j = 0, \ldots, k-1, \tag{23}$$

and the constants γ_s are defined recurrently by

$$\sum_{i=0}^{s} \frac{1}{i+1} \gamma_{s-i} = 1, \quad s = 0, 1, \ldots. \tag{24}$$

It is a D-stable explicit method with $m = 1$ (cf. Theorem 3) and its local error satisfies

$$l(y(x); h) = \gamma_k y^{(k+1)}(\xi) h^{k+1}, \tag{25}$$

where ξ is a suitable point from $[x, x + kh]$. The order of the Adams-Bashforth method is, consequently, k. The coefficients of the Adams-Bashforth methods are presented in Tab. 25.4 for first six values of k.

TABLE 25.4.

ν	0	1	2	3	4	5
$\beta_\nu^{(1)}$	1					
$2\beta_\nu^{(2)}$	−1	3				
$12\beta_\nu^{(3)}$	5	−16	23			
$24\beta_\nu^{(4)}$	−9	37	−59	55		
$720\beta_\nu^{(5)}$	251	−1274	2616	−2774	1901	
$1440\beta_\nu^{(6)}$	−425	2627	−6798	9482	−7673	4227

The *Adams-Moulton method* is given by

$$y_{n+k} - y_{n+k-1} = h\sum_{j=0}^{k} \beta_j^{(k)} f(x_{n+j}, y_{n+j}), \tag{26}$$

where

$$\beta_j^{(k)} = (-1)^{k-j} \sum_{s=k-j}^{k} \binom{s}{k-j} \gamma_s^*, \quad j = 0, \ldots, k, \tag{27}$$

and the constants γ_s^* are defined by

$$\gamma_0^* = 1, \quad \sum_{i=0}^{s} \frac{1}{i+1} \gamma_{s-i}^* = 0, \quad s = 1, 2, \ldots . \tag{28}$$

TABLE 25.5

ν	0	1	2	3	4	5
$2\beta_\nu^{(1)}$	1	1				
$12\beta_\nu^{(2)}$	−1	8	5			
$24\beta_\nu^{(3)}$	1	−5	19	9		
$720\beta_\nu^{(4)}$	−19	106	−264	646	251	
$1440\beta_\nu^{(5)}$	27	−173	482	−798	1427	475

We have again obtained a D-stable, now implicit method with $m = 1$. Its local discretization error satisfies, like in the case of the Adams-Bashforth method,

$$l(y(x); h) = \gamma^*_{k+1} y^{(k+2)}(\xi) h^{k+2} . \tag{29}$$

The Adams-Moulton method is thus of order $k + 1$. In Tab. 25.5, its coefficients are presented for $k = 0, \ldots, 5$.

(b) Methods of Numerical Differentiation. Backward Difference Methods

These methods are based on the interpolation of the values of the function sought followed by differentiation. They are given by

$$\sum_{j=0}^{k} \alpha_j^{(r)} y_{n+j} = h \beta_{k-r}^{(r)} f(x_{n+k-r}, y_{n+k-r}) , \quad r = 0, \ldots, k . \tag{30}$$

Here

$$\alpha_j^{(k)} = \frac{(-1)^{k-j} \sum_{s=k-j}^{k} \delta_{rs} \binom{s}{k-j}}{\sum_{s=1}^{k} \delta_{rs}} , \quad j = 0, \ldots, k-1 , \tag{31}$$

$$\alpha_k^{(k)} = 1 ,$$

$$\beta_{k-r}^{(r)} = \frac{1}{\sum_{s=1}^{k} \delta_{rs}} \tag{32}$$

and the constants δ_{rs} are defined by

$$\delta_{0s} = \frac{1}{s} , \quad s = 1, \ldots, k ,$$

$$\delta_r = \delta_{r-1,s} - \delta_{r-1,s-1} , \quad s = 1, \ldots, k , \quad r = 1, \ldots, k ,$$

where we put $\delta_{r0} = 0$ for $r = 0, \ldots, k$. The method is implicit for $r = 0$; for other values of r it is explicit. The local discretization error of all these methods is given by

$$l(y(x); h) = \delta_{r,k+1} y^{(k+1)}(\xi) h^{k+1} , \tag{33}$$

and thus they are of order k.

These methods are also called the *backward difference methods* and those of their subset characterized by $r = 0$ and $k \leqq 6$ are rather often used in practice

in connection with the solution of stiff differential systems (about stiff differential systems see Remark 25.5.2). Their coefficients can be found in Tab. 25.6.

TABLE 25.6

k	β_k	α_6	α_5	α_4	α_3	α_2	α_1	α_0
1	1						1	−1
2	$\frac{2}{3}$					1	$-\frac{4}{3}$	$\frac{1}{3}$
3	$\frac{6}{11}$				1	$-\frac{18}{11}$	$\frac{9}{11}$	$-\frac{2}{11}$
4	$\frac{12}{25}$			1	$-\frac{48}{25}$	$\frac{36}{25}$	$-\frac{16}{25}$	$\frac{3}{25}$
5	$\frac{60}{137}$		1	$-\frac{300}{137}$	$\frac{300}{137}$	$-\frac{200}{137}$	$\frac{75}{137}$	$-\frac{12}{137}$
6	$\frac{60}{147}$	1	$-\frac{360}{147}$	$\frac{450}{147}$	$-\frac{400}{147}$	$\frac{225}{147}$	$-\frac{72}{147}$	$\frac{10}{147}$

TABLE 25.7

n	method (34)		Adams-Bashforth method	
	approximate solution	error	approximate solution	error
0	5·000000	0	5·000000	0
1	4·652200	0	4·652200	0
2	4·354907	0·003373	4·355881	0·004347
⋮				
63	1·159287	−0·048115	1·206536	0·001252
64	1·246457	0·048115	1·199552	0·001210
65	1·141986	−0·049670	1·192825	0·001170
⋮				
127	0·599828	−0·425689	1·025688	0·000171
128	1·409513	0·384790	1·024888	0·000166
129	0·568993	−0·454960	1·024114	0·000161
⋮				
191	−1·606091	−2·609506	1·003441	0·000026
192	1·827522	0·824211	1·003335	0·000025
193	−1·679211	−2·682419	1·003233	0·000024

Example 2. Use the two-step method

$$y_{n+2} = y_n + 2hf(x_{n+1}, y_{n+1}) \tag{34}$$

of order 2 with two essential roots (cf. Theorem 3) and with growth parameters equal to 1 and -1, respectively, for solving the differential equation $y' = 1 - y^2$ with the initial condition $y(0) = 5$. For $h = 1/64$, the results are shown in Tab. 25.7 which contains also the approximate solution computed by the two-step Adams-Bashforth method, for comparison. The initial condition y_1 was taken in both cases from the exact solution. One can see from the table that the oscillations mentioned in the comment to Theorem 3 destroy the approximate solution computed by (34) very rapidly. On the other hand, the results obtained by the Adams-Bashforth method are fully satisfactory.

25.5. The Use and Comparison of Runge-Kutta and Linear Multistep Methods. Predictor-Corrector Methods

One of the most difficult problems which has to be solved when using Runge-Kutta or linear multistep method in practice is the problem of choosing a suitable integration step. This choice cannot be based on a priori error bounds of the total discretization error introduced in § 25.3 and § 25.4. Namely, even if we ignore the difficulties connected with gaining them, they are almost always so pessimistic that they imply needlessly small integration steps. This fact had also been the reason why we studied the asymptotic formula for the total discretization error. However, the deferred approach to the limit procedure (cf. (25.3.13) and the comment to Theorem 25.3.3) based on asymptotic error estimate is not too much recommendable, either. As a matter of fact, we have to repeat the calculations from the very beginning twice in this procedure which can be rather inefficient. For that reason the commonly accepted approach to the step size control is at present as follows: The integration step is chosen in such a way that (i) the local discretization error is small and (ii) the accumulation of local errors is not dangerous. The latter property is achieved if the number $z = h\lambda$, where h is the integration step and λ is the estimate of $\partial f/\partial y$, lies in the interval of absolute stability of the method used. To define the *interval of absolute stability* we say that it is formed of such numbers $z = h\lambda$ that any approximate solution y_n of the differential equation $y' = \lambda y$, computed by the method under consideration used with the integration step h, converges to zero as $n \to \infty$.

To obtain a realistic estimate of the local discretization error of a Runge-Kutta method of order p, one can use the formula

$$l(y(x_n); h) = \frac{1}{2^{p+1} - 1}(y_{n+1} - y^*_{n+1}), \tag{1}$$

where y_{n+1} is the approximate solution at the point x_{n+1} obtained using the integration step h and y^*_{n+1} is the approximation at the same point computed by the

same method with step $2h$. Formula (1) is a parallel to formula (25.3.13) and was derived on base of similar considerations.

The local error of the Runge-Kutta-Fehlberg method may be estimated by the procedure described in the text following the formula (25.3.23).

The theoretical formulae for estimates of local error are in the case of a linear multistep method substantially simpler than in the case of Runge-Kutta methods (cf., for example, (25.4.8), (25.4.25), (25.4.29)). Namely, for obtaining such an estimate it is sufficient to know the error constant and to estimate the $(p+1)$-st derivative of the solution. This is the reason why they can be used in some cases. More often, however, the Milne formula which will be described later or an analogue of (1) are used.

To find out that $h\partial f/\partial y$ lies in the interval of absolute stability of the given method we certainly must know this interval and, naturally, we must be able to estimate $\partial f/\partial y$. Thus, the interval of absolute stability is one of the important characteristics of the investigated method. For any of the methods introduced till now, this interval is of the type $(\alpha, 0)$ where $\alpha \leqq 0$. This fact means that if we solve any differential equation $y' = \lambda y$ with a positive λ, the error grows for any integration step. This is not as serious as it seems at the first glance. Namely, the exact solution also grows in this case and the error is proportional to it. Such situation is quite acceptable.

TABLE 25.8

	p	α	C_{p+1}
Runge-Kutta	2	-2	
	3	$-2{\cdot}51$	
	4	$-2{\cdot}78$	
Adams-Bashforth	2	-1	$\frac{5}{12}$
	3	$-\frac{6}{11}$	$\frac{3}{8}$
	4	$-\frac{3}{10}$	$\frac{251}{720}$
Adams-Moulton	2	$-\infty$	$-\frac{1}{12}$
	3	-6	$-\frac{1}{24}$
	4	-3	$-\frac{19}{720}$

The values of the above parameter α for some particular methods are presented in Tab. 25.8, where p denotes the order of the corresponding method. In the case of Adams methods the error constant as an important characteristic of a multistep method (cf. (25.4.8)) is also introduced. The interval of absolute stability

of a Runge-Kutta method depends only on its order p and is independent of its particular form for $p \leqq 4$. The intervals of absolute stability of Adams-Moulton methods are substantially larger and their error constants are substantially smaller than those of Adams-Bashforth methods. Thus, for example, the implicit three-step Adams-Moulton method of order four has the interval of absolute stability 10 times larger and the error constant about 13 times smaller than the four-step Adams--Bashforth method possessing the same order. These considerations, whose results are also typical for more general comparisons between explicit and implicit methods, so favour implicit methods that explicit linear multistep methods are seldom used on their own; they do, however, play an important role in predictor-corrector pairs.

If we intend to use an implicit linear k-step method to solve the given initial value problem we must solve for y_{n+k} the equation (25.4.1) at each step. In general, this equation is nonlinear. Nevertheless, we know that the unique solution for y_{n+k} exists and can be approached arbitrarily closely by the iteration (25.4.3) provided that the condition (25.4.2) is satisfied. Each step of the iteration (25.4.3) clearly involves an evaluation of the value of the function f. Our concern thus is to keep to minimum the number of the applications of the iteration (25.4.3) — particularly so when the evaluation of f at given values of its arguments is time consuming. We would therefore like to make the initial guess $y_{n+k}^{(0)}$ as accurate as possible. We do this by using a separate explicit method to estimate y_{n+k} and taking this predicted value for the initial guess $y_{n+k}^{(0)}$. The explicit method is called, in this connection, the *predictor*, and the implicit method (25.4.3) the *corrector*.

We can now proceed in one of two different ways. The first consists in continuing the iteration (25.4.3) until the iterates have converged (in practice, until some criterion such as $\left|y_{n+k}^{(s+1)} - y_{n+k}^{(s)}\right| < \varepsilon$, where ε is a pre-assigned tolerance, say, of the order of the local round-off error, is satisfied). We then regard the value $y_{n+k}^{(s+1)}$ so obtained as an acceptable approximation of the exact solution y_{n+k} of (25.4.1). Since each iteration corresponds to one application of the corrector, we call this mode of operation of the predictor-corrector method *correcting to convergence*. In this mode, we cannot tell in advance how many iterations will be necessary, that is, how many function evaluations will be required at each step. On the other hand, the accepted value $y_{n+k}^{(s+1)}$ being independent of the initial guess $y_{n+k}^{(0)}$, the local discretization error and the interval of absolute stability of the overall method are precisely those of the corrector alone; the properties of the predictor are of no importance.

In the alternative approach, which is once again motivated by the desire to restrict the number of function evaluations per step, we stipulate in advance the number m of applications of the corrector at each step. It is exactly this manner of using predictor-corrector pairs which is often called the *predictor-corrector method*.

Let us describe it in more detail. It turns out to be advantageous if the predictor and the corrector are separately of the same order, and this requirement may well make it necessary for the stepnumber of the predictor to be greater than that of the corrector. The notationally simplest way to deal with this contingency is to let both predictor and corrector have the same stepnumber k, but in the case of the corrector, to relax the condition that not both α_0 and β_0 shall vanish. Thus, let the linear multistep method used as predictor be the method

$$\sum_{j=0}^{k} \alpha_j^* y_{n+j} = h \sum_{j=0}^{k-1} \beta_j^* f(x_{n+j}, y_{n+j}), \quad \alpha_k^* = 1 \tag{2}$$

and the implicit method used as corrector the method

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(x_{n+j}, y_{n+j}), \quad \alpha_k = 1. \tag{3}$$

Further, let m be a fixed positive integer. By *predictor-corrector method* we call the procedure in which, for the approximate solution at the point x_{n+k}, the number $y_{n+k}^{(m)}$ is taken which is computed from

$$\begin{gathered} y_{n+k}^{(0)} + \sum_{j=0}^{k-1} \alpha_j^* y_{n+j}^{(m)} = h \sum_{j=0}^{k-1} \beta_j^* f_{n+j}^{(m-1)}, \\ f_{n+k}^{(s)} = f(x_{n+k}, y_{n+k}^{(s)}), \\ y_{n+k}^{(s+1)} + \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{(m)} = h\beta_k f_{n+k}^{(s)} + h \sum_{j=0}^{k-1} \beta_j f_{n+j}^{(m-1)}, \\ s = 0, \ldots, m-1, \end{gathered} \tag{4}$$

or from

$$\begin{gathered} y_{n+k}^{(0)} + \sum_{j=0}^{k-1} \alpha_j^* y_{n+j}^{(m)} = h \sum_{j=0}^{k-1} \beta_j^* f_{n+j}^{(m)}, \\ f_{n+k}^{(s)} = f(x_{n+k}, y_{n+k}^{(s)}), \\ y_{n+k}^{(s+1)} + \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{(m)} = h\beta_k f_{n+k}^{(s)} + h \sum_{j=0}^{k-1} \beta_j f_{n+j}^{(m)}, \\ s = 0, \ldots, m-1, \\ f_{n+k}^{(m)} = f(x_{n+k}, y_{n+k}^{(m)}). \end{gathered} \tag{5}$$

The only difference between these two modes of the predictor-corrector method consists in using different values of f at that points at which the approximate

solution has already been computed. In (4), we use the values $f_r^{(m-1)}$ whilst in (5) the values $f_r^{(m)}$. Thus, the mode given by (5) calls for one more function evaluation per step than the mode (4). Let P indicate an application of the predictor, C a single application of the corrector and E an evaluation of f in terms of known values of its arguments. Then the procedure given by (4) with $m = 1$ is denoted by PEC (predict, evaluate, correct), with $m = 2$ by $PECEC = P(EC)^2$, generally by $P(EC)^m$. The procedure given by (5) is denoted by $PECE$, $P(EC)^2E$, ... From this symbolical notation one sees on the first glance how many evaluations of the right-hand term of the given differential equation are necessary for one step of the method.

Note that for $m \to \infty$, the results of computing with either of the above modes tend to those given by the mode of correcting to convergence.

If m is finite, then the predictor-corrector method does not belong, in fact, to the class of linear multistep methods. Nevertheless, the following theorem holds:

Theorem 1. *Let the predictor-corrector method, where the predictor has the order p^* and the error constant $C^*_{p^*+1}$ and the corrector the order p and the error constant C_{p+1}, be used in the mode $P(EC)^mE$ or $P(EC)^m$, where p^*, p, and m are integers, $p^* \geqq 0$, $p \geqq 1$, and $m \geqq 1$. If $p^* > p$, then the main part of the local error $l(y(x); h)$ of this algorithm is equal to the main part of the local error of the corrector, i.e.,*

$$l(y(x); h) = C_{p+1}y^{(p+1)}(x)h^{p+1} + O(h^{p+2}). \tag{6}$$

If $p^* = p - q$, where $0 < q \leqq p$, and if $m \geqq q + 1$, the main part of the local error of the algorithm is again the same as that of the corrector, if $m = q$ it is of the same order as that of the corrector but both errors are not identical, and, finally, if $m \leqq q - 1$ it is of the form $Kh^{p-q+m+1} + O(h^{p-q+m+2})$.

Thus, for example, if $m = 1$, that is, if we iterate only once, the order of the local discretization error of the combined method is equal to that of the corrector even though the order of the predictor is by one less than the order of the corrector.

We see from Theorem 1 that there is little to be gained by choosing a predictor--corrector method for which $p^* > p$; it would normally have an unnecessarily large stepnumber and the higher accuracy of the predictor would not be reflected in the local truncation error of the overall method. In fact, it seems that it would be advantageous to choose a method for which $p^* = p - m$. However, it turns out that when $p^* = p$ it is possible to estimate the main part of the local truncation error of the predictor-corrector method (which, as follows from Theorem 1, then coincides with that of the corrector) without estimating higher derivatives of the exact solution $y(x)$. This technique was originated by W.E.Milne, and we will refer to it as *Milne's device.* It consists in computing the main part of the local error of

the predictor corrector method according to the formula

$$C_{p+1}y^{(p+1)}(x_n)h^{p+1} = \frac{C_{p+1}}{C^*_{p+1} - C_{p+1}}\left(y^{(m)}_{n+k} - y^{(0)}_{n+k}\right). \tag{7}$$

The interval of absolute stability of the predictor-corrector method is, in general, different from that of the corrector and it depends on the mode in which the combined method is used.

Thus, for example, the interval of absolute stability of the combination of four--step Adams-Bashforth method of the fourth order as the predictor and the three--step fourth order Adams-Moulton method as the corrector is $(-1{\cdot}25, 0)$ in the PECE mode and only $(-0{\cdot}16, 0)$ for the PEC mode. This feature is general, in essence, and the PEC mode has usually substantially worse stability properties than the PECE mode.

At present the most practicable predictor-corrector methods are based on Adams methods used in the PECE mode.

REMARK 1. The application of linear k-step methods and predictor-corrector methods to *systems of linear differential equations* of the first order is not different from their use in the case of one differential equation. Again it is sufficient to suppose that the corresponding quantities are vectors and not scalars. But the problems of the stability are rather different. Namely, the philosophy of controlling the stepsize involved the claim that the product $z = \lambda h$, where λ is an estimate for $\partial f/\partial y$, would lie in the interval of absolute stability of the given method. In the case of a system of equations, the derivative $\partial f/\partial y$ is a matrix and one must take its eigenvalue for λ. Since this number is complex, in general, we must deal with the domains of absolute stability instead of intervals. The *domain of absolute stability* is introduced in the same way as the interval of absolute stability was, one only starts from the differential equation $y' = \lambda y$, where λ is complex.

REMARK 2. When solving *stiff differential system* — a typical example of such a differential system is the system $\boldsymbol{y'} = \boldsymbol{Ay}$ such that the eigenvalues λ_j of $\boldsymbol{A}$ have negative real parts and that the ratio $\max|\lambda_j|/\min|\lambda_j|$ is a large number (even of order 10^6 in some practical situations) — the limitations on the stepsize implied by the stability may be extremely restrictive. Thus, for solving stiff differential systems, such methods that their domain of stability contains the whole left-hand half-plane of the complex plane (the so-called *A-stable methods*) or at least the infinite angle $\{z;\ -\alpha < \pi - \arg z < \alpha\}$ (*$A(\alpha)$-stable methods*) are mostly convenient. All such methods are necessarily implicit. The one-step Adams-Moulton method and the methods of numerical differentiation from Tab. 25.6 with $k = 1$ and $k = 2$ may serve as examples of A-stable methods. The remaining methods from this table form examples of $A(\alpha)$-stable methods with $\alpha = 88°$, $73°$, $51°$, and $18°$. Note also

that the method of successive approximations for solving the corresponding equations is not suitable when solving a stiff-system. The Newton method (cf. § 31.4) can rather be recommended. The calculations are then extremely laborious.

25.6. Extrapolation Methods. Richardson's Extrapolation, Gragg's Method

Extrapolation methods for the numerical solution of differential equations take their origin in the application of the general idea of *Richardson's extrapolation* which can be used also in many other branches of numerical analysis. In many situations in numerical analysis we wish to evaluate a number φ_0 but we are not able to do it directly (for example, we are to compute the value of the definite integral of a function the primitive of which cannot be represented by means of elementary functions or we are to compute the value of the solution of a differential equation which is not integrable by elementary functions). What we are only able to do is to compute an approximation $\varphi(h)$ of φ_0 where h is a positive parameter (typically the steplength) and where $\varphi(h) \to \varphi_0$ as $h \to 0$. The main idea of Richardson's extrapolation consists in the following considerations: Our task is, in fact, to evaluate φ at the point $h = 0$. This is not possible directly so, instead of it, we choose a finite number of h's,

$$h_0 > h_1 > h_2 > \cdots > h_M\,, \tag{1}$$

construct a function $\tilde{\varphi}(h)$ which interpolates $\varphi(h)$ at the points (1), i.e., for which

$$\tilde{\varphi}(h_s) = \varphi(h_s)\,, \quad s = 0,\ \ldots,\ M\,, \tag{2}$$

holds and take then the number $\tilde{\varphi}(0)$ for the new approximation.

If $\varphi(h)$ possesses an asymptotic expansion of the form

$$\varphi(h) = \varphi_0 + \varphi_1 h + \varphi_2 h^2 + \ldots\,, \tag{3}$$

where the coefficients $\varphi_0, \varphi_1, \ldots$ are independent of h, the function $\tilde{\varphi}(h)$ is the usual interpolation polynomial of degree M determined by the conditions (2). In this case, $\tilde{\varphi}(0)$ can be computed recurrently in such a way that we put

$$a_{0s} = \varphi(h_s)\,, \quad s = 0,\ 1,\ \ldots\,, \tag{4}$$

and, successively for $m = 1, 2, \ldots$, compute the numbers a_{ms} from

$$a_{ms} = \frac{h_s/h_{m+s}}{(h_s/h_{m+s}) - 1} a_{m-1,s+1} - \frac{1}{(h_s/h_{m+s}) - 1} a_{m-1,s}\,, \quad s = 0,\ 1,\ \ldots\,. \tag{5}$$

Then we have

$$\tilde{\varphi}(0) = a_{M0}\,. \tag{6}$$

If the asymptotic expansion for $\varphi(h)$ has the form

$$\varphi(h) = \varphi_0 + \varphi_2 h^2 + \varphi_4 h^4 + \dots \tag{7}$$

(this is rather frequent situation in practice), we take an even polynomial for $\tilde{\varphi}(h)$. The computational scheme is similar to that discussed above, only the recurrence (5) has to be replaced by

$$a_{ms} = \frac{(h_s/h_{m+s})^2}{(h_s/h_{m+s})^2 - 1} a_{m-1,s+1} - \frac{1}{(h_s/h_{m+s})^2 - 1} a_{m-1,s}\,. \tag{8}$$

The computations according to (4), (5) or (4), (8), respectively, can be arranged in the form of the so-called *T-scheme* as indicated in Tab. 25.9. To compute any element of this scheme means (see (5) or (8), respectively) to form a linear combination of the element lying in the same row and the preceding column and the element lying directly above the latter.

TABLE 25.9

a_{00}					
a_{01}	a_{10}				
a_{02}	a_{11}	a_{20}			
a_{03}	a_{12}	a_{21}	a_{30}		
$\vdots$	$\vdots$	$\vdots$	$\vdots$		
a_{0M}	$a_{1,M-1}$	$a_{2,M-2}$	$a_{3,M-3}$	$\cdots$	a_{M0}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$

For the columns of the T-scheme we have

$$a_{ms} = \varphi_0 + O(h_s^{m+1}) \tag{9}$$

or

$$a_{ms} = \varphi_0 + O(h_s^{2m+2})\,, \tag{10}$$

respectively, according to the validity of (3) or (7). Any further column converges to the exact value more rapidly than the preceding one and the diagonal again more rapidly than any column. The extrapolation based on the expansion (7) is, evidently, more effective than that based on (3).

REMARK 1. Romberg's quadrature formula (cf. § 13.13) took its origin exactly in the way just described from the trapezoidal rule used successively for $h = h_s = (b-a)/2^s$.

If we want to use the just described extrapolation method for initial value problems we can proceed as follows: For a given fixed numerical method (linear multistep, Runge-Kutta etc.), let $y(x; h)$ denote the approximation at the point x, given by the numerical method with steplength h, to the theoretical solution $y(x)$ of the initial value problem (25.2.1), (25.2.2). We intend to use the polynomial extrapolation to furnish approximations to $y(x)$ at the basic points $x_j = x_0 + jH$, $j = 0, 1, \ldots$, where H is the basic steplength. (For a given problem and required accuracy, H will typically be large as compared with the appropriate steplengths for the previously discussed methods.) We first choose a steplength $h_0 = H/N_0$, where N_0 is a positive integer (possibly 1), and apply the numerical method N_0 times starting from $x = x_0$ to obtain an approximation $y(x_1; h_0)$ to the theoretical solution $y(x_1)$. A second steplength $h_1 = H/N_1$, N_1 a positive integer greater than N_0, is chosen, and the method is applied N_1 times, again starting from x_0, to yield the approximation $y(x_1; h_1)$. Putting, in general, $h_s = H/N_s$ for $s = 0, \ldots, M$, where N_s is an increasing sequence of positive integers, and proceeding in this fashion, we obtain the sequence of approximations $y(x_1; h_s)$, $s = 0, \ldots, M$, to $y(x_1)$. (In practice, M is typically in the range 4 to 7.) Provided that there exists, for the given numerical method, an asymptotic expansion of the form

$$y(x; h) = y(x) + A_1 h + A_2 h^2 + \ldots , \tag{11}$$

then we can set $a_{0s} = y(x_1; h_s)$ and apply the process of polynomial extrapolation based on (5). The equation (8), of course, replaces (5) in the case when the numerical method possesses an asymptotic expansion of the form

$$y(x; h) = y(x) + A_2 h^2 + A_4 h^4 + \ldots . \tag{12}$$

We then take a_{M0} on the diagonal of the T-scheme for our final approximation to $y(x_1)$. To obtain a numerical solution at the next basic point $x_2 = x_0 + 2H$, we apply the whole of the above procedure to the new initial value problem $y' = f(x, y)$, $y(x_1) = a_{M0}$. Further repetitions of this process yield the approximate solution at all basic points.

The most often used starting method for the just described extrapolation procedure is *Gragg's method* (also called the *modified mid-point method*). Its algorithm is defined as follows:

$$\begin{gathered} h_s = H/N_s, \quad N_s \text{ even}, \\ y_0 = y(x_0), \\ y_1 = y_0 + h_s f(x_0, y_0), \\ y_{m+2} = y_m + 2h_s f(x_{m+1}, y_{m+1}), \quad m = 0, \ldots, N_s - 1, \\ y(x_1; h_s) = \tfrac{1}{4} y_{N_s+1} + \tfrac{1}{2} y_{N_s} + \tfrac{1}{4} y_{N_s-1} . \end{gathered} \tag{13}$$

Two popular choices of the sequence N_s are $\{2, 4, 6, 8, 12, 16, 24, \dots\}$, generally $N_s = 2N_{s-2}$ (with the exception of the first three terms) and $\{2, 4, 8, 16, 32, 64, \dots\}$, generally $N_s = 2N_{s-1}$. For Gragg's method, the asymptotic formula (12) is valid so that the extrapolation is performed according to (8).

Numerical experience shows that the extrapolation method based on Gragg's method is sufficiently efficient especially in the case that we are interested in the solution only at relatively few points.

B. BOUNDARY VALUE PROBLEMS

25.7. Shooting Method

This method, like the methods which will be described in § 25.8, is based on the idea of transforming the boundary value problem into one or several initial value problems and to solve this initial value problems by means of the methods of Part A. Thus, the solution of initial value problems is supposed to be an elementary operation in this connection.

To describe the shooting method, let us investigate the system (25.1.1) of m differential equations with a general boundary condition (25.1.5), i.e., the system

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad x \in [a, b], \tag{1}$$

with the condition

$$\boldsymbol{r}(\boldsymbol{y}(a), \boldsymbol{y}(b)) = \boldsymbol{o}. \tag{2}$$

In the algorithm of the shooting method we first choose a vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)^T$ arbitrarily and solve the system (1) with the initial condition

$$\boldsymbol{y}(a) = \boldsymbol{\alpha}. \tag{3}$$

The solution obtained in this way depends, consequently, on the parameter $\boldsymbol{\alpha}$ and we denote it, therefore, by $\boldsymbol{y}(x; \boldsymbol{\alpha})$. Further, solve the (generally nonlinear) system of equations

$$\boldsymbol{F}(\boldsymbol{\alpha}) = \boldsymbol{o}, \tag{4}$$

where the vector function $\boldsymbol{F}$ is defined by

$$\boldsymbol{F}(\boldsymbol{\alpha}) = \boldsymbol{r}(\boldsymbol{\alpha}, \boldsymbol{y}(b; \boldsymbol{\alpha})). \tag{5}$$

If the solution of (5) is denoted by $\boldsymbol{\alpha}^*$, the solution of the given boundary value problem is found as the solution of (1) with the initial condition

$$\boldsymbol{y}(a) = \boldsymbol{\alpha}^* . \tag{6}$$

The important part of the shooting method is, thus, the solution of the nonlinear system (4). If we use the Newton method (cf. § 31.4) for this purpose we construct a sequence of approximations according to the formula

$$\boldsymbol{\alpha}^{(i+1)} = \boldsymbol{\alpha}^{(i)} - [\mathrm{D}\boldsymbol{F}(\boldsymbol{\alpha}^{(i)})]^{-1}\boldsymbol{F}(\boldsymbol{\alpha}^{(i)}) , \tag{7}$$

where the symbol $\mathrm{D}\boldsymbol{F}$ denotes the (Gâteau) differential of $\boldsymbol{F}$, i.e., the matrix $\{\partial F_j/\partial\alpha_k\}$. Hence, to perform one iteration, one must evaluate $\boldsymbol{F}$ at the point $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(i)}$, evaluate the functional matrix $\mathrm{D}\boldsymbol{F}(\boldsymbol{\alpha}^{(i)})$ and solve the system of linear algebraic equations

$$\mathrm{D}\boldsymbol{F}(\boldsymbol{\alpha}^{(i)})(\boldsymbol{\alpha}^{(i+1)} - \boldsymbol{\alpha}^{(i)}) = -\boldsymbol{F}(\boldsymbol{\alpha}^{(i)}) . \tag{8}$$

To evaluate $\boldsymbol{F}$ at $\boldsymbol{\alpha}^{(i)}$ means to compute the vector $\boldsymbol{y}(b; \boldsymbol{\alpha}^{(i)})$, i.e., to solve the differential system (1) with the initial condition (3). The functional matrix $\mathrm{D}\boldsymbol{F}(\boldsymbol{\alpha})$ is defined by

$$\mathrm{D}\boldsymbol{F}(\boldsymbol{\alpha}) = \mathrm{D}_{\boldsymbol{u}}\boldsymbol{r}(\boldsymbol{\alpha}, \boldsymbol{y}(b; \boldsymbol{\alpha})) + \mathrm{D}_{\boldsymbol{v}}\boldsymbol{r}(\boldsymbol{\alpha}, \boldsymbol{y}(b; \boldsymbol{\alpha}))\mathrm{D}_{\boldsymbol{\alpha}}\boldsymbol{y}(b; \boldsymbol{\alpha}) , \tag{9}$$

where

$$\begin{aligned} \mathrm{D}_{\boldsymbol{u}}\boldsymbol{r}(\boldsymbol{u}, \boldsymbol{v}) &= \left(\frac{\partial r_j(\boldsymbol{u}, \boldsymbol{v})}{\partial u_k}\right) , \\ \mathrm{D}_{\boldsymbol{v}}\boldsymbol{r}(\boldsymbol{u}, \boldsymbol{v}) &= \left(\frac{\partial r_j(\boldsymbol{u}, \boldsymbol{v})}{\partial v_k}\right) , \end{aligned} \tag{10}$$

and the matrix $\mathrm{D}_{\boldsymbol{\alpha}}\boldsymbol{y}(b; \boldsymbol{\alpha}) = (\partial y_j(b; \boldsymbol{\alpha})/\partial\alpha_k)$ is the solution of the matrix differential equation

$$(\mathrm{D}_{\boldsymbol{\alpha}}\boldsymbol{y}(x; \boldsymbol{\alpha}))' = \mathrm{D}_{\boldsymbol{y}}\boldsymbol{f}(x, \boldsymbol{y}(x; \boldsymbol{\alpha}))\mathrm{D}_{\boldsymbol{\alpha}}\boldsymbol{y}(x; \boldsymbol{\alpha}) \tag{11}$$

with the initial condition

$$\mathrm{D}_{\boldsymbol{\alpha}}\boldsymbol{y}(a; \boldsymbol{\alpha}) = \boldsymbol{I} . \tag{12}$$

The prime denotes here the differentiation with respect to x and $\boldsymbol{I}$ is the identity matrix of order m.

The matrix $\mathrm{D}\boldsymbol{F}(\boldsymbol{\alpha})$ is usually approximated by the matrix

$$\Delta\boldsymbol{F}(\boldsymbol{\alpha}) = \Delta_1\boldsymbol{F}(\boldsymbol{\alpha}), \ \ldots, \ \Delta_m\boldsymbol{F}(\boldsymbol{\alpha}) , \tag{13}$$

where the k-th column of $\Delta \boldsymbol{F}(\boldsymbol{\alpha})$ is defined by

$$\Delta_k \boldsymbol{F}(\boldsymbol{\alpha}) = \frac{\boldsymbol{F}(\alpha_1, \ldots, \alpha_{k-1}, \alpha_k + \Delta\alpha_k, \alpha_{k+1}, \ldots, \alpha_m)}{\Delta\alpha_k}. \tag{14}$$

Taking into account the definition of $\boldsymbol{F}$ we see that the vector $\Delta_k \boldsymbol{F}(\boldsymbol{\alpha})$ is determined by the vectors

$$\boldsymbol{y}(b; \alpha_1, \ldots, \alpha_{k-1}, \alpha_k + \Delta\alpha_k, \alpha_{k+1}, \ldots, \alpha_m) \text{ and } \boldsymbol{y}(b; \alpha_1, \ldots, \alpha_m),$$

i.e., we again compute it by solving initial value problems.

Hence, to compute one Newton iteration, it is necessary to solve $m + 1$ initial value problems for the original system and a system of m linear algebraic equations.

In the case of a linear system

$$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{f} \tag{15}$$

with linear boundary conditions

$$\boldsymbol{U}\boldsymbol{y}(a) + \boldsymbol{V}\boldsymbol{y}(b) = \boldsymbol{c}, \tag{16}$$

the function $\boldsymbol{F}$ has the form

$$\boldsymbol{F}(\boldsymbol{\alpha}) = \boldsymbol{U}\boldsymbol{\alpha} + \boldsymbol{V}\boldsymbol{y}(b; \boldsymbol{\alpha}) - \boldsymbol{c}, \tag{17}$$

where $\boldsymbol{y}(x; \boldsymbol{\alpha})$ is the solution of (15) satisfying the initial condition (3). But it is possible to write this solution as

$$\boldsymbol{y}(x; \boldsymbol{\alpha}) = \boldsymbol{\Phi}_a(x)\boldsymbol{\alpha} + \boldsymbol{y}(x; \boldsymbol{o}), \tag{18}$$

where $\boldsymbol{\Phi}_a$ is the fundamental matrix of the system (15) for the point a, that is, the matrix which satisfies $\boldsymbol{\Phi}_a' = \boldsymbol{A}\boldsymbol{\Phi}_a$, $\boldsymbol{\Phi}_a(a) = \boldsymbol{I}$. The solution of the system (4) is then given by

$$\boldsymbol{\alpha}^* = [\boldsymbol{U} + \boldsymbol{V}\boldsymbol{\Phi}_a(b)]^{-1} [\boldsymbol{c} - \boldsymbol{V}\boldsymbol{y}(b; \boldsymbol{o})] \tag{19}$$

and the solution $\boldsymbol{y}(x; \boldsymbol{\alpha}^*)$ of the original boundary value problem by

$$\boldsymbol{y}(x; \boldsymbol{\alpha}^*) = \boldsymbol{\Phi}_a(x)\boldsymbol{\alpha}^* + \boldsymbol{y}(x; \boldsymbol{o}). \tag{20}$$

Hence, in the linear case, the solution is obtained by solving $m + 1$ initial value problems (for the m columns of the fundamental matrix $\boldsymbol{\Phi}_a$ and for the vector $\boldsymbol{y}(x; \boldsymbol{o})$) and by solving a system of linear algebraic equations.

REMARK 1. In the case that the boundary conditions are separated we choose the parameter α a priori in such a way that, say, the left-hand boundary condition is satisfied. The system of linear algebraic equations has then a smaller order.

Example 1. Solve the system

$$\begin{aligned} {}^1y' &= {}^2y\,, \\ {}^2y' &= 100\,{}^1y \end{aligned} \tag{21}$$

with the linear separated boundary conditions ${}^1y(0) = {}^1y(10) = 1$.

The function ${}^1y$ is, in this case, given by

$${}^1y(x) = \frac{e^{100} - 1}{e^{100} - e^{-100}} e^{-10x} + \frac{1 - e^{-100}}{e^{100} - e^{-100}} e^{10x}$$

and the function ${}^2y$ is its derivative. The initial conditions for ${}^1y$ and ${}^2y$ at $x = 0$ are

$${}^1y(0) = 1\,, \quad {}^2y(0) = -10 + 20\frac{1 - e^{-100}}{e^{100} - e^{-100}}\,.$$

It can be seen from here that the value of ${}^2y$ at the point 0 differs from -10 only starting from 43rd decimal place. Thus, if we compute in floating point with, say, 15 decimal places of mantissa (the actual computers very rarely compute more precisely) we compute the number $-10(1+\varepsilon)$, where ε is of order 10^{-15} in the best of cases, as the approximation of the exact value ${}^2y(0)$. Solving the system (21) with the initial conditions ${}^1y(0) = 1$, ${}^2y(0) = -10(1+\varepsilon)$ we obtain

$${}^1y(x) = -\tfrac{1}{2}\varepsilon e^{10x} + (1 + \tfrac{1}{2}\varepsilon)e^{-10x}\,.$$

If we have, for example, $\varepsilon = 10^{-16}$, we get

$${}^1y(10) \approx \tfrac{1}{2}10^{-16}e^{100} \approx 1{\cdot}34\,.\,10^{27}\,,$$

and we should obtain 1.

This simple example shows that the initial value problems arising in the shooting method may be very sensitive to small changes in initial conditions and, at the same time, that this phenomenon may be the worse the longer the interval, in which the solution is sought, is.

The *multishooting method*, the principles of which will be described now, removes in a high degree most of the above mentioned difficulties connected with the shooting method. The cost which we must pay for it is a great increase of the number of arithmetic operations. In order to describe the multishooting method, investigate the boundary value problem (1), (2), and let $x_0, \ldots, x_n$ be such points of $[a, b]$

for which $a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$. Denote by $\boldsymbol{y}(x; x_k, \boldsymbol{\alpha}_k)$ the solution of (1) satisfying the initial condition $\boldsymbol{y}(x_k) = \boldsymbol{\alpha}_k = (\alpha_{k1}, \ldots, \alpha_{km})^{\mathrm{T}}$. If we find $n+1$ vectors $\boldsymbol{\alpha}_k$, $k = 0, \ldots, n$, such that

$$\boldsymbol{y}(x_{k+1}; x_k, \boldsymbol{\alpha}_k) = \boldsymbol{\alpha}_{k+1}, \quad k = 0, \ldots, n-1, \tag{22}$$

and

$$\boldsymbol{r}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_n) = \boldsymbol{o}, \tag{23}$$

then the function $\boldsymbol{y} = \boldsymbol{y}(x)$ defined by the formula

$$\boldsymbol{y}(x) = \boldsymbol{y}(x; x_k, \boldsymbol{\alpha}_k) \text{ for } x \in [x_k, x_{k+1}]$$

is the solution of the given boundary value problem.

The crucial point of the multishooting method thus is the solution of the system (22), (23) of $(n+1)m$ (generally nonlinear) equations for $(n+1)m$ unknowns. If we again use Newton's method and again approximate the corresponding derivatives by difference quotients as in the simple shooting method, we must solve $(n+1)m$ initial value problems for the original differential equation and a linear algebraic system of $(n+1)m$ unknowns in order to perform one iteration. The special form of the system (22), (23) allows to reduce the linear algebraic system to a system of order m. Nevertheless, the multishooting method still remains to be extremely laborious. However, it is just only this method which gives a chance to get any results at all in some situations.

25.8. Methods of the Transfer and Normalized Transfer of Boundary Conditions

In these methods, we again want to transform the original boundary value problem into a sequence of initial value problems. In contrast to the shooting method, these methods work only in the case of linear equations with linear separated boundary conditions.

Thus, let $\boldsymbol{A}$ be an $m \times m$ matrix the entries of which are continuous functions and let $\boldsymbol{f}$ be an m-dimensional vector with continuous components. Investigate the differential equation

$$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{f} \tag{1}$$

with the boundary conditions

$$\boldsymbol{V}_a\boldsymbol{y}(a) = \boldsymbol{v}_a, \quad \boldsymbol{V}_b\boldsymbol{y}(b) = \boldsymbol{v}_b, \tag{2}$$

where $\boldsymbol{V}_a$ and $\boldsymbol{V}_b$ are $m_a \times m$ and $m_b \times m$ matrices and $\boldsymbol{v}_a$ and $\boldsymbol{v}_b$ are m_a- and m_b-dimensional vectors, respectively.

The *method of the transfer of boundary conditions* is based on the following theorems.

Theorem 1. *Let the $m_a \times m$ and $m_b \times m$ matrices $\boldsymbol{R}_a(x)$ and $\boldsymbol{R}_b(x)$, respectively, satisfy the differential equation*

$$\boldsymbol{R}_a' = -\boldsymbol{R}_a \boldsymbol{A} \tag{3}$$

with the initial condition

$$\boldsymbol{R}_a(a) = \boldsymbol{V}_a\,, \tag{4}$$

and the differential equation

$$\boldsymbol{R}_b' = -\boldsymbol{R}_b \boldsymbol{A} \tag{3'}$$

with the initial condition

$$\boldsymbol{R}_b(b) = \boldsymbol{V}_b\,, \tag{4'}$$

respectively. Further, let the m_a- and m_b-dimensional vectors $\boldsymbol{r}_a$ and $\boldsymbol{r}_b$, respectively, satisfy the differential equation

$$\boldsymbol{r}_a' = \boldsymbol{R}_a \boldsymbol{f} \tag{5}$$

with the initial condition

$$\boldsymbol{r}_a(a) = \boldsymbol{v}_a \tag{6}$$

and

$$\boldsymbol{r}_b' = \boldsymbol{R}_b \boldsymbol{f} \tag{5'}$$

with the initial condition

$$\boldsymbol{r}_b(b) = \boldsymbol{v}_b\,. \tag{6'}$$

Then any vector $\boldsymbol{y}$, which is the solution of (1) in $[a, b]$ and which, moreover, satisfies the first or the second of the boundary conditions (2) fulfils

$$\boldsymbol{R}_a(x)\boldsymbol{y}(x) = \boldsymbol{r}_a(x), \quad x \in [a, b]\,, \tag{7}$$

or

$$\boldsymbol{R}_b(x)\boldsymbol{y}(x) = \boldsymbol{r}_b(x), \quad x \in [a, b]\,, \tag{7'}$$

respectively.

The assertion of Theorem 1 can be expressed by words in such a way that any solution of (1) which, moreover, satisfies one linear condition of the type (2) satisfies the condition of the same type at any point of the given interval. Thus, using this theorem, we can transfer any condition of the type (2) to any point of $[a, b]$.

Theorem 2. *Let x_0 be any point from $[a, b]$. If the boundary value problem* (1), (2) *has a solution $\boldsymbol{y}$, then the vector $\boldsymbol{k} = \boldsymbol{y}(x_0)$ solves the system of linear algebraic equations*

$$\begin{bmatrix} \boldsymbol{R}_a(x_0) \\ \boldsymbol{R}_b(x_0) \end{bmatrix} \boldsymbol{k} = \begin{bmatrix} \boldsymbol{r}_a(x_0) \\ \boldsymbol{r}_b(x_0) \end{bmatrix}, \tag{8}$$

where the matrices $\boldsymbol{R}_a$, $\boldsymbol{R}_b$ and the vectors $\boldsymbol{r}_a$, $\boldsymbol{r}_b$ are defined in Theorem 1. *Conversely, if the system* (8) *has a solution $\boldsymbol{k}^*$, then the solution of* (1) *with the initial condition $\boldsymbol{y}(x_0) = \boldsymbol{k}^*$ is the solution of the original boundary value problem. If the boundary value problem* (1), (2) *has a unique solution, then the algebraic system* (8) *also has a unique solution, and conversely.*

Thus, solving the algebraic system (8), we can obtain the initial conditions which are satisfied by the solution of the original boundary value problem; its solution is then computed by solving an initial value problem. Since the matrices $\boldsymbol{R}_a$, $\boldsymbol{R}_b$ and the vectors $\boldsymbol{r}_a$, $\boldsymbol{r}_b$ are also computed by solving initial value problems we have transformed the original boundary value problem into a sequence of initial value problems. Using this approach in practice, one usually puts $x_0 = b$ or $x_0 = a$ since then $\boldsymbol{R}_b$ and $\boldsymbol{r}_b$ or $\boldsymbol{R}_a$ and $\boldsymbol{r}_a$ need not be constructed.

Since we have $\boldsymbol{R}_a = \boldsymbol{V}_a\boldsymbol{\Phi}_a^{-1}$ and $\boldsymbol{R}_b = \boldsymbol{V}_b\boldsymbol{\Phi}_b^{-1}$, where $\boldsymbol{\Phi}_a$ and $\boldsymbol{\Phi}_b$ are fundamental matrices of (1) for the points a and b, respectively, we solve very similar differential equations (cf. (25.7.18)) in the just described method of the transfer of boundary condition as in the shooting method. The problems connected with the practical realization of the method of the transfer of boundary condition are thus similar to those in the case of the shooting method. We will clarify them by a simple example:

Example 1. Solve the differential equation

$$-\left[(1+x)y'\right]' + qy = \left[q + \pi^2(1+x)\right]\sin \pi x - \pi \cos \pi x\,,$$

with the boundary conditions $y(0) = y(1) = 0$.

This problem is equivalent to the problem of solving the system

$$\begin{aligned} {}^1y' &= \frac{1}{1+x}\,{}^2y\,, \\ {}^2y' &= q\,{}^1y - \left[q + \pi^2(1+x)\right]\sin \pi x + \pi \cos \pi x \end{aligned}$$

with the boundary conditions

$${}^1y(0) = {}^1y(1) = 0\,.$$

The exact solution is ${}^1y = y = \sin \pi x$, ${}^2y = (1+x)y'$. In Tab. 25.10 we find the results for $q = 100$ and $q = 500$. The necessary initial value problems were solved by

the standard Runge-Kutta method with the steplength $h = 0{\cdot}025$. It is apparent from this table that the results, especially in the case $q = 500$, are completely

TABLE 25.10

x	$q = 100$		$q = 500$	
	approximate solution	error	approximate solution	error
0·0	0·000275	0·000275	163·484	163·484
0·1	0·309114	0·000097	18·3247	18·0157
0·2	0·587819	0·000034	2·77973	2·19195
0·3	0·809028	0·000011	1·09989	0·290869
0·4	0·951059	0·000003	0·992730	0·041673
0·5	1·000000	0	1·006387	0·006387
0·6	0·951056	0	0·952091	0·001034
0·7	0·809017	0	0·809189	0·000172
0·8	0·587785	0	0·587813	0·000028
0·9	0·309017	0	0·309021	0·000003
1·0	0	0	0	0

TABLE 25.11

x	$q = 100$		$q = 500$	
	approximate solution	error	approximate solution	error
0·0	0	0	0	0
0·1	0·309016	−0·000001	0·309012	−0·000005
0·2	0·587781	−0·000004	0·587774	−0·000011
0·3	0·809011	−0·000005	0·809006	−0·000011
0·4	0·951039	−0·000017	0·950562	−0·000495
0·5	0·999936	−0·000064	0·997559	−0·002441
0·6	0·950989	−0·000068	0·953125	0·002068
0·7	0·808945	−0·000072	0·710938	−0·098080
0·8	0·587387	−0·000398	0·625000	0·037215
0·9	0·307861	−0·001156	0	−0·309018
1·0	−0·000244	−0·000244	−1·000000	−1·000000

unsatisfactory. To provide a possibility of comparison, the results obtained by the simple shooting method are also presented (Tab. 25.11). This method gives similarly bad results, too.

The *method of the normalized transfer of boundary conditions* which will be described now reduces these problems substantially. In comparison with the ordinary transfer of boundary condition, this modified procedure need not be feasible even in some of cases when the original problem has a unique solution.

Before formulating the assertion on which the method is based we introduce the following notation. Write the matrix $\boldsymbol{V}_a$ from the boundary condition (2) in the form

$$\boldsymbol{V}_a = \left[\boldsymbol{V}_a^{(1)}, \boldsymbol{V}_a^{(2)}\right], \tag{9}$$

where $\boldsymbol{V}_a^{(1)}$ is the $m_a \times m_a$ matrix formed from the first m_a columns of $\boldsymbol{V}_a$. Similarly, put

$$\boldsymbol{V}_b = \left[\boldsymbol{V}_b^{(1)}, \boldsymbol{V}_b^{(2)}\right], \tag{10}$$

where $\boldsymbol{V}_b^{(1)}$ is the $m_b \times m_a$ matrix formed from the first m_a columns of $\boldsymbol{V}_b$. (Note that it is not necessary that $m_a = m_b$.) Finally, write $\boldsymbol{A}$ and $\boldsymbol{f}$ from (1) in the form

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11}, & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21}, & \boldsymbol{A}_{22} \end{bmatrix}, \quad \boldsymbol{f} = \begin{bmatrix} \boldsymbol{f}_1 \\ \boldsymbol{f}_2 \end{bmatrix}, \tag{11}$$

where $\boldsymbol{A}_{11}$ is the $m_a \times m_a$ matrix formed from the first m_a rows and columns of $\boldsymbol{A}$ and the types of the matrices $\boldsymbol{A}_{12}$, $\boldsymbol{A}_{21}$, and $\boldsymbol{A}_{22}$ are $m_a \times (m - m_a)$, $(m - m_a) \times m_a$, and $(m - m_a) \times (m - m_a)$, respectively; $\boldsymbol{f}_1$ is then the m_a-dimensional vector formed from the first m_a components of $\boldsymbol{f}$.

Theorem 3. *Let $\boldsymbol{y}$ solve the differential equation (1) in $[a, b]$ and let it satisfy the first boundary condition (2) in which the matrix $\boldsymbol{V}_a^{(1)}$ is nonsingular. Further, let the matrix differential equation*

$$\boldsymbol{G}' = \boldsymbol{A}_{11}\boldsymbol{G} - \boldsymbol{G}\boldsymbol{A}_{22} - \boldsymbol{G}\boldsymbol{A}_{21}\boldsymbol{G} + \boldsymbol{A}_{12} \tag{12}$$

with the initial condition

$$\boldsymbol{G}(a) = -(\boldsymbol{V}_a^{(1)})^{-1}\boldsymbol{V}_a^{(2)} \tag{13}$$

have a solution in the whole interval $[a, b]$. Then

$$\boldsymbol{y}_1(x) = \boldsymbol{G}(x)\boldsymbol{y}_2(x) + \boldsymbol{g}(x) \tag{14}$$

holds for any $x \in [a, b]$, where we put $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)^{\mathrm{T}}$, $\boldsymbol{y}_1$ being the m_a-dimensional vector formed from the first m_a components of $\boldsymbol{y}$, and where

$$\boldsymbol{g}' = (\boldsymbol{A}_{11} - \boldsymbol{G}\boldsymbol{A}_{21})\boldsymbol{g} + \boldsymbol{f}_1 - \boldsymbol{G}\boldsymbol{f}_2 \tag{15}$$

with the initial condition

$$\boldsymbol{g}(a) = (\boldsymbol{V}_a^{(1)})^{-1}\boldsymbol{v}_a. \tag{16}$$

REMARK 1. The difference between Theorem 1 and Theorem 3 consists in the fact that Theorem 3 gives the possibility to transfer the left-hand boundary condition in the "normalized" form, i.e., in the form where the identity matrix stands for the first m_a components of $\boldsymbol{R}_a$.

REMARK 2. The assumption on the existence of the solution of the differential equation (12) in the whole interval $[a, b]$ cannot be removed. Namely, (12) is a nonlinear differential equation so that existence of its solution is guaranteed by the general existence theorem (cf. § 17.2) only locally.

Theorem 4. *Let the assumptions of Theorem* 3 *be satisfied and let the boundary value problem* (1), (2) *have a unique solution. Then the system of linear algebraic equations*

$$\left[\boldsymbol{V}_b^{(2)} + \boldsymbol{V}_b^{(1)}\boldsymbol{G}(b)\right]\boldsymbol{k} = \boldsymbol{v}_b - \boldsymbol{V}_b^{(1)}\boldsymbol{g}(b) \tag{17}$$

has a unique solution $\boldsymbol{k}^*$ *and the component* $\boldsymbol{y}_2$ *of the solution of the given boundary value problem satisfies the differential equation*

$$\boldsymbol{y}_2' = (\boldsymbol{A}_{22} + \boldsymbol{A}_{21}\boldsymbol{G})\boldsymbol{y}_2 + \boldsymbol{f}_2 + \boldsymbol{A}_{21}\boldsymbol{g} \tag{18}$$

with the initial condition

$$\boldsymbol{y}_2(b) = \boldsymbol{k}^*.$$

The algorithm of the method of the normalized transfer of boundary condition is based on Theorems 3 and 4 and is described as follows:

(i) We solve the system of differential equations (12) with the initial conditions (13) and the system (15) with the initial conditions (16).

(ii) The component $\boldsymbol{y}_2$ of the solution sought is found as the solution of the system (18) with the initial conditions at the point b computed from the algebraic system (17).

(iii) The component $\boldsymbol{y}_1$ of the solution (if we are interested in it) is then computed from (14).

Hence, the method of the normalized transfer of boundary condition consists in solving two systems of differential equations, one of them being nonlinear, with the initial conditions prescribed at the point a (i.e., we solve these equations from the left to the right) and one system of differential equations with the initial conditions given at the point b (this system is thus solved from the right to the left).

REMARK 3. The alternative name of the just described procedure is the *invariant imbedding method.*

Example 2. Solve, by the method of the normalized transfer of boundary condition, the boundary value problem from Example 1.

The results are summarized in Tab 25.12. Again, the necessary initial value problems were solved by the standard Runge-Kutta method with the steplength $h = 0{\cdot}025$. We really see that they are substantially more reasonable than in the case of the method of the ordinary transfer or in the case of the shooting method.

TABLE 25.12

x	$q = 100$		$q = 500$	
	approximate solution	error	approximate solution	error
0·0	0	0	0	0
0·1	0·309012	−0·000005	0·309007	−0·000010
0·2	0·587778	−0·000007	0·587770	−0·000015
0·3	0·809009	−0·000008	0·808999	−0·000018
0·4	0·951048	−0·000009	0·951037	−0·000019
0·5	0·999989	−0·000011	0·999979	−0·000021
0·6	0·951045	−0·000012	0·951038	−0·000018
0·7	0·809009	−0·000008	0·809007	−0·000010
0·8	0·587784	−0·000001	0·587780	−0·000005
0·9	0·309018	0·000001	0·309016	−0·000001
1·0	0	0	0	0

25.9. Finite Difference Method

This method is very popular and represents, in principle, a universal method for solving boundary value problems not only for ordinary differential equations but especially for partial differential equations. For this reason, a separate Chap. 27 is devoted to this method. Since, in the quoted chapter, the emphasis is rather on linear than on one-dimensional problems we describe here, on a simple example, the way in which the finite difference method can be used for solving a nonlinear boundary value problem.

Consider the differential equation

$$y'' = f(x, y, y') \tag{1}$$

in $[a, b]$ with simple boundary conditions

$$y(a) = \gamma_1, \quad y(b) = \gamma_2 . \tag{2}$$

If we suppose that the right-hand term of (1) satisfies the conditions

$$0 < p_0 \leqq f_y(x, y, z) \leqq P_0, \quad |f_z(x, y, z)| \leqq Q_0 \tag{3}$$

(the symbols f_y and f_z denote the partial derivatives with respect to y or z) in a sufficiently large domain of variables x, y and z, then the boundary value problem (1), (2) has a solution.

The basic idea of the finite difference method consists in choosing, in the interval $[a, b]$ in which the solution is sought, a finite set of points called the *mesh* (*grid*) – most often the mesh is formed by equidistant points

$$x_k = a + kh, \quad k = 0, \ldots, n, \tag{4}$$

where $h = (b-a)/n$ is called the *mesh-size* and n is a positive integer – and, further, in replacing the derivatives occurring in the differential equation and, if necessary, those in the boundary conditions by difference quotients at these points. By a difference quotient we mean here any linear combination of the values of the function at different points which approximates the considered derivative. Neglecting the errors in the approximation of the derivatives, we obtain a system of equations for the unknown values of the solution at the mesh-points.

If we proceed in the case of the boundary value problem (1), (2) in the indicated way, i.e., if we replace the first and the second derivatives of y at the point x_k by the quotients $[y(x_{k+1}) - y(x_{k-1})]/(2h)$ and $[y(x_{k+1}) - 2y(x_k) + y(x_{k-1})]/h^2$ (which approximate the corresponding derivatives with errors proportional to h^2), respectively, we obtain the system of equations

$$\begin{gathered} \frac{y_{k-1} - 2y_k + y_{k+1}}{h^2} = f\left(x_k, y_k, \frac{y_{k+1} - y_{k-1}}{2h}\right), \quad k = 1, \ldots, n-1, \\ y_0 = \gamma_1, \quad y_n = \gamma_2, \end{gathered} \tag{5}$$

where y_k denotes the approximate solution at the point x_k. The error of the approximation computed from the system (5) is proportional to h^2 similarly as the errors in the approximation of derivatives in the original differential equation.

If we compare this situation with that in the case of initial value problems, there is a substantial difference: Here, the system (5) which represents the finite dimensional replacement of the original problem, does not form the algorithm for solving the original problem yet. It is not so before we indicate the actual algorithm for solving that system. One possibility is to use Newton's method which was already recommended in the case of the shooting method. Taking into account the

special form of the system (5), we can also use the simple iteration scheme

$$(1+\omega)y_k^{(i+1)} = \frac{1}{2}\left(y_{k-1}^{(i)} + y_{k+1}^{(i)}\right) + \omega y_k^{(i)} - \frac{1}{2}h^2 f\left(x_k, y_k^{(i)}, \frac{y_{k+1}^{(i)} - y_{k-1}^{(i)}}{2h}\right), \quad k = 1, \ldots, n-1, \qquad y_0^{(i+1)} = \gamma_1, \quad y_n^{(i+1)} = \gamma_2 . \tag{6}$$

If the parameter ω is chosen in such a way that

$$\omega \geqq \tfrac{1}{2}h^2 P_0 , \tag{7}$$

the convergence is guaranteed.

The presented example sufficiently illustrates how to proceed in the case of other boundary value problems. In the linear case, the arising system will naturally be also linear and, moreover, its matrix will be a band matrix with the bandwidth independent of the mesh-size h. The solution of such system by the Gaussian elimination method is then not too time-consuming, namely, the number of operations is proportional to the number of equations (cf. § 30.5).

Example 1. Solve the differential equation

$$y'' - (1 + 2\tan^2 x)y = 0$$

with the boundary conditions

$$y(0) = 1, \quad y(1) = \frac{1}{\cos 1}$$

TABLE 25.13

x	$h = 1/20$		$h = 1/40$	
	approximate solution	error	approximate solution	error
0·10	1·005182	0·000161	1·005613	0·000040
0·20	1·020651	0·000312	1·020417	0·000079
0·30	1·047206	0·000454	1·046866	0·000114
0·40	1·086288	0·000583	1·085851	0·000147
0·50	1·140190	0·000696	1·139669	0·000175
0·60	1·212411	0·000783	1·211825	0·000197
0·70	1·308284	0·000824	1·307666	0·000207
0·80	1·436106	0·000782	1·435521	0·000197
0·90	1·609299	0·000573	1·608870	0·000144

by the method (5). One can find the results for the mesh-sizes $h = 1/20$ and $h = 1/40$ in Tab. 25.13. We see from it that using the latter mesh-size, the error is approximately a quarter of that in the case of the former mesh-size. Thus, the error is really proportional to h^2.

25.10. The Eigenvalue Problem

In this paragraph, we mention very briefly the problems connected with the computation of eigenvalues and eigenfunctions of differential operators. We restrict ourselves to a simple, nevertheless important, problem called the *Sturm-Liouville problem.* This problem consists in finding all such values of the parameter λ (each of them being called *eigenvalue*) for which the differential equation

$$-[p(x)y']' + q(x)y = \lambda r(x)y\,, \quad x \in [a, b]\,, \tag{1}$$

with homogeneous boundary conditions

$$\begin{aligned} -\alpha_1 p(a)y'(a) + \beta_1 y(a) &= 0\,, \\ \alpha_2 p(b)y'(b) + \beta_2 y(b) &= 0\,, \end{aligned} \tag{2}$$

has a nontrivial (i.e., not identically vanishing) solution. This solution is then called the *eigenfunction* corresponding to the eigenvalue λ.

Theorem 1. *Let the coefficients p, q, and f of the differential equation* (1) *be continuous, p continuously differentiable and let*

$$p(x) \geqq p_0 > 0, \quad q(x) \geqq 0, \quad r(x) \geqq r_0 > 0, \tag{3}$$

where p_0 and r_0 are constants. Further let all the coefficients α_i and β_i in the boundary conditions be non-negative and let, moreover, $\alpha_i + \beta_i > 0$ for $i = 1, 2$. Then there exists an infinite sequence $0 \leqq \lambda_1 \leqq \lambda_2 \leqq \ldots$ of non-negative eigenvalues of the Sturm-Liouville problem (1), (2). *The corresponding eigenfunctions u_i can be chosen in such a way that they are orthogonal in $[a, b]$ with the weight r, i.e., that*

$$(u_i, u_j)_r \equiv \int_a^b r(x)u_i(x)u_j(x)\,dx = 0 \tag{4}$$

for $i \neq j$. Further, any function $u \in H^1$ which satisfies the boundary conditions (2) *can be written in the form of a generalized Fourier series*

$$u(x) = \sum_{i=0}^{\infty} \frac{(u, u_i)_r}{(u_i, u_i)_r} u_i(x) \tag{5}$$

converging uniformly in $[a, b]$.

The symbol H^1 denotes here the Sobolev space $W_2^{(1)}$ from § 22.4, which can be in the one-dimensional case defined also as the set of absolute continuous functions the derivatives of which are square integrable in $[a, b]$.

REMARK 1. The assumptions of Theorem 1 concerning the smoothness of p, q, and r can be substantially weakened.

For the approximation of eigenvalues and eigenfunctions of the Sturm-Liouville problem (1), (2), we can use more or less obvious modifications of almost all methods which were described in §§ 25.7, 25.8, and 25.9.

Thus, for example, if we write (1) as the system of first-order equations

$$\begin{bmatrix} {}^1y \\ {}^2y \end{bmatrix}' = \begin{bmatrix} 0, & q - \lambda r \\ 1/p, & 0 \end{bmatrix} \begin{bmatrix} {}^1y \\ {}^2y \end{bmatrix} \tag{6}$$

with the boundary conditions

$$\begin{aligned} -\alpha_1 {}^1y(a) + \beta_1 {}^2y(a) = 0\,, \\ \alpha_2 {}^1y(b) + \beta_2 {}^2y(b) = 0 \end{aligned} \tag{7}$$

(we set $py' = {}^1y$ and $y = {}^2y$ here), it follows from Theorem 25.8.1 that

$$-p(x)z(x)y'(x) + p(x)z'(x)y(x) = 0 \tag{8}$$

for any $x \in [a, b]$, where z is the solution of the differential equation

$$-[p(x)z']' + [q(x) - \lambda r(x)]\, z = 0 \tag{9}$$

with the initial conditions

$$z(a) = \alpha_1, \quad z'(a) = \beta_1/p(a)\,. \tag{10}$$

Namely, the matrix $\boldsymbol{R}_a$ is in this case of the form $[R_a^{(1)}, R_a^{(2)}]$, where $R_a^{(1)}$ and $R_a^{(2)}$ are scalar functions, and $R_a^{(1)} = -z(x)$ and $R_a^{(2)} = p(x)z'(x)$, where z is the solution of (9) with the initial conditions (10), and the function r_a obviously vanishes.

If we substitute $x = b$ in (8), we obtain

$$-p(b)z(b)y'(b) + p(b)z'(b)y(b) = 0\,, \tag{11}$$

and this equation has to hold simultaneously with the second boundary condition in (2), i.e., with the equation

$$\alpha_2 p(b)y'(b) + \beta_2 y(b) = 0\,. \tag{12}$$

Thus, the value of λ is determined from the condition that the system of linear equations (11), (12) would have a nontrivial solution, i.e., from the condition

$$\Delta(\lambda) = 0\,, \tag{13}$$

where

$$\Delta(\lambda) = \alpha_2 p(b) z'(b) + \beta_2 z(b) \tag{14}$$

is the determinant of this system. Choosing the value of λ in the differential equation (1), we are able to compute the values of Δ at different points λ. Hence, for solving (13), it is possible to use, for example, the regula falsi method (the method of false position, cf. § 31.4).

The method of the normalized transfer of boundary conditions can be used in a completely similar way.

The finite difference method will be described, for the sake of simplicity, only for the special case that we have $\alpha_1 = \alpha_2 = 0$ in the boundary conditions. If we put $p(x)y'(x) = z(x)$ in the differential equation (1) at $x = x_k$, $k = 1, \ldots$ $\ldots, n-1$ ($x_k = a + kh$, $h = (b-a)/n$), and if we approximate the derivative $[p(x)y'(x)]' = z'(x)$ by the quotient

$$[z(x+h/2) - z(x-h/2)]/h = [p(x+h/2)y'(x+h/2) - p(x-h/2)y'(x-h/2)]/h$$

and the derivatives $y'(x+h/2)$ and $y'(x-h/2)$ by the quotients

$$[y(x+h) - y(x)]/h \quad \text{and} \quad [y(x) - y(x-h)]/h\,,$$

respectively, we obtain the system

$$\begin{gathered} -p(x_k - h/2)y_{k-1} + \left[p(x_k - h/2) + p(x_k + h/2) + h^2 q(x_k)\right] y_k - \\ -p(x_k + h/2)y_{k+1} = \lambda h^2 r(x_k) y_k\,, \quad k = 1, \ldots, n-1\,, \\ y_0 = y_n = 0 \end{gathered} \tag{15}$$

which can be written in the matrix form

$$\boldsymbol{A}\boldsymbol{y} = \lambda \boldsymbol{B}\boldsymbol{y}\,. \tag{16}$$

Thus, to find an approximation of an eigenvalue of the Sturm-Liouville problem (1), (2) means to find such λ that the system of linear algebraic equations (16) has a nontrivial solution. This problem is called the *generalized eigenvalue problem* and it is, in the case that both matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are singular, rather difficult (cf. § 30.16). In our special situation, it can be simply transformed into an ordinary matrix eigenvalue problem by premultiplying (16) by $\boldsymbol{B}^{-1}$. This is

possible since the matrix $\boldsymbol{B}$ is obviously nonsingular (it is, in fact, diagonal with positive entries on the main diagonal).

The approximate solution of an eigenvalue problem can be based also on variational principles which utilize some minimal properties of eigenvalues. We now describe some of them.

Let the Sturm-Liouville problem with the boundary conditions

$$y(a) = y(b) = 0 \tag{17}$$

be given. If we introduce a functional R defined by

$$R(u) = \frac{\int_a^b \{p(x)[u'(x)]^2 + q(x)u^2(x)\}\,\mathrm{d}x}{\int_a^b r(x)u^2(x)\mathrm{d}x} \equiv \frac{[u, u]}{(u, u)_r} \tag{18}$$

and called the *Rayleigh quotient* we are able to formulate the following theorem:

Theorem 2. *Let p, q, and r satisfy* (3). *Then*

$$\lambda_1 = \min_{\substack{u \in H_0^1 \\ u \neq 0}} R(u) \tag{19}$$

and

$$\lambda_j = \min_{\substack{u \in H_0^1,\, u \neq 0 \\ (u, u_i)_r = 0,\, i = 1, \ldots, j-1}} R(u)\,. \tag{20}$$

The symbol H_0^1 denotes the subspace of those functions of the Sobolev space H^1 which satisfy the boundary conditions (17).

Theorem 2 enables us to estimate the least eigenvalue of the Sturm-Liouville problem (1), (2) from above in such a way that the minimum in (19) is taken over a suitable finite-dimensional subspace of H_0^1.

The computation of other eigenvalues according to Theorem 2 is not too convenient since it is necessary to minimize the Rayleigh quotient on the class of such functions which are orthogonal to all the eigenfunctions which correspond to the preceding eigenvalues. For this reason, another variational principle known as the *Courant minimax principle* is often more useful. This principle will be formulated in the following theorem:

Theorem 3. *Let $w_1, \ldots, w_{j-1}$ be any linearly independent functions from L_2. Further, let ($w_1, \ldots, w_{j-1}$ being fixed) $V(w_1, \ldots, w_{j-1})$ be the space of such functions $u \in H_0^1$ for which we have $(u, w_i)_r = 0$ for $i = 1, \ldots, j-1$. Finally, let*

$$M(w_1, \ldots, w_{j-1}) = \min_{u \in V(w_1, \ldots, w_{j-1})} R(u)\,. \tag{21}$$

Then

$$\lambda_j = \max_{w_1, \ldots, w_{j-1}} M(w_1, \ldots, w_{j-1}) . \tag{22}$$

On the basis of this theorem, the procedure of approximating eigenvalues of the given Sturm-Liouville problem can be as follows:

(i) We choose a finite dimensional subspace D_N of H_0^1. Let $\Phi_1, \ldots, \Phi_N$ be the basis in D_N and put

$$R^*(\alpha_1, \ldots, \alpha_N) = R\left(\sum_{i=1}^{N} \alpha_i \Phi_i\right) ; \tag{23}$$

(ii) we find the extremal values of the function R^* (of N real variables $\alpha_1, \ldots, \alpha_N$).

This last problem leads to the eigenvalue problem (16), where $\boldsymbol{A}$ and $\boldsymbol{B}$ are $N \times N$ matrices with entries a_{ij} and b_{ij} given by the formulae $a_{ij} = [\Phi_i, \Phi_j]$ and $b_{ij} = (\Phi_i, \Phi_j)_r$, respectively. The matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric and positive definite so that the eigenvalue problem can be solved without any difficulties by means of the methods of Chap. 30. As a result, we obtain N real eigenvalues which approximate the first N eigenvalues of the Sturm-Liouville problem (1), (17). If the eigenvectors corresponding to the eigenvalue λ_i, $i = 1, \ldots, N$, are denoted by $\boldsymbol{\alpha}^{(i)} = (\alpha_1^{(i)}, \ldots, \alpha_N^{(i)})^{\mathrm{T}}$, then the functions

$$u_i(x) = \sum_{k=1}^{N} \alpha_k^{(i)} \Phi_k(x) \tag{24}$$

are the approximations of the corresponding eigenfunctions. Usually, a finite element space (cf. Chap. 24) is chosen for the space D_N. Then the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are sparse matrices and this fact positively influences the efficiency of the algorithm.

REMARK 2. Theorems 2 and 3 hold even in the case of a substantially more general problem

$$My = \lambda N y ,$$

where M is a differential operator of order $2m$ defined by

$$My = \sum_{k=0}^{m} (-1)^k \left[p_k(x) y^{(k)}\right]^{(k)}$$

and N a $2n$-th order differential operator given by

$$Ny = \sum_{k=0}^{n} (-1)^k \left[n_k(x) y^{(k)}\right]^{(k)} ,$$

with the nonseparated boundary conditions of the form

$$\sum_{k=0}^{m-1} \left[\alpha_{jk} y^{(k)}(a) + \beta_{jk} y^{(k)}(b)\right] = 0\,, \quad j = 1, \ldots, 2m\,.$$

Naturally, the definition of the Rayleigh quotient has to be modified properly. See § 17.17, where also a simple method can be found giving estimates for the first eigenvalue λ_1 *from below.*

26. SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS BY INFINITE SERIES (BY THE FOURIER METHOD)

By KAREL REKTORYS

References: [79], [99], [125], [223], [277], [290], [339], [369], [385], [436], [465], [486], [494].

In this chapter we show some typical examples of the so-called *Fourier method* for solving boundary-value problems, otherwise known as the *method of separation of variables* or the *product method.* This method consists — roughly speaking — in assuming the solution of the given boundary-value problem in the form

$$u = \sum_{n=1}^{\infty} a_n u_n \,, \tag{1}$$

where a_n are constants to be determined and u_n are functions satisfying the given differential equation and some of the given boundary conditions. Each of the functions u_n is assumed to be in the form of a product of functions of one variable only. The method will be thoroughly explained in § 26.1; in the remaining paragraphs of this chapter we shall proceed more briefly.

All functions and constants considered in this chapter are assumed to be real.

26.1. Equation of a Vibrating String

A function $u(x, t)$ is to be found, satisfying the differential equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \qquad (0 < x < l,\, t > 0)\,, \tag{2}$$

continuous in the closed domain $\overline{\Omega}\,(0 \leqq x \leqq l,\, t \geqq 0)$ and satisfying the following initial and boundary conditions:

$$u(x, 0) = \varphi_0(x)\,, \tag{3}$$

$$\frac{\partial u}{\partial t}(x, 0) = 0\,, \tag{4}$$

$$u(0, t) = 0\,, \tag{5}$$

$$u(l, t) = 0\,. \tag{6}$$

The continuity required implies that

$$\varphi_0(0) = 0\,, \quad \varphi_0(l) = 0\,. \tag{7}$$

The physical meaning of the problem is the following: A string of length l with its ends kept fixed (conditions (5), (6)) is put into the (initial) position with the amplitude $\varphi_0(x)$ and, after being released at the time $t = 0$, begins to vibrate. (Thus, at this time $t = 0$, its points have zero velocity; condition (4).) The amplitude $u(x, t)$ at each point x $(0 \leqq x \leqq l)$ and for all t $(t \geqq 0)$ is to be determined. (Other problems also lead to the solution of equation (2), for example longitudinal vibration of bars, twisting of bars etc.)

Let us assume the solution of the problem in the form

$$u(x, t) = \sum_{n=1}^{\infty} a_n u_n(x, t)\,, \tag{8}$$

where each of the functions $u_n(x, t)$ has the following properties:

a) it is of the form $u_n(x, t) = X_n(x)\,T_n(t)$, (9)

b) satisfies equation (2),

c) satisfies conditions (4), (5), (6), (10)

d) is not identically equal to zero.

Thus, each partial sum of the series (8) will satisfy equation (2) and conditions (4), (5), (6). By a suitable choice of the coefficients a_n, we shall satisfy the condition (3).

Since $u_n(x, t)$ satisfies equation (2), on putting (9) into (2) and dividing by the product $X_n T_n$ (supposing $X_n T_n \neq 0$) we find that

$$\frac{X_n''(x)}{X_n(x)} = \frac{T_n''(t)}{T_n(t)}\,. \tag{11}$$

This equation is to be satisfied for all x, t of the region Ω $(0 < x < l,\ t > 0)$, since (2) is to be satisfied everywhere in Ω. The left-hand side of (11) is independent of t, being a function of x only, but it is also independent of x, since it is equal in Ω to the right-hand side of (11), and this right-hand side is independent of x. Thus, the left-hand and the right-hand sides of (11) are both equal to a (common) constant. Let us denote this constant by $-\lambda_n$. Then, equation (11) yields

$$X_n'' + \lambda_n X_n = 0\,, \tag{12}$$

$$T_n'' + \lambda_n T_n = 0\,. \tag{13}$$

A more detailed analysis shows that equations (12), (13) must be satisfied even at (isolated) zeros of the function $X_n(x)$, or $T_n(t)$, respectively, so that the assumption $X_n\,T_n \neq 0$ is not essential.

To equation (12), boundary conditions

$$X_n(0) = 0\,, \quad X_n(l) = 0 \tag{14}$$

should be assigned. In fact, the function $u_n(x, t)$ is to be determined so as to fulfill conditions (5), (6) and to be of the form $u_n(x, t) = X_n(x)T_n(t)$. Since $T_n(t)$ must not be identically zero (condition d)), the conditions

$$u_n(0, t) = 0\,, \quad u_n(l, t) = 0$$

yield the conditions (14).

It follows from the same condition d) that we have to find such a solution of the problem (12), (14) which is not identically zero. It can be easily verified that $\lambda_n < 0$ or $\lambda_n = 0$ in (12) yields only the solution $X_n(x) \equiv 0$. If $\lambda_n > 0$, then the general integral of (12) is

$$X_n = C_n \cos \sqrt{(\lambda_n)}x + D_n \sin \sqrt{(\lambda_n)}x\,. \tag{15}$$

The first of conditions (14) yields $C_n = 0$, so that

$$X_n = D_n \sin \sqrt{(\lambda_n)}x\,;$$

from the second of them it then follows (since we must have $D_n \neq 0$) that

$$\lambda_n = \frac{n^2\pi^2}{l^2}\,, \quad n \text{ being an integer.} \tag{16}$$

It is sufficient to consider

$$n = 1, 2, 3, \ldots$$

only, since for $n = 0$ we get the zero solution and for $n = -1, -2, -3, \ldots$ we do not obtain anything new. If we choose $D_n = 1$, we get

$$X_n(x) = \sin \frac{n\pi x}{l}\,. \tag{17}$$

Similarly, using (16), we find that non-zero solutions of equation (13), satisfying condition (4), are

$$T_n(t) = \cos \frac{n\pi t}{l}\,.$$

Thus each of the functions

$$u_n(x, t) = \sin \frac{n\pi x}{l} \cos \frac{n\pi t}{l} \qquad (n = 1, 2, 3, \ldots) \tag{18}$$

(multiplied, eventually, by an arbitrary constant) is a solution of (2), (4), (5) and (6).

Let $u(x, t)$ be of the form (8). Since, for $t = 0$, $\cos(n\pi t/l) = 1$, we have

$$u_n(x, 0) = \sin \frac{n\pi x}{l},$$

and the condition (3) becomes

$$\sum_{n=1}^{\infty} a_n \sin \frac{n\pi x}{l} = \varphi_0(x). \tag{19}$$

The series on the left-hand side of (19) is the Fourier sine series of the function $\varphi_0(x)$, so that (see § 16.3)

$$a_n = \frac{2}{l} \int_0^l \varphi_0(x) \sin \frac{n\pi x}{l} \, dx \qquad (n = 1, 2, 3, \dots) \tag{20}$$

(thus the a_n's are the Fourier coefficients of the function $\varphi_0(x)$ with respect to the system of the functions $\sin(n\pi x/l)$).

Thus, if the solution of the given problem can be expressed in the form (8), then u_n and a_n are given by (18) and (20), respectively.

If, instead of (4), the condition

$$\frac{\partial u}{\partial t}(x, 0) = \varphi_1(x) \tag{21}$$

is prescribed, then

$$u(x, t) = \sum_{n=1}^{\infty} \sin \frac{n\pi x}{l} \left(a_n \cos \frac{n\pi t}{l} + b_n \sin \frac{n\pi t}{l} \right), \tag{22}$$

where a_n are given by (20) and b_n by

$$b_n = \frac{2}{n\pi} \int_0^l \varphi_1(x) \sin \frac{n\pi x}{l} \, dx. \tag{23}$$

(If $\varphi_1 \equiv 0$, we get, of course, the previous result.)

As mentioned above, if (8) is the solution of the boundary-value problem (2)–(6), then the u_n are given by (18), and the a_n by (20). In order that the series (8) (with u_n and a_n determined in this way) may be in fact the solution of the problem, it is sufficient that the function $\varphi_0(x)$ have two continuous derivatives in $[0, l]$ and that (7) and $\varphi_0''(0) = \varphi_0''(l) = 0$ be fulfilled. If the condition (4) is replaced by (21),

then, in addition, the function $\varphi_1(x)$ is supposed to have a continuous derivative in $[0, l]$ and $\varphi_1(0) = \varphi_1(l) = 0$.

In applications, the functions $\varphi_0(x)$ and $\varphi_1(x)$ do not always have all the required properties. For example φ_0 and φ_1 are functions such that φ_0, φ_1, φ_0', φ_1' are continuous in $[0, l]$ and equal to zero for $x = 0$ and $x = l$. In this case the series (22) represents the *generalized solution* of the problem (Definition 18.5.3). That is to say, this series is uniformly convergent in the whole domain $\overline{\Omega}$ $(0 \leqq x \leqq l, t \geqq 0)$ and each of its partial sums satisfies equation (2) in Ω. But this last assertion may not be true for the sum of this series.

However, from the engineering point of view, this fact is not of particular importance. Indeed, in practice, we take only a finite number of terms in (8), say

$$v_k(x, t) = \sum_{n=1}^{k} a_n u_n(x, t) .$$

This function satisfies all the conditions (2) – (6) except condition (3) which is satisfied only approximately. In this way, we obtain, in fact, exact solution of a rather different problem, with the initial function

$$v_k(x, 0) = \sum_{n=1}^{k} a_n \sin \frac{n\pi x}{l}$$

substituted for the given function

$$\varphi_0(x) = \sum_{n=1}^{\infty} a_n \sin \frac{n\pi x}{l} .$$

If k is sufficiently large, the function $v_k(x, 0)$ differs only "slightly" from the function $\varphi_0(x)$ and, from the practical point of view, the function $v_k(x, t)$ is usually an acceptable approximation of the solution $u(x, t)$ of the given problem.

The same remark holds for all problems treated in this chapter.

If instead of equation (2) we are considering the equation

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} \qquad (a > 0)$$

then (using the substitution $t = \tau/a$) we obtain the solution

$$u(x, t) = \sum_{n=1}^{\infty} \sin \frac{n\pi x}{l} \left(a_n \cos \frac{n\pi a t}{l} + b_n \sin \frac{n\pi a t}{l} \right) ,$$

where the a_n are given by (20), and the b_n by

$$b_n = \frac{2}{n\pi a} \int_0^l \varphi_1(x) \sin \frac{n\pi x}{l} \, dx .$$

26.2. Potential Equation and Stationary Heat-Conduction Equation

Let us find the solution of the equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \qquad (0 < x < a, \, 0 < y < b) \tag{1}$$

continuous in the rectangle $\overline{\Omega}$ $(0 \leqq x \leqq a, \, 0 \leqq y \leqq b)$ and satisfying the boundary conditions

$$u(x, 0) = f(x) \qquad (\text{with } f(0) = 0, \, f(a) = 0), \tag{2}$$

$$u(x, b) = 0, \tag{3}$$

$$u(0, y) = 0, \tag{4}$$

$$u(a, y) = 0. \tag{5}$$

(This is the problem of finding the stationary temperature field in an infinite right prism whose cross-section is a rectangle and whose three faces are kept at zero temperature and the fourth one at the temperature $f(x)$; or a similar problem for the potential.)

As in § 26.1, the solution is supposed to be of the form

$$u(x, y) = \sum_{n=1}^{\infty} a_n u_n(x, y),$$

where each of the functions $u_n(x, y)$ is of the form $X_n(x) Y_n(y)$, satisfies equation (1) and the boundary conditions (3), (4), (5), and is not identically zero. Putting $X_n Y_n$ for u_n into (1), we get

$$\frac{X_n''}{X_n} + \frac{Y_n''}{Y_n} = 0$$

so that

$$\frac{X_n''(x)}{X_n(x)} = -k_n, \quad \frac{Y_n''(y)}{Y_n(y)} = k_n, \quad k_n > 0. \tag{6}$$

The general solution of the first of equations (6) is

$$X_n(x) = C_1 \sin \sqrt{(k_n)}\, x + C_2 \cos \sqrt{(k_n)}\, x ;$$

conditions (4), (5) yield $X_n(0) = 0$, $X_n(a) = 0$; thus $C_2 = 0$ and

$$k_n = \frac{n^2\pi^2}{a^2} \qquad (n = 1, 2, 3, \dots). \tag{7}$$

The general integral of the second of equations (6) is (using (7))

$$Y_n(y) = D_1 \sinh \frac{n\pi y}{a} + D_2 \cosh \frac{n\pi y}{a}. \tag{8}$$

It is convenient to choose D_1 and D_2 in (8) in such a way that for $y = 0$ we shall have $Y_n = 1$. Further, for $y = b$ it follows from (3) that $Y_n = 0$. These two conditions are satisfied by the function

$$Y_n(y) = \frac{\sinh \dfrac{n\pi(b-y)}{a}}{\sinh \dfrac{n\pi b}{a}}.$$

If, in addition, we put

$$a_n = \frac{2}{a}\int_0^a f(x) \sin \frac{n\pi x}{a}\,\mathrm{d}x,$$

then if, for example, $f(x)$ and $f'(x)$ are continuous functions in $[0, a]$, the required solution of the problem (1) – (5) is

$$u(x, y) = \sum_{n=1}^{\infty} a_n \sin \frac{n\pi x}{a} \frac{\sinh \dfrac{n\pi(b-y)}{a}}{\sinh \dfrac{n\pi b}{a}}. \tag{9}$$

If the boundary conditions (3), (4), (5) are not homogeneous (i.e. if temperature is prescribed also on the remaining three faces of the prism), we obtain the solution by superposition of solutions of four problems of this type.

26.3. Heat Conduction in Rectangular Regions

If we solve the equation

$$\frac{\partial u}{\partial t} = a\frac{\partial^2 u}{\partial x^2} \qquad (0 < x < l,\ t > 0)$$

with the boundary conditions

$$u(0, t) = 0, \quad u(l, t) = 0, \quad u(x, 0) = f(x)$$

by the Fourier method, we get

$$u(x, t) = \sum_{n=1}^{\infty} a_n \sin \frac{n\pi x}{l} \, e^{-a(n^2\pi^2/l^2)t}, \tag{1}$$

where

$$a_n = \frac{2}{l} \int_0^l f(x) \sin \frac{n\pi x}{l} \, dx.$$

Similarly, the solution of the equation

$$\frac{\partial u}{\partial t} = a \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) \qquad (0 < x < l_1, \, 0 < y < l_2, \, t > 0)$$

with the boundary conditions

$$u(x, y, 0) = f(x, y), \quad u(x, 0, t) = 0, \quad u(x, l_2, t) = 0,$$
$$u(0, y, t) = 0, \quad u(l_1, y, t) = 0,$$

is

$$u(x, y, t) = \sum_{m,n=1}^{\infty} A_{mn} \sin \frac{m\pi x}{l_1} \sin \frac{n\pi y}{l_2} \, e^{-a[(\pi^2/l_1^2)m^2 + (\pi^2/l_2^2)n^2]t}, \tag{2}$$

where

$$A_{mn} = \frac{4}{l_1 l_2} \int_0^{l_1} \int_0^{l_2} f(x, y) \sin \frac{m\pi x}{l_1} \sin \frac{n\pi y}{l_2} \, dx \, dy.$$

(The convergence of the series (2) is understood in the sense of Remark 16.3.10: We say that the series (2) has the sum $u(x_0, y_0, t_0)$ at the point (x_0, y_0, t_0), if corresponding to every $\varepsilon > 0$, there exists an n_0 such that for every pair of numbers M, N for which simultaneously $M > n_0$, $N > n_0$, the relation

$$\left| \sum_{m=1}^{M} \sum_{n=1}^{N} A_{mn} \sin \frac{m\pi x_0}{l_1} \sin \frac{n\pi y_0}{l_2} \, e^{-a[(\pi^2/l_1^2)m^2 + (\pi^2/l_2^2)n^2]t_0} - u(x_0, y_0, t_0) \right| < \varepsilon$$

holds.)

If $f(x)$ has a continuous derivative in $[0, l]$ and $f(0) = f(l) = 0$, or if $f(x, y)$ has continuous first partial derivatives in the rectangle $\overline{\Omega}$ $(0 \leqq x \leqq l_1, \, 0 \leqq y \leqq l_2)$ and $f(x, y) = 0$ on the sides of this rectangle, then the series (1), or (2), respectively, give in fact the solution of the problem in question.

Similar results hold for the three-dimensional case.

26.4. Heat Conduction in an Infinite Circular Cylinder. Application of Bessel Functions

Let us solve the problem

$$\frac{\partial u}{\partial t} = k\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r}\right) \qquad (0 \leqq r < c,\ t > 0), \tag{1}$$

$$u(c, t) = 0, \tag{2}$$

$$u(r, 0) = f(r). \tag{3}$$

(This is the problem of heat conduction in an infinite circular cylinder of radius c, the surface of which is kept at zero temperature and the initial temperature of which is independent of φ.)

The solution $u(r, t)$ is also supposed to be axially symmetric (i.e. independent of φ) and to be expressible as an infinite series, the terms of which are of the form

$$R(r)\,T(t) \tag{4}$$

and satisfy equation (1) and the condition (2). Putting (4) into (1), we get

$$\frac{T'}{kT} = \frac{1}{R}\left(R'' + \frac{R'}{r}\right). \tag{5}$$

As in § 26.1, we see that the left-hand and the right-hand sides of (5) are both equal to a (common) negative constant, say $-\lambda^2$. Thus

$$rR'' + R' + \lambda^2 rR = 0, \tag{6}$$

$$T' + k\lambda^2 T = 0. \tag{7}$$

Using the substitution

$$\lambda r = z \quad \text{or} \quad \frac{\mathrm{d}R}{\mathrm{d}r} = \lambda\frac{\mathrm{d}R}{\mathrm{d}z}, \quad \frac{\mathrm{d}^2R}{\mathrm{d}r^2} = \lambda^2\frac{\mathrm{d}^2R}{\mathrm{d}z^2},$$

we get from (6)

$$z\frac{\mathrm{d}^2R}{\mathrm{d}z^2} + \frac{\mathrm{d}R}{\mathrm{d}z} + zR = 0, \tag{8}$$

which is the Bessel equation of order $n = 0$ (§ 17.15); its solution is the function

$$R(z) = \mathrm{J}_0(z)$$

(§ 16.4), so that the solution of (6) is (for a fixed λ)

$$R(r) = \mathrm{J}_0(\lambda r)\,. \tag{9}$$

Since the functions of the form (4) must satisfy the boundary condition (2) (for all t), it is necessary for λ to be such that

$$\mathrm{J}_0(\lambda c) = 0\,. \tag{10}$$

According to Theorem 16.4.8 and Remark 16.4.8, equation (10) has real roots only and of these only the positive roots λ_n ($n = 1, 2, \dots$) need be considered. For a fixed λ_n, the function (9) becomes

$$R_n(r) = \mathrm{J}_0(\lambda_n r)$$

while

$$T_n(t) = \mathrm{e}^{-k\lambda_n^2 t}$$

according to (7).

Let us introduce the constants a_n by

$$a_n = \frac{2}{c^2 \mathrm{J}_1^2(\lambda_n c)} \int_0^c r \mathrm{J}_0(\lambda_n r) f(r)\,\mathrm{d}r$$

(see Theorem 16.4.9), where J_1 is the Bessel function of index 1. Then if, for example, $f(r)$ is continuous and has a piecewise continuous derivative in $[0, c]$, and $f(c) = 0$, the solution of the problem (1), (2), (3) is

$$u(r, t) = \sum_{n=1}^{\infty} a_n \mathrm{J}_0(\lambda_n r)\,\mathrm{e}^{-k\lambda_n^2 t}\,.$$

26.5. Deflection of a Rectangular Simply Supported Plate

Let us solve the equation

$$\Delta^2 w \equiv \frac{\partial^4 w}{\partial x^4} + 2\frac{\partial^4 w}{\partial x^2 \partial y^2} + \frac{\partial^4 w}{\partial y^4} = \frac{q_0}{D} \sin\frac{\pi x}{a} \sin\frac{\pi y}{b} \tag{1}$$

in the rectangle $\Omega\,(0 < x < a,\ 0 < y < b)$ with the boundary conditions

$$w = 0\,, \quad \frac{\partial^2 w}{\partial x^2} = 0 \quad \text{for } x = 0\,,\ x = a\,, \tag{2}$$

$$w = 0\,, \quad \frac{\partial^2 w}{\partial y^2} = 0 \quad \text{for } y = 0\,,\ y = b\,. \tag{3}$$

(This is the problem of deflection of a rectangular simply supported plate with loading

$$q = q_0 \sin \frac{\pi x}{a} \sin \frac{\pi y}{b} .)$$

Let us assume the solution in the form

$$w = C \sin \frac{\pi x}{a} \sin \frac{\pi y}{b} . \tag{4}$$

The boundary conditions (2) and (3) are then satisfied. If we put (4) into (1), we get

$$\pi^4 \left(\frac{1}{a^2} + \frac{1}{b^2} \right)^2 C = \frac{q_0}{D} ;$$

hence

$$w = \frac{q_0}{\pi^4 D \left(\frac{1}{a^2} + \frac{1}{b^2} \right)^2} \sin \frac{\pi x}{a} \sin \frac{\pi y}{b} . \tag{5}$$

If the loading is

$$q = q_0 \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} \tag{6}$$

(m, n being positive integers), we obtain similarly

$$w = \frac{q_0}{\pi^4 D \left(\frac{m^2}{a^2} + \frac{n^2}{b^2} \right)^2} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} . \tag{7}$$

Now, let

$$q = f(x, y) = \sum_{m,n=1}^{\infty} a_{mn} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} ,$$

where

$$a_{mn} = \frac{4}{ab} \int_0^a \int_0^b f(x, y) \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} \, dx \, dy$$

(see Theorem 16.3.5). Then

$$w = \frac{1}{\pi^4 D} \sum_{m,n=1}^{\infty} \frac{a_{mn}}{\left(\frac{m^2}{a^2} + \frac{n^2}{b^2} \right)^2} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} .$$

In particular, if $q = q_0 = \text{const.}$, we get

$$w = \frac{16 q_0}{\pi^6 D} \sum_{m,n=1,3,5,\ldots} \frac{\sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}}{mn \left(\frac{m^2}{a^2} + \frac{n^2}{b^2} \right)^2} .$$

REMARK 1. In this section, a non-homogeneous equation with homogeneous (zero) boundary conditions has been solved. Problems with nonhomogeneous boundary conditions are often encountered in applications. In many such cases it is preferable to look for the solution in the form $w = w_1 + w_2$, where w_1 is a function satisfying the given boundary conditions (but not, in general, the equation in question), while the function w_2 satisfies homogeneous boundary conditions and the equation with a non-zero right-hand side. In this simple way, the problem may often be modified to a form suitable for application of the Fourier method.

27. SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS BY THE FINITE-DIFFERENCE METHOD

By EMIL VITÁSEK

References: [13], [21], [39], [89], [151], [152], [155], [180], [204], [252], [314], [321], [333], [390], [392], [460], [461], [474], [475], [480].

The *finite-difference method* is a very popular method for numerical solution of partial differential equations of all types, in essence, which occur in applications of mathematics. The finite-element method, which rapidly develops at present, diminished a little its significance. Nevertheless, many technical problems formulated by means of partial differential equations are up to now solved by this method.

27.1. Basic Idea of the Finite-Difference Method

The basic idea of the finite-difference method is very simple: The domain in which the solution of the given differential equation is sought is subdivided by a *net* (*grid*) with a finite number of *mesh points* and the derivatives at each mesh point are replaced by *finite-difference approximations*. By such an approximation, we mean here a linear combination of the values of the considered function at the given point and at some neighbouring points. In fact, the finite-difference approximation arises in such a way that the function is replaced, in the neighbourhood of the given point, by an interpolation polynomial and the derivatives are computed from this polynomial. If, for example, a polynomial of the second degree is constructed in such a manner that it coincides with the given function u at the points 0, 1 and 2 (see Fig. 27.1), then on the assumption that the function u has continuous

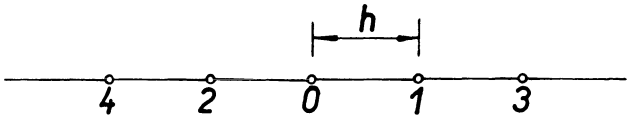


Fig. 27.1.

derivatives up to the fourth order at least, we have

$$\frac{\partial^2 u}{\partial x^2} = \frac{u_2 - 2u_0 + u_1}{h^2} - R_0$$

where

$$R_0 = \frac{2h^2}{4!} \frac{\partial^4 u(\theta h)}{\partial x^4}, \quad -1 < \theta < 1,$$

and the finite-difference approximation to the second derivative of the function u at the point 0 is $(u_2 - 2u_0 + u_1)/h^2$. If the given function is interpolated from five points, one obtains (under similar assumptions concerning the smoothness of the function u)

$$\frac{\partial^2 u}{\partial x^2} = \frac{-u_4 + 16u_2 - 30u_0 + 16u_1 - u_3}{12h^4} + R_1$$

where

$$R_1 = \frac{h^4}{90} \frac{\partial^6 u(\theta h)}{\partial x^6}, \quad -1 < \theta < 1,$$

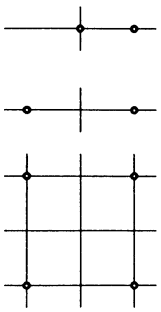
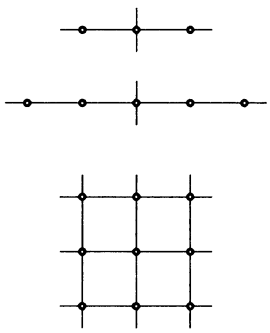
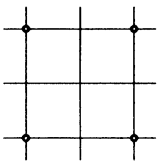
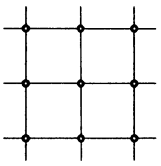
i.e., a formula whose order of accuracy is higher. In both the introduced cases, the derivative is expressed in terms of the corresponding finite-difference approximation and a *remainder* which is neglected in further investigations. Similar situation occurs in the case of other derivatives. The finite-difference formulae for the most frequent partial derivatives of a function of two variables are presented in Tab. 27.1, where $u(ih, kh) = u_{ik}$ and the symbol $O(h^p)$ denotes that the error is of order h^p. This means that the error is smaller in magnitude than Mh^p, where M is a constant, for sufficiently small h. Any formula of Tab. 27.1 is valid if the function u satisfies some smoothness assumptions. Thus, for example, for the first formula of this table, the boundedness of the second derivative is sufficient, for the second and third formula, the boundedness of the third derivative, etc.

If the derivatives are replaced by finite-difference expressions as indicated, one obtains a system of n (in general non-linear) equations for determining the approximate values of the unknown function at n different points of the net. This system of equations is then solved by appropriate numerical methods.

From this description of the basic idea of the finite-difference method one sees that this method can be applied to the solution of very different types of differential equations. Note already at this place that when solving some special types of differential equations (a typical example is a partial differential equation of parabolic type) it is often necessary to use special kinds of nets in which the time mesh-size depends on the space mesh-size (see, e.g., Tab. 27.3 and Remark 27.7.1).

In practice, the finite-difference method is now used mainly for linear equations, since in this case the corresponding system of finite-difference equations is also linear, and for systems of linear algebraic equations (which are, moreover, of a special

TABLE 27.1

Derivative	Scheme	Approximate formula
$\frac{\partial u}{\partial x}$		$\frac{\partial u_{ik}}{\partial x} = \frac{u_{i+1,k} - u_{ik}}{h} + O(h)$ $\frac{\partial u_{ik}}{\partial x} = \frac{u_{i+1,k} - u_{i-1,k}}{2h} + O(h^2)$ $\frac{\partial u_{ik}}{\partial x} =$ $= \frac{u_{i+1,k+1} - u_{i-1,k+1} + u_{i+1,k-1} - u_{i-1,k-1}}{4h} +$ $+O(h^2)$
$\frac{\partial^2 u}{\partial x^2}$		$\frac{\partial^2 u_{ik}}{\partial x^2} = \frac{u_{i+1,k} - 2u_{ik} + u_{i-1,k}}{h^2} + O(h^2)$ $\frac{\partial^2 u_{ik}}{\partial x^2} = \frac{1}{12h^2}(-u_{i+2,k} + 16u_{i+1,k} - 30u_{ik} +$ $+ 16u_{i-1,k} - u_{i-2,k}) + O(h^4)$ $\frac{\partial^2 u_{ik}}{\partial x^2} = \frac{1}{3h^2}(u_{i+1,k+1} - 2u_{i,k+1} + u_{i-1,k+1} +$ $+ u_{i+1,k} - 2u_{ik} + u_{i-1,k} + u_{i+1,k-1} -$ $- 2u_{i,k-1} + u_{i-1,k-1}) + O(h^2)$
$\frac{\partial^2 u}{\partial x \partial y}$		$\frac{\partial^2 u_{ik}}{\partial x \partial y} = \frac{1}{4h^2}(u_{i+1,k+1} - u_{i+1,k-1} - u_{i-1,k+1} +$ $+ u_{i-1,k-1}) + O(h^2)$
$\frac{\partial^4 u}{\partial x^4}$		$\frac{\partial^4 u_{ik}}{\partial x^4} = \frac{1}{h^4}(u_{i+2,k} - 4u_{i+1,k} + 6u_{ik} -$ $- 4u_{i-1,k} + u_{i-2,k}) + O(h^2)$
$\frac{\partial^4 u}{\partial x^2 \partial y^2}$		$\frac{\partial^4 u_{ik}}{\partial x^2 \partial y^2} = \frac{1}{h^4}(u_{i+1,k+1} + u_{i-1,k+1} + u_{i+1,k-1} +$ $+ u_{i-1,k-1} - 2u_{i+1,k} - 2u_{i-1,k} -$ $- 2u_{i,k+1} - 2u_{i,k-1} + 4u_{ik}) + O(h^2)$

form, namely, their matrices are sparse) many efficient methods of numerical solution have been established (see Chap. 30). In the non-linear case the finite-difference method is also often used but theoretical difficulties then occur connected with the

questions of convergence of the approximate solution to the exact solution and practical difficulties in solving systems of non-linear equations as well (cf. § 31.5).

27.2. Principal Types of Nets

The finite-difference method is mainly applied to two-dimensional cases. We therefore introduce the most frequently used types of plane nets.

(a) Rectangular Nets. Nets of this type are now most widely used. They can be divided into

(i) *irregular nets*,

(ii) *regular rectangular nets*,

(iii) *square nets*.

(i) *Irregular nets* are formed by different rectangles. They are used in order to simplify the formulation of boundary conditions (with such nets we ensure that

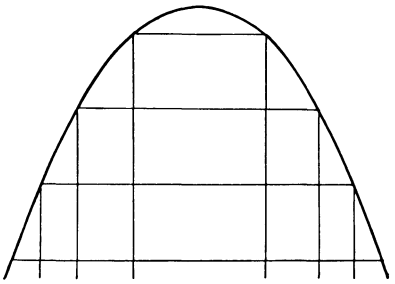


Fig. 27.2.

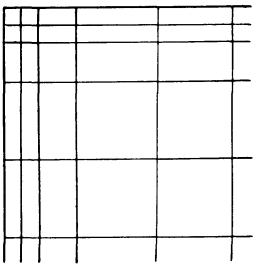


Fig. 27.3.

the mesh points lie on the boundary of the given domain (see Fig. 27.2 and § 27.5)) and to refine nets (see Fig. 27.3 and § 27.3).

(ii) *Regular rectangular nets* are formed by congruent rectangles. They play an important role when solving partial differential equations of parabolic and hyperbolic type. They can be also advantageous in some special situations. If we solve, for example, the differential equation

$$k_1 \frac{\partial^2 u}{\partial x^2} + k_2 \frac{\partial^2 u}{\partial y^2} = 0,$$

where k_1 and k_2 are different constants we can use such rectangular net which leads to the simplest computation scheme (with equal coefficients).

(iii) *Square nets* are formed by equal squares. Nets of this type are the most frequently used ones especially in the case of elliptic problems. The main reason is that the corresponding difference formulae are simple.

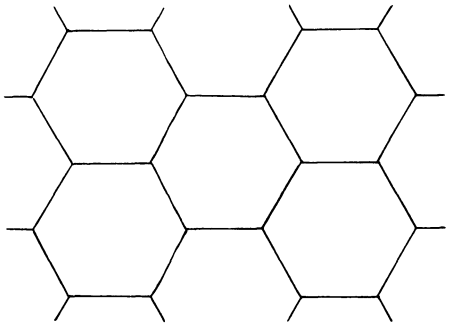


Fig. 27.4.

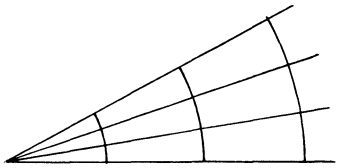


Fig. 27.5.

(b) Hexagonal and Triangular Nets (Fig. 27.4). This type is seldom used.

(c) Polar Nets (Fig. 27.5). This type is also very seldom used. Such nets are sometimes convenient in special domains as, for example, in sectors of a circle etc. The difference formulae are rather complicated.

27.3. Refinement of Nets

As we have seen above, the accuracy of approximation depends on the density of the net. But the refinement of net greatly increases the amount of computation. For this reason it is convenient to refine the net only in those regions where we are interested in higher accuracy. The easiest method of refinement of a net consists in using irregular nets (see § 27.2). Moreover, square nets can be refined by means of diagonal nets, as is seen from Fig. 27.6.

27.4. Finite-Difference Formulae for the Most Frequently Occurring Operators

1. *Poisson's equation*

$$\Delta u \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y)$$

see Tab. 27.2.

2. *The heat conduction equation*

$$\frac{\partial^2 u}{\partial x^2} = a^2 \frac{\partial u}{\partial t}, \quad \Delta u = a^2 \frac{\partial u}{\partial t}$$

see Tab. 27.3 and 27.4, respectively.

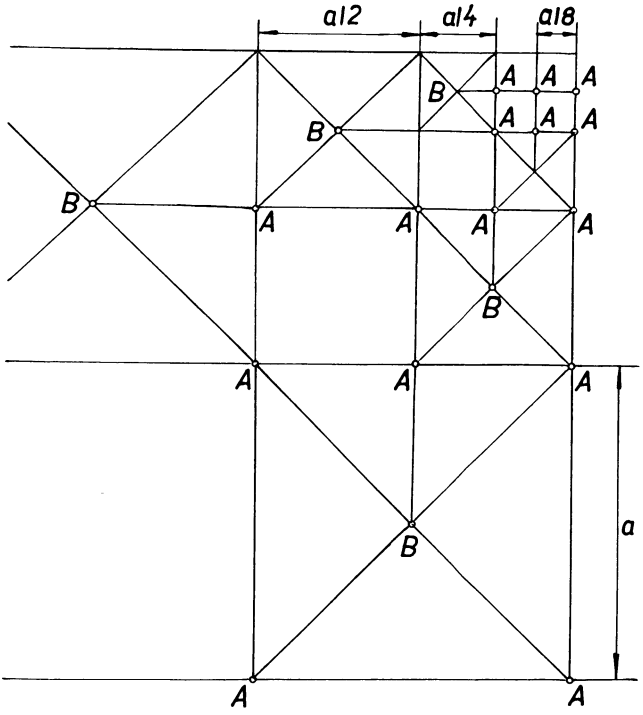


Fig. 27.6.

3. *The biharmonic equation*

$$\Delta\Delta u \equiv \frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} = f(x, y)$$

see Tab. 27.5.

4. *The wave equation*

$$a^2 \frac{\partial^2 u}{\partial x^2} = \frac{\partial u^2}{\partial t^2}, \quad a^2 \Delta u = \frac{\partial^2 u}{\partial t^2}$$

see Tab. 27.6.

In Tab. 27.3 – 27.6 the order of accuracy indicates that the error, when applying the corresponding finite-difference operator to a sufficiently smooth function, is of order h^p. The symbol $\tau^q + h^p$ has a similar meaning.

27.5. Formulation of Boundary Conditions

(a) Boundary Conditions which Do Not Contain Derivatives (the values of the function to be found are given on the boundary of the domain considered). Essentially, two methods are used:

TABLE 27.2

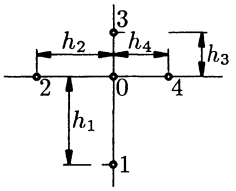
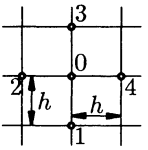
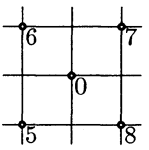
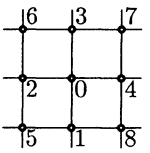
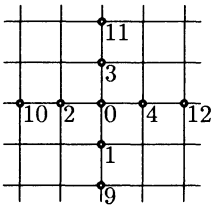
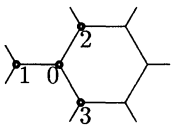
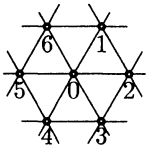
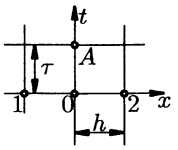
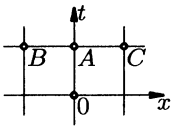
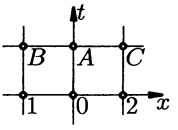
Scheme	Difference equation	Order of accuracy
	$\frac{2}{h_1+h_3}\left(\frac{u_1-u_0}{h_1}+\frac{u_3-u_0}{h_3}\right)+ + \frac{2}{h_2+h_4}\left(\frac{u_2-u_0}{h_2}+\frac{u_4-u_0}{h_4}\right)=f_0$	h $(h_i \leqq h)$
	$u_0=\frac{1}{4}(u_1+u_2+u_3+u_4)-\frac{h^2 f_0}{4}$	h^2
	$u_0=\frac{1}{4}(u_5+u_6+u_7+u_8)-\frac{h^2 f_0}{2}$	h^2
	$u_0=\frac{1}{20}\left[4(u_1+u_2+u_3+u_4)+(u_5+u_6+ + u_7+u_8)\right]-\frac{3}{10}h^2 f_0-\frac{1}{40}h^4\Delta f_0- -\frac{1}{1200}h^6\Delta\Delta f_0-\frac{1}{600}h^6\frac{\partial^4 f_0}{\partial x^2\partial y^2}$	h^6
	$u_0=\frac{1}{60}\left[16(u_1+u_2+u_3+u_4)- -(u_9+u_{10}+u_{11}+u_{12})\right]-\frac{12h^2 f_0}{60}$	h^4
	$u_0=\frac{1}{3}(u_1+u_2+u_3)-\frac{h^2 f_0}{4}$	h
	$u_0=\frac{1}{6}(u_1+u_2+u_3+u_4+u_5+u_6)- -\frac{1}{4}h^2 f_0-\frac{1}{64}h^4\Delta f_0$	h^4

TABLE 27.3

Scheme	Relation between h and τ	Difference equation	Order of accuracy
	$\tau \leqq \frac{a^2h^2}{2}$	$u_A = \frac{\tau}{a^2h^2}u_1 + \left(1 - \frac{2\tau}{a^2h^2}\right)u_0 + \frac{\tau}{a^2h^2}u_2$	$\tau + h^2$
	arbitrary	$-\frac{\tau}{a^2h^2}u_B + \left(1 + \frac{2\tau}{a^2h^2}\right)u_A - \frac{\tau}{a^2h^2}u_C = u_0$	$\tau + h^2$
	arbitrary	$-\frac{\tau}{2a^2h^2}u_B + \left(1 + \frac{\tau}{a^2h^2}\right)u_A - \frac{\tau}{2a^2h^2}u_C = \frac{\tau}{2a^2h^2}u_1 + \left(1 - \frac{\tau}{a^2h^2}\right)u_0 + \frac{\tau}{2a^2h^2}u_2$	$\tau^2 + h^2$

(i) *Collatz's method of linear interpolation.* Only regular mesh points are used and the boundary condition is transferred to the mesh point nearest to the boundary by linear interpolation or extrapolation (see Fig. 27.7 below):

$$u_A = \frac{\sigma}{1+\sigma}u_B + \frac{1}{1+\sigma}\varphi(C),$$

where $\varphi(C)$ is the value of the given function defined on the boundary at the point C.

(ii) *The use of irregular nets* in such a manner that the mesh points lie on the boundary of the given domain (see Example 27.7.1).

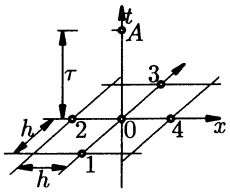
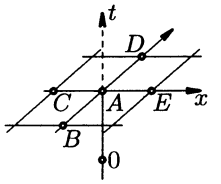
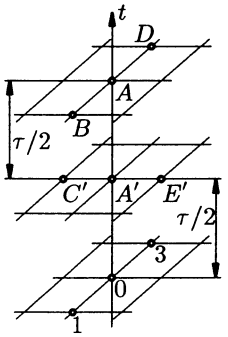
(b) Boundary Conditions Containing Derivatives. This case is treated in a similar way. The derivatives in the boundary conditions are usually linearly interpolated and replaced by finite-differences.

Let us demonstrate this procedure in the case where a linear combination of the value and the normal derivative of the unknown function is given on the boundary:

$$\frac{\partial u}{\partial n} = -k(u - \varphi),$$

where n is the outward normal, φ is the given function, and k is a positive constant.

TABLE 27.4

Scheme	Relation between h and τ	Difference equation	Order of accuracy
	$\tau \leqq \frac{a^2h^2}{4}$	$u_A = \frac{\tau}{a^2h^2}(u_1+u_2+u_3+u_4)+ +\left(1-\frac{4\tau}{a^2h^2}\right)u_0$	$\tau + h^2$
	arbitrary	$\left(1+\frac{4\tau}{a^2h^2}\right)u_A - -\frac{\tau}{a^2h^2}(u_B+u_C+u_D+u_E)=u_0$	$\tau + h^2$
	arbitrary	$\left(1+\frac{\tau}{a^2h^2}\right)u_{A'} - \frac{\tau}{2a^2h^2}(u_{C'}+u_{E'}) = = \left(1-\frac{\tau}{a^2h^2}\right)u_0 + \frac{\tau}{2a^2h^2}(u_1+u_3)$ $\left(1+\frac{\tau}{a^2h^2}\right)u_A - \frac{\tau}{2a^2h^2}(u_B+u_D) = = \left(1-\frac{\tau}{a^2h^2}\right)u_{A'} + \frac{\tau}{2a^2h^2}(u_{C'}+u_{E'})$	$\tau^2 + h^2$

Let us express the normal derivative by means of finite-differences in two ways (see Fig. 27.8):

$$\frac{u_1-u_0}{\delta} \approx \frac{\partial u}{\partial n}, \quad \frac{u_0-u_A}{\eta} \approx \frac{\partial u}{\partial n},$$

and substitute in the boundary condition. We obtain two equations

$$\frac{u_1-u_0}{\delta} = -k(u_0-\varphi_0),$$
$$\frac{u_0-u_A}{\eta} = -k(u_0-\varphi_0).$$

The value u_A at the irregular point of the net is computed by linear interpolation from the values u_2 and u_3:

$$u_A = \frac{(h-\varepsilon)u_2+\varepsilon u_3}{h}.$$

TABLE 27.5

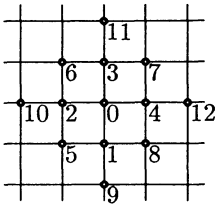
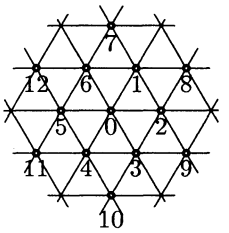
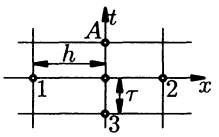
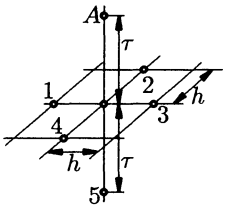
Scheme	Difference equation	Order of accuracy
	$u_0 = \frac{1}{20}[8(u_1+u_2+u_3+u_4)- - 2(u_5+u_6+u_7+u_8)- - (u_9+u_{10}+u_{11}+u_{12})] + \frac{1}{20}f_0h^4$	h^2
	$u_0 = \frac{1}{12}[3(u_1+u_2+u_3+u_4+u_5+u_6)- - (u_7+u_8+u_9+u_{10}+u_{11}+u_{12})] + \frac{3}{64}f_0h^4$	h^2

TABLE 27.6

Equation	Scheme	Difference equation	Order of accuracy
$a^2\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial t^2}$	 $\tau = h/a$	$u_A = u_1 + u_2 - u_3$	h^2
$a^2\Delta u = \frac{\partial^2 u}{\partial t^2}$	 $\tau = h/(a\sqrt{2})$	$u_A = \frac{1}{2}(u_1+u_2+u_3+u_4- - 2u_5)$	h^2

Eliminating u_0 and u_A from the last three equations, we obtain, at the point 1, the equation

$$u_1 = \frac{(1-k\delta)(h-\varepsilon)}{(1+k\eta)h}u_2 + \frac{(1-k\delta)\varepsilon}{(1+k\eta)h}u_3 + k\frac{\delta+\eta}{1+k\eta}\varphi_0.$$

The treatment in other cases is similar. Moreover, in some cases special formulae which guarantee higher accuracy are used.

27.6. Error Estimates

The problem of *error estimates* when solving partial differential equations by the finite-difference method is rather complicated. A rough orientation in the total discretization error is given by the *local discretization* (*truncation*) *error*, i.e., by

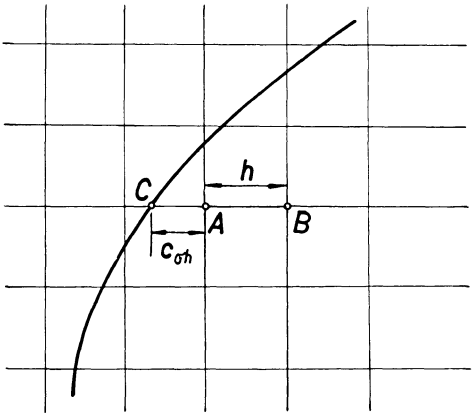


Fig. 27.7.

the error which is performed by replacing differential operators by difference ones. The majority of *a priori estimates* (i.e., the estimates which can be formed before the beginning of the computation) which can be found in the literature (see, e.g., [21], [152]) are very complicated (they depend, among others, on the possibility

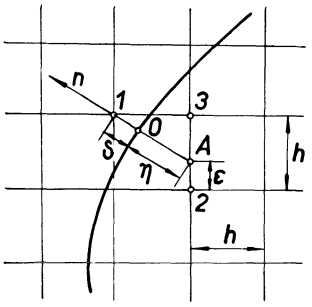


Fig. 27.8.

of bounding high derivatives of the unknown solution). Moreover, these estimates always have to count with the worst possible case, and for that reason, they are very pessimistic, i.e., they are many times greater than the actual error. Thus, the value of such estimates is very limited.

Practically, it is possible to obtain a realistic assessment of the error by the *deferred approach to the limit procedure* which was already used in Chap. 25 (cf., for example, formulae (25.2.16) and (25.3.13)). The basic idea of this method is as follows:

Let us suppose that the order of the error is p (p is usually the order of accuracy to which the derivatives are approximated by the corresponding finite-difference formulae, see tables in § 27.4), i.e., let us suppose that there exists a function $\alpha(x, y$ (independent of h) such that

$$\varepsilon_h(x, y) = u_h(x, y) - u(x, y) = \alpha(x, y)h^p + O(h^{p+1})$$

holds, where u is the exact solution and u_h the approximate solution computed when using the mesh size h. Then one obtains the simple formula

$$\varepsilon_h(x, y) = \frac{2^p}{2^p - 1}\left[u_h(x, y) - u_{h/2}(x, y)\right] + O(h^{p+1})$$

for the error ε_h, where $u_{h/2}$ denotes the approximate solution gained in the net with the half mesh size $h/2$. To obtain this estimate one must solve the given problem twice and the formula which has been introduced is the typical example of an *a posteriori estimate.*

27.7. Examples. Laplace's Equation. Heat-Conduction Equation. Biharmonic Equation

Example 1. Let us solve the problem of a *stationary temperature field* in the plane half-circular plate bounded by the straight line $y = 0$ and by the half of the circumference $y = (36 - x^2)^{1/2}$ if heat is transferred on the linear part of the boundary into a medium of known temperature (given by a function $f(x)$) and on the remainder of the boundary, a constant temperature $u = 0$ is given.

This problem yields the Laplace differential equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

in the domain $x^2 + y^2 < 36$, $y > 0$ and the boundary condition

$$\frac{\partial u}{\partial n}(x, 0) = -k\left[u(x, 0) - f(x)\right],$$

where n is the outward normal and k is a positive constant, on the linear part of the boundary, and

$$u\left(x, (36 - x^2)^{1/2}\right) = 0$$

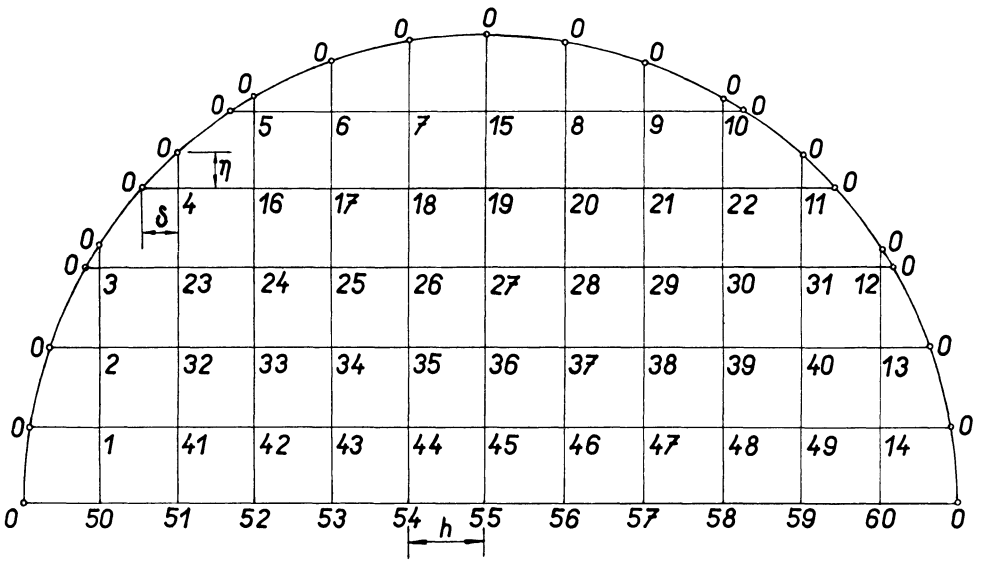


Fig. 27.9.

on the remaining part of the boundary (see Fig. 27.9).

Let us cover the domain $x^2 + y^2 < 36$, $y > 0$ with a net consisting of the system of lines $x = \pm k$, $y = l$, $k, l = 0, 1, \ldots, 6$ (in our special case, $h = 1$). At all points denoted by zeros in Fig. 27.9, the condition $u = 0$ holds.

At the mesh points 15–49 we use the second formula of Tab. 27.2 and obtain, for instance at the point 30, the equation

$$u_{30} = \tfrac{1}{4}(u_{39} + u_{29} + u_{22} + u_{31})$$

and similarly at other mesh points of this type.

At the mesh points 1–14 (i.e. at the mesh points adjacent to irregular points of the net) we use the first formula of Tab. 27.2 and obtain, for instance for the point 4,

$$u_4 = \frac{\eta\delta}{(h+\eta)(\delta+\eta)}u_{23} + \frac{\eta\delta}{(h+\delta)(\delta+\eta)}u_{16}.$$

At mesh points 50–60 we formulate the boundary condition by the procedure described in § 27.5 (where we put $\delta = \varepsilon = 0$, $\eta = h$). The typical equation (written for the mesh point 55) is

$$u_{55} = \frac{u_{45} + khf_{55}}{1 + kh}.$$

Proceeding in this way, we obtain a system of 60 equations for 60 unknowns and this system is then solved by some numerical method for special (sparse) systems of linear algebraic equations (cf. Chap. 30).

Example 2. Let us compute the temperature $u(x, t)$ of a rod of length L, if its temperature at $t = 0$ is $u(x, 0) = p(x)$, where $p(x)$ is a given function, and if the temperature at its ends is zero: $u(0, t) = u(L, t) = 0$. Assuming that heat is conducted in the direction of the x-axis only, the problem is described by the equation

$$\frac{\partial^2 u}{\partial x^2} = a^2 \frac{\partial u}{\partial t}$$

in the domain $0 < x < L$, $t > 0$ (a^2 being a constant characterizing the heat properties of the rod) with the initial condition

$$u(x, 0) = p(x)$$

and boundary conditions

$$u(0, t) = u(L, t) = 0.$$

Let a positive integer M be chosen and let $h = L/M$. Further let $\tau = \beta a^2 h^2$, where β is an arbitrary number for which $0 < \beta \leqq 1/2$, and let us construct, in the domain concerned, a rectangular net formed by a system of parallel lines $x = lh$, $l = 0, \ldots, M$, $t = k\tau$, $k = 0, 1, \ldots$. Let our differential equation be replaced by a

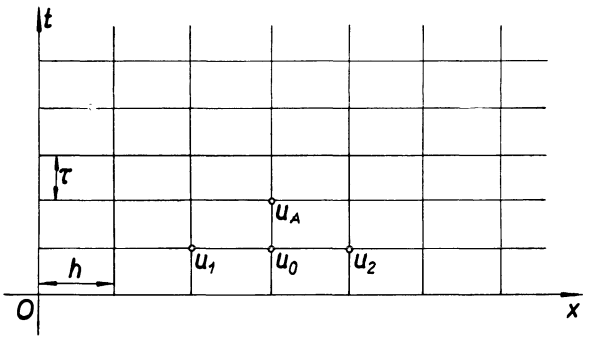


Fig. 27.10.

finite-difference equation according to the first formula of Tab. 27.3 (see Fig. 27.10):

$$u_A = \beta u_1 + (1 - 2\beta)u_0 + \beta u_2.$$

The values in the zero time row (i.e. for $t = 0$) are known from the initial condition. From the above formula, the values of the unknown function at inner mesh points of the first time row are computed; at the boundary mesh points the values are known from the boundary conditions (they are zeros). Similarly we obtain the approximation to the unknown function for any required time row.

Since the values of the function sought are explicitely given in terms of its already computed values, the described computational scheme is called the *explicit scheme.*

REMARK 1. The restricting condition $\beta = \tau/(a^2h^2) \leqq 1/2$ is essential. It can be easily shown that the values of the approximate solution in the k-th time row are linear combinations of terms of the form

$$\left(1 - 4\beta \sin^2 \frac{l\pi h}{2L}\right)^k \sin \frac{l\pi x}{L}, \quad l = 1, \ldots, M-1.$$

If $\beta > 1/2$, then the expression

$$\left|1 - 4\beta \sin^2 \frac{l\pi h}{2L}\right|$$

is greater than unity for those values of l for which $\sin(l\pi h/2L)$ differs sufficiently little from unity so that the above terms tend to infinity with increasing k (i.e., if the net is refined). Thus, the values of the computed function also tend to infinity which contradicts the physical nature of the problem.

REMARK 2. The second and third scheme of Tab. 27.3 are called the *implicit schemes* and they are, at the first glance, not practicable since we must solve a system of linear algebraic equations to obtain the solution in any time row, supposing that the solution in the preceding row has already been computed. However, this system has a tridiagonal matrix so that the number of operations, which we need to solve it, is proportional to the number of equations, i.e., to the number $1/h$ (see Chap. 30). Thus, the number of operations necessary to apply both the implicit methods mentioned above is proportional to the number $1/(\tau h)$. The number of operations of the second method of Tab. 27.3 does not increase, consequently, faster than in the case of the explicit method when the net is refined. Even more convenient is the situation in the case of the third formula from the mentioned table called the *Crank-Nicolson formula.* Namely, the local error of this method is $\tau^2 + h^2$ so that it is reasonable to put in it $\tau = O(h)$ and not $\tau = O(h^2)$ as in the case of methods with local errors of order $\tau + h^2$. Both the mentioned implicit methods are at least as efficient as the explicit method and, moreover, τ can be chosen completely independent of h.

REMARK 3. The numbers of operations necessary for the practical realization of methods from Tab. 27.4 which are used for solving the heat-conduction equation in two space variables are, in their turn, proportional to the expressions $1/h^4$, $1/(\tau h^4)$, and $1/(\tau h^2)$, supposing that all the systems of linear algebraic equations, if they occur, are solved by the Gaussian elimination method. An apparently advantageous number of operations of the third method known under the name of the *alternating directions method* follows from the fact that we proceed, in the computation, in half

time steps and solve, in each of these halfsteps, $1/h$ linear systems with tridiagonal matrices for $1/h$ unknowns.

Example 3. Let us determine the deflection of a square plate of the side $2a$ loaded by a uniform loading and clamped on the boundary.

This problem yields the differential equation

$$\Delta\Delta u \equiv \frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} = f, \quad f = \frac{P}{N},$$

where P is the load on the unit area and N is a constant depending on the elastic properties of the plate; the boundary conditions are

$$\begin{aligned} u &= 0 \quad \text{for} \quad x = \pm a, \quad y = \pm a, \\ \frac{\partial u}{\partial x} &= 0 \quad \text{for} \quad x = \pm a, \\ \frac{\partial u}{\partial y} &= 0 \quad \text{for} \quad y = \pm a. \end{aligned}$$

In order to write the finite-difference equations at a mesh point (x, y) in the form given in Tab. 27.5, the values at the mesh points $(x + h, y)$, $(x - h, y)$, $(x, y + h)$, $(x, y - h)$, $(x + h, y + h)$, $(x - h, y + h)$, $(x + h, y - h)$, $(x - h, y - h)$, $(x + 2h, y)$, $(x - 2h, y)$, $(x, y + 2h)$ and $(x, y - 2h)$ are necessary. At boundary mesh points, the solution is known (it is zero). But it is necessary to know the values of the function

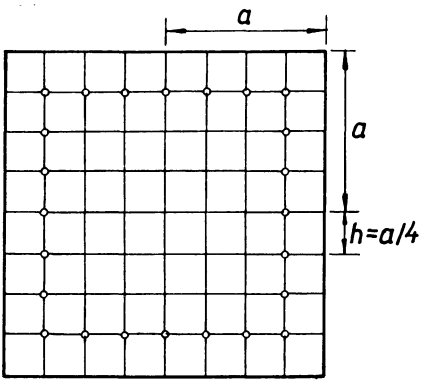


Fig. 27.11.

u at the mesh points, adjacent to the boundary mesh points, denoted in Fig. 27.11 by small circles. These values can be obtained, generally, by linear interpolation from the known boundary values of $\partial u/\partial x$ and $\partial u/\partial y$. In the actual case, they are also equal to zero.

27.8. General Scheme of the Finite-Difference Method

In this paragraph, we formulate the finite-difference method abstractly and introduce basic convergence theorems for a general linear boundary value problem.

Let a domain $\Omega \subset E_N$ with boundary S be given and let us investigate the differential equation

$$Lu = f \quad \text{in } \Omega \tag{1}$$

with boundary conditions

$$l_i u = \varphi_i \quad \text{on } S_i, \quad i = 1, \ldots, s. \tag{2}$$

Here u is the function sought, f a given function, L a linear differential operator, S_i some parts of the boundary S, φ_i functions defined on S_i and l_i linear operators mapping the function u defined on Ω onto the functions defined on S_i. Note, that it is not necessary for S_i's to be disjoint for different i as well as it is not necessary that the union of S_i is the whole boundary S of the given domain.

Example 1. Let us consider the wave equation in one space variable,

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial t^2}$$

in the domain $\Omega = \{(x, t);\ 0 < x < 1,\ 0 < t < T\}$ with the initial conditions

$$u(x, 0) = p(x), \quad \frac{\partial u(x, 0)}{\partial t} = q(x)$$

(p, q are given functions) and the boundary conditions

$$u(0, t) = u(1, t) = 0.$$

In this case, we have $s = 4$, $S_1 = S_2 = \{(x, t);\ 0 \leqq x \leqq 1,\ t = 0\}$, $S_3 = \{(x, t);\ x = 0,\ 0 < t < T\}$, $S_4 = \{(x, t);\ x = 1,\ 0 < t < T\}$, $l_1 u = u(x, 0)$, $l_2 u = \dfrac{\partial u}{\partial t}(x, 0)$, $l_3 u = u(0, t)$, $l_4 u = u(1, t)$.

Further, let a finite set of points from the closed set $\overline{\Omega}$ (the bar denotes the closure) be given for any $h > 0$. This set is called the *net* (*grid, mesh*) and will be denoted by $\overline{\Omega}^{(h)}$. Let L_h be a linear operator mapping the function $u^{(h)}$ defined on the net $\overline{\Omega}^{(h)}$ on a function $L_h u^{(h)}$ defined on a proper subset $\Omega^{(h)}$ of $\overline{\Omega}^{(h)}$. The elements of the set $\Omega^{(h)}$ are called the *inner or interior mesh points of the net* $\overline{\Omega}^{(h)}$. Finally, let $l_i^{(h)}$ $(i = 1, \ldots, s)$ be a linear operator mapping the function $u^{(h)}$ defined on $\overline{\Omega}^{(h)}$ on a function $l_i^{(h)} u^{(h)}$ defined on $S_i^{(h)} \subset \overline{\Omega}^{(h)}$, $S_i^{(h)} \cap \Omega^{(h)} = \emptyset$ and $\Lambda_i^{(h)}$ an operator which maps the function φ_i restricted on $S_{0i}^{(h)}$, $S_{0i}^{(h)}$ being the finite subset

of S_i, on a function $\Lambda_i^{(h)}\varphi_i$ defined on $S_i^{(h)}$. The elements of $\bigcup_i S_i^{(h)}$ are called the *boundary mesh points* and we suppose that

$$\bigcup_{i=1}^{s} S_i^{(h)} \cup \Omega^{(h)} = \overline{\Omega}^{(h)}.$$

To solve the problem (1), (2) by the finite-difference method means to solve the equation

$$L_h u^{(h)} = f, \tag{3}$$

called the *difference equation*, with right-hand term being defined on $\Omega^{(h)}$ and equal there to the values of the right-hand term of (1), with the boundary conditions

$$l_i^{(h)} u^{(h)} = \Lambda_i^{(h)}\varphi_i, \quad i = 1, \ldots, s. \tag{4}$$

Any of equations (4) represents a finite number of equations for the mesh points from the sets $S_i^{(h)}$ and (3) a finite number of equations for the mesh points from $\Omega^{(h)}$. Thus, (3) and (4) form a system of linear algebraic equations for determining the values of $u^{(h)}$ at the mesh points from $\overline{\Omega}^{(h)}$.

REMARK 1. The operators $\Lambda_i^{(h)}$ in (4) represent the way of transforming the boundary conditions at the mesh points of $\overline{\Omega}^{(h)}$. The right-hand term of the original equation can be transformed to the mesh in a similarly general manner. We have restricted ourselves to the above special case (namely, when the given function f is mapped on the function defined on $\Omega^{(h)}$ by its function values) since it is most often used in practice.

Example 2. When formulating the Dirrichlet boundary conditions by means of linear interpolation, we have $s = 1$, the set $S_1^{(h)}$ is the set of mesh points of the type A from Fig. 27.7, the set $S_{01}^{(h)}$ is the set of points of the type C from the same figure, and

$$\left(l_i^{(h)} u^{(h)}\right)(A) = u^{(h)}(A) - \frac{\sigma}{1+\sigma} u^{(h)}(B),$$

$$\left(\Lambda^{(h)}\varphi\right)(A) = \frac{1}{1+\sigma}\varphi(C)$$

(see § 27.5).

Let U be a normed space (§ 22.4) of functions defined on $\overline{\Omega}$ and such that the expressions Lu, and $l_i u$ are defined for any $u \in U$. Further, let F and Φ_i, $i = 1, \ldots, s$, respectively, be normed spaces of functions defined on Ω and on S_i, respectively, and such that $Lu \in F$ and $l_i u \in \Phi_i$, $i = 1, \ldots, s$, respectively, for any $u \in U$. Let $U^{(h)}$ be a normed space of functions defined on $\overline{\Omega}^{(h)}$, $F^{(h)}$ a normed space of functions defined on $\Omega^{(h)}$ and $\Phi_i^{(h)}$, $i = 1, \ldots, s$, normed spaces of functions defined on $S_i^{(h)}$,

and let, for $u^{(h)} \in U^{(h)}$ and $\varphi_i \in \Phi_i$, $L_h u^{(h)} \in F^{(h)}$, $l_i^{(h)} u^{(h)} \in \Phi_i^{(h)}$, $\Lambda_i^{(h)} \varphi_i \in \Phi_i^{(h)}$. Let us suppose that any functions $u \in U$ and $f \in F$, respectively, are as functions restricted on $\overline{\Omega}^{(h)}$ and $\Omega^{(h)}$, respectively, elements of the spaces $U^{(h)}$ and $F^{(h)}$, respectively, so that the expressions $L_h u$ and $l_i^{(h)} u$ have sense. Let us finally suppose that the norms defined in the spaces just introduced satisfy

$$\begin{gathered} \|u\|_{U^{(h)}} \to \|u\|_U, \quad \|f\|_{F^{(h)}} \to \|f\|_F, \\ \left\|\Lambda_i^{(h)} \varphi_i\right\|_{\Phi_i^{(h)}} \to \|\varphi_i\|_{\Phi_i} \end{gathered} \tag{5}$$

for $h \to 0$.

Definition 1. We say that the *difference equation* (3) *with the boundary conditions* (4) *approximates the differential equation* (1) *with the boundary conditions* (2) if, for any $u \in U$ and for $h \to 0$,

$$\begin{gathered} \|Lu - L_h u\|_{F^{(h)}} \to 0, \\ \left\|\Lambda_i^{(h)}(l_i u) - l_i^{(h)} u\right\|_{\Phi_i^{(h)}} \to 0. \end{gathered} \tag{6}$$

Definition 2. We say that the *approximation of the differential equation* (1) *with the boundary conditions* (2) *by the difference equation* (3) *with the boundary conditions* (4) *is of order* p, if there exist, for any function $u \in U$, constants M and M_i such that

$$\begin{gathered} \|Lu - L_h u\|_{F^{(h)}} \leqq M h^p, \\ \left\|\Lambda_i^{(h)}(l_i u) - l_i^{(h)} u\right\|_{\Phi_i^{(h)}} \leqq M_i h^p \end{gathered} \tag{7}$$

for any sufficiently small h.

Definition 3. We say that the *difference equation* (3) *with the boundary conditions* (4) *is well-posed* (*or stable with respect to input data*) if it has, for any sufficiently small h and for any right-hand terms f and φ_i, the unique solution and if, moreover, there exist such constants N and N_i that

$$\left\|u^{(h)}\right\|_{U^{(h)}} \leqq N \left\|L_h u^{(h)}\right\|_{F^{(h)}} + \sum_{i=1}^{s} N_i \left\|l_i^{(h)} u^{(h)}\right\|_{\Phi_i^{(h)}} \tag{8}$$

for any $u^{(h)} \in U^{(h)}$ and any sufficiently small h.

REMARK 2. The fact that the problem (3), (4) is well-posed means that its solution depends continuously on the input data, i.e., on the right-hand terms of (3) and (4).

Definition 4. We say that the *difference equation* (3) *is stable with respect to the right-hand term* if the equation (3) with the homogeneous boundary conditions

$$l_i^{(h)} u^{(h)} = 0, \quad i = 1, \ldots, s, \tag{9}$$

has, for sufficiently small h, exactly one solution and if

$$\|u^{(h)}\|_{U(h)} \leqq N\|L_h u^{(h)}\|_{F^{(h)}}. \tag{10}$$

We say that (3) is *stable with respect to boundary conditions* if the homogeneous equation

$$L_h u^{(h)} = 0 \tag{11}$$

with the boundary conditions (4) has the unique solution for any sufficiently small h and if

$$\|u^{(h)}\|_{U^{(h)}} \leqq \sum_{i=1}^{s} N_i \|l_i^{(h)} u^{(h)}\|_{\Phi_i^{(h)}}. \tag{12}$$

The stability with respect only to some of boundary conditions is defined in an obvious way.

Theorem 1. *Let $u \in U$ be the solution of the differential equation* (1) *with the boundary conditions* (2). *Let the difference equation* (3) *with the boundary conditions* (4) *approximate the equation* (1) *with the boundary conditions* (2). *Further, let the problem* (3), (4) *be well-posed. Then*

$$\lim_{h\to 0} \|u^{(h)} - u\|_{U^{(h)}} = 0. \tag{13}$$

If, moreover, the approximation is of order p then

$$\|u^{(h)} - u\|_{U^{(h)}} \leqq h^p \Big(MN + \sum_{i=1}^{s} M_i N_i\Big). \tag{14}$$

REMARK 3. The assumptions of this theorem can be, in special situations, weakened. For example, if some boundary condition is formulated exactly, i.e., if $S_i^{(h)} \subset S_i$, $l_i^{(h)} = l_i$, and $\Lambda_i^{(h)} \varphi_i = \varphi_i$ for some index i, it is not necessary to require the continuous dependence on this condition.

REMARK 4. To assert that the finite dimensional problem (3), (4) approximates the original problem (1), (2) is usually simple and it is very often sufficient to use the Taylor formula in a trivial way. The investigation of well-posedness is, in contrast, substantially more difficult. Some general properties of the solution like the maximum principle, the monotonicity of the corresponding matrices, etc. may be very useful here.

Theorem 2. *Let the assumptions of Theorem* 1 *be satisfied and let there exist functions ψ and ψ_i independent of h and such that, for the given solution of the differential equation* (1) *with the boundary conditions* (2),

$$\begin{aligned} &\lim_{h\to 0} \|h^{-p}(Lu - L_h u) - \psi\|_{F^{(h)}} = 0, \\ &\lim_{h\to 0} \Big\|h^{-p}\big[\Lambda_i^{(h)}(l_i u) - l_i^{(h)} u\big] - \Lambda_i^{(h)} \psi_i\Big\|_{\Phi_i^{(h)}} = 0 \end{aligned} \tag{15}$$

hold. Let, further, in some class V on which the difference equation (3) *with the boundary conditions* (4) *approximates the differential equation* (1) *with the boundary conditions* (2), *there exists a solution of the boundary value problem*

$$Lw = \psi, \quad l_i w = \psi_i, \quad i = 1, \ldots, s. \tag{16}$$

Then

$$\lim_{h \to 0} \left\| h^{-p}(u^{(h)} - u) - w \right\|_{U^{(h)}} = 0. \tag{17}$$

REMARK 5. Theorem 2 forms the theoretical basis for using the deferred approach to limit procedure (cf. § 27.6).

REMARK 6. All considerations in this paragraph have been performed for the case of a net the geometry of which is characterized by a single parameter h. It is more or less clear that this is not substantial and that the introduced theorems hold also in the case of irregular nets.

REMARK 7. A similar abstract scheme of the finite-difference method can be constructed, without any substantial difficulties, also for non-linear problems. The verification of general assumptions for actual cases is here, however, extremely complicated, as a rule.

28. INTEGRAL TRANSFORMS (OPERATIONAL CALCULUS)

By JINDŘICH NEČAS

References: [7], [12], [28], [33], [68], [77], [110], [116], [118], [119], [134], [148], [238], [319], [325], [326], [374], [398], [435], [455] [467], [478].

For solving certain types of ordinary differential equations, particularly those with constant coefficients, and certain types of partial differential equations (e.g. equation of heat conduction, of string and diaphragm vibrations, etc.) in special domains, transform methods may be advantageously used. From among them, the Laplace-Carson transform is formally identical with the operational calculus, and the finite Fourier transform leads to the expansion of a function in a Fourier series.

28.1. One-Dimensional Infinite Transforms (the Laplace, Fourier, Mellin, Hankel Transforms)

By each of integral transforms, given in Tab. 28.1 below (the transforms (1)–(8)), to every function $f(t)$ (the so-called *original*) from some class of functions, there is assigned a certain function $F(p)$ (the so-called *image of the function* $f(t)$). For example, the so-called *Laplace image* (i.e. the image by the Laplace transform (1)) of the function $f(t) = \mathrm{e}^{3t}$ is

$$F(p) = \int_0^\infty \mathrm{e}^{3t}\, \mathrm{e}^{-pt}\, \mathrm{d}t = \int_0^\infty \mathrm{e}^{(3-p)t}\, \mathrm{d}t = \left[\frac{1}{3-p}\, \mathrm{e}^{(3-p)t}\right]_0^\infty = \frac{1}{p-3} \quad (\mathrm{Re}\, p > 3). \tag{9}$$

(The Laplace image of a function $f(t)$ is frequently denoted by $\mathcal{L}\{f(t)\}$. Thus, in our example we have $F(p) = \mathcal{L}\{\mathrm{e}^{3t}\} = 1/(p-3)$.)

In the transforms (5), (6), (8), we have $t \in [0, +\infty)$, $p \geqq 0$; in the transform (4), $t \in (-\infty, +\infty)$ and p is real; in the transforms (1), (2), (7), and (3), $t \in [0, +\infty)$, and $t \in (-\infty, +\infty)$, respectively, p being a complex number (not an arbitrary one; its choice depends on the function $f(t)$).

TABLE 28.1

Transform	Image	Inversion formula
Laplace	$F(p)=\int_0^\infty f(t)\,e^{-pt}\,dt$ (1)	$f(t)=\frac{1}{2\pi i}\int_{x-i\infty}^{x+i\infty} F(p)\,e^{pt}\,dp$ (1')
Laplace-Carson	$F(p)=p\int_0^\infty f(t)\,e^{-pt}\,dt$ (2)	$f(t)=\frac{1}{2\pi i}\int_{x-i\infty}^{x+i\infty} \frac{F(p)}{p}\,e^{pt}\,dp$ (2')
Bilateral-Laplace	$F(p)=\int_{-\infty}^\infty f(t)\,e^{-pt}\,dt$ (3)	$f(t)=\frac{1}{2\pi i}\int_{x-i\infty}^{x+i\infty} F(p)\,e^{pt}\,dp$ (3')
Fourier	$F(p)=\int_{-\infty}^\infty f(t)\,e^{-ipt}\,dt$ (4)	$f(t)=\frac{1}{2\pi}\int_{-\infty}^\infty F(p)\,e^{ipt}\,dp$ (4')
Fourier-Cosine	$F(p)=\sqrt{\left(\frac{2}{\pi}\right)}\int_0^\infty f(t)\cos pt\,dt$ (5)	$f(t)=\sqrt{\left(\frac{2}{\pi}\right)}\int_0^\infty F(p)\cos pt\,dp$ (5')
Fourier-Sine	$F(p)=\sqrt{\left(\frac{2}{\pi}\right)}\int_0^\infty f(t)\sin pt\,dt$ (6)	$f(t)=\sqrt{\left(\frac{2}{\pi}\right)}\int_0^\infty F(p)\sin pt\,dp$ (6')
Mellin	$F(p)=\int_0^\infty f(t)t^{p-1}dt$ (7)	$f(t)=\frac{1}{2\pi i}\int_{x-i\infty}^{x+i\infty} F(p)t^{-p}dp$ (7')
Hankel	$F(p)=\int_0^\infty J_\nu\left(2\sqrt{(pt)}\right)f(t)dt$ (8)	$f(t)=\int_0^\infty J_\nu\left(2\sqrt{(pt)}\right)F(p)dp$ (8')
	(J_ν is the Bessel function of the first kind, $\nu>-1$)	

The improper integrals are understood in the usual sense, for example,

$$\int_0^\infty f(t)\,e^{-pt}\,dt=\lim_{b\to+\infty}\int_0^b f(t)\,e^{-pt}\,dt,$$

etc. In order to guarantee the convergence of these integrals, the functions $f(t)$ and the numbers p must have certain properties. For example, integral (9) (see the preceding page) is convergent for those complex numbers p which satisfy the inequality $\mathrm{Re}\,p>3$. Thus, the Laplace image of the function e^{3t} is a function of a complex variable p, defined in the half-plane $\mathrm{Re}\,p>3$ of the Gaussian plane (i.e. in the half-plane $x>3$, if we set $p=x+iy$; for $\mathrm{Re}\,p\leqq 3$ the function $1/(p-3)$ is *not* the Laplace image of the function e^{3t}). If, in addition, the function $f(t)$ has certain special properties, then also the corresponding image $F(p)$ has certain particular properties (see § 28.3).

In the integral transforms of Tab. 28.1, the following conditions are imposed on the function $f(t)$: $f(t)$ is absolutely integrable

(i) in every finite interval $0\leqq a\leqq t\leqq b<+\infty$ in case of transforms (1) and (2),

(ii) in every finite interval $-\infty<a\leqq t\leqq b<+\infty$ in case (3),

(iii) in the interval $(-\infty,+\infty)$ in case (4),

(iv) in the interval $[0, +\infty)$ in cases (5) and (6),

(v) in every finite interval $0 < a \leqq t \leqq b < +\infty$ in case (7).

In the case of the transform (8) we assume the function $f(t)t^{\nu/2}$ to be absolutely integrable in every finite interval $0 \leqq a \leqq t \leqq b < +\infty$.

By an absolutely integrable function $g(t)$ in an interval (a, b) (or $[a, b]$) we understand a function for which both the integrals $\int_a^b g(t)\,\mathrm{d}t$ and $\int_a^b |g(t)|\,\mathrm{d}t$ are convergent. Furthermore, we assume that the functions $f(t)$ are such that

(i) in the case of transforms (1) and (2) a constant $\sigma \geqq -\infty$ exists such that the integral (1) is convergent for $\mathrm{Re}\, p > \sigma$ (in the case (9) we had $\sigma = 3$; not every function possesses the property just mentioned; for example, for $f(t) = \mathrm{e}^{t^2}$, the integral (1) is divergent for *every* p);

(ii) in cases (3) and (7) numbers $\sigma_1 \geqq -\infty$, $\sigma_2 \leqq +\infty$ exist such that for $\sigma_1 < \mathrm{Re}\, p < \sigma_2$ the integral (3) or (7) is convergent, respectively.

The inversion formulae (1') to (8') assign the original to the image provided certain assumptions are satisfied (§ 28.3). The integral (1') means

$$\lim_{\omega \to +\infty} \frac{1}{2\pi \mathrm{i}} \int_{x-\omega \mathrm{i}}^{x+\omega \mathrm{i}} F(p)\, \mathrm{e}^{pt}\, \mathrm{d}p$$

and may also be written in the form

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} F(x + \mathrm{i}y)\, \mathrm{e}^{(x+\mathrm{i}y)t}\, \mathrm{d}y,$$

and similarly in the other cases.

By the Hankel transform, the transform

$$F(v) = \int_0^{\infty} u\, \mathrm{J}_\nu(vu)\varphi(u)\, \mathrm{d}u$$

(which follows from (8) by the substitution $\sqrt{(2p)} = v$, $\sqrt{(2t)} = u$, $\varphi(u) = f(\frac{1}{2}u^2)$) is sometimes understood.

Instead of the Laplace and Fourier transform, one often speaks of the *Laplace* and *Fourier integral*, respectively.

When solving differential equations, we use transforms for reducing the number of independent variables in the differential equation under consideration. If an ordinary differential equation is solved by means of an integral transform, an algebraic equation is obtained; using transforms for solving partial differential equations, the number of independent variables is reduced by one. The type of transform used depends on the equation under consideration and on the corresponding domain of definition. (A thorough treatment of these problems may be found, e.g. in [435], [467]).

The Laplace transform is the most frequently employed owing to the simple relationship which exists between the Laplace image of a function $f(t)$ and of its derivative $f'(t)$, i.e.

$$\int_0^\infty f'(t)\,\mathrm{e}^{-pt}\,\mathrm{d}t = p\,F(p) - f(0). \tag{10}$$

This relationship follows from the theorem on integration by parts provided certain obvious assumptions are satisfied. If we assume that $f(0) = 0$, then to the operation of differentiation of the original there corresponds the algebraic operation of multiplying the image $F(p)$ by p. Upon this fact the Heaviside operational calculus may be theoretically based. To the operation of multiplying by p there corresponds the differentiation of the original, to the operation $1/p$ the integration of the original; hence, p is the inverse operator of $1/p$. This idea has led J. Mikusiński to the definition of the differentiation operator independently of the integral transform methods; this has been well developed and adapted for applications in [325].

Similarly, using the Fourier transform and assuming absolute integrability of the functions $f(t)$, $f'(t)$, $f''(t)$, etc., in the interval $(-\infty, +\infty)$ (this assumption yields $f(t) \to 0$ for $t \to \pm\infty$, etc.), we obtain

$$\int_{-\infty}^\infty f'(t)\,\mathrm{e}^{-\mathrm{i}pt}\,\mathrm{d}t = \left[f(t)\,\mathrm{e}^{-\mathrm{i}pt}\right]_{-\infty}^{\infty} + \mathrm{i}p\int_{-\infty}^\infty f(t)\,\mathrm{e}^{-\mathrm{i}pt}\,\mathrm{d}t = \mathrm{i}pF(p), \tag{11}$$

$$\int_{-\infty}^\infty f''(t)\,\mathrm{e}^{-\mathrm{i}pt}\,\mathrm{d}t = \left[f'(t)\,\mathrm{e}^{-\mathrm{i}pt}\right]_{-\infty}^{\infty} + \mathrm{i}p\int_{-\infty}^\infty f'(t)\,\mathrm{e}^{-\mathrm{i}pt}\,\mathrm{d}t = (\mathrm{i}p)^2F(p) = -p^2F(p), \tag{12}$$

etc.

28.2. Applications of the Laplace and Fourier Transforms to the Solution of Differential Equations. Examples

Example 1. Let us find the current $i(t)$ in a circuit consisting of an inductance L and resistance R, provided $i(t) = 0$ for $t = 0$ and an electromotive force E is applied at the time $t = 0$.

The circuit is governed by the differential equation

$$L\frac{\mathrm{d}i}{\mathrm{d}t} + Ri = E. \tag{1}$$

We multiply equation (1) by the factor e^{-pt} and integrate between the limits 0 and $+\infty$. Assuming the function $i(t)$ bounded and $\mathrm{Re}\,p > 0$, and writing $J(p)$ for $\int_0^\infty i(t)\,\mathrm{e}^{-pt}\,\mathrm{d}t$, we have

$$L\int_0^\infty \mathrm{e}^{-pt}\frac{\mathrm{d}i}{\mathrm{d}t}\mathrm{d}t = L\left[i\,\mathrm{e}^{-pt}\right]_0^\infty + Lp\int_0^\infty i\,\mathrm{e}^{-pt}\,\mathrm{d}t = Lp\,J(p),$$

because $i(0) = 0$ and $\lim_{t\to+\infty} i\,\mathrm{e}^{-pt} = 0$. (Formulae (28.1.10) could also be applied directly, of course.) Further, $\int_0^\infty E\,\mathrm{e}^{-pt}\,\mathrm{d}t = E/p$. Thus, for the image $J(p)$ we get from (1) an algebraic equation,

$$Lp\,J(p) + R\,J(p) = \frac{E}{p}.$$

Hence,

$$J(p) = \frac{E}{p}\frac{1}{R+Lp} = \frac{E}{R}\left(\frac{1}{p} - \frac{1}{p+R/L}\right).$$

Using tables of transform pairs (e.g. the first and sixth pairs in Tab. 28.2 below) we find that $1/p$ is the image of the function $f(t) = 1$, and $1/(p+R/L)$ is the image of the function $\mathrm{e}^{-(R/L)t}$; consequently, for the desired solution we get

$$i(t) = \frac{E}{R}\left(1 - \mathrm{e}^{-(R/L)t}\right).$$

Example 2. Let us find the temperature distribution $u(t, x)$ in a semi-infinite rod with insulated surface, the end of which is kept in a basin with constant temperature, i.e. $u(t, 0) = U_0 = \text{const}$. Let the initial temperature be zero, i.e. $u(0, x) = 0$, and assume that $\lim_{x\to\infty} u(x, t) = 0$.

The function u satisfies the differential equation

$$\frac{\partial^2 u}{\partial x^2} = k\frac{\partial u}{\partial t}, \tag{2}$$

where k is a positive constant determined by conductivity, specific heat and specific mass of the rod. We multiply equation (2) by the factor e^{-pt} with $\mathrm{Re}\,p > 0$ and integrate it between the limits 0 and $+\infty$. We thus obtain

$$\int_0^\infty \frac{\partial u}{\partial t}(t, x)\,\mathrm{e}^{-pt}\,\mathrm{d}t = \left[u(t, x)\,\mathrm{e}^{-pt}\right]_0^\infty + p\int_0^\infty u(t, x)\,\mathrm{e}^{-pt}\,\mathrm{d}t = pU(p, x),$$

where $U(p, x) = \int_0^\infty u(t, x)\,\mathrm{e}^{-pt}\,\mathrm{d}t$ is the image of $u(t, x)$. Further assuming that differentiation under the integral sign is permitted (Theorem 13.9.9; see also [77], p. 167), we obtain

$$\int_0^\infty \frac{\partial^2 u}{\partial x^2}(t, x)\,\mathrm{e}^{-pt}\,\mathrm{d}t = \frac{\mathrm{d}^2 U}{\mathrm{d}x^2}(p, x).$$

Thus, for the image $U(p, x)$ we get from (2) the ordinary differential equation

$$\frac{\mathrm{d}^2 U}{\mathrm{d}x^2}(p, x) - kpU(p, x) = 0, \tag{3}$$

with boundary conditions

$$U(p, 0) = \int_0^\infty U_0 \, \mathrm{e}^{-pt} \, \mathrm{d}t = \frac{U_0}{p}, \quad \lim_{x \to +\infty} U(p, x) = 0.$$

For the general solution of the differential equation (3) we have

$$U(p, x) = A(p) \, \mathrm{e}^{\sqrt{(kp)}x} + B(p) \, \mathrm{e}^{-\sqrt{(kp)}x}.$$

The boundary conditions yield

$$U(p, x) = \frac{U_0}{p} \, \mathrm{e}^{-\sqrt{(kp)}x}.$$

In tables of transforms we find that the function $\mathrm{e}^{-\sqrt{(kp)}x}/p$ is the image of the function $\operatorname{erfc}\big(x\sqrt{(k)}/(2\sqrt{(t)})\big)$ so that for the desired solution we get

$$u(t, x) = U_0 \operatorname{erfc} \frac{x\sqrt{k}}{2\sqrt{t}} = \frac{2}{\sqrt{\pi}} U_0 \int_{\frac{x\sqrt{k}}{2\sqrt{t}}}^\infty \mathrm{e}^{-u^2} \, \mathrm{d}u = U_0 \left[1 - \frac{2}{\sqrt{\pi}} \int_0^{\frac{x\sqrt{k}}{2\sqrt{t}}} \mathrm{e}^{-u^2} \, \mathrm{d}u \right].$$

Example 3. We have to find the steady temperature distribution $u(x, y)$ in the upper half-plane Ω of the xy plane, i.e. in

$$\Omega = \{ (x, y) \, ; \, y > 0 \},$$

supposed the function u assumes the known values $g(x)$ on the boundary S of Ω (on the x-axis),

$$u(x, 0) = g(x). \tag{4}$$

Let us assume that the function $g(x)$ is absolutely integrable in the interval $(-\infty, +\infty)$ and that so are the (unknown) functions

$$u, \quad \frac{\partial u}{\partial x}, \quad \frac{\partial^2 u}{\partial x^2}, \quad \frac{\partial^2 u}{\partial y^2}$$

for every $y > 0$ fixed.

The function u satisfies the differential equation

$$\Delta u \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \qquad \text{in } \Omega. \tag{5}$$

Let us use the Fourier transform in the variable x (thus keeping y fixed). Denoting

$$\int_{-\infty}^\infty u(x, y) \, \mathrm{e}^{-\mathrm{i}px} \, \mathrm{d}x = U(p, y), \tag{6}$$

we have (cf. (28.1.12)).

$$\int_{-\infty}^{\infty} \frac{\partial^2 u}{\partial x^2}(x, y)\, \mathrm{e}^{-\mathrm{i}px}\, \mathrm{d}x = -p^2 U(p, y).$$

Further, differentiating (6) two times with respect to y, we get, under obvious assumptions,

$$\int_{-\infty}^{\infty} \frac{\partial^2 u}{\partial y^2}(x, y)\, \mathrm{e}^{-\mathrm{i}px}\, \mathrm{d}x = \frac{\mathrm{d}^2 U}{\mathrm{d}y^2}(p, y).$$

The Fourier transform, applied to equation (5) and to the condition (4) thus yields

$$\frac{\mathrm{d}^2 U}{\mathrm{d}y^2} - p^2 U = 0, \quad y > 0, \tag{7}$$

$$U(p, 0) = G(p), \tag{8}$$

where

$$G(p) = \int_{-\infty}^{\infty} g(x)\, \mathrm{e}^{-\mathrm{i}px}\, \mathrm{d}x \tag{9}$$

is the Fourier transform of the function $g(x)$.

The general integral of (7) is

$$U(p, y) = C_1(p)\, \mathrm{e}^{|p|y} + C_2(p)\, \mathrm{e}^{-|p|y}. \tag{10}$$

However, $\mathrm{e}^{|p|y}$ cannot be a Fourier transform of any absolutely integrable function $u(x)$ (for any fixed $y > 0$). In fact, if $\int_{-\infty}^{\infty} |u(x)|\, \mathrm{d}x = a$, then

$$\left| \int_{-\infty}^{\infty} u(x)\, \mathrm{e}^{-\mathrm{i}px}\, \mathrm{d}x \right| \leqq \int_{-\infty}^{\infty} |u(x)|\, \mathrm{d}x = a = \text{const}.$$

(because $|\mathrm{e}^{-\mathrm{i}px}| = 1$ for every x and every real p), while $\mathrm{e}^{|p|y} \to +\infty$ for $|p| \to +\infty$ if $y > 0$. So the first term on the right-hand side of (10) should be dropped out. From (8) we then obtain

$$U(p, y) = G(p)\, \mathrm{e}^{-|p|y}. \tag{11}$$

To find the original $u(x, y)$ to the image (11), let us apply the Convolution Theorem 28.3.7, setting there $F(p) = \mathrm{e}^{-|p|y}$. Using tables, or applying directly (28.1.4'), we get, first,

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(p)\, \mathrm{e}^{\mathrm{i}px}\, \mathrm{d}p = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}px}\, \mathrm{e}^{-|p|y}\, \mathrm{d}p = \frac{1}{\pi} \frac{y}{x^2 + y^2}, \quad y > 0. \tag{12}$$

(In details:

$$\int_{-\infty}^{\infty} e^{ipx} e^{-|p|x}\, dp = \int_0^{\infty} e^{ipx} e^{-py}\, dp + \int_{-\infty}^{0} e^{ipx} e^{py}\, dp =$$
$$= \int_0^{\infty} e^{ipx} e^{-py}\, dp + \int_0^{\infty} e^{-ipx} e^{-py}\, dp =$$
$$= 2\int_0^{\infty} e^{-py} \cos px\, dp = \frac{2y}{x^2+y^2}, \quad y > 0;$$

we have used the substitution $p = -q$ in $\int_{-\infty}^{0}$ and denoted then again the variable by p; further, the Euler formula $e^{iz} + e^{-iz} = 2\cos z$ has been applied, and finally formula (9) in § 13.10.) Thus, by the Convolution Theorem 28.3.7, we obtain the general formula for the solution to the problem (5), (4) in the form

$$u(x, y) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{y}{(x-v)^2 + y^2} g(v)\, dv.$$

28.3. Some Results of Fundamental Importance. Tables

From the point of view of applications of transform methods in practice and even from the theoretical point of view the following two problems are of basic importance:

1. to decide whether or not a given function $F(p)$ is the image of some function $f(t)$ (under the transform considered);
2. to find the original for a given image.

Let us present, first, some typical properties of images:

Laplace transform: $F(p)$ is a holomorphic function in the half-plane $\mathrm{Re}\, p > \sigma$. (We use the notation from § 28.1; if $\sigma = -\infty$, $F(p)$ is holomorphic in the entire plane.)

The same is true for the Laplace-Carson transform.

In the case of the bilateral Laplace transform the image $F(p)$ is holomorphic in the strip $\sigma_1 < \mathrm{Re}\, p < \sigma_2$.

In the case of the Fourier transform the image $F(p)$ is defined only for real p, and $F(p)$ is a continuous function. (In the case of the cosine and sine transforms p is restricted to have only non-negative values.)

The images of Mellin transforms have the same properties as those given above for the bilateral Laplace transform.

In the case of the Hankel transform the image $F(p)$ is defined only for non-negative p.

The problems 1 and 2 just stated were unsolved until recently. Let us start with some results for the Laplace transform.

Theorem 1. *The condition*

$$\sup_{\delta>\gamma} \int_{-\infty}^{\infty} |F(\delta + i\tau)|^2 \, d\tau < \infty$$

is necessary and sufficient for a function $F(p)$, holomorphic in the half-plane $\operatorname{Re} p > \gamma$, to be the Laplace image of a function $f(t)$ satisfying the inequality

$$\int_0^{\infty} |f(t)|^2 e^{-2\gamma t} \, dt < \infty.$$

Theorem 2. *If $F(p)$ is the Laplace image of an original $f(t)$, then*

$$f(t) = \frac{d}{dt} \lim_{\omega \to +\infty} \frac{1}{2\pi i} \int_{\gamma - i\omega}^{\gamma + i\omega} F(p) \frac{e^{pt}}{p} dp. \tag{1}$$

(More accurately, equation (1) holds for almost every t in the interval $[0, \infty)$, i.e. with the possible exception of points which constitute a set of measure zero.) *The integration is performed along a straight line $\operatorname{Re} p = \gamma$ with $\gamma > 0$, which lies in the domain of definition of the function $F(p)$.* (Thus, it suffices to take γ sufficiently large.)

In (1) the differentiation may sometimes be performed under the integral sign. Sufficient conditions:

Theorem 3. *Let $\int_0^{\infty} |f(t)| e^{-\alpha t} \, dt < \infty$ for $\alpha \geqq \gamma$. Let $f(t)$ be a function of bounded variation in a neighbourhood of a point t $(t \geqq 0)$. Then we have*

$$\lim_{\omega \to +\infty} \frac{1}{2\pi i} \int_{\alpha - i\omega}^{\alpha + i\omega} e^{pt} F(p) \, dp = \begin{cases} \dfrac{f(t+0) + f(t-0)}{2} & \text{for } t > 0, \\ \dfrac{f(+0)}{2} & \text{for } t = 0, \\ 0 & \text{for } t < 0. \end{cases}$$

Here, $f(t \pm 0)$ denotes the limit from the right and the left, respectively, of the function $f(t)$ at the point t. If $f(t)$ is continuous at the point t, then

$$\lim_{\omega \to +\infty} \frac{1}{2\pi i} \int_{\alpha - i\omega}^{\alpha + i\omega} e^{pt} F(p) \, dp = f(t).$$

In practice the originals are calculated by Theorem 3 from the integral

$$\frac{1}{2\pi i} \int_{\alpha - i\omega}^{\alpha + i\omega} e^{pt} F(p) \, dp.$$

A formal application of the inverse transform may lead to wrong results (see e.g. [118], p. 193).

For calculation of the original from a given image the Residue Theorem (Theorem 20.5.1) is often used:

Theorem 4. *Let $F(p)$ be the Laplace image of a function $f(t)$ and let*

$$f(t) = \lim_{\omega \to +\infty} \frac{1}{2\pi i} \int_{\gamma - i\omega}^{\gamma + i\omega} e^{pt} F(p)\, dp.$$

Let the function $F(p)$ be holomorphic in the complex plane except for poles p_1, p_2, ... (e.g., let $F(p)$ be a rational function) *interior to the half-plane $\operatorname{Re} p < \gamma$. Let a sequence of arcs C_n (which do not pass through the poles) be such that each C_n meets the straight line $\operatorname{Re} p = \gamma$ at two points $\gamma + i\beta_n$ and $\gamma - i\beta_n$ and lies in the half-plane $\operatorname{Re} p \leqq \gamma$ while each arc C_n together with the line segment with end points $\gamma - i\beta_n$ and $\gamma + i\beta_n$ form the boundary of a region Ω_n which contains exactly the poles $p_1, p_2, \ldots, p_n$* (see Fig. 28.1). *Further, let $\beta_n \to +\infty$ as $n \to \infty$ and*

$$\lim_{n \to \infty} \int_{C_n} F(p)\, e^{pt}\, dp = 0.$$

Then

$$f(t) = \sum_{n=1}^{\infty} \operatorname*{res}_{p=p_n} \left[F(p)\, e^{pt} \right], \tag{2}$$

where $\operatorname{res}_{p=p_n} [F(p)\, e^{pt}]$ denotes the residue of the function $F(p)\, e^{pt}$ at the pole $p = p_n$.

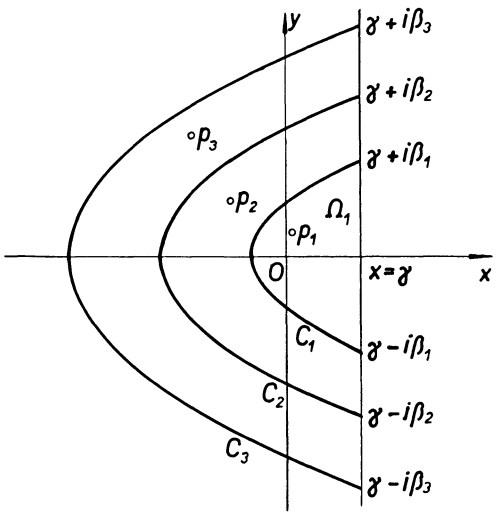


Fig. 28.1.

With the aid of this and similar theorems the originals corresponding to various images can be established. In particular, the calculation of the original of every rational function can be reduced to the application of formula (2) (with a finite number of terms on its right-hand side).

Example 1. Let us find the original of the function $\omega/(p^2+\omega^2)$.

Using (2) and formulae for the calculation of residues, we get

$$f(t) = \operatorname*{res}_{p=i\omega}\left[\frac{\omega\, e^{pt}}{p^2+\omega^2}\right] + \operatorname*{res}_{p=-i\omega}\left[\frac{\omega\, e^{pt}}{p^2+\omega^2}\right] = \frac{e^{i\omega t}}{2i} - \frac{e^{-i\omega t}}{2i} = \sin\omega t,$$

in accordance with Tab. 28.2.

Extensive tables of transform pairs are given, e.g., in [118]. In such tables the image $F(p)$ is always given first (see Tab. 28.2).

To tables there is usually attached the so-called *grammar*, which summarizes the basic rules governing relationship between originals and images. In Tab. 28.3 below a sample of a grammar for Laplace transform is given.

TABLE 28.3

Image	Original
$\frac{1}{\alpha}F(\frac{p}{\alpha})$	$f(\alpha t)$
$pF(p) - f(0)$	$f'(t)$
$p^2F(p) - pf(0) - f'(0)$	$f''(t)$
$\frac{F(p)}{p}$	$\int_0^t f(\tau)\,d\tau$
$\int_p^\infty F(r)\,dr$	$\frac{f(t)}{t}$
$F(p-p_0)$	$e^{p_0 t} f(t)$
$F(p)G(p)$	$\int_0^t f(\tau)\, g(t-\tau)\,d\tau$

Sufficient conditions for validity of the inversion formula for the Fourier transform are stated in the following assertion:

Theorem 5. *Let $f(t)$ be the original of a Fourier image $F(p)$. If $f(t)$ has bounded variation in a neighbourhood of a point t, then we have*

$$\frac{f(t+0)+f(t-0)}{2} = \lim_{\omega\to+\infty}\frac{1}{2\pi}\int_{-\omega}^{\omega} F(p)\, e^{ipt}\,dp.$$

TABLE 28.2

$F(p) = \int_0^\infty f(t)\,\mathrm{e}^{-pt}\,\mathrm{d}t$	Original $f(t)$
$\frac{1}{p}$	1
$\frac{1}{p^2}$	t
$\frac{1}{p^{n+1}}$	$\frac{t^n}{n!}$, n a nonnegative integer
$\frac{1}{\sqrt{p}}$	$\frac{1}{\sqrt{(\pi t)}}$
$\frac{1}{p^{\nu+1}}$	$\frac{t^\nu}{\Gamma(\nu+1)}$, $\nu > -1$; for the Γ function see § 13.11
$\frac{1}{p+a}$	e^{-at}
$\frac{1}{(p+a)^2}$	$t\,\mathrm{e}^{-at}$
$\frac{\omega}{p^2+\omega^2}$	$\sin \omega t$
$\frac{p}{p^2+\omega^2}$	$\cos \omega t$
$\frac{a}{p^2-a^2}$	$\sinh at$
$\frac{p}{p^2-a^2}$	$\cosh at$

If, in addition, $f(t)$ is continuous at the point t, then

$$\lim_{\omega \to +\infty} \frac{1}{2\pi} \int_{-\omega}^{\omega} F(p)\,\mathrm{e}^{\mathrm{i}pt}\,\mathrm{d}p = f(t).$$

If $f(t)$ is the original for the bilateral Laplace transform with image denoted by $F(p)$ and if

$$\int_{-\infty}^{\infty} \left| f(t)\,\mathrm{e}^{-p_0 t} \right| \mathrm{d}t < \infty$$

for some $p_0 = x_0 + \mathrm{i}y_0$, then the function $f(t)\,\mathrm{e}^{-x_0 t}$ is the original for the Fourier transform whose Fourier image is $F(x_0 + \mathrm{i}y)$.

TABLE 28.2 (continued)

$F(p) = \int_0^\infty f(t)\,\mathrm{e}^{-pt}\,\mathrm{d}t$	Original $f(t)$
$\dfrac{\mathrm{e}^{-ap}}{p}$	$\begin{cases} 0 & \text{for } t < a \\ 1 & \text{for } t \geqq a,\ a \geqq 0 \end{cases}$
$\dfrac{\mathrm{e}^{-ap}}{\sqrt{p}}$	$\begin{cases} 0 & \text{for } t < a \\ \dfrac{1}{\sqrt{[\pi(t-a)]}} & \text{for } t \geqq a,\ a \geqq 0 \end{cases}$
$\dfrac{\mathrm{e}^{-\frac{a}{p}}}{p}$	$\mathrm{J}_0\left(2\sqrt{(at)}\right)$
	for Bessel functions see § 16.4
$\dfrac{\mathrm{e}^{-\frac{a}{p}}}{\sqrt{p}}$	$\dfrac{1}{\sqrt{(\pi t)}}\cos\left(2\sqrt{(at)}\right)$
$\dfrac{\mathrm{e}^{-a\sqrt{p}}}{p}$	$\operatorname{erfc}\dfrac{a}{2\sqrt{t}} = 1 - \operatorname{erf}\dfrac{a}{2\sqrt{t}} =$
	$= \dfrac{2}{\sqrt{\pi}}\int_{a/(2\sqrt{t})}^{\infty} \mathrm{e}^{-u^2}\,\mathrm{d}u =$
	$= 1 - \dfrac{2}{\sqrt{\pi}}\int_0^{a/(2\sqrt{t})} \mathrm{e}^{-u^2}\,\mathrm{d}u, \quad a \geqq 0$
$\dfrac{\cos\frac{1}{p}}{\sqrt{p}}$	$\dfrac{\cos\sqrt{(2t)}\,\cosh\sqrt{(2t)}}{\sqrt{(\pi t)}}$
$\dfrac{\sin\frac{1}{p}}{\sqrt{p}}$	$\dfrac{\sin\sqrt{(2t)}\,\sinh\sqrt{(2t)}}{\sqrt{(\pi t)}}$
$\ln\dfrac{p+a}{p}$	$\dfrac{1-\mathrm{e}^{-at}}{t}$

If $f(t)$ is an original for the Mellin transform with the corresponding image denoted by $F(p)$ and if further

$$\int_0^\infty \left|f(t)t^{p_0-1}\right|\ dt < \infty$$

for some $p_0 = x_0 + \mathrm{i}y_0$, then $f(\mathrm{e}^{-u})\,\mathrm{e}^{-x_0 u}$ is the original for the Fourier transform with the Fourier image $F(x_0 + \mathrm{i}y)$. (The substitution $t = \mathrm{e}^{-u}$ has been used.)

We will add some elementary properties of the *Fourier transform* (integrability of the functions in question is always assumed):

Theorem 6. *If*

$$\int_{-\infty}^{\infty} |f(x)|\,\mathrm{d}x < \infty,$$

then for

$$F(p) = \int_{-\infty}^{\infty} f(t)\,\mathrm{e}^{-\mathrm{i}pt}\,\mathrm{d}t$$

the "Riemann-Lebesgue lemma"

$$\lim_{|p|\to\infty} F(p) = 0$$

is valid.

Theorem 7 (*Fourier Transform to the Convolution*). *Let*

$$\int_{-\infty}^{\infty} |f(t)|\,\mathrm{d}t < \infty, \quad \int_{-\infty}^{\infty} |g(t)|\,\mathrm{d}t < \infty.$$

Then the function

$$h(t) = \int_{-\infty}^{\infty} f(t-u)\,g(u)\,\mathrm{d}u = \int_{-\infty}^{\infty} f(t)\,g(t-u)\,\mathrm{d}u$$

is absolutely integrable as well, i.e.

$$\int_{-\infty}^{\infty} |h(t)|\,\mathrm{d}t < \infty,$$

and

$$H(p) = F(p)G(p).$$

Theorem 8. *Let*

$$\int_{-\infty}^{\infty} |f(t)|\,\mathrm{d}t < \infty, \quad \int_{-\infty}^{\infty} |f'(t)|\,\mathrm{d}t < \infty.$$

Then

$$\int_{-\infty}^{\infty} f'(t)\,\mathrm{e}^{-\mathrm{i}pt}\,\mathrm{d}t = \mathrm{i}p\int_{-\infty}^{\infty} f(t)\,\mathrm{e}^{-\mathrm{i}pt}\,\mathrm{d}t = \mathrm{i}pF(p).$$

(Cf. (28.1.11).)

Theorem 9. *If*

$$\int_{-\infty}^{\infty} |f(t)|\,\mathrm{d}t < \infty, \quad \int_{-\infty}^{\infty} |F(p)|\,\mathrm{d}p < \infty,$$

then we have (for almost all t)

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(p)\, \mathrm{e}^{\mathrm{i}pt}\, \mathrm{d}p.$$

Theorem 10. *If*

$$\int_{-\infty}^{\infty} |f(t)|\, \mathrm{d}t < \infty \quad \text{and} \quad |f(t)| \leqq M$$

(f is bounded for (almost) all t), then

$$\int_{-\infty}^{\infty} |f(t)|^2\, \mathrm{d}t = \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(p)|^2\, \mathrm{d}p.$$

We shall finish with a short table of Fourier cosine and Fourier sine transforms, see Tab. 28.4 and Tab. 28.5 below. In practice, we meet these tables much more often than those for the Fourier transform, obviously because of the following facts:

Owing to the well-known Euler relation $\mathrm{e}^{-\mathrm{i}z} = \cos z - \mathrm{i} \sin z$, we get, for an even function (thus satisfying $f(-t) = f(t)$ for (almost) all t)

$$\int_{-\infty}^{\infty} f(t)\, \mathrm{e}^{-\mathrm{i}pt}\, \mathrm{d}t = 2 \int_{0}^{\infty} f(t) \cos pt\, \mathrm{d}t$$

and for an odd function ($f(-t) = -f(t)$)

$$\int_{-\infty}^{\infty} f(t)\, \mathrm{e}^{-\mathrm{i}pt}\, \mathrm{d}t = -2\mathrm{i} \int_{0}^{\infty} f(t) \sin pt\, \mathrm{d}t.$$

Moreover, every $f(t)$ may be (uniquely) decomposed into an even and an odd function:

$$f(t) = \frac{1}{2}[f(t) + f(-t)] + \frac{1}{2}[f(t) - f(-t)].$$

28.4. Two-Dimensional and Multidimensional Transforms

Definition 1. The *two-dimensional Laplace transform* assigns to an original $f(x, y)$, which is absolutely integrable in every rectangle $0 \leqq t_1 \leqq a < \infty$, $0 \leqq \leqq t_2 \leqq b < \infty$, the image

$$F(u, v) = \int_{0}^{\infty} \int_{0}^{\infty} f(t_1, t_2)\, \mathrm{e}^{-ut_1 - vt_2}\, \mathrm{d}t_1\, \mathrm{d}t_2,$$

where u, v are complex numbers.

TABLE 28.4

Original $f(t)$	$F(p) = \int_0^\infty f(t) \cos pt \, dt$
$f(at), \quad a > 0$	$\frac{1}{a} F\left(\frac{p}{a}\right)$
$\begin{cases} 1 & \text{for } 0 < t \leqq a \\ 0 & \text{for } t > a \end{cases}$	$\frac{\sin ap}{p}$
$\frac{1}{\sqrt{t}}$	$\sqrt{\left(\frac{\pi}{2}\right)} \frac{1}{\sqrt{p}}$
$\frac{1}{t^\nu}, \quad 0 < \nu < 1$	$\sin \frac{\pi\nu}{2} \Gamma(1-\nu) p^{\nu-1};$ for the function Γ see § 13.11
$e^{-at}, \quad a > 0$	$\frac{a}{a^2 + p^2}$
$\frac{e^{-at}}{t^{1-\nu}}, \quad a > 0, \quad \nu > 0$	$\Gamma(\nu)(a^2 + p^2)^{-\nu/2} \cos\left(\nu \arctan \frac{p}{a}\right)$
$e^{-at^2}, \quad a > 0$	$\frac{1}{2} \sqrt{\left(\frac{\pi}{a}\right)} e^{-\frac{p^2}{4a}}$
$\frac{\sin at}{t}, \quad a > 0$	$\begin{cases} \frac{\pi}{2} & \text{for } p < a \\ \frac{\pi}{4} & \text{for } p = a \\ 0 & \text{for } p > a \end{cases}$
$\frac{\sin at}{t^{1-\nu}}, \quad a > 0, \quad -1 < \nu < 1$	$\frac{\pi}{4} \left[\cos \frac{\pi\nu}{2} \Gamma(1-\nu)\right]^{-1} \cdot$ $\cdot \left[(p+a)^{-\nu} - \operatorname{sign}(p-a)\lvert p-a\rvert^{-\nu}\right]$
$e^{-bt} \sin at, \quad a > 0, \quad b > 0$	$\frac{a+p}{2}\left[b^2 + (a+p)^2\right]^{-1} + \frac{a-p}{2}\left[b^2 + (a-p)^2\right]^{-1}$
$e^{-bt} \cos at, \quad a > 0, \quad b > 0$	$\frac{b}{2}\left\{\left[b^2 + (a-p)^2\right]^{-1} + \left[b^2 + (a+p)^2\right]^{-1}\right\}$
$\ln\left(1 + \frac{a^2}{t^2}\right)$	$\frac{\pi}{p}(1 - e^{-ap})$
$\frac{\sinh at}{\cosh bt}, \quad 0 < a < b$	$\frac{\pi}{2b} \frac{\sin \frac{\pi a}{b}}{\cos \frac{\pi a}{b} + \cosh \frac{\pi p}{b}}$

TABLE 28.5

Original $f(t)$	$F(p) = \int_0^\infty f(t) \sin pt \, dt$
$f(at), \quad a > 0$	$\frac{1}{a} F\left(\frac{p}{a}\right)$
$\begin{cases} 1 & \text{for } 0 < t \leqq a \\ 0 & \text{for } t > a \end{cases}$	$\frac{1 - \cos ap}{p}$
$\frac{1}{\sqrt{t}}$	$\sqrt{\left(\frac{\pi}{2}\right)} \frac{1}{\sqrt{p}}$
$\frac{1}{t}$	$\frac{\pi}{2}$
$\frac{1}{t^\nu}, \quad 0 < \nu < 2 \quad \nu \neq 1$	$\cos \frac{\pi\nu}{2} \Gamma(1-\nu) p^{\nu-1}$
$\begin{cases} 0 & \text{for } 0 < t \leqq a \\ (t^2 - a^2)^{-\nu - 1/2} & \text{for } t > a, \quad -\frac{1}{2} < \nu < \frac{1}{2} \end{cases}$	$\frac{\nu\pi}{2} \Gamma\left(\frac{1}{2} - \nu\right) \left(\frac{p}{2a}\right)^\nu J_\nu(ap)$
$\frac{1}{t(a^2 + t^2)}, \quad a > 0$	$\frac{\pi}{2a^2}(1 - e^{-ap})$
$(a + it)^\nu - (a - it)^\nu, \quad a > 0, \quad \nu < 0$	$\pi i \left[\Gamma(-\nu)\right]^{-1} p^{-\nu-1} e^{-ap}$
$e^{-at}, \quad a > 0$	$\frac{p}{a^2 + p^2}$
$\frac{e^{-at}}{\sqrt{t}}, \quad a > 0$	$\sqrt{\left(\frac{\pi}{2}\right)} \frac{\sqrt{\left[\sqrt{(a^2 + p^2)} - a\right]}}{\sqrt{(a^2 + p^2)}}$
$\frac{e^{-at}}{t^{1-\nu}}, \quad a > 0, \quad \nu > 0$	$\frac{\Gamma(\nu)}{(a^2 + p^2)^{\nu/2}} \sin\left(\nu \arctan \frac{p}{a}\right)$
$\frac{\cos at}{t}, \quad a > 0$	$\begin{cases} \frac{\pi}{2} & \text{for } p > a \\ \frac{\pi}{4} & \text{for } p = a \\ 0 & \text{for } p < a \end{cases}$
$\frac{\cos at}{t^{1-\nu}}, \quad a > 0, \quad -1 < \nu < 1, \quad \nu \neq 0$	$\frac{\Gamma(\nu)}{2} \sin \frac{\pi\nu}{2} \left[\frac{1}{(p+a)^\nu} + \frac{\text{sign}(p-a)}{\lvert p - a \rvert^\nu}\right]$
$\begin{cases} \arcsin t & \text{for } 0 < t \leqq 1 \\ 0 & \text{for } t > 1 \end{cases}$	$\frac{\pi}{2p}\left[J_0(p) - \cos p\right]$

A more detailed treatment of two-dimensional Laplace transform may be found in [478].

Definition 2. The *n-dimensional Fourier transform* assigns to an original $f(t_1, t_2, \ldots, t_n)$ satisfying

$$\underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{n\text{-tuple}}|f(t_1, \ldots, t_n)|\mathrm{d}t_1 \ldots \mathrm{d}t_n < +\infty$$

the image

$$F(p_1, \ldots, p_n) = \underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{n\text{-tuple}} f(t_1, \ldots, t_n)\, \mathrm{e}^{-\mathrm{i}(p_1 t_1 + \ldots + p_n t_n)}\, \mathrm{d}t_1 \ldots \mathrm{d}t_n,$$

where $p_1, \ldots, p_n$ are arbitrary real numbers.

28.5. One-Dimensional Finite Transforms

A one-dimensional finite transform of a function $f(t)$ in one variable assigns the Fourier coefficients of $f(t)$ to this function. The application of a one-dimensional finite transform leads to the expansion of a function in Fourier series, the application in partial differential equations to the Fourier method. More about these transforms may be found in [467].

29. APPROXIMATE SOLUTION OF FREDHOLM'S INTEGRAL EQUATIONS

By KAREL REKTORYS

References: [22], [109], [252], [276], [323], [387], [415], [434], [460].

Equations with *degenerate kernels* are solved according to § 19.2. For the *boundary element method* see § 24.7. Further methods can be found e.g. in Baker, C.T.H.: The Numerical Treatment of Integral Equations. Oxford, Clarendon Press 1976.

For the *theory* of integral equations see Chap. 19.

29.1. Successive Approximations (Iterations)

Suppose we have a Fredholm equation (see § 19.1)

$$f(x) - \lambda \int_a^b K(x, s) f(s)\, \mathrm{d}s = g(x). \tag{1}$$

Let $g \in L_2(a, b)$, $K \in L_2(Q)$, where $Q = (a, b) \times (a, b)$. Denote

$$\int_a^b \int_a^b |K(x, s)|^2 \, \mathrm{d}x \, \mathrm{d}s = B^2 \quad (B > 0),$$

$$\int_a^b |g(x)|^2 \, \mathrm{d}x = D^2 \quad (D > 0).$$

Let us construct the sequence of functions

$$\begin{aligned} f_0(x) &= g(x), \\ f_1(x) &= g(x) + \lambda \int_a^b K(x, s) f_0(s)\, \mathrm{d}s, \\ f_2(x) &= g(x) + \lambda \int_a^b K(x, s) f_1(s)\, \mathrm{d}s, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ f_{n+1}(x) &= g(x) + \lambda \int_a^b K(x, s) f_n(s)\, \mathrm{d}s, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{aligned} \tag{2}$$

Theorem 1. *If*

$$\int_a^b |K(x, s)|^2 \, ds \leqq C, \tag{3}$$

where the constant C is the same for all $x \in [a, b]$, and $|\lambda| < 1/B$, then the sequence $f_0(x), f_1(x), f_2(x), \ldots$ converges uniformly in $[a, b]$ to the (unique) solution $f(x)$ of equation (1). (In particular, if the functions $g(x)$ and $K(x, s)$ are continuous in $[a, b]$ and $\overline{Q} = [a, b] \times [a, b]$, respectively, then $f(x)$ is continuous in $[a, b]$.) The absolute value of the difference between $f_n(x)$ and $f(x)$ does not exceed

$$D\sqrt{(C)}\,\frac{B^n|\lambda|^{n+1}}{1 - B|\lambda|}. \tag{4}$$

Example 1 ([323]).

$$f(x) - 0{\cdot}1\int_0^1 K(x, s)f(s)\,ds = 1, \quad K(x, s) = \begin{cases} x & \text{for} \quad 0 \leqq x \leqq s, \\ s & \text{for} \quad s \leqq x \leqq 1. \end{cases} \tag{5}$$

We easily find

$$B = \frac{1}{\sqrt{6}}, \quad C = \frac{1}{3}, \quad D = 1, \quad \lambda = 0{\cdot}1.$$

By (3), the successive approximations are convergent. If we take $n = 2$, then by (4) the error attains at most the value

$$1 \cdot \sqrt{\left(\frac{1}{3}\right)} \cdot \frac{\frac{1}{6} \cdot 0{\cdot}1^3}{1 - (1/\sqrt{(6)}) \cdot 0{\cdot}1} \doteq 0{\cdot}0001. \tag{6}$$

We have (by (2))

$$f_0(x) = 1, \; f_1(x) = 1 + \tfrac{1}{10}x - \tfrac{1}{20}x^2, \; f_2(x) = 1 + \tfrac{31}{300}x - \tfrac{1}{20}x^2 - \tfrac{1}{600}x^3 + \tfrac{1}{2400}x^4.$$

If we take $f(x) \approx f_2(x)$, then by (6) the error in the whole interval [0, 1] does not exceed 0·0001.

REMARK 1. For the *resolvent* of an integral equation see § 19.4. Cf. also Example 19.4.2.

29.2. Approximate Solution of Integral Equations Making Use of Quadrature Formulae

Using numerical integration (§ 13.14), we replace an integral by a sum:

$$\int_a^b h(x)\,dx \approx (b - a)\sum_{k=1}^{n} C_k h(x_k).$$

Here, x_k and C_k are points and constants, respectively, defined by the given quadrature formula. For example, if we apply to the evaluation of the integral

$$\int_0^1 h(x)\,\mathrm{d}x$$

the trapezoidal rule for three points ($n = 3$), then

$$\begin{aligned} &x_1 = 0, \quad x_2 = \tfrac{1}{2}, \quad x_3 = 1, \\ &C_1 = \tfrac{1}{4}, \quad C_2 = \tfrac{1}{2}, \quad C_3 = \tfrac{1}{4}. \end{aligned} \tag{1}$$

In the given integral equation

$$f(x) - \int_a^b K(x, s) f(s)\,\mathrm{d}s = g(x) \tag{2}$$

let us replace the integral by a sum, choosing a certain quadrature formula with dividing points $s_1, s_2, \ldots, s_n$. Let the same partition be chosen in the interval $[a, b]$ for the variable x. Denote the corresponding dividing points by $x_1, x_2, \ldots, x_n$. Let us write down equation (2), for each x_k, replacing the integral by the corresponding sum. We obtain

$$\begin{aligned} &f(x_1) \approx g(x_1) + (b-a)\sum_{k=1}^{n} C_k K(x_1, s_k) f(s_k), \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &f(x_n) \approx g(x_n) + (b-a)\sum_{k=1}^{n} C_k K(x_n, s_k) f(s_k). \end{aligned} \tag{3}$$

The partition of the interval $[a, b]$ being the same for s as for x, we have $f(x_1) = f(s_1)$, etc. If we write in (3) the equality sign instead of $\approx$, we obtain a system of n equations for n unknowns $f(x_1), f(x_2), \ldots, f(x_n)$ (see Example 1 below).

REMARK 1. Since the sign $\approx$ in the system (3) has been replaced by the equality sign, the resulting values $f(x_1), f(x_2), \ldots, f(x_n)$ are not the exact values of the unknown function $f(x)$ at the points $x_1, x_2, \ldots, x_n$. Thus let us denote them by $f_n(x_1), f_n(x_2), \ldots, f_n(x_n)$ (specifying thus their dependence on the number n of dividing points of the chosen quadrature) and construct the function

$$f_n(x) = g(x) + (b-a)\sum_{k=1}^{n} C_k K(x, s_k) f_n(s_k)$$

which represents an approximation of the required solution $f(x)$ in the interval $[a, b]$. In particular, this function attains, for $x = x_1, x = x_2, \ldots, x = x_n$, the values $f_n(x_1), f_n(x_2), \ldots, f_n(x_n)$, respectively, obtained by solution of the system

(3) (with $=$ written instead of $\approx$). Now, it can be shown that the sequence of functions $f_n(x)$ tends, with increasing n, to the exact solution $f(x)$ of the given equation (2) (under certain assumptions on the smoothness of the kernel and of the function g), provided there exists exactly one solution of this equation. For details see e.g. [252], where an estimate of the error may also be found.

Example 1. Let us consider the equation

$$f(x) - \int_0^1 (x - s) f(s)\,\mathrm{d}s = x^2. \tag{4}$$

Theorem 29.1.1 implies that equation (4) is (uniquely) solvable, since the kernel is bounded and

$$\lambda = 1, \quad B^2 = \int_0^1 \int_0^1 (x - s)^2\,\mathrm{d}x\,\mathrm{d}s < 1, \quad \text{hence} \quad \lambda < 1/B.$$

Let us choose the above-mentioned numerical quadrature (1). For $x_1 = 0$, $x_2 = \frac{1}{2}$, $x_3 = 1$ we write down approximate equations which arise from equation (4) if we replace the integral by a sum:

$$f(0) - \int_0^1 (-s) f(s)\,\mathrm{d}s = 0, \quad f\left(\tfrac{1}{2}\right) - \int_0^1 \left(\tfrac{1}{2} - s\right) f(s)\,\mathrm{d}s = \tfrac{1}{4},$$
$$f(1) - \int_0^1 (1 - s) f(s)\,\mathrm{d}s = 1.$$

Hence, for the values of the approximate solution at the points 0, $\frac{1}{2}$, 1, the relations

$$\begin{aligned} f(0) &- \left[\tfrac{1}{4} \cdot 0 \cdot f(0) + \tfrac{1}{2} \cdot \left(-\tfrac{1}{2}\right) \cdot f\left(\tfrac{1}{2}\right) + \tfrac{1}{4} \cdot (-1) \cdot f(1)\right] = 0, \\ f\left(\tfrac{1}{2}\right) &- \left[\tfrac{1}{4} \cdot \tfrac{1}{2} \cdot f(0) + \tfrac{1}{2} \cdot 0 \cdot f\left(\tfrac{1}{2}\right) + \tfrac{1}{4} \cdot \left(-\tfrac{1}{2}\right) \cdot f(1)\right] = \tfrac{1}{4}, \\ f(1) &- \left[\tfrac{1}{4} \cdot 1 \cdot f(0) + \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot f\left(\tfrac{1}{2}\right) + \tfrac{1}{4} \cdot 0 \cdot f(1)\right] = 1. \end{aligned} \tag{5}$$

hold. This is the system (3). After simplification we obtain

$$\begin{gathered} f(0) + \tfrac{1}{4} f\left(\tfrac{1}{2}\right) + \tfrac{1}{4} f(1) = 0, \quad -\tfrac{1}{8} f(0) + f\left(\tfrac{1}{2}\right) + \tfrac{1}{8} f(1) = \tfrac{1}{4}, \\ -\tfrac{1}{4} f(0) - \tfrac{1}{4} f\left(\tfrac{1}{2}\right) + f(1) = 1. \end{gathered} \tag{6}$$

The solution is

$$f(0) = -\tfrac{19}{72}, \quad f\left(\tfrac{1}{2}\right) = \tfrac{7}{72}, \quad f(1) = \tfrac{69}{72}. \tag{7}$$

This example is an illustrative one only, for equation (4) may obviously be solved as an equation with a degenerate kernel (§ 19.2). The exact solution is $f(x) = x^2 + \frac{3}{13}x - \frac{17}{78}$. The values at the points $x = 0$, $x = \frac{1}{2}$, $x = 1$ are

$$f(0) = -\tfrac{17}{78}, \quad f\left(\tfrac{1}{2}\right) = \tfrac{11 \cdot 5}{78}, \quad f(1) = \tfrac{79}{78}. \tag{8}$$

The difference between (7) and (8) is rather high, for we have made use of a very rough quadrature. The result may be considerably improved by choosing a more precise formula for numerical integration (see § 13.13) and a finer partition of the given interval.

The above method may also be applied to the approximate evaluation of eigenvalues.

29.3. Replacement of the Kernel by a Degenerate Kernel

We replace the given kernel $K(x, s)$ by a "close" degenerate kernel $k(x, s)$ and solve the equation with this kernel. For example, the relation

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

holds. Hence we replace the equation

$$f(x) - \int_0^{\pi/2} \sin xs \, f(s)\, \mathrm{d}s = g(x)$$

by the following equation with degenerate kernel:

$$\varphi(x) - \int_0^{\pi/2} \left(xs - \frac{x^3 s^3}{3!} + \frac{x^5 s^5}{5!} \right) \varphi(s)\, \mathrm{d}s = g(x).$$

If the given equation is uniquely solvable and if $k(x, s)$ is sufficiently close to $K(x, s)$, then the corresponding equation with the degenerate kernel $k(x, s)$ is also uniquely solvable and the solution $\varphi(x)$ of the new equation is sufficiently close to the solution $f(x)$ of the original equation. For an exact theorem and an estimate of error see, for example [252].

29.4. The Galerkin Method (Method of Moments) and the Ritz Method

Let $\varphi_1(x), \varphi_2(x), \dots$ constitute a complete (not necessarily orthogonal) system of (linearly independent) functions in $L_2(a, b)$ (see Remarks 16.2.14, 16.2.15). Let us look for the approximate solution of the equation

$$f(x) - \int_a^b K(x, s) f(s)\, \mathrm{d}s = g(x) \tag{1}$$

in the form

$$f_n(x) = g(x) + \sum_{i=1}^{n} c_i \varphi_i(x), \tag{2}$$

where the coefficients c_i are defined by the conditions

$$\begin{aligned} &\int_a^b \left[f_n(x) - \int_a^b K(x, s) f_n(s)\, \mathrm{d}s - g(x) \right] \varphi_1(x)\, \mathrm{d}s = 0, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &\int_a^b \left[f_n(x) - \int_a^b K(x, s) f_n(s)\, \mathrm{d}s - g(x) \right] \varphi_n(x)\, \mathrm{d}s = 0. \end{aligned} \tag{3}$$

The expression in square brackets is thus orthogonal to the functions $\varphi_1(x)$, $\varphi_2(x), \ldots, \varphi_n(x)$. System (3) is a system of n linear equations for n unknowns $c_1, c_2, \ldots, c_n$. Solving this system, we obtain an approximate solution. If the system of functions $\varphi_1(x), \varphi_2(x), \ldots$ is orthonormal (Definition 16.2.4), then it can be proved (see [252]) that $f_n(x)$ coincides with the approximate solution obtained by the method of § 29.3, the kernel $K(x, s)$ being replaced by the kernel

$$K_n(x, s) = \sum_{i=1}^{n} u_i(s) \varphi_i(x), \quad \text{where} \quad u_i(s) = \int_a^b K(t, s) \varphi_i(t)\, \mathrm{d}t.$$

According to § 29.3 this fact may be used to prove convergence. In particular, if equation (1) is uniquely solvable and if $K_n(x, s)$ tends uniformly in the square $a \leqq x \leqq b$, $a \leqq s \leqq b$ to $K(x, s)$, the sequence of functions $f_n(x)$ tends uniformly in $[a, b]$ to $f(x)$.

System (3) is obtained also if the *Ritz method* is applied to the approximate solution of integral equations with *symmetric* kernels. Let $K(x, s) = K(s, x)$ in (1) (we suppose that the kernel is real) and let the approximate solution be assumed in the form (2) again, where, now, the constants c_i are such that, n being fixed, the function (2) gives minimal value to the functional

$$I(\varphi) \equiv \int_a^b \varphi^2(x)\, \mathrm{d}x - \int_a^b \int_a^b K(x, s) \varphi(x) \varphi(s)\, \mathrm{d}x\, \mathrm{d}s - 2 \int_a^b \varphi(x) g(x)\, \mathrm{d}x$$

(for the function minimizing the functional $I(\varphi)$ is the solution of the given integral equation). If we substitute $f_n(x)$ for $\varphi(x)$ (and $f_n(s)$ for $\varphi(s)$), then the functional I becomes a function of $c_1, c_2, \ldots, c_n$. If we set the derivatives

$$\frac{\partial I}{\partial c_1}, \frac{\partial I}{\partial c_2}, \ldots, \frac{\partial I}{\partial c_n}$$

equal to zero (the condition for a minimum), we obtain equations for the unknown constants $c_1, c_2, \ldots, c_n$. These equations are identical with the equations given by the system (3). For more details and for examples see [252].

29.5. Application of the Ritz Method to Approximate Determination of the First Characteristic Value of an Equation with a Symmetric Kernel

Let $K(x, s)$ be a real symmetric kernel. Then, in accordance with Theorem 19.3.3, the relation

$$\left|\frac{1}{\lambda_1}\right| = \max_{\varphi \in L_2(a,b)} \left| \int_a^b \int_a^b K(x, s)\varphi(x)\varphi(s)\,dx\,ds \right| \tag{1}$$

holds, where

$$\int_a^b \varphi^2(x)\,dx = 1. \tag{2}$$

Let $\psi_1(x), \psi_2(x), \ldots$ be a complete sequence of functions in $L_2(a, b)$. Let us put, in (1),

$$\varphi = \sum_{i=1}^{n} a_i \psi_i(x). \tag{3}$$

If we denote

$$\int_a^b \psi_i(x)\psi_k(x)\,dx = (\psi_i, \psi_k),$$

$$\int_a^b \int_a^b K(x, s)\psi_i(x)\psi_k(s)\,dx\,ds = A_{ik} \quad (A_{ik} = A_{ki}),$$

then the problem of finding an approximate value of $|1/\lambda_1|$ leads to the problem of finding the maximum of the expression

$$\left| \sum_{i,k=1}^{n} A_{ik} a_i a_k \right| \tag{4}$$

under the condition

$$\sum_{i,k=1}^{n} a_i a_k (\psi_i, \psi_k) = 1. \tag{5}$$

This problem can then be conveniently solved by the method of Lagrange's multipliers. (Cf. Theorem 12.12.3. For details see for example [323].)

If we determine $a_1, a_2, \ldots, a_n$ and substitute them into (4), we obtain the approximate value of $|1/\lambda_1|$; the process converges as $n \to \infty$ to the exact value $|1/\lambda_1|$. If, moreover, there exists only one characteristic function corresponding to λ_1 (up to a multiplicative constant), then (3), with the above-mentioned constants $a_1, a_2, \ldots, a_n$, is an approximation to this function. For details see [323].

We frequently meet the following cases:

1. The functions $\psi_k(x)$ constitute an orthonormal system. Then the condition (5) reads

$$\sum_{k=1}^{n} a_k^2 = 1.$$

2. The quadratic form

$$\sum_{i,k=1}^{n} A_{ik} a_i a_k \quad (A_{ik} = A_{ki}) \tag{6}$$

is positive definite, hence (6) is everywhere positive (with the exception of the case $a_1 = a_2 = \ldots = a_n = 0$) so that the sign of the absolute value in (4) can be omitted. It can be shown that in this case the maximum of the form (4) is given by the greatest of the roots of the equation

$$\begin{vmatrix} A_{11} - \sigma, & A_{12}, & \ldots, & A_{1n} \\ A_{21}, & A_{22} - \sigma, & \ldots, & A_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ A_{n1}, & A_{n2}, & \ldots, & A_{nn} - \sigma \end{vmatrix} = 0. \tag{7}$$

The form (6) is always (i.e. for each n) positive definite if the kernel $K(x, s)$ is positive, i.e. if for each function $\varphi(x) \not\equiv 0$ the integral

$$\int_a^b \int_a^b K(x, s)\varphi(x)\varphi(s)\,\mathrm{d}x\,\mathrm{d}s$$

is positive. Then all the characteristic functions are positive as well and the evaluated maximum gives directly an approximate value for $1/\lambda_1$.

Example 1. Suppose we have the equation

$$f(x) - \lambda \int_0^1 K(x, s) f(s)\,\mathrm{d}s = 0$$

with the kernel

$$K(x, s) = \begin{cases} \frac{1}{2}x(2 - s) & (x \leqq s), \\ \frac{1}{2}s(2 - x) & (x \geqq s). \end{cases} \tag{8}$$

For the sequence $\psi_k(x)$ we choose the orthonormal sequence

$$\psi_k(x) = \sqrt{(2)} \sin k\pi x\,, \quad k = 1, 2, \ldots\,.$$

If we take $n = 2$, we obtain

$$A_{11} = \frac{2}{\pi^2}, \quad A_{12} = A_{21} = -\frac{1}{2\pi^2}, \quad A_{22} = \frac{1}{2\pi^2}$$

so that the problem (4), (5) is transformed into the problem of finding the maximum of the function (of two variables)

$$\frac{1}{\pi^2}\left(2a_1^2 - a_1a_2 + \frac{1}{2}a_2^2\right) \tag{9}$$

under the condition

$$a_1^2 + a_2^2 = 1.$$

(We need not write the sign of the absolute value in (9), since the expression (9) is always positive, except for $a_1 = 0$ and $a_2 = 0$ simultaneously. The kernel (8) can be shown to be positive.)

Equation (7) becomes

$$\begin{vmatrix} \frac{2}{\pi^2} - \sigma, & -\frac{1}{2\pi^2} \\ -\frac{1}{2\pi^2}, & \frac{1}{2\pi^2} - \sigma \end{vmatrix} = 0.$$

The greater of the roots has the value $\sigma \doteq 0{\cdot}218$. Hence

$$\lambda_1 \approx \frac{1}{0{\cdot}218} \doteq 4{\cdot}59.$$

In [323] a more precise value $\lambda_1 \doteq 4{\cdot}115$ has been determined in a slightly more complicated way. In the same book some other methods of approximate evaluation of characteristic values are discussed.

For the approximate solution of integral equations of the first kind see e.g. [415].

30. NUMERICAL METHODS IN LINEAR ALGEBRA

By JITKA SEGETHOVÁ AND KAREL SEGETH

References: [56], [104], [121], [145], [173], [182], [199], [231], [344], [373], [376], [378], [405], [410], [433], [446], [451], [452], [459], [460], [461], [475], [496], [497], [498].

A. SOLUTION OF SYSTEMS OF LINEAR ALGEBRAIC EQUATIONS

The concepts of system of m linear equations for n unknowns, matrix of the system, right-hand side vector, solution vector and augmented matrix were introduced in § 1.18. In this chapter we will be concerned with the numerical solution of systems with *real coefficients* and *real right-hand sides.* Most methods, however, can be applied to systems with complex coefficients and right-hand sides as well.

The statements on existence and uniqueness of solutions of linear algebraic systems are also presented in § 1.18. In what follows — except for § 30.4 — we will solve nonsingular systems for which $m = n$, i.e. *systems with a nonsingular square matrix of order* n. Such a matrix has nonzero determinant and these systems have one and only one solution (Theorem 1.18.7).

Numerical methods for solving systems can be divided into two large groups, namely the *direct* and the *iterative methods.* Direct methods yield — if all computations were carried out without roundoff — the true solution of the system after a finite number of arithmetic operations that is known in advance. They are discussed in § 30.1 and § 30.5. Iterative methods start with some initial approximation to the solution and construct a sequence of approximations that converges to the true solution. Methods of this kind are presented in § 30.3, § 30.6 and § 30.7. We are concerned with the practical choice of a method suitable for solving a given system in § 30.9. We also refer to the software for solving systems there, since we do not expect the reader to solve systems of more than three equations "by hand" or with the help of pocket calculator. *All examples in Part* A *of Chap.* 30 *are of illustrative nature, and are calculated for systems of at most three equations and without roundoff.*

30.1. Gaussian Elimination and LU Factorization

The system of n linear algebraic equations for n unknowns to be solved is written in the form

$$\sum_{j=1}^{n} a_{ij}x_j = b_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

or in the matrix form (cf. Remark 1.25.3)

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{2}$$

where

$$\boldsymbol{A} = \begin{bmatrix} a_{11}, & a_{12}, & \ldots, & a_{1n} \\ a_{21}, & a_{22}, & \ldots, & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1}, & a_{n2}, & \ldots, & a_{nn} \end{bmatrix} \tag{3}$$

is a nonsingular square matrix of order n and

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

are the column vector of unknowns and the column right-hand side vector.

REMARK 1. According to Theorem 1.18.7, the solution of the system can be expressed by the explicit formula (Cramer's rule)

$$x_j = \frac{D_j}{D}, \quad j = 1, 2, \ldots, n, \tag{4}$$

where $D = \det \boldsymbol{A}$ is the determinant of the matrix (3) and D_j is the determinant of the matrix that results from (3) after replacing its j-th column by the right-hand side column $\boldsymbol{b}$. Determinant is the sum of $n!$ terms, each of which is the product of certain n matrix entries (Definition 1.17.1). Therefore, the computation of the solution of the system (1) from the formula (4) requires, for $n > 3$, so many arithmetic operations that it is impracticable.

Before describing the Gaussian elimination, let us introduce two important concepts (cf. Definition 1.26.3).

Definition 1. The square matrix

$$\begin{bmatrix} a_{11}, & a_{12}, & \dots, & a_{1n} \\ 0, & a_{22}, & \dots, & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0, & 0, & \dots, & a_{nn} \end{bmatrix}$$

having zeros below the main (principal) diagonal is called an *upper triangular matrix.* The square matrix

$$\begin{bmatrix} a_{11}, & 0, & \dots, & 0 \\ a_{21}, & a_{22}, & \dots, & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}, & a_{n2}, & \dots, & a_{nn} \end{bmatrix}$$

having zeros above the diagonal is called a *lower triangular matrix.*

The essence of the *Gaussian elimination* consists in transforming the system (1) into an equivalent system (possessing the same solution) whose matrix is upper triangular. This system can be easily solved. The elimination procedure is the same as in § 1.18. Operations, that we use and that lead to equivalent systems, are multiplying an equation by a nonzero number and adding an equation multiplied by a number to another equation of the system. The same operations with the rows of the augmented matrix correspond to these operations with equations (Theorem 1.18.6).

We start with the first row of the augmented matrix. Assuming $a_{11} \neq 0$, we multiply this row by the number a_{21}/a_{11} and subtract it from the second row. After this modification, the entry in the second row and first column is zero. In turn, we further multiply the first row (for $i = 3, \dots, n$) by the number a_{i1}/a_{11} and subtract it from the i-th row. This step being carried out, the first column of the augmented matrix is zero below the diagonal, i.e. zero except for the entry a_{11}. In the resulting system (the *first derived system*), the first equation did not change while the other equations now are

$$\sum_{j=2}^{n} a_{ij}^{(1)} x_j = b_i^{(1)}, \quad i = 2, \dots, n.$$

The unknown x_1 is eliminated from these equations.

Further steps similarly use the k-th row to transform the augmented matrix successively into the form where the k-th column is zero below the diagonal ($k = 2, 3, \dots, n-1$). The unknowns $x_2, x_3, \dots, x_{n-1}$ are thus eliminated step by step and we obtain the *second, third,* and finally the $(n-1)$*-st derived system.* The result

is the equivalent system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1,n-1}x_{n-1} + a_{1n}x_n &= b_1\,, \\ a_{22}^{(1)}x_2 + \ldots + a_{2,n-1}^{(1)}x_{n-1} + a_{2n}^{(1)}x_n &= b_2^{(1)}\,, \\ \ddots \quad \vdots \qquad\qquad \vdots \qquad\quad \vdots & \\ a_{n-1,n-1}^{(n-2)}x_{n-1} + a_{n-1,n}^{(n-2)}x_n &= b_{n-1}^{(n-2)}\,, \\ a_{nn}^{(n-1)}x_n &= b_n^{(n-1)} \end{aligned} \tag{5}$$

with an upper triangular matrix that has the same solution as the system (1). This solution is now calculated by the *backsubstitution procedure*: We compute x_n from the last equation of (5), substitute its value in the last but one equation and compute x_{n-1} from it, etc. Finally, we substitute the values of the unknowns $x_n, x_{n-1}, \ldots, x_2$ having already been found in the first equation and compute x_1 from it.

To solve the system requires of order n^3 arithmetic operations.

REMARK 2. Obviously, the algorithm fails if some of the entries $a_{11}, a_{22}^{(1)}, \ldots$ $\ldots, a_{n-1,n-1}^{(n-2)}$ appearing in the denominator is zero. If $a_{kk}^{(k-1)} = 0$ it is sufficient to find such a row between the $(k+1)$-st and the n-th row of the augmented matrix of the $(k-1)$-st derived system, which has a nonzero entry in the k-th column, and then interchange this row and the k-th row. Such a row has to exist provided that the original matrix $\boldsymbol{A}$ is nonsingular.

In practical computation, we must take into account also the influence of roundoff that will be discussed later in § 30.3. A detailed analysis shows that, when dividing by the number $a_{kk}^{(k-1)}$ in the k-th row, it is advantageous that this number be as large as possible. Now we can employ the possibility to interchange equations of the system (rows of the augmented matrix) and take for $a_{kk}^{(k-1)}$ such an entry in the "rest" of the k-th column which has maximal magnitude. This entry is called a *pivot*. In the k-th step of the elimination we thus first find an index p such that

$$|a_{pk}^{(k-1)}| = \max_{i=k,\ldots,n} |a_{ik}^{(k-1)}|\,.$$

The entry $a_{pk}^{(k-1)}$ is the pivot and we interchange the p-th and the k-th row. Then we continue with eliminating the unknown x_k. The algorithm modified in this way is called the *Gaussian elimination with (partial) pivoting*. Notice that the algorithm cannot fail for a nonsingular matrix $\boldsymbol{A}$ (cf. Remark 2) since a nonzero pivot can be found in each step.

REMARK 3. *Complete pivoting* is carried out in the k-th step in such a way that we find indices p and q for which

$$|a_{pq}^{(k-1)}| = \max_{i,j=k,\ldots,n} |a_{ij}^{(k-1)}|.$$

The entry $a_{pq}^{(k-1)}$ is now the pivot and we interchange not only the p-th and the k-th row but also the q-th and the k-th column. The complete pivoting requires more operations than the partial pivoting. The interchange of columns of the matrix of the system represents also the interchange (renumbering) of the corresponding unknowns and this must be taken into account in the elimination procedure (cf. Remark 1.18.1). Therefore, the complete pivoting is used rather rarely.

Definition 2. A real symmetric matrix $\boldsymbol{A}$ is called *positive definite* (cf. Remark 1.29.3) if $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} > 0$ for any nonzero column vector $\boldsymbol{x}$. ($\boldsymbol{x}^{\mathrm{T}}$ is the transpose of $\boldsymbol{x}$, i.e. the row vector $\boldsymbol{x}^{\mathrm{T}} = (x_1, \ldots, x_n)$, cf. Definition 1.16.3. The notation $\boldsymbol{x}'$ is also used.)

REMARK 4. Pivoting is not necessary for suppressing the roundoff error in the Gaussian elimination if the matrix of the system is positive definite. For such a matrix, the algorithm is even feasible without pivoting (cf. Remark 2).

Example 1. Solve the system

$$\begin{aligned} 9x_1 + 3x_2 &= 15, \\ -3x_1 - 3x_2 + 6x_3 &= 3, \\ 6x_1 + 8x_2 + 3x_3 &= 7 \end{aligned} \tag{6}$$

with a nonsingular matrix by the Gaussian elimination with partial pivoting. Since the coefficient 9 of the first equation has maximal magnitude of all the coefficients in the first column, it is the pivot and we need not interchange rows. After the first elimination step, we obtain the first derived system

$$\begin{aligned} 9x_1 + 3x_2 &= 15, \\ -2x_2 + 6x_3 &= 8, \\ 6x_2 + 3x_3 &= -3. \end{aligned}$$

In the second step, we look for the pivot in the second column and the second and third row. The coefficient of maximal magnitude is in the position (2, 3) and the pivot is thus equal to 6. We interchange the second and the third equation, obtaining the second derived system

$$\begin{aligned} 9x_1 + 3x_2 &= 15, \\ 6x_2 + 3x_3 &= -3, \\ 7x_3 &= 7 \end{aligned}$$

with an upper triangular matrix after the second step. By backsubstitution we successively compute

$$x_3 = 7/7 = 1\,,$$
$$x_2 = (-3 - 3\,.\,1)/6 = -1\,,$$
$$x_1 = (15 + 3\,.\,1)/9 = 2\,.$$

REMARK 5. The Gaussian elimination can be used to solve several systems of equations having the same matrix and different right-hand sides. The operations we carried out with the single right-hand side are now performed with all the right--hand sides (the augmented matrix is augmented with all the right-hand sides) and the backsubstitution is performed with each right-hand side individually. A certain disadvantage consists in the fact that all the right-hand sides have to be known before the computation starts.

The *LU factorization* is a method algorithmically equivalent to the Gaussian elimination. It is based on the factorization of the matrix $\boldsymbol{A}$ of the system (2) into the product of two factors,

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}\,, \tag{7}$$

where $\boldsymbol{L}$ is a lower triangular matrix with 1's on the diagonal and $\boldsymbol{U}$ is an upper triangular matrix. The name of the method is derived from the first letters of words "lower" and "upper". The factorization (7) is unique and can be calculated e.g. from the formulae

$$\begin{aligned}
u_{11} &= a_{11}\,,\\
l_{i1} &= a_{i1}/u_{11}\,, \quad i = 2, \ldots, n\,,\\
u_{1r} &= a_{1r}\,,\\
u_{ir} &= a_{ir} - \sum_{j=1}^{i-1} l_{ij}u_{jr}\,, \quad i = 2, \ldots, r\,,\\
l_{ir} &= \frac{1}{u_{rr}}\left(a_{ir} - \sum_{j=1}^{r-1} l_{ij}u_{jr}\right), \quad i = r+1, \ldots, n\,,\\
&\qquad\qquad \text{for } r = 2, \ldots, n\,.
\end{aligned} \tag{8}$$

The computed entries of $\boldsymbol{L}$ and $\boldsymbol{U}$ are stored in place of those entries of $\boldsymbol{A}$ that are no longer needed for the computation. Pivoting, which minimizes the accumulation of roundoff errors, is also advisable in the LU factorization (cf. Remark 2). Like in the Gaussian elimination, no actual interchange of rows (or columns) in the computer storage is performed in the LU factorization. It is sufficient to store the indices corresponding to the interchange.

It can be easily shown that the entries l_{ij} of $\boldsymbol{L}$ are (for $i > j$) the coefficients $a_{ij}^{(j-1)}/a_{jj}^{(j-1)}$ used for multiplying the j-th row to be subtracted from the i-th row in the j-th step of elimination. $\boldsymbol{U}$ is the upper triangular matrix of the system (5) resulting from the elimination.

If we know the factors (7) we solve the system (2) in two steps. We rewrite the system in the form

$$\boldsymbol{Ax} = \boldsymbol{LUx} = \boldsymbol{b}. \tag{9}$$

The process of solving the system

$$\boldsymbol{Ly} = \boldsymbol{b} \tag{10}$$

with an auxiliary unknown vector $\boldsymbol{y}$ is called the *forward substitution*, the process of solving the system

$$\boldsymbol{Ux} = \boldsymbol{y} \tag{11}$$

is the backsubstitution we know from the Gaussian elimination. Substituting (10) and (11) in (9), we can prove that the vector $\boldsymbol{x}$ solves the original system. Discussing the Gaussian elimination, we saw that the backsubstitution (11) was very simple. The forward substitution (10) is quite analogous but we start solving the system with the first equation.

Example 2. Using the formulae (8), we find that the LU factorization (7) of the matrix $\boldsymbol{A}$ of the system (6) from Example 1 has the form

$$\begin{bmatrix} 9, & 3, & 0 \\ 6, & 8, & 3 \\ -3, & -3, & 6 \end{bmatrix} = \begin{bmatrix} 1, & 0, & 0 \\ \frac{2}{3}, & 1, & 0 \\ -\frac{1}{3}, & -\frac{1}{3}, & 1 \end{bmatrix} \begin{bmatrix} 9, & 3, & 0 \\ 0, & 6, & 3 \\ 0, & 0, & 7 \end{bmatrix}.$$

Notice that the second and the third row of the matrix $\boldsymbol{A}$ on the left-hand side are interchanged. This corresponds to the interchange of these rows during elimination caused by the partial pivoting in Example 1.

REMARK 6. The advantage of the LU factorization is apparent especially in case of solving several systems with the same matrix and different right-hand sides if not all of them are known before the computation starts, but arise e.g. during some iterative process. The factorization (7) is then computed only once, and the systems (10) and (11) are solved for each right-hand side individually. The factorization can be carried out even without knowing the right-hand side and requires of order n^3 operations. The solution of the systems (10) and (11) needs only of order n^2 operations for each right-hand side.

REMARK 7. If $\boldsymbol{A}$ is symmetric, we can construct the factorization (*Choleski factorization*)

$$\boldsymbol{A} = \boldsymbol{LL}^{\mathrm{T}} \tag{12}$$

instead of (7). The entries of $\boldsymbol{L}$ are real if $\boldsymbol{A}$ is positive definite. This factorization saves arithmetic operations as well as computer storage since it is sufficient to calculate the entries of $\boldsymbol{L}$, and $\boldsymbol{L}^{\mathrm{T}}$ is its transpose (cf. Definition 1.16.3). The diagonal of $\boldsymbol{L}$ need not now consist of only 1's. The factorization

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\mathrm{T}}, \tag{13}$$

where $\boldsymbol{L}$ is again a lower triangular matrix with 1's on the diagonal and $\boldsymbol{D}$ is a diagonal matrix, is even more advantageous for a symmetric matrix. In addition to the two systems (10) and (11) we then also solve a system with the diagonal matrix $\boldsymbol{D}$, which is a very easy task.

30.2. Computation of the Determinant and the Inverse Matrix

We stated in § 30.1 that the computer calculation of the determinant of a nonsingular matrix $\boldsymbol{A}$ of order n with the help of Definition 1.17.1 is practically infeasible even for small n since it requires an enormous amount of arithmetic operations. This calculation, however, can be performed with advantage if the LU factorization of $\boldsymbol{A}$ is used. The determinant of a product of matrices equals the product of the determinants of these matrices (Theorem 1.25.4). If we thus find the LU factorization

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U} \tag{1}$$

in accord with § 30.1, we have

$$\det \boldsymbol{A} = \det \boldsymbol{L} \,.\, \det \boldsymbol{U} = \det \boldsymbol{U}, \tag{2}$$

since the determinant of a triangular matrix equals the product of its diagonal entries (Theorem 1.26.5) and $\boldsymbol{L}$ has 1's on its diagonal. The value of $\det \boldsymbol{U}$ can be calculated with the help of the same theorem.

If we used pivoting when computing the LU factorization, i. e., if we interchanged some rows of the matrix of the system, each such interchange causes a change of the sign of the determinant (Theorem 1.17.4). During the LU factorization, we thus have to keep track of such interchanges and change the sign of $\det \boldsymbol{U}$ in (2) if the total number of interchanges is odd.

Example 1. By the formula (2), calculate the determinant of the matrix

$$\boldsymbol{A} = \begin{bmatrix} 9, & 3, & 0 \\ -3, & -3, & 6 \\ 6, & 8, & 3 \end{bmatrix}$$

of the system from Example 30.1.1. Since

$$\det \boldsymbol{U} = \begin{vmatrix} 9, & 3, & 0 \\ 0, & 6, & 3 \\ 0, & 0, & 7 \end{vmatrix} = 378$$

according to Example 30.1.2 and since we once interchanged rows during the LU factorization, we have $\det \boldsymbol{A} = -378$.

The LU factorization can also be employed to compute the inverse $\boldsymbol{A}^{-1}$ of a nonsingular square matrix $\boldsymbol{A}$.

Definition 1. A square matrix $\boldsymbol{A}^{-1}$ of order n is called the *inverse of a nonsingular square matrix* $\boldsymbol{A}$ of the same order if $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$ where

$$\boldsymbol{I} = \begin{bmatrix} 1, & 0, & \dots, & 0 \\ 0, & 1, & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0, & 0, & \dots, & 1 \end{bmatrix}$$

is the *identity matrix* (cf. Definition 1.25.5).

The formula $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$ from the definition of $\boldsymbol{A}^{-1}$ means that if $\boldsymbol{z}_k$ is the k-th column of $\boldsymbol{A}^{-1}$, then

$$\boldsymbol{A}\boldsymbol{z}_k = \boldsymbol{e}_k\,, \tag{3}$$

where $\boldsymbol{e}_k$ is the k-th column of the identity matrix $\boldsymbol{I}$ (whose all components are equal to 0 except for the k-th one which equals 1). Therefore, the procedure for the computation of $\boldsymbol{A}^{-1}$ consists in solving the system (3) successively for $k = 1, \dots, n$. Finally we form the inverse $\boldsymbol{A}^{-1}$ from the column vectors $\boldsymbol{z}_k$ calculated. The LU factorization with partial pivoting is advantageous for solving these systems since we once compute the factors $\boldsymbol{L}$ and $\boldsymbol{U}$, and carry out only the forward substitution and backsubstitution for each system (3) (Remark 30.1.6). The Gaussian elimination with n right-hand parts $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ (which are all known beforehand) yields the same result (Remark 30.1.5).

REMARK 1. According to Remark 1.25.3, we can calculate the solution of the linear algebraic system

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{4}$$

by the formula

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}\,. \tag{5}$$

The computation of the inverse $\boldsymbol{A}^{-1}$, however, requires the same number of arithmetic operations as the solution of n linear algebraic systems (with the same matrix). Moreover, we need of order n^2 additional operations to substitute in the

formula (5). Therefore, it usually does not pay to solve a nonsingular system (4) by this procedure.

Further methods for the computation of the inverse matrix are based on partitioning the matrix $\boldsymbol{A}$ followed by the computation of inverses to several matrices of lower orders (see, e.g., [378]). In particular cases, also the *Sherman-Morrison* or the *Woodbury formula* [376] can be applied with advantage. Provided that we know $\boldsymbol{A}^{-1}$, these formulae express how the inverse changes if we change $\boldsymbol{A}$ in a definite way.

30.3. Roundoff Error. Iterative Improvement of the Solution

No matter what method is used, the numerical solution of a linear algebraic system is, in general, influenced by roundoff errors that may accumulate in the course of the computation. The *backward analysis* [496] is used to study the error caused by roundoff in algebraic problems. It means for our problem that we represent the computation with roundoff as a true computation (without roundoff) of the solution of a perturbed system, i.e. a system with a perturbed matrix as well as right-hand side. These perturbations of the matrix of the system and the right-hand side can be estimated and are, as a rule, relatively small. But even a small change of the matrix or right-hand side can cause a large change of the solution. This depends on certain properties of the matrix of the system we will try to characterize briefly.

Example 1. The solution of the system

$$\begin{aligned} 2x + \quad & 4y = 6, \\ 4x + 8{\cdot}00001y &= 12{\cdot}00001 \end{aligned}$$

is $x = 1$, $y = 1$. If this system is perturbed only a little, e.g.

$$\begin{aligned} 2x + 4{\cdot}00001y &= 5{\cdot}99998, \\ 4x + 8{\cdot}00003y &= 11{\cdot}99994, \end{aligned}$$

its solution is $x = 7$, $y = -2$.

We will introduce the concepts of norm of a vector and a matrix that enable us to "measure the magnitude" of vectors and matrices in a certain sense. (We present only some of the possible ways of introducing the norm, cf. Example 22.4.3, Remark 22.5.5, and Remarks 1 and 2.)

Definition 1. We denote the *norm of a vector* $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ by $\|\boldsymbol{x}\|$ and put (*Euclidean norm*)

$$\|\boldsymbol{x}\| = \sqrt{\left(\sum_{i=1}^{n} x_i^2\right)} .$$

We denote the *norm of an m by n matrix* $\boldsymbol{A} = (a_{ij})$ by $\|\boldsymbol{A}\|$ and put (*spectral norm*)

$$\|\boldsymbol{A}\| = \sqrt{(\varrho(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}))},$$

where $\varrho(\boldsymbol{B})$ is the *spectral radius of a square matrix* B of order n,

$$\varrho(\boldsymbol{B}) = \max_{i=1,\ldots,n} |\lambda_i|,$$

and λ_i are eigenvalues of $\boldsymbol{B}$ (cf. Definition 1.28.3).

Some bounds for the spectral radius are given in § 30.10.

REMARK 1. Vectors can be viewed as elements of a linear normed space, matrices as linear operators defined on this space and mapping it again onto a linear normed space of vectors. The norm of elements of a linear normed space, the norm of operators, and their properties are discussed from a general point of view in § 22.4 and § 22.5.

In linear algebra, norms of vectors and matrices different from those introduced in Definition 1 are often used, too. For example,

$$\|\boldsymbol{x}\| = \sum_{i=1}^{n} |x_i|$$

is called the *sum norm of a vector* $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ and

$$\|\boldsymbol{x}\| = \max_i |x_i|$$

is called the *uniform* or *maximum norm of a vector* $\boldsymbol{x}$ (see [145]). The following norms of an m by n matrix $\boldsymbol{A} = (a_{ij})$ correspond to these norms of vectors:

$$\|\boldsymbol{A}\| = \max_k \sum_{i=1}^{m} |a_{ik}|,$$

$$\|\boldsymbol{A}\| = \max_i \sum_{k=1}^{n} |a_{ik}|,$$

respectively.

REMARK 2. We use the norms introduced in Definition 1 in this chapter, until otherwise stated. It can, however, be proved by functional analytic tools that all norms of vectors and matrices (possessing the properties presented in § 22.4) are equivalent. This means that if $\|\boldsymbol{x}\|_1$ and $\|\boldsymbol{x}\|_2$ are two norms of a vector $\boldsymbol{x}$ then there are positive constants c and C such that $c\|\boldsymbol{x}\|_1 \leqq \|\boldsymbol{x}\|_2 \leqq C\|\boldsymbol{x}\|_1$ holds for

any vector $\boldsymbol{x}$. Similarly, if $\|\boldsymbol{A}\|_1$ and $\|\boldsymbol{A}\|_2$ are two norms of a matrix $\boldsymbol{A}$, then there are positive constants d and D such that $d\|\boldsymbol{A}\|_1 \leqq \|\boldsymbol{A}\|_2 \leqq D\|\boldsymbol{A}\|_1$ holds for any matrix $\boldsymbol{A}$. The equivalence of norms makes possible to generalize many statements of this chapter.

Definition 2. The number $\kappa(\boldsymbol{A}) = \left\|\boldsymbol{A}^{-1}\right\| \|\boldsymbol{A}\|$ is called the (*spectral*) *condition number of a nonsingular matrix* $\boldsymbol{A}$.

If the condition number $\kappa(\boldsymbol{A})$ is large, then a small perturbation of a matrix entry or a right-hand side component causes a large change of the solution of the system. On the contrary, if the condition number is small (but it is always $\kappa(\boldsymbol{A}) \geqq 1$), then only small changes of the solution correspond to small perturbations of the matrix or the right-hand side. The matrix of the system is called *ill-conditioned* in the former case and *well-conditioned* in the latter one. It is hard to quantify the concepts of "large condition number" or "ill-conditioned matrix". The practical importance of these concepts consists rather in the fact that we can decide which of two given matrices is "worse" or "better" conditioned.

Example 2. The condition numbers of matrices of the systems in Example 1 can be calculated with help of Definitions 2 and 1, the required eigenvalues being found as roots of characteristic polynomials (§ 30.10). The result is that both the condition numbers are approximately equal to $5 \cdot 10^6$, i.e., both the matrices are ill-conditioned. In this connection, notice that the matrix

$$\begin{bmatrix} 2, & 4 \\ 4, & 8 \end{bmatrix}$$

is singular.

Already in § 30.1, we tried to avoid an unnecessary loss of accuracy of the result due to roundoff and we employed pivoting to this end. The use of more accurate arithmetic (e.g. double precision) also helps to reduce the influence of roundoff errors.

If we have computed the solution of a system by the LU factorization, we can improve the accuracy for this solution, expending a relatively small number of additional operations (of order n^2). Denote by $\boldsymbol{x}_0$ the computed numerical solution of the system

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{1}$$

Further let $\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$ be the *residual*. If we could solve the system

$$\boldsymbol{A}\boldsymbol{y} = \boldsymbol{r}_0 \tag{2}$$

exactly, then the vector $\boldsymbol{x}_0 + \boldsymbol{y}$ would be the true solution of the system (1) since $\boldsymbol{A}(\boldsymbol{x}_0 + \boldsymbol{y}) = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{A}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{r}_0 = \boldsymbol{b}$. Let us solve the system (2) numerically. As

we know the LU factorization of the matrix $\boldsymbol{A}$, it suffices to carry out the forward substitution and backsubstitution (cf. § 30.1). Denote by $\boldsymbol{y}_0$ the computed solution of the system (2). If $\boldsymbol{A}$ is not very ill-conditioned, the vector $\boldsymbol{x}_1 = \boldsymbol{x}_0 + \boldsymbol{y}_0$ is an improved, more accurate solution. We can continue this procedure as long as the norm of the residual decreases.

In general, we start with the initial approximation $\boldsymbol{x}_0$ and compute successively

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} + \boldsymbol{y}_{k-1}\,,$$

where $\boldsymbol{y}_{k-1}$ is the numerically computed solution of the system $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{r}_{k-1}$ and

$$\boldsymbol{r}_{k-1} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_{k-1} \tag{3}$$

is the numerically computed residual. Since we expect components of the residual $\boldsymbol{r}_{k-1}$ to be small and since they are computed as difference of the corresponding components of the vectors $\boldsymbol{b}$ and $\boldsymbol{A}\boldsymbol{x}_{k-1}$ of almost the same magnitude, it is important to compute the residual (3) in double precision. If $\boldsymbol{A}$ is not very ill-conditioned, then 2 or 3 just described iteration steps are enough to get an improved solution. The procedure is inefficient for very ill-conditioned matrices. In agreement with Definition 2, the difference of the true and computed solution of the system with such a matrix may be large even if the corresponding residual is small.

30.4. Singular Value Decomposition. Solution of Systems with Singular and Rectangular Matrices

We first present a statement known in linear algebra (cf. [145]).

Theorem 1. *Let $\boldsymbol{A}$ be a real m by n matrix, $m \geqq n$. Then there exist an m by n matrix $\boldsymbol{U} = (u_{ij})$, a diagonal matrix $\boldsymbol{W} = \operatorname{diag}(w_i)$ of order n with nonnegative diagonal entries w_i, and a square matrix $\boldsymbol{V} = (v_{ij})$ of order n such that*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{W}\boldsymbol{V}^{\mathrm{T}}\,. \tag{1}$$

The columns of $\boldsymbol{U}$ are mutually orthonormal vectors (i.e.

$$\sum_{i=1}^{m} u_{ik}u_{il} = \delta_{kl}\,, \quad k = 1, \ldots, n\,, \quad l = 1, \ldots, n\,,$$

where δ_{kl} is the Kronecker delta,

$$\delta_{kl} = \begin{cases} 1 & \text{if } l = k\,, \\ 0 & \text{if } l \neq k\,, \end{cases}$$

cf. Remark 8.1.3) *and $\boldsymbol{V}$ is an orthogonal matrix* (cf. Theorem 1.25.12).

REMARK 1. The formula (1) is called the *singular value decomposition of the matrix* $\boldsymbol{A}$. The matrix $\boldsymbol{W}$ is determined uniquely except for permutations of its diagonal entries.

Diagonal entries w_i of $\boldsymbol{W}$ are called *singular values of the matrix* $\boldsymbol{A}$.

Theorem 1 plays a very important role in numerical practice since there is a sophisticated and numerically stable algorithm for the computation of singular value decomposition by G. H. Golub and C. Reinsch [460] (see also [376], [446], [498]). Squares of singular values of the matrix $\boldsymbol{A}$ are eigenvalues of the square matrix $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ of order n. The algorithm is, in principle, based on methods for the computation of eigenvalues and eigenvectors from Part B of this chapter. Several important applications of Theorem 1 are presented later in this paragraph.

Example 1. We can easily verify that

$$\boldsymbol{A} = \begin{bmatrix} 2\sqrt{3}, & -4\sqrt{2}, & 2 \\ -\sqrt{6}, & -2, & -\sqrt{2} \\ \sqrt{6}, & 2, & \sqrt{2} \\ 2\sqrt{3}, & 4\sqrt{2}, & 2 \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{1}{\sqrt{2}}, & 0, & -\frac{1}{\sqrt{2}} \\ 0, & -\frac{1}{\sqrt{2}}, & 0 \\ 0, & \frac{1}{\sqrt{2}}, & 0 \\ \frac{1}{\sqrt{2}}, & 0, & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 4\sqrt{6}, & 0, & 0 \\ 0, & 2\sqrt{6}, & 0 \\ 0, & 0, & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2}, & -\frac{\sqrt{2}}{\sqrt{3}}, & \frac{1}{2\sqrt{3}} \\ \frac{1}{\sqrt{2}}, & \frac{1}{\sqrt{3}}, & \frac{1}{\sqrt{6}} \\ \frac{1}{2}, & 0, & -\frac{\sqrt{3}}{2} \end{bmatrix} = \boldsymbol{U}\boldsymbol{W}\boldsymbol{V}^{\mathrm{T}},$$

where the matrices $\boldsymbol{U}$, $\boldsymbol{W}$ and $\boldsymbol{V}$ possess the properties from Theorem 1. The singular values of $\boldsymbol{A}$ are $4\sqrt{6}$, $2\sqrt{6}$ and 0; they are square roots of the eigenvalues 96, 24 and 0 of the square matrix $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ of order 3.

REMARK 2. Computing the singular value decomposition (1) numerically, we do not obtain, due to roundoff errors, exact zero singular values where zero values should be if we carried out the computation without roundoff. Therefore, if a numerically computed singular value is negligible as compared with the maximal singular value (e.g. if their ratio is less than about 10^{-6} in single precision), we consider this small singular value to be zero.

Theorem 2. *The rank of an m by n matrix $\boldsymbol{A}$, $m \geqq n$, is equal to the number of its nonzero singular values.*

REMARK 3. Theorem 2 can also be used to determine the rank of an m by n matrix $\boldsymbol{A}$ when $m < n$. In this case, we construct the singular value decomposition of $\boldsymbol{A}^{\mathrm{T}}$ since the matrices $\boldsymbol{A}$ and $\boldsymbol{A}^{\mathrm{T}}$ have the same rank by Theorem 1.16.2.

If we have computed singular values numerically, we always have to decide whether they are zero or not in accordance with Remark 2. The determination

of rank by Remark 1.16.2 (transformation of a matrix into triangular form) performed with roundoff can give a completely false result.

Example 2. The rank of the matrix $\boldsymbol{A}$ from Example 1 equals 2 as $\boldsymbol{A}$ has two nonzero singular values.

We will show the application of singular value decomposition in detail for square matrices. Recall that all singular values of a nonsingular matrix are positive according to Theorem 2.

Theorem 3. *Let $\boldsymbol{A}$ be a nonsingular square matrix of order n. The condition number of $\boldsymbol{A}$* (Definition 30.3.2) *is then*

$$\kappa(\boldsymbol{A}) = \frac{w_{\max}}{w_{\min}},$$

where $w_{\max}$ and $w_{\min}$ are the maximal and minimal singular value of $\boldsymbol{A}$.

Theorem 4. *Let $\boldsymbol{A}$ be a nonsingular square matrix of order n. Then $\boldsymbol{U}$ in the singular value decomposition* (1) *of $\boldsymbol{A}$ is a square orthogonal matrix of order n and*

$$\boldsymbol{A}^{-1} = \boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{U}^{\mathrm{T}}, \tag{2}$$

where $\boldsymbol{W}^{-1} = \operatorname{diag}(1/w_i)$ is the inverse of $\boldsymbol{W}$.

The formula (2) is a simple consequence of Theorem 1.25.7 and Definition 1.25.7 of orthogonal matrix. Therefore, if we know the singular value decomposition of $\boldsymbol{A}$, we can solve the linear algebraic system

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{3}$$

in such a way that we calculate

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} = \boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{b}. \tag{4}$$

In case of a nonsingular matrix $\boldsymbol{A}$, this procedure requires more arithmetic operations than the application of methods of § 30.1 (cf. also Remark 30.2.1) but it can give a satisfactory result even when $\boldsymbol{A}$ is ill-conditioned.

If $\boldsymbol{A}$ is singular, Frobenius Theorem 1.18.1 states that there are two possible cases when the system (3) is to be solved:

1. The rank h of $\boldsymbol{A}$ equals the rank of the augmented matrix of the system (3). Then the system (3) has infinitely many solutions and all its solutions can be expressed (by Theorem 1.18.5) as the sum of a solution of the system (3) and

an arbitrary linear combination of all independent solutions of the homogeneous system

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\,. \tag{5}$$

The system (5) has $n - h$ linearly independent solutions (Theorem 1.18.4).

2. The rank h of $\boldsymbol{A}$ is less than the rank of the augmented matrix. Then the system (3) has no solution.

If $\boldsymbol{A}$ is singular, then at least one of its singular values is zero, the matrix $\boldsymbol{W}$ is singular, and $\boldsymbol{W}^{-1}$ does not exist. In this case, we cannot use (4) to compute the solution. We will show how to generalize the concept of inverse matrix in such a case.

Definition 1. Let $\boldsymbol{A}$ be a square matrix with the singular value decomposition (1). Put

$$\boldsymbol{W}^{\mathrm{I}} = \operatorname{diag}(p_i)\,,$$

where

$$\begin{aligned} p_i &= 1/w_i \quad \text{for } w_i \neq 0\,, \\ p_i &= 0 \qquad \text{for } w_i = 0\,. \end{aligned} \tag{6}$$

The matrix $\boldsymbol{A}^{+} = \boldsymbol{V}\boldsymbol{W}^{\mathrm{I}}\boldsymbol{U}^{\mathrm{T}}$ is called the *pseudoinverse* (the *Moore-Penrose generalized inverse*) *of* $\boldsymbol{A}$.

We have thus obtained a certain analogue of the formula (2). If $\boldsymbol{A}$ is nonsingular, then $\boldsymbol{A}^{-1} = \boldsymbol{A}^{+}$ since $\boldsymbol{W}^{-1} = \boldsymbol{W}^{\mathrm{I}}$. If $\boldsymbol{A}$ is singular, we use also the second line of (6) when constructing the matrix $\boldsymbol{W}^{\mathrm{I}}$. If this construction is carried out numerically, then we must decide whether a singular value w_i is zero or not in virtue of Remark 2, i.e. we put $p_i = 0$ if, for example, $w_i/w_{\max} < 10^{-6}$ in single precision.

If $\boldsymbol{A}$ is singular, we rewrite (4) in the form

$$\boldsymbol{x} = \boldsymbol{A}^{+}\boldsymbol{b} = \boldsymbol{V}\boldsymbol{W}^{\mathrm{I}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{b}\,. \tag{7}$$

Practical application of this formula is shown in the following two theorems.

Theorem 5. *If the rank of* $\boldsymbol{A}$ *equals the rank of the augmented matrix of the system* (3), *then the solution* $\boldsymbol{x}$ *computed by* (7) *possesses the property that* $\|\boldsymbol{x}\| \leqq \|\boldsymbol{y}\|$ *holds for any solution* $\boldsymbol{y}$ *of* (3). *Any solution of the homogeneous system* (5) *is equal to a linear combination of those columns of* $\boldsymbol{V}$ *from the singular value decomposition of* $\boldsymbol{A}$ *that correspond to zero singular values of* $\boldsymbol{A}$.

Theorem 6. *If the rank of* $\boldsymbol{A}$ *is less than the rank of the augmented matrix of* (3), *then the vector* $\boldsymbol{x}$ *computed by* (7) *possesses the property that* $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\| \leqq \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{y}\|$ *holds for any vector* $\boldsymbol{y}$.

The formula (7) thus gives answer in both the cases occurring when we solve the system with a square singular matrix. Moreover, we even need not examine which of the two cases takes place. In the former case, the formula expresses that of infinitely many solutions which has minimal norm. All solutions of (3) can also be obtained by the singular value decomposition since the columns of $\boldsymbol{V}$ corresponding to zero singular values are linearly independent solutions of the homogeneous system (5). In the latter case, no solution of (3) exists but the formula (7) yields a "solution" that satisfies the equations of the system as well as possible in the sense of minimal norm of the residual, i.e. in the least-square sense.

Example 3. We will make use of the following singular value decomposition of a singular square matrix $\boldsymbol{A}$ of rank 1:

$$\boldsymbol{A} = \begin{bmatrix} -1, & 1 \\ 1, & -1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}}, & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}}, & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 2, & 0 \\ 0, & 0 \end{bmatrix} \begin{bmatrix} -\frac{1}{\sqrt{2}}, & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}}, & \frac{1}{\sqrt{2}} \end{bmatrix} = \boldsymbol{U}\boldsymbol{W}\boldsymbol{V}^{\mathrm{T}}. \tag{8}$$

The matrix $\boldsymbol{A}$ has singular values 2 and 0, and they are square roots of eigenvalues 4 and 0 of $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$. The matrix $\boldsymbol{W}^{\mathrm{I}}$ is now of the form

$$\boldsymbol{W}^{\mathrm{I}} = \begin{bmatrix} \frac{1}{2}, & 0 \\ 0, & 0 \end{bmatrix}.$$

Solve first the system

$$\begin{aligned} -x_1 + x_2 &= -1, \\ x_1 - x_2 &= 1 \end{aligned} \tag{9}$$

with the matrix $\boldsymbol{A}$. The augmented matrix has rank 1, too, and we thus have the first case. Substituting in (7), we obtain the solution $\boldsymbol{x} = \boldsymbol{V}\boldsymbol{W}^{\mathrm{I}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{b} = (\frac{1}{2}, -\frac{1}{2})^{\mathrm{T}}$ with minimal norm. The homogeneous system has the solution $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^{\mathrm{T}}$, which is found as the second column of $\boldsymbol{V}$ (the second row of $\boldsymbol{V}^{\mathrm{T}}$ in (8)) since the second diagonal entry of $\boldsymbol{W}$ is zero. The general solution of (9) is then $\boldsymbol{x} = (\frac{1}{2}, -\frac{1}{2})^{\mathrm{T}} + \alpha(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^{\mathrm{T}}$, where α is an arbitrary number.

Further solve the system

$$\begin{aligned} -x_1 + x_2 &= 0, \\ x_1 - x_2 &= 4 \end{aligned}$$

with the same matrix $\boldsymbol{A}$. The augmented matrix has rank 2, i.e. the full rank, and we thus have the second case. There exists no solution of the system. Substituting in (7), we obtain the "solution" $\boldsymbol{x} = \boldsymbol{V}\boldsymbol{W}^{\mathrm{I}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{b} = (1, -1)^{\mathrm{T}}$ which minimizes norm of the residual. In fact, $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} = (2, 2)^{\mathrm{T}}$.

REMARK 4. There are a lot of other methods for both the cases of the system with a singular square matrix but they are often numerically unstable. We therefore recommend the use of the singular value decomposition and the formula (7). We,

however, emphasize once more that small singular values have to be considered zero in numerical computation (Remark 2). Otherwise the application of (7) loses any sense.

The situation is similar when we solve systems with rectangular matrices, both in case $m < n$ (underdetermined system) and $m > n$ (overdetermined system), see [376]. Since we presented Theorem 1 on the singular value decomposition only for $m \geqq n$ we add $n-m$ zero rows to $\boldsymbol{A}$ before the computation of the singular value decomposition if $m < n$. We then use (7) in both the above mentioned cases and obtain the solution possessing the properties stated in Theorems 5 and 6. Remark 4 holds for systems with rectangular matrices, too.

30.5. Sparse Systems. Cyclic Reduction

In solving a great number of problems of numerical analysis, we meet linear algebraic systems whose matrices possess certain special properties (see, e.g., [145]). The matrices can have, for example, their nonzero entries ordered in some regular way (diagonal matrices, triangular matrices, see Definition 30.1.1, etc.) or the values of their nonzero entries show a certain sort of symmetry (e.g. Vandermonde, Toeplitz, or Hilbert matrices). Systems with *sparse matrices*, i.e. matrices having only a very small number of nonzero entries (e.g. three or five nonzero entries in each row), also occur very often. These systems arise, for example, from the discretization of boundary value problems for ordinary and partial differential equations by the finite difference and finite element methods (see Chapters 24, 25 and 27). Provided that the matrix of such a system has order n, it is considered sparse if the total number of its nonzero entries is proportional to n. (The order n of the system depends on $1/h$, i.e. on the reciprocal of the discretization parameter h.)

Special properties of the matrix of the system can often be employed to construct special algorithms that solve the system given very efficiently from the viewpoint of both the number of arithmetic operations and the storage required (see, e.g., [376]).

This paragraph is mainly concerned with sparse matrices. The application of iterative methods is often typical for solving sparse systems (§ 30.6 and § 30.7). There exist, however, direct methods very efficient for some classes of sparse matrices.

Definition 1. A square matrix of order n and of the form

$$\boldsymbol{A} = \begin{bmatrix} a_{11}, & a_{12}, & 0, & 0, & \dots, & 0 \\ a_{21}, & a_{22}, & a_{23}, & 0, & \dots, & 0 \\ 0, & a_{32}, & a_{33}, & a_{34}, & \dots, & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0, & \dots, & 0, & a_{n-1,n-2}, & a_{n-1,n-1}, & a_{n-1,n} \\ 0, & \dots, & 0, & 0, & a_{n,n-1}, & a_{nn} \end{bmatrix} \tag{1}$$

having at most three nonzero entries in each row (that are located on the main diagonal and its two "neighbouring" diagonals) is called *tridiagonal.*

Tridiagonal systems occur very often in problems of numerical analysis. Applying the LU factorization method from § 30.1 to the solution of the system

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{2}$$

with the tridiagonal matrix (1), we arrive at simple formulae for the factorization and forward substitution

$$\begin{aligned} \alpha_1 &= -\frac{a_{12}}{a_{11}}, \quad \beta_1 = \frac{b_1}{a_{11}}, \\ \alpha_i &= -\frac{a_{i,i+1}}{a_{i,i-1}\alpha_{i-1} + a_{ii}}, \quad i = 2, \ldots, n-1, \\ \beta_i &= \frac{b_i - a_{i,i-1}\beta_{i-1}}{a_{i,i-1}\alpha_{i-1} + a_{ii}}, \quad i = 2, \ldots, n, \end{aligned} \tag{3}$$

and for the backsubstitution

$$\begin{aligned} x_n &= \beta_n, \\ x_i &= \alpha_i x_{i+1} + \beta_i, \quad i = n-1, \ldots, 1. \end{aligned} \tag{4}$$

REMARK 1. We did not perform pivoting in the formulae (3) and (4) since interchanges of rows of $\boldsymbol{A}$ would destroy its tridiagonal structure. It may thus happen that division by zero occurs in the course of the computation even when $\boldsymbol{A}$ is nonsingular or that the result of the computation is unfavourably influenced by roundoff. The conditions for feasibility of the algorithm, which are usually satisfied for matrices arising from practical problems, are presented e.g. in [410]. An example of such condition is the positive definiteness of $\boldsymbol{A}$ (cf. Remark 30.1.4).

REMARK 2. The number of arithmetic operations needed to carry out the factorization algorithm for a tridiagonal matrix is of order n. The number of operations is thus proportional to the number of unknowns here while this number is of order n^3 for a general matrix (cf. Remark 30.1.6). Also storage requirements of the method are minimal: instead of n^2 entries of $\boldsymbol{A}$, we store only its three diagonals as three vectors and, in addition, two auxiliary vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Definition 2. A method for solving the system (2) of n equations is called *fast* if it needs of order at most $n \log n$ arithmetic operations to yield the solution.

REMARK 3. The logarithmic factor in the definition of the fast method influences the number of arithmetic operations weakly if n is large. It is also admitted in the definition of fast methods in other branches of numerical analysis (e.g. the fast Fourier transform).

REMARK 4. The factorization method (3) and (4) for a tridiagonal system is thus fast in the sense of Definition 2. The method can be generalized even to systems whose matrices have nonzero entries located on more than three diagonals (see, e.g., [410]).

Another important type of sparse matrices are bandmatrices.

Definition 3. A square matrix $\boldsymbol{A}$ is called a *bandmatrix of bandwidth* $2m+1$ if

$$a_{ij} = 0 \quad \text{for } |i-j| > m. \tag{5}$$

Example 1. The matrix

$$\begin{bmatrix} 6, & 0, & -1, & 0, & 0, & 0 \\ -1, & 6, & 0, & -1, & 0, & 0 \\ -1, & -1, & 6, & 0, & -1, & 0 \\ 0, & 0, & -1, & 6, & 0, & -1 \\ 0, & 0, & 0, & -1, & 6, & 0 \\ 0, & 0, & 0, & -1, & -1, & 6 \end{bmatrix}$$

of order 6 is a bandmatrix of bandwidth 5 (i.e., $m = 2$ in the formula (5)).

Example 2. Diagonal matrix is a bandmatrix of bandwidth 1 ($m = 0$), tridiagonal matrix is a bandmatrix of bandwidth 3 ($m = 1$).

REMARK 5. If $\boldsymbol{A}$ is a bandmatrix of order n, then the triangular matrices $\boldsymbol{L}$ and $\boldsymbol{U}$ computed by the LU factorization (§ 30.1) are also bandmatrices with the same m (i.e., $l_{ij} = 0$ for $i - j > m$ and $u_{ij} = 0$ for $j - i > m$). This property is used to construct special LU factorization algorithms that require of order $m^2 n$ arithmetic operations and storage of the size $(2m+1)n$ when the system (2) is solved.

Implementation of the LU factorization method for sparse matrices, whose nonzero entries are placed in no regular pattern, is somewhat more difficult. In the course of the factorization, nonzero entries may appear in the matrices $\boldsymbol{L}$ and $\boldsymbol{U}$ in places where the corresponding entry of $\boldsymbol{A}$ is zero. Therefore, methods which perform suitable permutations of rows (and possibly also columns) of $\boldsymbol{A}$ and are capable of minimizing (to some extent) this fill-in are applied before the factorization step (see, e.g., [145], [373], [459]). As a consequence, the number of required operations is minimized, too. Moreover, it is necessary to arrange in a proper way that only nonzero entries of $\boldsymbol{A}$, $\boldsymbol{L}$ and $\boldsymbol{U}$ are stored (see, e.g., [373], [460]).

Tridiagonal matrices and bandmatrices are a simple example where sparsity can be exploited. A further standard type are e.g. *profile matrices* [145] and many others [376].

REMARK 6. Methods that are fast in the sense of Definition 2 are undoubtedly very advantageous in numerical practice. We have so far met only one fast direct method, the LU factorization of a tridiagonal matrix (and its generalization in Remark 4). In practical problems with bandmatrices, however, the number m (determining the bandwidth) usually depends on the number n of equations and then the resulting LU factorization is not fast. For example, the discretization of a two-dimensional boundary value problem for an elliptic differential equation typically leads to $m = \sqrt{n}$ and the total number of arithmetic operations is then of order n^2 (Remark 5).

A fast direct method, called the *cyclic reduction*, can be constructed if we can exploit, in addition to sparsity of the matrix, certain special properties of the values of nonzero entries of this matrix.

Let the matrix $\boldsymbol{A}$ of the system (2) be block tridiagonal (cf. Definition 1.26.1 and Definition 1) and let the system be of the form

$$\begin{bmatrix} \boldsymbol{W} & -\boldsymbol{I} & & & \\ -\boldsymbol{I} & \boldsymbol{W} & -\boldsymbol{I} & & \\ & \ddots & \ddots & \ddots & \\ & & -\boldsymbol{I} & \boldsymbol{W} & -\boldsymbol{I} \\ & & & -\boldsymbol{I} & \boldsymbol{W} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_{N-1} \\ \boldsymbol{x}_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_{N-1} \\ \boldsymbol{b}_N \end{bmatrix}, \tag{6}$$

where $\boldsymbol{W}$ is a square matrix of order M, $\boldsymbol{I}$ is the identity matrix of order M, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N$ are M-component vectors. The system thus has $n = MN$ equations. Let $N = 2^{s+1} - 1$, where s is a positive integer. Let us apply the block Gaussian elimination to the solution of the system (6) and carry it out systematically in such a way that the number of block equations is roughly halved in each step (see, e.g., [446] for detailed discussion). We arrive at the following algorithm:

Put

$$\boldsymbol{W}_i = \prod_{r=1}^{2^i} \left(\boldsymbol{W} - 2\cos\frac{(2r-1)\pi}{2^i}\boldsymbol{I} \right), \quad i = 0, \ldots, s, \tag{7}$$

and, successively for $i = 1, \ldots, s$, compute the vectors

$$\boldsymbol{b}_j^{(i)} = \boldsymbol{b}_{2j-1}^{(i-1)} + \boldsymbol{W}_{i-1}\boldsymbol{b}_{2j}^{(i-1)} + \boldsymbol{b}_{2j+1}^{(i-1)}, \quad j = 1, \ldots, 2^{s+i-1} - 1, \tag{8}$$

where

$$\boldsymbol{b}_j^{(0)} = \boldsymbol{b}_j, \quad j = 1, \ldots, 2^{s+1} - 1.$$

Subscripts of matrices and superscripts of vectors denote the step of the method.

The backsubstitution starts with solving the system of M equations

$$\boldsymbol{W}_s \boldsymbol{x}_1^{(s)} = \boldsymbol{b}_1^{(s)}. \tag{9}$$

Further put, for $i = s-1, s-2, \ldots, 0$,

$$\boldsymbol{x}_{2j}^{(i)} = \boldsymbol{x}_j^{(i+1)}, \quad j = 1, \ldots, 2^{s-i} - 1,$$
$$\boldsymbol{x}_0^{(i)} = \boldsymbol{x}_{2^{s+i-1}}^{(i)} = \boldsymbol{0}$$

and successively solve systems of M equations

$$\boldsymbol{W}_i \boldsymbol{x}_{2j+1}^{(i)} = \boldsymbol{x}_{2j}^{(i)} + \boldsymbol{b}_{2j+1}^{(i)} + \boldsymbol{x}_{2j+2}^{(i)}, \quad j = 0, \ldots, 2^{s-i} - 1. \tag{10}$$

The resulting vectors $\boldsymbol{x}_1^{(0)}, \ldots, \boldsymbol{x}_{2^{s+1}-1}^{(0)}$ form the solution of the system (6).

The feature substantial for devising this algorithm is that the diagonal blocks of the matrix $\boldsymbol{A}$ are mutually identical and the nonzero off-diagonal blocks as well. The algorithm is fast if $\boldsymbol{W}$ satisfies a proper additional assumption. For example, if $\boldsymbol{W}$ is tridiagonal, if we multiply by $\boldsymbol{W}_i$ in (8) with the help of the factorization (7), and if we also solve the systems (9) and (10) using this factorization (in the way similar to (30.1.9)), we obtain an algorithm requiring of order $MN \log N$ arithmetic operations. It is thus fast by Definition 2.

REMARK 7. The above presented version of the cyclic reduction algorithm is numerically unstable. A stable modification is derived, e.g., in [451].

REMARK 8. The cyclic reduction method can be modified in such a way that it can also be applied to systems somewhat more general than (6) (see, e.g., [452]). A combination of the cyclic reduction and the discrete Fourier transform (implemented by the fast Fourier transform) is called the FACR method and the number of arithmetic operations required is of the same order as with the cyclic reduction. The constant at the term $MN \log N$ in the formula for the number of operations is, however, less for the FACR method (see, e.g., [451]).

30.6. Iterative Methods. One-Point Iteration, the Jacobi and Gauss-Seidel Methods, Successive Overrelaxation. Conjugate Gradient Method

Iterative methods for solving linear algebraic systems have become a classical part of numerical analysis and are treated in vast literature (see, e.g., [145], [378], [475]). We will present only some basic methods and show later in § 30.7 how a proper combination of iterative and direct methods, the preconditioning, may lead to the acceleration of convergence of iterative methods.

One-point matrix iterative methods for solving the system of n equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{1}$$

consist in the following procedure: We choose an arbitrary initial approximation $\boldsymbol{x}_0$ and compute further approximations by the formula

$$\boldsymbol{x}_{k+1} = \boldsymbol{B}_k\boldsymbol{x}_k + \boldsymbol{C}_k\boldsymbol{b}, \quad k = 0, 1, \ldots,$$

where $\boldsymbol{B}_k$ and $\boldsymbol{C}_k$ are square matrices of order n constructed in an appropriate way from $\boldsymbol{A}$. A new approximation $\boldsymbol{x}_{k+1}$ is thus computed only from the last preceding approximation $\boldsymbol{x}_k$. If the method converges, then

$$\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}_t, \tag{2}$$

where $\boldsymbol{x}_t$ is the true solution of (1) and the limit (2) is to be understood in virtue of certain metric (see Definition 22.2.2), e.g.

$$\lim_{k\to\infty} \|\boldsymbol{x}_k - \boldsymbol{x}_t\| = 0$$

(see Remark 22.4.3 and Definition 30.3.1).

If we use an iterative method, we must always be interested in the conditions for convergence and verify them for a particular given matrix. Moreover, any practical computation can involve only a finite number of arithmetic operations. It would thus be suitable to stop the iterative process in the k-th step if $\|\boldsymbol{x}_k - \boldsymbol{x}_t\| < \varepsilon$ for some tolerance ε chosen in advance. We, however, do not know the true solution $\boldsymbol{x}_t$ and thus choose termination criteria mostly in the form

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| < \varepsilon$$

or

$$\|\boldsymbol{r}_{k+1}\| < \varepsilon,$$

where $\boldsymbol{r}_{k+1} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_{k+1}$ is the residual, or

$$|(\boldsymbol{r}_{k+1} - \boldsymbol{r}_k)^{\mathrm{T}}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)| < \varepsilon,$$

where $\boldsymbol{p}^{\mathrm{T}}\boldsymbol{q} = p_1q_1 + \cdots + p_nq_n$ is the inner product of two column vectors $\boldsymbol{p}$ and $\boldsymbol{q}$.

Definition 1. The iterative procedure

$$\boldsymbol{x}_{k+1} = \boldsymbol{B}\boldsymbol{x}_k + \boldsymbol{C}\boldsymbol{b}, \quad k = 0, 1, \ldots, \tag{3}$$

where $\boldsymbol{x}_0$ is an arbitrary initial approximation, and $\boldsymbol{B}$ and $\boldsymbol{C}$ are square matrices of order n (independent of k), is called a *stationary one-point matrix iterative method* for solving the system (1). We say that this method is *consistent* with the system (1) if

$$\boldsymbol{C}\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{I}. \tag{4}$$

$\boldsymbol{B}$ is called the *iteration matrix.*

Consistency of the method ensures that the convergent sequence of approximations $\boldsymbol{x}_k$ tends to the true solution $\boldsymbol{x}_t$. Really, the iteration formula (3) is transformed into an identity after substituting $\boldsymbol{x}_t$ both for $\boldsymbol{x}_k$ and $\boldsymbol{x}_{k+1}$.

Example 1. Let us derive a one-point iteration. We begin with the system (1) given and add $\boldsymbol{x}$ to both its sides. We gradually obtain

$$\begin{aligned} \boldsymbol{x} + \boldsymbol{A}\boldsymbol{x} &= \boldsymbol{x} + \boldsymbol{b}, \\ \boldsymbol{x} &= (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} + \boldsymbol{b}. \end{aligned} \tag{5}$$

Put now $\boldsymbol{x}_{k+1}$ instead of $\boldsymbol{x}$ on the left-hand side of (5) and $\boldsymbol{x}_k$ instead of $\boldsymbol{x}$ on the right-hand side. We arrive at

$$\boldsymbol{x}_{k+1} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x}_k + \boldsymbol{b}, \tag{6}$$

i.e. $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{A}$ and $\boldsymbol{C} = \boldsymbol{I}$. The method (6) is consistent since $\boldsymbol{B}$ and $\boldsymbol{C}$ obviously satisfy the condition (4).

Theorem 1. *The sequence of approximations $\boldsymbol{x}_k$ produced by the consistent stationary one-point matrix iterative method* (3) *converges to the true solution of* (1) *for an arbitrary initial approximation $\boldsymbol{x}_0$ if and only if $\varrho(\boldsymbol{B}) < 1$, where $\varrho(\boldsymbol{B})$ is the spectral radius of the iteration matrix $\boldsymbol{B}$* (Definition 30.3.1).

REMARK 1. The rate of convergence increases with decreasing $\varrho(\boldsymbol{B})$.

Theorem 2. *Let $\|\boldsymbol{B}\|$ be any norm of the matrix $\boldsymbol{B}$* (see Definition 30.3.1 and Remark 30.3.1). *If $\|\boldsymbol{B}\| < 1$, then $\varrho(\boldsymbol{B}) < 1$.*

Theorems 1 and 2 thus provide a condition sufficient for the convergence of a consistent stationary one-point matrix iterative method: It is sufficient that an arbitrary norm of the iteration matrix be less than 1.

We will present several particular examples of stationary methods. Let us rewrite the matrix $\boldsymbol{A}$ of the system (1) in the form

$$\boldsymbol{A} = \boldsymbol{D} - \boldsymbol{E} - \boldsymbol{F},$$

where $\boldsymbol{D}$ is a diagonal matrix, $\boldsymbol{E}$ is a lower triangular matrix with zero diagonal, and $\boldsymbol{F}$ is an upper triangular matrix with zero diagonal. If all diagonal entries of $\boldsymbol{A}$ are nonzero we can substitute for $\boldsymbol{A}$ in (1), obtaining

$$\boldsymbol{D}\boldsymbol{x} = (\boldsymbol{E} + \boldsymbol{F})\boldsymbol{x} + \boldsymbol{b},$$

from where the consistency follows, and further

$$\boldsymbol{x} = \boldsymbol{D}^{-1}(\boldsymbol{E} + \boldsymbol{F})\boldsymbol{x} + \boldsymbol{D}^{-1}\boldsymbol{b}.$$

Finally we have the iteration formula

$$\boldsymbol{x}_{k+1} = \boldsymbol{D}^{-1}(\boldsymbol{E} + \boldsymbol{F})\boldsymbol{x}_k + \boldsymbol{D}^{-1}\boldsymbol{b}$$

known as the *Jacobi method.* For individual components we obtain (the components of $\boldsymbol{x}_k$ are denoted by $x_i^{(k)}$ and similarly for $\boldsymbol{x}_{k+1}$)

$$x_i^{(k+1)} = -\frac{1}{a_{ii}} \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)} + \frac{b_i}{a_{ii}}, \quad i = 1, \ldots, n, \tag{7}$$

which means that the i-th component of $\boldsymbol{x}_{k+1}$ is calculated from the i-th equation of (1), in which we have substituted the components of $\boldsymbol{x}_k$ for all the other components of $\boldsymbol{x}$.

REMARK 2. By Theorem 1, the Jacobi method converges for an arbitrary initial approximation if and only if $\varrho(\boldsymbol{D}^{-1}(\boldsymbol{E} + \boldsymbol{F})) < 1$. A sufficient condition that can be more easily verified for a given matrix $\boldsymbol{A}$ follows from Theorem 2.

If we again calculate the i-th component of $\boldsymbol{x}_{k+1}$ from the i-th equation like in (7) but use now the components $x_j^{(k+1)}$, $j < i$, of the new approximation that have just been computed, we obtain the *Gauss-Seidel method*

$$x_i^{(k+1)} = -\frac{1}{a_{ii}} \left(\sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) + \frac{b_i}{a_{ii}}, \quad i = 1, \ldots, n, \tag{8}$$

or, in matrix notation,

$$\boldsymbol{x}_{k+1} = (\boldsymbol{D} - \boldsymbol{E})^{-1}\boldsymbol{F}\boldsymbol{x}_k + (\boldsymbol{D} - \boldsymbol{E})^{-1}\boldsymbol{b}.$$

An analogue of Remark 2 holds here, too, but with the iteration matrix $(\boldsymbol{D} - \boldsymbol{E})^{-1}\boldsymbol{F}$ of the Gauss-Seidel method. In practice, the following conditions for convergence are verified more easily (see, e.g., [475] for further criteria).

Theorem 3. *Let $\boldsymbol{A}$ be a positive definite matrix* (Definition 30.1.2). *Then the Gauss-Seidel method converges for an arbitrary initial approximation.*

Definition 2. We say that a square matrix $\boldsymbol{A} = (a_{ij})$ of order n is *diagonally dominant* if there are positive numbers $z_1, \ldots, z_n$ such that

$$|a_{ii}|z_i > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|z_j\,, \quad i = 1, \ldots, n\,. \tag{9}$$

Theorem 4. *If $\boldsymbol{A}^{\mathrm{T}}$ is diagonally dominant, then $\boldsymbol{A}$ is diagonally dominant as well.*

The condition from Definition 2 can thus be verified either for $\boldsymbol{A}$ or for its transpose $\boldsymbol{A}^{\mathrm{T}}$, whichever is easier.

Theorem 5. *If $\boldsymbol{A}$ is diagonally dominant, then both the Jacobi and the Gauss-Seidel methods converge for an arbitrary initial approximation.*

The *successive overrelaxation* (SOR) *method* is a generalization of the Gauss-Seidel method. It includes, in addition, a real parameter ω. For some classes of matrices, the SOR method converges for a certain range of values of ω and it is (at least theoretically) possible to find an optimal value ω_0 that minimizes the spectral radius of the iteration matrix (and maximizes the rate of convergence), see, e.g., [145], [475].

Example 2. Solve the system

$$\begin{bmatrix} 4, & -1, & 0 \\ -1, & 4, & -1 \\ 0, & -1, & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -4 \\ 0 \\ 4 \end{bmatrix} \tag{10}$$

by the Gauss-Seidel iterative method. We can easily verify that the matrix $\boldsymbol{A}$ of (10) satisfies the inequalities (9) of Definition 2 if we put $z_1 = z_2 = z_3 = 1$. $\boldsymbol{A}$ is thus diagonally dominant and, by Theorem 5, the Gauss-Seidel method converges to the true solution $\boldsymbol{x}_t$ of (10) for any initial approximation. (The matrix given is symmetric and it can be proved [145] to be positive definite. The convergence then follows from Theorem 3 as well.)

The iteration formula (8) for the system (10) is

$$\begin{aligned} x_1^{(k+1)} &= \tfrac{1}{4}x_2^{(k)} - 1\,, \\ x_2^{(k+1)} &= \tfrac{1}{4}x_1^{(k+1)} + \tfrac{1}{4}x_3^{(k)}\,, \\ x_3^{(k+1)} &= \tfrac{1}{4}x_2^{(k+1)} + 1\,. \end{aligned}$$

Putting $\boldsymbol{x}_0 = (0, 0, 0)^{\mathrm{T}}$, we obtain by successive substitutions that

$$\boldsymbol{x}_1 = (-1, -\tfrac{1}{4}, \tfrac{15}{16})^{\mathrm{T}},\ \boldsymbol{x}_2 = (-\tfrac{17}{16}, -\tfrac{1}{32}, \tfrac{127}{128})^{\mathrm{T}},\ \boldsymbol{x}_3 = (-\tfrac{129}{128}, -\tfrac{1}{256}, \tfrac{1023}{1024})^{\mathrm{T}},\ \text{etc.}$$

Convergence of the approximate solution to the exact solution $\boldsymbol{x}_t = (-1, 0, 1)^T$ is very fast in case of the system (10).

Conjugate direction methods represent a very broad class of nonstationary iterative methods. From this class, we will present the *conjugate gradient method*, which is very often used. See, e.g., [104] for further methods.

Let us solve the system (1) with a positive definite matrix $\boldsymbol{A}$ (Definition 30.1.2). We introduce the functional F by

$$F(\boldsymbol{x}) = \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} \tag{11}$$

for all real n-component vectors. Then it can be shown that this functional assumes its (unique) minimum for $\boldsymbol{x} = \boldsymbol{x}_t$, where $\boldsymbol{x}_t$ is the true solution of (1). The conjugate gradient method belongs to procedures based on minimizing the functional (11). It is devised in such a way that the minimum of F is reached (if we computed without roundoff) after at most n steps. It can thus be considered as a direct method as well.

The iteration formulae are as follows: We choose an initial approximation $\boldsymbol{x}_0$, put $\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$ and $\boldsymbol{p}_0 = \boldsymbol{r}_0$, and successively compute (for $k = 0, 1, \dots$)

$$\begin{aligned}
\alpha_k &= \frac{\boldsymbol{r}_k^T\boldsymbol{r}_k}{\boldsymbol{p}_k^T\boldsymbol{A}\boldsymbol{p}_k}, \\
\boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \alpha_k\boldsymbol{p}_k, \\
\boldsymbol{r}_{k+1} &= \boldsymbol{r}_k - \alpha_k\boldsymbol{A}\boldsymbol{p}_k, \\
\beta_k &= \frac{\boldsymbol{r}_{k+1}^T\boldsymbol{r}_{k+1}}{\boldsymbol{r}_k^T\boldsymbol{r}_k}, \\
\boldsymbol{p}_{k+1} &= \boldsymbol{r}_{k+1} + \beta_k\boldsymbol{p}_k.
\end{aligned}$$

The process cannot be continued if $\boldsymbol{r}_k = \boldsymbol{0}$ occurs for $k < n$. Since $\boldsymbol{r}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k$ is the residual, the zero residual means that $\boldsymbol{x}_k$ is the true solution, though.

REMARK 3. In the formulae of the conjugate gradient method, the matrix $\boldsymbol{A}$ of the system (1) solved occurs only in the product $\boldsymbol{A}\boldsymbol{p}_k$ and, when evaluating the initial residual, in the product $\boldsymbol{A}\boldsymbol{x}_0$. The method is thus advantageous for large sparse matrices $\boldsymbol{A}$. It is sufficient to code a subprogram for the evaluation of the product $\boldsymbol{A}\boldsymbol{y}$, where $\boldsymbol{y}$ is a given vector, and $\boldsymbol{A}$ even need not be stored.

Some other iterative methods possess this property, too. For example, in the iteration formula (3), the vector $\boldsymbol{x}_k$ is multiplied by the iteration matrix $\boldsymbol{B}$. If the matrix $\boldsymbol{A}$ of (1) is sparse, then $\boldsymbol{B}$ may be sparse as well.

REMARK 4. The application of the conjugate gradient method is advantageous if we obtain a sufficiently accurate approximation to the solution already after

a small number of steps (far less than the order n of the matrix of the system). The rate of convergence of the method depends on the condition number $\kappa(\boldsymbol{A})$ (Definition 30.3.2) and increases with decreasing $\kappa(\boldsymbol{A})$. The best efficiency of the method is thus reached for large well-conditioned matrices.

REMARK 5. In iterative methods, the accuracy of the result is also influenced by roundoff error. For example, the conjugate gradient method does not give, in general, the true solution after n steps if roundoff is involved. The loss of accuracy may be large for ill-conditioned matrices.

30.7. Preconditioned Iterative Methods. Incomplete Factorization

In § 30.6 we discussed the convergence of iterative methods. We will show in this paragraph how to increase the rate of convergence (to precondition an iterative method) at the cost of solving an auxiliary system of n linear algebraic equations in each step of the iterative method.

Let us again solve the system of n equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{1}$$

Choose a nonsingular matrix $\boldsymbol{P}$ of order n with the following two properties:

P1. The system

$$\boldsymbol{P}\boldsymbol{x} = \boldsymbol{c} \tag{2}$$

can be solved by a fast direct method (Definition 30.5.2).

P2. $\boldsymbol{P}$ is in some sense close to the matrix $\boldsymbol{A}$ of the system (1) solved.

REMARK 1. Put

$$\boldsymbol{E} = \boldsymbol{P} - \boldsymbol{A}, \tag{3}$$

then "closeness" can mean that $\|\boldsymbol{E}\|$ is small or that the product $\boldsymbol{P}^{-1}\boldsymbol{A}$ is "close" to the identity matrix, i.e., that $\left\|\boldsymbol{I} - \boldsymbol{P}^{-1}\boldsymbol{A}\right\|$ is small.

Let us first consider stationary iterative methods. We will proceed like in Example 30.6.1 but add the vector $\boldsymbol{P}\boldsymbol{x}$ to both sides of (1). We obtain

$$\boldsymbol{P}\boldsymbol{x} = (\boldsymbol{P} - \boldsymbol{A})\boldsymbol{x} + \boldsymbol{b}$$

and, finally, the iteration formula

$$\boldsymbol{P}\boldsymbol{x}_{k+1} = (\boldsymbol{P} - \boldsymbol{A})\boldsymbol{x}_k + \boldsymbol{b}. \tag{4}$$

The formula (4) is a system of equations with the matrix $\boldsymbol{P}$ and a known right--hand side. The new approximation $\boldsymbol{x}_{k+1}$ is obtained as the solution of this system. According to Property P1, the system (4) can be solved by a fast direct method.

To analyse the convergence of the method, we rewrite (4) in the form (30.6.3), i.e.

$$\boldsymbol{x}_{k+1} = \boldsymbol{P}^{-1}(\boldsymbol{P} - \boldsymbol{A})\boldsymbol{x}_k + \boldsymbol{P}^{-1}\boldsymbol{b}.$$

We thus have $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{P}^{-1}\boldsymbol{A}$ and, by Theorem 30.6.1, the method (4) converges if $\varrho(\boldsymbol{I} - \boldsymbol{P}^{-1}\boldsymbol{A}) < 1$. The rate of convergence increases with decreasing $\varrho(\boldsymbol{I} - \boldsymbol{P}^{-1}\boldsymbol{A})$ (Property P2, cf. Remark 30.6.1). The iterative method represented by (4) is called a *preconditioned iterative method*, the matrix $\boldsymbol{P}$ is called a *preconditioner*.

Example 1. If only a small change of $\boldsymbol{A}$ (Property P2) leads to $\boldsymbol{P}$ which satisfies the assumptions for the cyclic reduction method (§ 30.5), we can use this matrix $\boldsymbol{P}$ for preconditioning. Theoretical bound for $\|\boldsymbol{I} - \boldsymbol{P}^{-1}\boldsymbol{A}\|$ is usually not available. We will show another, very often used, choice of $\boldsymbol{P}$ in the conclusion of this paragraph.

REMARK 2. The choice ideal from the viewpoint of Property P1 would be $\boldsymbol{P} = \boldsymbol{I}$. This, however, is no preconditioning since (4) transforms into (30.6.6). From the viewpoint of Property P2, the ideal choice is $\boldsymbol{P} = \boldsymbol{A}$. But $\boldsymbol{P}$ has to possess also Property P1, i.e., we must be able to solve (4) "fast", and there is no reason to solve a system with the matrix $\boldsymbol{A}$ by an iterative method if a fast direct method can be applied.

We can precondition e.g. the conjugate gradient method in a similar way, too. Let $\boldsymbol{P}$ be a positive definite matrix (it is thus symmetric, too, see Definition 30.1.2) and let it possess Properties P1 and P2. Then there exists a unique positive definite matrix $\boldsymbol{P}^{-1/2}$ such that $\boldsymbol{P}^{-1/2}\boldsymbol{P}^{-1/2} = \boldsymbol{P}^{-1}$. Let us apply the conjugate gradient method to the system

$$\boldsymbol{P}^{-1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2}(\boldsymbol{P}^{1/2}\boldsymbol{x}) = \boldsymbol{P}^{-1/2}\boldsymbol{b} \tag{5}$$

obtained from (1). Using simple substitutions, we can achieve that the result of the iterative process is the solution $\boldsymbol{x}$ sought.

We thus choose an arbitrary $\boldsymbol{x}_0$, put $\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$ and $\boldsymbol{p}_0 = \boldsymbol{P}^{-1}\boldsymbol{r}_0$, and successively compute (for $k = 0, 1, \dots$)

$$\alpha_k = \frac{\boldsymbol{r}_k^{\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{r}_k}{\boldsymbol{p}_k^{\mathrm{T}}\boldsymbol{A}\boldsymbol{p}_k},$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{p}_k,$$

$$\boldsymbol{r}_{k+1} = \boldsymbol{r}_k - \alpha_k\boldsymbol{A}\boldsymbol{p}_k,$$

$$\beta_k = \frac{\boldsymbol{r}_{k+1}^{\mathrm{T}} \boldsymbol{P}^{-1} \boldsymbol{r}_{k+1}}{\boldsymbol{r}_k^{\mathrm{T}} \boldsymbol{P}^{-1} \boldsymbol{r}_k},$$

$$\boldsymbol{p}_{k+1} = \boldsymbol{P}^{-1} \boldsymbol{r}_{k+1} + \beta_k \boldsymbol{p}_k .$$

As compared with the original method, we have to find the vector $\boldsymbol{y} = \boldsymbol{P}^{-1}\boldsymbol{r}_{k+1}$ in each step of the preconditioned method. We compute it as the solution of an auxiliary system $\boldsymbol{P}\boldsymbol{y} = \boldsymbol{r}_{k+1}$ (cf. Property P1). By the conjugate gradient method, we solve the system (5) with the matrix $\boldsymbol{P}^{-1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2} = \boldsymbol{I} - \boldsymbol{P}^{-1/2}\boldsymbol{E}\boldsymbol{P}^{-1/2} = \boldsymbol{I} - \tilde{\boldsymbol{E}}$ instead of (1) with the matrix $\boldsymbol{A}$. By definitions 30.3.2 and 30.3.1, the condition number of a nonsingular matrix equals

$$\kappa(\boldsymbol{P}^{-1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2}) = \frac{\lambda_{\max}(\boldsymbol{I} - \tilde{\boldsymbol{E}})}{\lambda_{\min}(\boldsymbol{I} - \tilde{\boldsymbol{E}})} .$$

It is thus close to 1 if $\lambda_{\max}(\tilde{\boldsymbol{E}})$ is small (Property P2), since $\lambda_{\min}(\boldsymbol{I} - \tilde{\boldsymbol{E}}) = 1 - \lambda_{\max}(\tilde{\boldsymbol{E}})$ (Theorem 30.12.2).

We will present a further often used fast direct method which can be used to precondition iterative methods, the incomplete factorization, in its simplest form.

In § 30.5 we showed that the LU factorization is fast for tridiagonal matrices (Remark 30.5.4). We further stated that the LU factorization is not fast for a bandmatrix whose bandwidth depends on the number n of equations (Remark 30.5.6). We will now show a certain LU factorization which is fast but at the cost of not being the exact (complete) factorization of the matrix of the system.

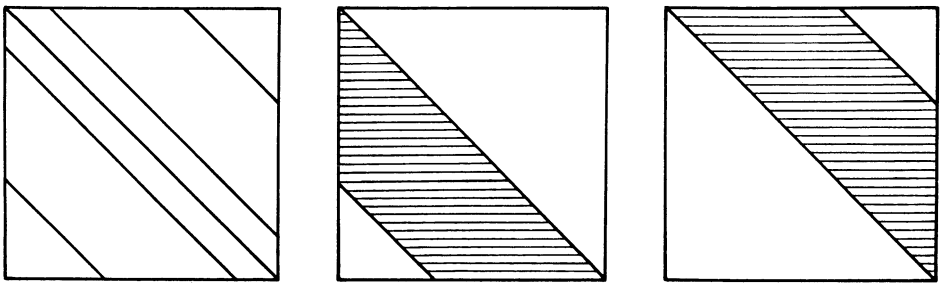


Fig. 30.1. LU factorization of a sparse matrix $\boldsymbol{A}$. Nonzero entries of $\boldsymbol{A}$ lie on the shown five diagonals only, nonzero entries of $\boldsymbol{L}$ and $\boldsymbol{U}$ in the indicated strips. The other entries of the matrices are zero.

Example 2. Let the matrix $\boldsymbol{A}$ of (1) be of order n and have the band structure schematically shown in Fig. 30.1, i.e., let it be sparse with nonzero entries placed on its five diagonals. Such a matrix arises, e.g., from the finite difference discretization of a two-dimensional boundary value problem for the Poisson equation on a rectangular grid (Chap. 27). Considering $\boldsymbol{A}$ as a bandmatrix of bandwidth $2m + 1$

(where m depends on n), we obtain, by the factorization from § 30.1, the triangular matrices $\boldsymbol{L}$ and $\boldsymbol{U}$ (Remark 30.5.5) which are filled-in. (They are also shown in Fig. 30.1.) In general, all entries in their bands may be nonzero.

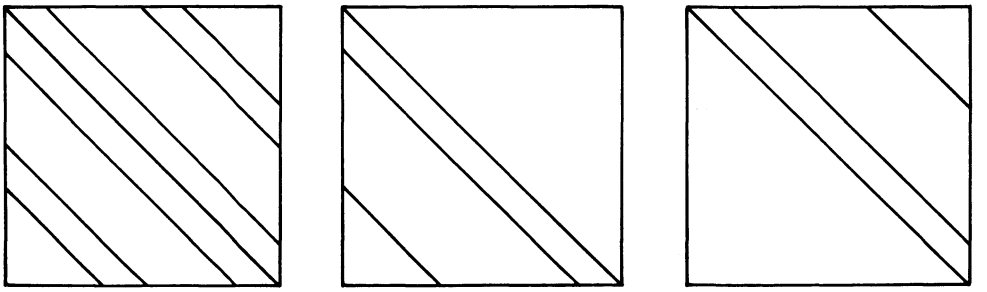


Fig. 30.2. Incomplete factorization. Nonzero entries of the factors $\tilde{\boldsymbol{L}}$ and $\tilde{\boldsymbol{U}}$ lie on the shown diagonals but the product $\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{U}}$ differs from the original matrix $\boldsymbol{A}$.

Conversely, we will now start with the desired form of the factors (Fig.30.2). Let nonzero entries of these factors (denoted by $\tilde{\boldsymbol{L}}$ and $\tilde{\boldsymbol{U}}$) be located only in places where there are nonzero entries of $\boldsymbol{A}$ (and only in the lower or upper triangle of the respective factor). Multiplying $\tilde{\boldsymbol{L}}$ and $\tilde{\boldsymbol{U}}$, we obtain a matrix $\tilde{\boldsymbol{A}}$ different from $\boldsymbol{A}$. In our case, nonzero entries of $\tilde{\boldsymbol{A}}$ are placed on its seven diagonals. The product $\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{U}}$ is called the *incomplete factorization of* $\boldsymbol{A}$.

The factors $\tilde{\boldsymbol{L}}$ and $\tilde{\boldsymbol{U}}$ are determined from $\boldsymbol{A}$ by the standard algorithm from § 30.1 but we compute only their entries having been declared nonzero in advance. The other entries of $\tilde{\boldsymbol{L}}$ and $\tilde{\boldsymbol{U}}$ involved in (30.1.8) are not computed but put equal to zero.

REMARK 3. The computation of the incomplete factorization in Example 2 requires of order n arithmetic operations, the solution of the systems $\tilde{\boldsymbol{L}}\boldsymbol{y} = \boldsymbol{c}$ and $\tilde{\boldsymbol{U}}\boldsymbol{x} = \boldsymbol{y}$ also of order n operations. Putting $\boldsymbol{P} = \tilde{\boldsymbol{A}} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{U}}$, we have the "complete" factorization of $\boldsymbol{P}$ and we can solve (4) by a fast direct method. The matrix $\boldsymbol{P}$ thus possesses Property P1 required for preconditioning.

REMARK 4. We usually know very little about the matrix $\boldsymbol{E} = \tilde{\boldsymbol{A}} - \boldsymbol{A} = \boldsymbol{P} - \boldsymbol{A}$ (where $\tilde{\boldsymbol{A}}$ is the matrix from Example 2) from the theoretical point of view (Property P2). Anyway, the preconditioning based on the incomplete factorization is used very often and its effect is, as a rule, very good. We may admit even more nonzero entries in $\tilde{\boldsymbol{L}}$ and $\tilde{\boldsymbol{U}}$ than in Example 2 to make the norm of $\boldsymbol{E}$ smaller. Similarly to Example 2, we can also factorize general sparse matrices incompletely. If $\boldsymbol{A}$ is symmetric, it is possible to construct an incomplete Choleski factorization analogous to (30.1.12) or (30.1.13).

30.8. Algebraic Multigrid Method

The *multigrid method* is a very general and efficient way of solving boundary value problems for partial differential equations and some other problems. It has recently been developed as a special procedure of the multilevel adaptive technique, which can be characterized as a combination of the process of problem discretization and the process of discrete problem solution, the latter usually representing the solution of a linear algebraic system where each component of the solution vector corresponds to a single grid point. We will only briefly mention the principle of the multigrid method. A systematic discussion can be found, e.g., in [56], [199].

Typical feature of the multigrid method is the discretization (by the finite difference or finite element method, see Chapters 24 and 27) on several regular grids (levels) with different grid spacings. The method has three basic components: 1. relaxation steps on the individual levels (iteration steps of classical iterative methods of § 30.6, e.g., the Gauss-Seidel method); 2. interpolation (prolongation) of values from coarse grid to a finer one; 3. restriction of values from fine grid to a coarser one.

There are a lot of particular strategies that combine the three components of the method. The most frequent procedure is called *cycling*. We start with the discretization of the problem and with an initial approximation to the solution on the finest grid. Several relaxation steps (that are applied to make the residual smaller, but are not used to solve the system as accurately as possible) are carried out on this grid and the residual (called the *defect*) is restricted to a coarser grid. The problem is discretized on this coarser grid, too, and a correction to the solution is computed here from the restricted defect. This correction is interpolated back to the finer grid and added to the approximate solution, and several relaxation steps again follow.

The system for the correction on the coarser grid is of the same kind as the original system for the solution on the finest grid. It is thus solved in such way that we pass to a further, even coarser grid. This gives rise to a recurrent procedure that passes from the finest grid to coarser and coarser grids and then again returns back to the finest grid. In this way, the core of the computation is transferred to coarser grids with small number of grid points, which is one of the reasons for the high efficiency of the multigrid method. The system of equations is actually solved only on the coarsest grid and this is usually done by a direct method (§ 30.1). Considered as a method for solving the system on the finest grid, the multigrid method is a fast method.

The multigrid method is thus related to a boundary value problem for a differential equation and solves the linear algebraic system obtained from discretization. The converse procedure is also possible. We can start with a (sparse) linear algebraic system, consider it as a discretization of a continuous boundary value problem,

and apply the multigrid method to its solution. In the computational procedure obtained, the sequence of discretization levels (grids) need not explicitly appear. The aim of the procedure is to solve a linear algebraic system and it is thus called the *algebraic multigrid method.* There are a great number of strategies here, too (see, e.g., [405]).

30.9. Choice of the Method. Basic Software

Matrices of the systems, that are most often met in practical problems, fall into one of two large classes:

1. Full matrices of moderate order (say, $n < 30$).

2. Sparse matrices of high and very high order (n equal to several hundreds or several thousands is no exception).

Matrices from the first class are mostly treated by direct methods of § 30.1. It is advantageous to employ a possible symmetry of the matrix. Suspecting the matrix to be ill-conditioned, we perform, in addition, about two steps of the iterative improvement from § 30.3. If we find out that the full matrix of the system is special in virtue of § 30.5, we apply the corresponding special method.

Our recommendation for sparse matrices from the second class is not so unambiguous. It is hard, for example, to store a matrix of order 1000 (including its zero entries) in the main memory, and thus the use of both direct and iterative methods is usually accompanied with certain programming tricks. If some fast direct method (§ 30.5) is applicable then it certainly pays. Efficient employment of the LU factorization to sparse matrices with nonzero entries placed at random (§ 30.5) has to consist of three steps: minimization of the fill-in (it is carried out once for a particular structure of zeros and nonzeros), factorization of the matrix (it is performed once for particular numerical values of entries), and solution of the system (it is carried out for every right-hand side), and it is relatively expensive. It is advantageous if the result of minimizing the fill-in can be exploited for more matrices of the same zero-nonzero structure or if we solve several systems that differ only in right-hand sides.

All iterative methods bring the risk of terminating the process too early and getting thus s solution which is not accurate enough. This may be, in some cases, an asset. If we are interested in a less accurate solution (e.g. in its two significant digits), we need not expend unnecessary labour on solving the system accurately by a direct method. We must be very cautious if we use iterative methods in which a parameter or parameters to be chosen occur. Taking a wrong value of parameter (substantially different from the optimal value), we may decelerate the convergence of the method.

Classical iterative methods from § 30.6 should never be used without preconditioning (§ 30.7). The incomplete factorization preconditioner can practically always be constructed and it requires few arithmetic operations on the whole. There are a lot of iterative methods for large sparse systems obtained from the discretization of boundary value problems for differential equations. These methods are derived with regard to the original continuous problem and are not discussed here (except for the multigrid method); see, e.g., [104]. Their special features make them, as a rule, very efficient. The multigrid method (§ 30.8) can be recommended without doubt.

To solve singular systems and systems with rectangular matrices, the application of the singular value decomposition from § 30.4 is advantageous. Numerical stability is the most important of its merits.

Any computer is equipped with some standard numerical software, which usually includes also programs or subprograms for solving linear algebraic systems. We are going to mention some more specialized program libraries and packages, most of which are written in FORTRAN.

Very good universal libraries available on the commercial basis are IMSL Library [231] and NAG Library [344] for mainframes, and personal computer programs from Numerical Recipes, the book [376] which contains all source programs in FORTRAN as well as Pascal (C is also available). High quality is the feature of specialized packages LINPACK [121] and SPARSPAK [173] (focused on sparse matrices) available directly from the authors. A great number of program implementations of various strategies in the multigrid method is the contents of the MUGTAPE84 collection (The Weizmann Institute of Science, Rehovot, Israel). The FISHPACK package [452] is the implementation of some versions of the cyclic reduction method. If the linear algebraic system solved results from the discretization of a boundary value problem for a differential equation, we can often use software for solving the boundary value problem (which includes the discretization and the solution of the system) and we thus need not be concerned with solving the system in particular.

New methods have recently been developed to be used on parallel computers (see, e.g., [461]).

In general, we can say that this part of Chap. 30 is to serve as a guide for the choice of a method when a particular problem is solved. It definitely pays to look for a suitable program in the software available. In rather exceptional cases, it is necessary to code one's own program.

B. COMPUTATION OF EIGENVALUES AND EIGENVECTORS OF MATRICES

The concepts of eigenvalue (characteristic value) and characteristic polynomial of a matrix were introduced in § 1.28. In this part of Chap. 30, we will be concerned with the computation of eigenvalues and eigenvectors (see § 30.10 for the definition) of a real square matrix. Notice that the eigenvalues may be complex even if the matrix itself is real.

Numerical methods for the computation of eigenvalues and eigenvectors are iterative in their nature. Most methods are based on the transformation (reduction) of the matrix given (§ 30.12, § 30.14) into a similar matrix (defined in § 1.28) whose eigenvalues can be calculated more easily. The reduction is usually performed as a sequence of certain elementary transformations. Some methods finally yielding the eigenvalues are presented in § 30.12 and § 30.13. Eigenvectors are computed either simultaneously with eigenvalues, or a posteriori, e.g. by the inverse iteration method (§ 30.15). Certain properties of the matrix given are often important for the choice of the method. For example, the fact that the matrix is symmetric (or Hermitian if it is complex), tridiagonal, etc. can be exploited.

Moreover, it is substantial what is the required result: We may compute, e.g., one or more eigenvalues (maximal, minimal, or lying in some interval), all eigenvalues, all eigenvalues as well as all eigenvectors, etc. The choice of a suitable method is the subject of § 30.17 where we also refer to software for computing eigenvalues and eigenvectors. Like in Part A of this chapter, we do not expect the reader to carry out the calculation for a matrix of order higher that 3 "by hand".

The topics of the bounds for eigenvalues (§ 30.10), the power method (§ 30.11), and the solution of a generalized eigenproblem (§ 30.16) are covered in this part of Chap. 30, too. All examples are of illustrative nature and are calculated without roundoff.

30.10. Bounds for Eigenvalues

The eigenproblem is defined as follows (cf. § 1.28):

Definition 1. Let $\boldsymbol{A}$ be a (complex, in general) square matrix of order n. If there exist a complex number λ and a vector $\boldsymbol{x} \neq \boldsymbol{0}$ (with complex components) such that

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}\,, \tag{1}$$

then λ is called an *eigenvalue of* $\boldsymbol{A}$ and $\boldsymbol{x}$ is an *eigenvector of* $\boldsymbol{A}$ belonging to this eigenvalue. The set of all eigenvalues of $\boldsymbol{A}$ is called the *spectrum of* $\boldsymbol{A}$.

REMARK 1. The equation (1) is equivalent to

$$(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{x} = \boldsymbol{0},$$

where $\boldsymbol{I}$ is the identity matrix of order n (Definition 30.2.1). This is a homogeneous linear algebraic system having a nonzero solution $\boldsymbol{x}$ if and only if the matrix $\boldsymbol{A} - \lambda \boldsymbol{I}$ of the system is singular, i.e. if and only if $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$ (§ 1.16, § 1.17, § 1.18). This determinant is a polynomial $p(\lambda)$ of degree n in variable λ.

Definition 2. The polynomial $p(\lambda) = \det(\boldsymbol{A} - \lambda \boldsymbol{I})$ is called the *characteristic polynomial of* $\boldsymbol{A}$ (cf. Definition 1.28.3).

REMARK 2. Each root of the characteristic polynomial is an eigenvalue of $\boldsymbol{A}$. The matrix $\boldsymbol{A}$ thus has exactly n eigenvalues (which need not be distinct). A multiple root is a multiple eigenvalue. A real matrix can also have imaginary eigenvalues. The computation of roots of the characteristic polynomial by methods of Chap. 31 can thus also serve to finding eigenvalues. This procedure is used only when the computation of the coefficients of this polynomial is not required to this end since this computation generally needs a great number of arithmetic operations (see, e.g., [378]).

The location of eigenvalues, i.e. the description of a domain in the complex plane where the eigenvalues lie, is often a useful information if we compute the roots of the characteristic polynomial or in some other situations.

Theorem 1 (*Gershgorin*). *Let* $\boldsymbol{A} = (a_{ij})$ *be a (complex, in general) square matrix of order* n. *For* $i = 1, \ldots, n$, *denote by* C_i *the disk with center at the point* a_{ii} *of the complex plane and with radius*

$$r_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| .$$

Then all eigenvalues of $\boldsymbol{A}$ *lie in the domain* D *that is the union of the Gershgorin disks* C_i, $i = 1, \ldots, n$.

Example 1. Two Gershgorin disks are associated with the matrix

$$\boldsymbol{A} = \begin{bmatrix} 1, & 2 \\ -1, & 1 \end{bmatrix}$$

from Example 1.28.1. One of them has center at the point 1 and radius 2, the other has center also at the point 1 but radius 1. Their union D is thus the larger of the two disks. We can easily verify that both the eigenvalues $\lambda_1 = 1 + \mathrm{i}\sqrt{2}$ and $\lambda_2 = 1 - \mathrm{i}\sqrt{2}$ lie in this disk.

REMARK 3. Theorem 1 also implies the bound for the spectral radius of the matrix,

$$\varrho(\boldsymbol{A}) \leqq \max_i \sum_{j=1}^{n} |a_{ij}| .$$

Note that $|\lambda_i| \leqq \varrho(\boldsymbol{A})$ holds for all eigenvalues λ_i of $\boldsymbol{A}$ by Definition 30.3.1.

Example 2. For the spectral radius of the matrix $\boldsymbol{A}$ from Example 1, we have the bound $\varrho(\boldsymbol{A}) \leqq \max(3, 2) = 3$. (Notice that $\varrho(\boldsymbol{A}) = |\lambda_1| = \sqrt{3}$ follows from Example 1.)

Theorem 2. *Let $\boldsymbol{A}$ be a square matrix. Then*

$$\varrho(\boldsymbol{A}) \leqq \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} ,$$

where $\|\boldsymbol{A}\boldsymbol{x}\|$ is norm of the vector $\boldsymbol{A}\boldsymbol{x}$.

Theorem 3. *Let $\boldsymbol{A}$ be a real positive definite matrix* (Definition 30.1.2). *Then*

$$\varrho(\boldsymbol{A}) = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}} \boldsymbol{x}} , \qquad \frac{1}{\varrho(\boldsymbol{A}^{-1})} = \min_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}} \boldsymbol{x}} .$$

30.11. Power Method

In this paragraph, we will present a method for computing the eigenvalue largest in magnitude. First we are going to recall some general concepts (needed also in the following paragraphs) and to present some properties of eigenvalues and eigenvectors.

REMARK 1. Two complex square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same order are called *similar* if there exists a nonsingular matrix $\boldsymbol{P}$ such that $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$ (Definition 1.28.1). Similar matrices possess the same eigenvalues (Theorem 1.28.5). If $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$ belonging to the eigenvalue λ, then $\boldsymbol{P}\boldsymbol{x}$ is an eigenvector of $\boldsymbol{B}$ belonging to the same eigenvalue.

Every complex square matrix $\boldsymbol{A}$ of order n is similar to a Jordan matrix $\boldsymbol{J}$ of order n (Theorem 1.28.8, Examples 1.28.2 to 1.28.4), where

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{J}_1 & & \\ & \ddots & \\ & & \boldsymbol{J}_s \end{bmatrix} \tag{1}$$

is a block diagonal matrix and each its diagonal block $\boldsymbol{J}_i$ (called the *Jordan block*, Definition 1.28.4) is an upper triangular matrix

$$\boldsymbol{J}_i = \begin{bmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{bmatrix}$$

having the eigenvalue λ_i on the diagonal and 1's above the diagonal, while the other entries of the Jordan block $\boldsymbol{J}_i$ are zero. An eigenvalue can appear in more Jordan blocks.

REMARK 2. While the number of eigenvalues of a matrix $\boldsymbol{A}$ of order n equals n (some of them may, however, coincide, i.e., be multiple) by Remark 30.10.2, the number of linearly independent eigenvectors of this matrix equals the number s from (1), $s \leqq n$. It is $s = n$ if and only if all the Jordan blocks are of order 1 and the corresponding matrix $\boldsymbol{J}$ is thus diagonal. Such a matrix $\boldsymbol{A}$ is called *nondefective*. In this case, all the eigenvectors form a *basis of an n-dimensional vector space* (a *complete system of eigenvectors*). This is true, for example, for a real symmetric matrix.

Definition 1. Let $|\lambda_1| \geqq |\lambda_2| \geqq \cdots \geqq |\lambda_n|$ hold for the eigenvalues λ_i of a matrix $\boldsymbol{A}$ of order n. If $|\lambda_1| > |\lambda_2|$, then λ_1 is called the *dominant eigenvalue*.

We will now present the *power method*. Let λ_1 be the dominant eigenvalue of $\boldsymbol{A}$. Choose a vector $\boldsymbol{v}_0$ and put

$$\boldsymbol{v}_{m+1} = \frac{1}{c_{m+1}} \boldsymbol{A}\boldsymbol{v}_m , \tag{2}$$

where the number c_{m+1} equals that component of the vector $\boldsymbol{A}\boldsymbol{v}_m$ which is maximal in magnitude. For a proper choice of $\boldsymbol{v}_0$ (see, e.g., [497]), we have

$$\lim_{m \to \infty} c_m = \lambda_1 \tag{3}$$

and the sequence of vectors $\boldsymbol{v}_m$ converges to the eigenvector $\boldsymbol{x}_1$.

The vector $\boldsymbol{v}_m$ is equal — apart from a number factor — to the product of the m-th power of $\boldsymbol{A}$ and $\boldsymbol{v}_0$ by (2). The name of the method comes from this fact.

REMARK 3. The normalization factor $1/c_{m+1}$ in (2) is necessary since the computation may otherwise end with overflow in several steps. Some other choices of c_{m+1} preserving the convergence of (3) as well as some other modifications of the formula (2) are also possible (see, e.g., [378]). In practice, the rather complicated conditions for convergence are, as a rule, not being verified but the behaviour of the iterative process itself indicates whether the method converges or not.

REMARK 4. We can see from (2) that the power method (as well as its other versions) exploits the matrix $\boldsymbol{A}$ given only for multiplying a vector by it. The method is thus both suitable and efficient for sparse matrices (§ 30.5) since it saves arithmetic operations as well as computer storage.

REMARK 5. The power method can be used to calculate a further (subdominant) eigenvalue and an eigenvector belonging to it if, reducing the matrix $\boldsymbol{A}$ given, we construct a matrix $\boldsymbol{W}$ possessing the same eigenvalues as $\boldsymbol{A}$ except for λ_1, which is replaced by zero. Various constructions of $\boldsymbol{W}$ are given, e.g., in [378]. The procedure can be repeated recurrently to calculate further and further eigenvalues of $\boldsymbol{A}$.

30.12. Jacobi Method

The *Jacobi method* is an iterative procedure for the calculation of all eigenvalues and eigenvectors of a real symmetric matrix. We will first present certain properties of real symmetric matrices.

Theorem 1. *Let $\boldsymbol{A}$ be a real symmetric matrix. Then its eigenvalues are also real and r eigenvectors belong to an eigenvalue of multiplicity r* (cf. Theorem 1.28.10).

Theorem 2. *Let $\boldsymbol{A}$ be a real symmetric matrix. Then there exist an orthogonal matrix $\boldsymbol{Q}$* (Definition 1.25.7) *and a diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^{\mathrm{T}}$. $\boldsymbol{D}$ is the Jordan matrix of $\boldsymbol{A}$, its diagonal entries d_i are the eigenvalues of $\boldsymbol{A}$, and the i-th column of $\boldsymbol{Q}$ is the eigenvector $\boldsymbol{x}_i$ belonging to the eigenvalue d_i* (cf. Theorem 1.28.12).

REMARK 1. Since $\boldsymbol{Q}^{\mathrm{T}} = \boldsymbol{Q}^{-1}$ by definition of the orthogonal matrix, the matrices $\boldsymbol{A}$ and $\boldsymbol{D}$ are similar (Remark 30.11.1).

Eigenvalues and eigenvectors of a real symmetric matrix $\boldsymbol{A}$ can be found as follows: We construct a sequence of orthogonal matrices $\boldsymbol{R}_k$ such that the sequence

$$\boldsymbol{T}_0 = \boldsymbol{A}, \quad \boldsymbol{T}_k = \boldsymbol{R}_k^{\mathrm{T}} \boldsymbol{A} \boldsymbol{R}_k, \quad k = 1, 2, \ldots, \tag{1}$$

converges to a diagonal matrix $\boldsymbol{D}$. The eigenvalues of $\boldsymbol{A}$ are then on the diagonal of $\boldsymbol{D}$ and matrices $\boldsymbol{R}_k$ converge to an orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{A}$ (Theorem 2).

We will construct the matrix $\boldsymbol{R}_k$ successively as the product of orthogonal matrices

$$\boldsymbol{R}_k = \boldsymbol{S}_1 \boldsymbol{S}_2 \ldots \boldsymbol{S}_k, \tag{2}$$

i.e.,

$$\boldsymbol{R}_0 = \boldsymbol{I}\,, \quad \boldsymbol{R}_k = \boldsymbol{R}_{k-1}\boldsymbol{S}_k\,, \quad k = 1, 2, \ldots,$$

because a product of orthogonal matrices is again an orthogonal matrix.

Since our goal is to reduce $\boldsymbol{A}$ to diagonal form, we choose $\boldsymbol{S}_k$ in the k-th step in such a way that some nonzero off-diagonal entry of $\boldsymbol{T}_{k-1}$ is transformed into a zero entry of $\boldsymbol{T}_k$. All matrices $\boldsymbol{T}_k$ are similar to $\boldsymbol{A}$ by (1) (Remark 1) and their eigenvalues thus coincide.

Put $\boldsymbol{T}_{k-1} = (t_{ij}^{(k-1)})$ and $\boldsymbol{T}_k = (t_{ij}^{(k)})$. Wishing to replace the entry $t_{pq}^{(k-1)}$ by zero, we put

$$\lambda = -t_{pq}^{(k-1)}\,, \quad \mu = \tfrac{1}{2}(t_{pp}^{(k-1)} - t_{qq}^{(k-1)})\,, \quad \nu = (\lambda^2 + \mu^2)^{1/2} \tag{3}$$

and determine the angle θ in such a way that

$$\cos\theta = \left(\frac{\nu + |\mu|}{2\nu}\right)^{1/2}\,, \quad \sin\theta = \frac{\lambda \operatorname{sgn}\mu}{2\nu\cos\theta}\,. \tag{4}$$

The angle θ itself need not be calculated as only its sine and cosine appear in the following formulae.

The entries of the orthogonal matrix $\boldsymbol{S}_k = (s_{ij}^{(k)})$ are given by

$$\begin{aligned} s_{pp}^{(k)} &= s_{qq}^{(k)} = \cos\theta\,, \\ s_{pq}^{(k)} &= -s_{qp}^{(k)} = \sin\theta\,, \\ s_{ii}^{(k)} &= 1 \quad \text{for } i \neq p \text{ and } i \neq q\,, \\ s_{ij}^{(k)} &= 0 \quad \text{for other } i \text{ and } j\,. \end{aligned} \tag{5}$$

$\boldsymbol{S}_k$ is called the *plane rotation matrix*. It differs from the identity matrix only in the p-th and q-th rows and columns,

$$\boldsymbol{S}_k = \begin{bmatrix} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & \cos\theta & & \ldots & & \sin\theta & & & \\ & & & & 1 & & & & & & \\ & & & \vdots & & \ddots & & \vdots & & & \\ & & & & & & 1 & & & & \\ & & & -\sin\theta & & \ldots & & \cos\theta & & & \\ & & & & & & & & 1 & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{bmatrix} \tag{6}$$

(where $p < q$). A straightforward calculation of the product $\boldsymbol{S}_k^{\mathrm{T}} \boldsymbol{S}_k$ really shows that $\boldsymbol{S}_k$ is orthogonal. Simultaneously with $t_{pq}^{(k-1)}$, the entry $t_{qp}^{(k-1)}$ is also replaced by zero in the transformation.

We can readily obtain formulae [378] that directly express the entries of

$$\boldsymbol{T}_k = \boldsymbol{S}_k^{\mathrm{T}} \boldsymbol{T}_{k-1} \boldsymbol{S}_k \tag{7}$$

by means of the indices p and q chosen ($p \neq q$), the quantity θ, and the entries of $\boldsymbol{T}_{k-1}$. In this transformation, only the p-th and q-th rows and columns change while

$$t_{ij}^{(k)} = t_{ij}^{(k-1)} \quad \text{for } i \neq p,\ i \neq q,\ j \neq p \quad \text{and} \quad j \neq q\,. \tag{8}$$

Since both $\boldsymbol{T}_{k-1}$ and $\boldsymbol{T}_k$ are symmetric matrices, it suffices to calculate only the entries of one triangle of $\boldsymbol{T}_k$. If we are also interested in eigenvectors, we have to compute the entries of $\boldsymbol{R}_k$ (by means of p and q, θ, and the entries of $\boldsymbol{R}_{k-1}$), too. All these formulae can easily be derived; they are given, e.g., in [378] and a special version minimizing the influence of roundoff is presented in [376].

We have thus shown how to perform a single step of the Jacobi method. An important task is now to find a strategy that guarantees the convergence of matrices $\boldsymbol{T}_k$ to the diagonal matrix $\boldsymbol{D}$. A quite natural strategy is to take the largest in magnitude off-diagonal entry of $\boldsymbol{T}_{k-1}$ for the entry $t_{pq}^{(k-1)}$ to be annihilated. But the entries replaced by zero in the previous steps may again become nonzero in a current step. The process is thus infinite and requires a new search for a suitable entry in every step.

The *threshold (cyclic) Jacobi method* is more efficient from the point of view of arithmetic operations. It consists in fixing a positive number (threshold), systematically and repeatedly searching through all off-diagonal entries, and successively annihilating those which are greater that the threshold chosen. This strategy is discussed in detail e.g. in [376].

Example 1. Carry out a step of the Jacobi method with the symmetric matrix

$$\boldsymbol{T}_0 = \boldsymbol{A} = \begin{bmatrix} 16, & 1, & -24 \\ 1, & 9, & 0 \\ -24, & 0, & 2 \end{bmatrix}.$$

The largest in magnitude off-diagonal entry of $\boldsymbol{T}_0$ is $t_{13}^{(0)} = t_{31}^{(0)} = -24$. We thus put $p = 1$, $q = 3$ and construct an orthogonal matrix $\boldsymbol{S}_1$ to annihilate the corresponding entry of the matrix $\boldsymbol{T}_1$ calculated by (7).

According to (3) and (4), we obtain $\lambda = 24$, $\mu = 7$, $\nu = 25$, $\cos\theta = 0{\cdot}8$ and $\sin\theta = 0{\cdot}6$. Finally,

$$\boldsymbol{S}_1 = \begin{bmatrix} 0{\cdot}8, & 0{\cdot}0, & 0{\cdot}6 \\ 0{\cdot}0, & 1{\cdot}0, & 0{\cdot}0 \\ -0{\cdot}6, & 0{\cdot}0, & 0{\cdot}8 \end{bmatrix}$$

by (5). Multiplying the matrices in (7), we have

$$\boldsymbol{T}_1 = \boldsymbol{S}_1^{\mathrm{T}} \boldsymbol{T}_0 \boldsymbol{S}_1 = \begin{bmatrix} 34{\cdot}0, & 0{\cdot}8, & 0{\cdot}0 \\ 0{\cdot}8, & 9{\cdot}0, & 0{\cdot}6 \\ 0{\cdot}0, & 0{\cdot}6, & -16{\cdot}0 \end{bmatrix} .$$

Really, the entries $t_{13}^{(1)}$ and $t_{31}^{(1)}$ are zero. We can easily verify (8), which states in our case that $t_{22}^{(1)} = t_{22}^{(0)} = 9$. The entries of the first and third rows and columns have changed and the zero entries $t_{23}^{(0)} = t_{32}^{(0)}$ have been replaced by nonzero entries $t_{23}^{(1)} = t_{32}^{(1)} = 0{\cdot}6$. The sum of squares of the diagonal entries has increased. Since the order of $\boldsymbol{A}$ is very low ($n = 3$) we obtained numbers very close to the eigenvalues of $\boldsymbol{A}$ on the diagonal of $\boldsymbol{T}_1$ already after the first step of the Jacobi method. (The eigenvalues of $\boldsymbol{A}$ are (rounded) 34·03, 8·99 and −16·01.)

30.13. LR and QR Methods

In this paragraph, we will discuss two iterative methods for calculating eigenvalues and eigenvectors of a general square matrix and show classes of matrices for which these methods are efficient. The reduction of a matrix to the form suitable for the application of the LR or QR methods is studied in the next paragraph.

We begin with the *LR method.* Putting $\boldsymbol{A}_1 = \boldsymbol{A}$, we construct a sequence of matrices $\boldsymbol{A}_k$ in such a way that we find the LU factorization of $\boldsymbol{A}$ in accord with § 30.1, i.e.

$$\boldsymbol{A}_k = \boldsymbol{L}_k \boldsymbol{U}_k , \tag{1}$$

and put

$$\boldsymbol{A}_{k+1} = \boldsymbol{U}_k \boldsymbol{L}_k , \quad k = 1, 2, \ldots . \tag{2}$$

No pivoting can be carried out in the LU factorization now and the existence of the factorization (1) is thus not guaranteed for a general matrix $\boldsymbol{A}$ (cf. Remarks 30.1.2 and 30.1.4). If the factorization does exist, then $\boldsymbol{A}_k$ and $\boldsymbol{A}_{k+1}$ are similar since $\boldsymbol{L}_k^{-1} \boldsymbol{A}_k \boldsymbol{L}_k = \boldsymbol{L}_k^{-1} \boldsymbol{L}_k \boldsymbol{U}_k \boldsymbol{L}_k = \boldsymbol{A}_{k+1}$ by virtue of (1) and (2). If the sequence of matrices $\boldsymbol{A}_k$ can be constructed, if $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$ holds, and if some further conditions are fulfilled (see, e.g., [497]), then the matrices $\boldsymbol{A}_k$ tend to an upper triangular matrix $\boldsymbol{U}$ having the eigenvalues of $\boldsymbol{A}$ on its diagonal. (The matrices $\boldsymbol{U}_k$ converge to the same matrix $\boldsymbol{U}$ and $\boldsymbol{L}_k$ converge to the identity matrix.)

The term *LR factorization* is traditionally used instead of LU factorization when the calculation of eigenvalues is involved. The letters L and R stand for the left and right (i.e. lower and upper) triangular matrices and appear also in the name of the method just discussed.

REMARK 1. The LR method converges for a very large class of matrices containing e.g. positive definite matrices (Definition 30.1.2). The method can be modified for these matrices in such a way that the LU factorization is replaced by the Choleski factorization (Remark 30.1.7).

Example 1. We will show a single step of the LR method on a simple example. We have

$$\boldsymbol{A}_1 = \boldsymbol{A} = \begin{bmatrix} 9, & 3, & 0 \\ 6, & 8, & 3 \\ -3, & -3, & 6 \end{bmatrix} = \begin{bmatrix} 1, & 0, & 0 \\ \frac{2}{3}, & 1, & 0 \\ -\frac{1}{3}, & -\frac{1}{3}, & 1 \end{bmatrix} \begin{bmatrix} 9, & 3, & 0 \\ 0, & 6, & 3 \\ 0, & 0, & 7 \end{bmatrix} = \boldsymbol{L}_1 \boldsymbol{U}_1$$

by Example 30.1.2. According to (2), we compute

$$\boldsymbol{U}_1 \boldsymbol{L}_1 = \begin{bmatrix} 9, & 3, & 0 \\ 0, & 6, & 3 \\ 0, & 0, & 7 \end{bmatrix} \begin{bmatrix} 1, & 0, & 0 \\ \frac{2}{3}, & 1, & 0 \\ -\frac{1}{3}, & -\frac{1}{3}, & 1 \end{bmatrix} = \begin{bmatrix} 11, & 3, & 0 \\ 3, & 5, & 3 \\ -\frac{7}{3}, & -\frac{7}{3}, & 7 \end{bmatrix} = \boldsymbol{A}_2 .$$

It would be premature in this case to draw conclusions on convergence from a single step of the iterative method. We, however, can check up that our calculation was correct. The matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ have the same characteristic polynomial $p(\lambda) = -\lambda^3 + 23\lambda^2 - 165\lambda + 378$.

We will discuss the efficiency of the LR method later. Now we will present another procedure, the *QR method*, which is based on a different factorization of the matrix $\boldsymbol{A}$ given.

Theorem 1. *To any real square matrix* $\boldsymbol{A}$*, there exist an orthogonal matrix* $\boldsymbol{Q}$ (Definition 1.25.7) *and an upper triangular matrix* $\boldsymbol{U}$ (*both of the same order as* $\boldsymbol{A}$) *such that*

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{U} . \tag{3}$$

The principal idea of the QR method is analogous to the previous method. We put $\boldsymbol{A}_1 = \boldsymbol{A}$ and construct a sequence of matrices $\boldsymbol{A}_k$ in such a way that we find the factorization (3) of $\boldsymbol{A}_k$, i.e.

$$\boldsymbol{A}_k = \boldsymbol{Q}_k \boldsymbol{U}_k , \tag{4}$$

with an orthogonal matrix $\boldsymbol{Q}_k$ and an upper triangular matrix $\boldsymbol{U}_k$. Further we put

$$\boldsymbol{A}_{k+1} = \boldsymbol{U}_k \boldsymbol{Q}_k , \quad k = 1, 2, \ldots . \tag{5}$$

The matrices $\boldsymbol{A}_k$ and $\boldsymbol{A}_{k+1}$ are similar since $\boldsymbol{Q}_k^{-1}\boldsymbol{A}_k\boldsymbol{Q}_k = \boldsymbol{Q}_k^{-1}\boldsymbol{Q}_k\boldsymbol{U}_k\boldsymbol{Q}_k = \boldsymbol{A}_{k+1}$ by (4) and (5).

The QR method converges for classes of matrices analogous to those for which the LR method does (see, e.g., [497]). The matrices $\boldsymbol{A}_k$ converge to an upper triangular matrix $\boldsymbol{U}$ having the eigenvalues of $\boldsymbol{A}$ on its diagonal. (The matrices $\boldsymbol{U}_k$ tend to the same matrix $\boldsymbol{U}$.)

The formula (4) is called the *QR factorization* where the letter R stands for the right (i.e. upper) triangular matrix. Analogously to Theorem 1, there also exists the factorization $\boldsymbol{A} = \boldsymbol{QL}$, where $\boldsymbol{Q}$ is an orthogonal and $\boldsymbol{L}$ a left (i.e. lower) triangular matrix. The *QL method* can thus be derived, too.

We will further show a suitable algorithm for constructing the QR factorization (3). It is based on the same idea as the construction of orthogonal matrices $\boldsymbol{S}_k$ in the Jacobi method (§ 30.12).

We rewrite (3) in the form $\boldsymbol{U} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}$ and construct the orthogonal matrix $\boldsymbol{Q}^{\mathrm{T}}$ as the product

$$\boldsymbol{Q}^{\mathrm{T}} = \boldsymbol{M}_{n-1}\boldsymbol{M}_{n-2}\dots\boldsymbol{M}_1\,, \tag{6}$$

$$\boldsymbol{M}_l = \boldsymbol{S}_{l,l+1}\dots\boldsymbol{S}_{ln}\,, \quad l = 1, \dots, n-1\,, \tag{7}$$

where $\boldsymbol{S}_{pq} = (s_{ij}^{(pq)})$ is an orthogonal matrix and n is the order of $\boldsymbol{A}$. Denote by

$$\boldsymbol{W} = \boldsymbol{S}_{pq}\boldsymbol{V} \tag{8}$$

the transformation of a matrix $\boldsymbol{V}$ performed by $\boldsymbol{S}_{pq}$ and annihilating the entry v_{qp}. Put

$$\begin{aligned} \tan\theta &= \frac{v_{qp}}{v_{pp}} \quad \text{for } v_{pp} \neq 0\,, \\ \theta &= \frac{\pi}{2} \quad \text{for } v_{pp} = 0\,, \end{aligned} \tag{9}$$

$$\begin{aligned} s_{pp}^{(pq)} &= s_{qq}^{(pq)} = \cos\theta\,, \\ s_{pq}^{(pq)} &= -s_{qp}^{(pq)} = \sin\theta\,, \\ s_{ii}^{(pq)} &= 1 \quad \text{for } i \neq p \text{ and } i \neq q\,, \\ s_{ij}^{(pq)} &= 0 \quad \text{for other } i \text{ and } j\,. \end{aligned} \tag{10}$$

The matrix $\boldsymbol{S}_{pq}$ is of the form (30.12.6) and only the p-th and q-th rows of $\boldsymbol{V}$ are changed by the multiplication in (8). No entry that has been replaced by zero can become nonzero in the further course of the computation.

Example 2. Let the matrix

$$\boldsymbol{A} = \begin{bmatrix} 9{\cdot}6, & -12{\cdot}0, & -3{\cdot}4 \\ 7{\cdot}2, & 1{\cdot}0, & 6{\cdot}2 \\ 0{\cdot}0, & 6{\cdot}0, & -1{\cdot}0 \end{bmatrix}$$

be given. Compute its QR factorization (3). According to (6) and (7), we put $\boldsymbol{Q}^{\mathrm{T}} = \boldsymbol{S}_{23}\boldsymbol{S}_{12}$ as $a_{31} = 0$. For $\boldsymbol{S}_{12}$, the quantity θ in (10) is determined by (9), i.e. $\tan\theta = a_{21}/a_{11} = 7{\cdot}2/9{\cdot}6 = 0{\cdot}75$. From this, we obtain an intermediate result, the product

$$\boldsymbol{S}_{12}\boldsymbol{A} = \begin{bmatrix} 0{\cdot}8, & 0{\cdot}6, & 0{\cdot}0 \\ -0{\cdot}6, & 0{\cdot}8, & 0{\cdot}0 \\ 0{\cdot}0, & 0{\cdot}0, & 1{\cdot}0 \end{bmatrix} \boldsymbol{A} = \begin{bmatrix} 12, & -9, & 1 \\ 0, & 8, & 7 \\ 0, & 6, & -1 \end{bmatrix} = \tilde{\boldsymbol{A}}.$$

Similarly we get $\tan\theta = \tilde{a}_{32}/\tilde{a}_{22} = 6/8 = 0{\cdot}75$ for $\boldsymbol{S}_{23}$ and

$$\boldsymbol{S}_{23} = \begin{bmatrix} 1{\cdot}0, & 0{\cdot}0, & 0{\cdot}0 \\ 0{\cdot}0, & 0{\cdot}8, & 0{\cdot}6 \\ 0{\cdot}0, & -0{\cdot}6, & 0{\cdot}8 \end{bmatrix}.$$

Finally we compute the products $\boldsymbol{U} = \boldsymbol{S}_{23}\boldsymbol{S}_{12}\boldsymbol{A} = \boldsymbol{S}_{23}\tilde{\boldsymbol{A}}$ and $\boldsymbol{Q} = \boldsymbol{S}_{12}^{\mathrm{T}}\boldsymbol{S}_{23}^{\mathrm{T}}$ and arrive at

$$\boldsymbol{A} = \begin{bmatrix} 0{\cdot}80, & -0{\cdot}48, & 0{\cdot}36 \\ 0{\cdot}60, & 0{\cdot}64, & -0{\cdot}48 \\ 0{\cdot}00, & 0{\cdot}60, & 0{\cdot}80 \end{bmatrix} \begin{bmatrix} 12, & -9, & 1 \\ 0, & 10, & 5 \\ 0, & 0, & -5 \end{bmatrix} = \boldsymbol{QU}, \tag{11}$$

i.e. the factorization of the matrix $\boldsymbol{A}$ given into an orthogonal matrix $\boldsymbol{Q}$ and an upper triangular matrix $\boldsymbol{U}$.

Definition 1. The square matrix

$$\begin{bmatrix} a_{11}, & a_{12}, & a_{13}, & \dots, & a_{1,n-2}, & a_{1,n-1}, & a_{1n} \\ a_{21}, & a_{22}, & a_{23}, & \dots, & a_{2,n-2}, & a_{2,n-1}, & a_{2n} \\ 0, & a_{32}, & a_{33}, & \dots, & a_{3,n-2}, & a_{3,n-1}, & a_{3n} \\ 0, & 0, & a_{43}, & \dots, & a_{4,n-2}, & a_{4,n-1}, & a_{4n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0, & 0, & 0, & \dots, & a_{n-1,n-2}, & a_{n-1,n-1}, & a_{n-1,n} \\ 0, & 0, & 0, & \dots, & 0, & a_{n,n-1}, & a_{nn} \end{bmatrix}$$

(which differs from an upper triangular matrix in one, in general nonzero, diagonal below the main diagonal) is said to be in *upper Hessenberg form.* The concept of matrix in *lower Hessenberg form* is introduced analogously.

REMARK 2. Both the LR and the QR methods require of order n^3 arithmetic operations in each step if $\boldsymbol{A}$ is a full square matrix of order n. When $\boldsymbol{A}$ is a bandmatrix (e.g. a tridiagonal matrix, see Definitions 30.5.3 and 30.5.1), then all matrices $\boldsymbol{A}_k$

are bandmatrices (or tridiagonal matrices) as well. Also, if $\boldsymbol{A}$ is in upper Hessenberg form, then all matrices $\boldsymbol{A}_k$ are in this form. Consequently, both the LR and the QR methods require now, in each step, a number of operations which is of order less than n^3. Some procedures for reducing a general matrix to a similar tridiagonal matrix or a similar matrix in Hessenberg form in a finite number of steps will be presented in § 30.14. Special modifications of the LR and QR methods for tridiagonal matrices are given e.g. in [378].

REMARK 3. The matrix $\boldsymbol{A} = \boldsymbol{A}_1$ from Example 1 is in lower Hessenberg form. After one step of the LR method, the matrix $\boldsymbol{A}_2$ is in the same form.

Example 3. We will show a single step of the QR method for the matrix $\boldsymbol{A} = \boldsymbol{A}_1$ from Example 2. We can exploit the QR factorization (11) of $\boldsymbol{A}$. Denoting the factors by $\boldsymbol{Q}_1$ and $\boldsymbol{U}_1$, i.e. $\boldsymbol{A}_1 = \boldsymbol{Q}_1 \boldsymbol{U}_1$, we compute by (5)

$$\boldsymbol{U}_1 \boldsymbol{Q}_1 = \begin{bmatrix} 12, & -9, & 1 \\ 0, & 10, & 5 \\ 0, & 0, & -5 \end{bmatrix} \begin{bmatrix} 0{\cdot}80, & -0{\cdot}48, & 0{\cdot}36 \\ 0{\cdot}60, & 0{\cdot}64, & -0{\cdot}48 \\ 0{\cdot}00, & 0{\cdot}60, & 0{\cdot}80 \end{bmatrix} =$$

$$= \begin{bmatrix} 4{\cdot}20, & -10{\cdot}92, & 9{\cdot}44 \\ 6{\cdot}00, & 9{\cdot}40, & -0{\cdot}80 \\ 0{\cdot}00, & -3{\cdot}00, & -4{\cdot}00 \end{bmatrix} = \boldsymbol{A}_2 .$$

Note that the matrix $\boldsymbol{A} = \boldsymbol{A}_1$ is in upper Hessenberg form. This form is preserved by the transformation and $\boldsymbol{A}_2$ is in the same form.

REMARK 4. The simultaneous computation of eigenvectors together with eigenvalues can be carried out (at the cost of additional operations whose number is again of order n^3) in both the LR and QR methods in the way which follows from Remark 30.11.1.

Theorem 2. *Let the matrix $\boldsymbol{A}$ have an eigenvalue λ and let k be an arbitrary (complex) number. Then the matrix $\boldsymbol{A} - k\boldsymbol{I}$ has the eigenvalue $\lambda - k$.*

REMARK 5. Theorem 2 can be used to accelerate the convergence of both LR and QR methods if we choose a proper number k and factorize the matrix $\boldsymbol{A} - k\boldsymbol{I}$ instead of $\boldsymbol{A}$. This procedure is called the *spectrum shifting.*

30.14. Reducing Matrices to Simpler Forms. The Givens and Householder Methods. The Lanczos and Wilkinson Methods

We presented the Jacobi method, that reduces a given symmetric matrix to a similar diagonal matrix, in § 30.12. Its disadvantage consists in the fact that it is

an iterative method. We will now show some procedures that reduce a given matrix to a similar matrix in simpler form in a finite number of steps. These procedures are important, first of all, as a preparatory phase of the computation of eigenvalues and eigenvectors, which is followed by the application of the LR or QR methods from § 30.13.

The *Givens method* is based on the same idea as the Jacobi method. Let $\boldsymbol{A}$ be a real symmetric matrix, $\boldsymbol{T}_0 = \boldsymbol{A}$. Like in (30.12.1) and (30.12.2), we construct the sequence of matrices

$$\boldsymbol{T}_k = \boldsymbol{S}_k^{\mathrm{T}} \boldsymbol{T}_{k-1} \boldsymbol{S}_k \,, \tag{1}$$

where $\boldsymbol{S}_k$ is an orthogonal matrix. In § 30.12, we chose $\boldsymbol{S}_k$ in the form (30.12.6) and θ in such a way that the entry $t_{pq}^{(k-1)}$ of $\boldsymbol{T}_{k-1} = (t_{ij}^{(k-1)})$ was annihilated. The matrices $\boldsymbol{T}_k$ and $\boldsymbol{T}_{k-1}$ differed only in the p-th and q-th rows and columns.

Choosing θ in (30.12.6) such that

$$\cos\theta = \alpha t_{rp}^{(k-1)} \,, \quad \sin\theta = -\alpha t_{rq}^{(k-1)} \,, \tag{2}$$

where

$$\alpha^{-1} = \left((t_{rp}^{(k-1)})^2 + (t_{rq}^{(k-1)})^2 \right)^{1/2} \,, \quad r \neq p \,, \quad r \neq q \,, \tag{3}$$

$t_{rq}^{(k-1)}$ is replaced by zero in the transformation (1) (and $t_{qr}^{(k-1)}$ as well due to symmetry). The matrices $\boldsymbol{T}_k$ and $\boldsymbol{T}_{k-1}$ differ only in the p-th and q-th rows and columns.

The orthogonal matrix $\boldsymbol{S}_k$ of the form (30.12.6) defined by (2) and (3) is thus determined by three indices p, q and r. Let us denote it by $\boldsymbol{S}_{pqr}$. If the transformations (1) employing matrices $\boldsymbol{S}_{pqr}$ are performed for a suitable choice of the indices p, q and r and in proper order, all the annihilated entries of $\boldsymbol{A}$ remain zero (in contrast to the Jacobi method). We arrive at a symmetric tridiagonal matrix $\boldsymbol{T}_N$ (but not a diagonal matrix) similar to $\boldsymbol{A}$ after a finite number N of the transformations (1).

Put thus $r = p - 1$, choose successively $p = 2, \ldots, n-1$, and, for each p, carry out the transformation (1) successively for $q = p+1, \ldots, n$. In this way, we finally obtain a symmetric tridiagonal matrix to which the methods of § 30.13 can efficiently be applied. Wishing to calculate also eigenvectors, we have to determine the product of all the transformation matrices $\boldsymbol{S}_{pqr}$ (cf. Remark 30.11.1) simultaneously with the reduction of $\boldsymbol{A}$.

Example 1. Use the Givens method to reduce a given symmetric full matrix $\boldsymbol{A} = \boldsymbol{T}_0$ of order 4 to tridiagonal form. In the first step, the formula (1) determines the matrix $\boldsymbol{T}_1 = \boldsymbol{S}_{231}^{\mathrm{T}} \boldsymbol{T}_0 \boldsymbol{S}_{231}$ with zero entries $t_{31}^{(1)}$ and $t_{13}^{(1)}$. The matrix

$\boldsymbol{T}_2 = \boldsymbol{S}_{241}^{\mathrm{T}} \boldsymbol{T}_1 \boldsymbol{S}_{241}$ from the second step has, in addition, zero entries $t_{41}^{(2)}$ and $t_{14}^{(2)}$. Let us have, e.g.,

$$\boldsymbol{T}_2 = \begin{bmatrix} 6, & 2, & 0, & 0 \\ 2, & -5, & 4, & 3 \\ 0, & 4, & 5, & 1 \\ 0, & 3, & 1, & 8 \end{bmatrix}.$$

The last step of the reduction,

$$\boldsymbol{T}_3 = \boldsymbol{S}_{342}^{\mathrm{T}} \boldsymbol{T}_2 \boldsymbol{S}_{342}\,, \tag{4}$$

remains to be performed. Its result will be $t_{42}^{(3)} = t_{24}^{(3)} = 0$. We thus have $p = 3$, $q = 4$ and $r = 2$, and find $\alpha^{-1} = \left((t_{23}^{(2)})^2 + (t_{24}^{(2)})^2\right)^{1/2} = (4^2 + 3^2)^{1/2} = 5$, $\cos\theta = \alpha t_{23}^{(2)} = 0{\cdot}8$, and $\sin\theta = -\alpha t_{24}^{(2)} = -0{\cdot}6$ from (2) and (3). The formula (30.12.6) now implies

$$\boldsymbol{S}_{342} = \begin{bmatrix} 1{\cdot}0, & 0{\cdot}0, & 0{\cdot}0, & 0{\cdot}0 \\ 0{\cdot}0, & 1{\cdot}0, & 0{\cdot}0, & 0{\cdot}0 \\ 0{\cdot}0, & 0{\cdot}0, & 0{\cdot}8, & -0{\cdot}6 \\ 0{\cdot}0, & 0{\cdot}0, & 0{\cdot}6, & 0{\cdot}8 \end{bmatrix}$$

and the result of multiplication in (4) is

$$\boldsymbol{T}_3 = \begin{bmatrix} 6{\cdot}00, & 2{\cdot}00, & 0{\cdot}00, & 0{\cdot}00 \\ 2{\cdot}00, & -5{\cdot}00, & 5{\cdot}00, & 0{\cdot}00 \\ 0{\cdot}00, & 5{\cdot}00, & 7{\cdot}04, & 1{\cdot}72 \\ 0{\cdot}00, & 0{\cdot}00, & 1{\cdot}72, & 5{\cdot}96 \end{bmatrix}.$$

The matrix $\boldsymbol{T}_3$ is tridiagonal.

The *Householder method* is used to treat the same problem, i.e., to reduce a real symmetric matrix $\boldsymbol{A}$ to a similar tridiagonal matrix. We put $\boldsymbol{A}_1 = \boldsymbol{A}$ and construct the sequence of matrices

$$\boldsymbol{A}_k = \boldsymbol{P}_k \boldsymbol{A}_{k-1} \boldsymbol{P}_k\,, \quad k = 2, \ldots, n-1\,, \tag{5}$$

where

$$\boldsymbol{P}_k = \boldsymbol{I} - 2\boldsymbol{v}_k \boldsymbol{v}_k^{\mathrm{T}} \tag{6}$$

is a symmetric orthogonal matrix of order n determined with the help of a suitable vector $\boldsymbol{v}_k$ such that $\boldsymbol{v}_k^{\mathrm{T}} \boldsymbol{v}_k = 1$. The choice of $\boldsymbol{P}_k$, which is called the *reflection matrix*, can be done in such a way that all the entries of $\boldsymbol{A}_k$ in rows $1, \ldots, k-1$ except for their "tridiagonal" entries are zero. The same holds for the entries of $\boldsymbol{A}_k$ in columns $1, \ldots, k-1$ by symmetry. Putting $\boldsymbol{A}_{k-1} = (a_{ij}^{(k-1)})$, $\boldsymbol{v}_k = (v_1^{(k)}, \ldots, v_n^{(k)})^{\mathrm{T}}$, and

$$S = \sum_{j=k}^{n} \left(a_{k-1,j}^{(k-1)}\right)^2\,, \tag{7}$$

we have

$$\boldsymbol{A}_{k-1} = \begin{bmatrix} a_{11}^{(k-1)} & a_{12}^{(k-1)} & & & \cdots & 0 \\ a_{21}^{(k-1)} & a_{22}^{(k-1)} & a_{23}^{(k-1)} & & & \vdots \\ & \ddots & \ddots & \ddots & & \\ & & & a_{k-2,k-2}^{(k-1)} & a_{k-2,k-1}^{(k-1)} \cdots & a_{k-2,n}^{(k-1)} \\ \vdots & & & a_{k-1,k-2}^{(k-1)} & a_{k-1,k-1}^{(k-1)} \cdots & a_{k-1,n}^{(k-1)} \\ & & & \vdots & \vdots & \vdots \\ 0 & \cdots & & a_{n,k-2}^{(k-1)} & a_{n,k-1}^{(k-1)} \cdots & a_{nn}^{(k-1)} \end{bmatrix} \tag{8}$$

and choose

$$\begin{aligned} v_j^{(k)} &= 0\,, \quad j = 1, \ldots, k-1\,, \\ \left(v_k^{(k)}\right)^2 &= \frac{1}{2}\left(1 \pm \frac{a_{k-1,k}^{(k-1)}}{\sqrt{S}}\right)\,, \\ v_j^{(k)} &= \pm \frac{a_{k-1,j}^{(k-1)}}{2v_k^{(k)}\sqrt{S}}\,, \quad j = k+1, \ldots, n\,, \end{aligned} \tag{9}$$

where the sign in (9) is chosen to maximize $|v_k^{(k)}|$. By (6), $\boldsymbol{P}_k$ can be written in the form

$$\boldsymbol{P}_k = \begin{bmatrix} \boldsymbol{I}, & \boldsymbol{0} \\ \boldsymbol{0}, & \boldsymbol{N} \end{bmatrix}\,, \tag{10}$$

where $\boldsymbol{I}$ is the identity matrix of order $k-1$ and $\boldsymbol{N}$ is a symmetric (full, in general) square matrix of order $n-k+1$. Since the position of zero entries of $\boldsymbol{A}_k$ is known it is sufficient to calculate only the nonzero ones when substituting into (5).

If we wish to compute also eigenvectors of $\boldsymbol{A}$ we have to determine the product of all the transformation matrices $\boldsymbol{P}_k$ simultaneously with the reduction of $\boldsymbol{A}$, too (cf. Remark 30.11.1).

Example 2. Use the Householder method to reduce a given symmetric full matrix $\boldsymbol{A} = \boldsymbol{A}_1$ of order 4 to tridiagonal form. By (5), $\boldsymbol{A}_2 = (a_{ij}^{(2)})$ is given by the formula $\boldsymbol{A}_2 = \boldsymbol{P}_2\boldsymbol{A}_1\boldsymbol{P}_2$, and has the form (8) with $k = 3$ and zero entries $a_{13}^{(2)}$, $a_{31}^{(2)}$, $a_{14}^{(2)}$ and $a_{41}^{(2)}$. Like in Example 1, let

$$\boldsymbol{A}_2 = \begin{bmatrix} 6, & 2, & 0, & 0 \\ 2, & -5, & 4, & 3 \\ 0, & 4, & 5, & 1 \\ 0, & 3, & 1, & 8 \end{bmatrix}\,.$$

It remains to carry out the last step of the transformation

$$\boldsymbol{A}_3 = \boldsymbol{P}_3\boldsymbol{A}_2\boldsymbol{P}_3\,, \tag{11}$$

resulting in $a_{24}^{(3)} = a_{42}^{(3)} = 0$.

From (9), we find the components of $\boldsymbol{v}_3 = (v_1^{(3)}, v_2^{(3)}, v_3^{(3)}, v_4^{(3)})^{\mathrm{T}}$, which determines $\boldsymbol{P}_3$ by means of (6). Substituting into (7), we first get $S = \left(a_{23}^{(2)}\right)^2 + \left(a_{24}^{(2)}\right)^2 = 25$ and further $v_1^{(3)} = v_2^{(3)} = 0$, $v_3^{(3)} = \sqrt{0{\cdot}9}$ and $v_4^{(3)} = \frac{1}{3}\sqrt{0{\cdot}9}$. Finally we compute by (6)

$$\boldsymbol{P}_3 = \begin{bmatrix} 1{\cdot}0, & 0{\cdot}0, & 0{\cdot}0, & 0{\cdot}0 \\ 0{\cdot}0, & 1{\cdot}0, & 0{\cdot}0, & 0{\cdot}0 \\ 0{\cdot}0, & 0{\cdot}0, & -0{\cdot}8, & -0{\cdot}6 \\ 0{\cdot}0, & 0{\cdot}0, & -0{\cdot}6, & 0{\cdot}8 \end{bmatrix},$$

which is a symmetric matrix of the form (10). The result of multiplication in (11) is

$$\boldsymbol{A}_3 = \begin{bmatrix} 6{\cdot}00, & 2{\cdot}00, & 0{\cdot}00, & 0{\cdot}00 \\ 2{\cdot}00, & -5{\cdot}00, & -5{\cdot}00, & 0{\cdot}00 \\ 0{\cdot}00, & -5{\cdot}00, & 7{\cdot}04, & -1{\cdot}72 \\ 0{\cdot}00, & 0{\cdot}00, & -1{\cdot}72, & 5{\cdot}96 \end{bmatrix}.$$

Example 3. In Examples 1 and 2, we reduced the matrices $\boldsymbol{T}_2$ and $\boldsymbol{A}_2$, where $\boldsymbol{T}_2 = \boldsymbol{A}_2$, to tridiagonal form. The resulting matrix $\boldsymbol{T}_3$ is similar to $\boldsymbol{T}_2$, the resulting matrix $\boldsymbol{A}_3$ is similar to $\boldsymbol{A}_2$. Multiplying $\boldsymbol{A}_3$ by the symmetric orthogonal matrix

$$\boldsymbol{K} = \begin{bmatrix} 1, & 0, & 0, & 0 \\ 0, & 1, & 0, & 0 \\ 0, & 0, & -1, & 0 \\ 0, & 0, & 0, & 1 \end{bmatrix}$$

from the left as well as from the right, we can verify that $\boldsymbol{T}_3$ and $\boldsymbol{A}_3$ are similar, i.e., that $\boldsymbol{K}\boldsymbol{A}_3\boldsymbol{K} = \boldsymbol{T}_3$.

REMARK 1. The number of arithmetic operations needed to apply both the Givens and the Householder method is of order n^3, the number of operations for the Householder method, however, is about one half as compared with the Givens method.

Up to now we have been concerned with reducing real symmetric matrices. We will conclude this paragraph with a brief survey of methods for the reduction of nonsymmetric matrices.

The *Lanczos method*, consisting in the construction of two finite sequences of vectors (see, e.g., [378]), reduces a general nonsymmetric matrix to a similar tridiagonal matrix.

REMARK 2. If a matrix is tridiagonal, three-term recurrent formulae can be derived (see [378]) to yield the value of the characteristic polynomial $p(\lambda)$ at any argument λ and also the value of the derivative $p'(\lambda)$ of this polynomial. Since

similar matrices possess the same characteristic polynomial (Theorem 1.28.5) we can, after reduction to tridiagonal form, look for the eigenvalues of a matrix with the help of some method for finding roots of a polynomial from Chap. 31 without computing coefficients of the characteristic polynomial (cf. Remark 30.10.2).

The fact, that each transformation in both the Givens and the Householder methods always annihilates two entries symmetric about the diagonal, is a mere consequence of the symmetry of the matrix $\boldsymbol{A}$ given. Applying any of the two methods to a nonsymmetric matrix, we obtain a matrix in Hessenberg form (Definition 30.13.1) as a result. This form is suitable for using the methods of § 30.13 as well.

The *Wilkinson method* (see [378]), based on the Gaussian elimination (§ 30.1), is often employed to reduce a nonsymmetric matrix to upper Hessenberg form. The matrices of derived systems, that appear in the elimination process, however, are not similar to the given matrix. We have to perform the elimination steps not only with rows, but also with columns, to obtain a similar matrix. Pivoting can be applied. The Wilkinson method is finite and requires of order n^3 arithmetic operations.

REMARK 3. Roundoff errors may accumulate considerably and destroy the result if the matrix being reduced is nonsymmetric. The influence of roundoff is diminished by *balancing*, i.e. the procedure which transforms the matrix given into a similar matrix whose i-th row and i-th column have approximately the same norm (see, e.g., [376], [498]).

30.15. Inverse Iteration Method

The *inverse iteration method* can be used to compute one or, successively, several eigenvectors if they belong to eigenvalues which are already known with sufficient accuracy. At the same time, the method can improve accuracy of the eigenvalue itself (see [376]).

Let λ^* be an approximation to the eigenvalue λ of $\boldsymbol{A}$ and let an eigenvector $\boldsymbol{x}$ belongs to λ, i.e. $\boldsymbol{Ax} = \lambda\boldsymbol{x}$. We can readily verify that $\boldsymbol{x}$ is then an eigenvector of $(\boldsymbol{A} - \lambda^*\boldsymbol{I})^{-1}$, belonging to the eigenvalue $1/(\lambda - \lambda^*)$. The vector $\boldsymbol{x}$ is computed by the power method (§ 30.11) applied to $(\boldsymbol{A} - \lambda^*\boldsymbol{I})^{-1}$ since, for λ^* sufficiently close to λ, its eigenvalue $1/(\lambda - \lambda^*)$ is so large that it is dominant.

We thus choose a vector $\boldsymbol{v}_0$ such that $\boldsymbol{v}_0^{\mathrm{T}}\boldsymbol{v}_0 = 1$ and compute by (30.11.2)

$$c_{m+1}(\boldsymbol{A} - \lambda^*\boldsymbol{I})\boldsymbol{v}_{m+1} = \boldsymbol{v}_m\,, \tag{1}$$

where the number c_{m+1} is a normalization factor chosen in such a way that $\boldsymbol{v}_{m+1}^{\mathrm{T}}\boldsymbol{v}_{m+1} = 1$ and $\boldsymbol{v}_{m+1}^{\mathrm{T}}\boldsymbol{v}_m > 0$.

If conditions for the convergence of the power method are fulfilled, then the sequence of vectors $\boldsymbol{v}_m$ tends to the eigenvector $\boldsymbol{x}$. The rate of convergence of the method is high if the difference $|\lambda - \lambda^*|$ is small. In practice, 2 or 3 iteration steps suffice.

REMARK 1. Every step of the inverse iteration involves the solution of the linear algebraic system (1) for the unknown vector $\boldsymbol{v}_{m+1}$. The matrix of (1) is very ill--conditioned and it is even singular for $\lambda^* = \lambda$ (Remark 30.10.1). This fact, however, does not influence the accuracy of the evaluation of components of the eigenvector in the iterative process (1).

REMARK 2. If the matrix of the system (1) is full and if the LU factorization is applied to the solution of the system, every step of the inverse iteration method requires of order n^2 operations and the factorization (performed only once) needs of order n^3 additional operations (Remark 30.1.6). For this reason, the method is mostly applied to matrices having been reduced to tridiagonal or Hessenberg form where the number of operations required is of lower order.

30.16. Generalized Eigenproblem

In many applications, a problem more general, than that formulated in Definition 30.10.1, arises.

Definition 1. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be (complex, in general) square matrices of order n. If there exist a complex number λ and a vector $\boldsymbol{x} \neq \boldsymbol{0}$ (with complex components) such that

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{B}\boldsymbol{x} \tag{1}$$

then we say that λ is an *eigenvalue of the problem* (1) and $\boldsymbol{x}$ is its *eigenvector* belonging to this eigenvalue.

REMARK 1. If at least one of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is nonsingular, then the generalized eigenproblem can be transformed to a standard eigenproblem from Definition 30.10.1. If $\boldsymbol{A}$ is nonsingular, the equation (1) is equivalent to

$$(\boldsymbol{A}^{-1}\boldsymbol{B})\boldsymbol{x} = \mu \boldsymbol{x} \tag{2}$$

and $\lambda = 1/\mu$. If $\boldsymbol{B}$ is nonsingular, then (1) is equivalent to

$$(\boldsymbol{A}\boldsymbol{B}^{-1})\boldsymbol{y} = \lambda \boldsymbol{y} \tag{3}$$

and $\boldsymbol{x} = \boldsymbol{B}^{-1}\boldsymbol{y}$. The eigenproblems (2) and (3) can be solved by the methods shown in the preceding paragraphs.

If both the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are singular, we can solve the generalized eigenproblem e.g. by the *QZ method* (see [446], [461]) which is a certain analogue of the QR method.

30.17. Choice of the Method. Basic Software

It is rather difficult to give unique recommendations for the choice of a proper method for the computation of eigenvalues since this choice depends on a lot of circumstances. Properties of the matrix given and quantities to be calculated are the principal factors we already mentioned in the introduction to Part B of this chapter.

Most methods we presented fail if a real matrix has a pair of complex conjugate eigenvalues, some fail for multiple eigenvalues. These cases can be detected since our methods then do not converge if applied in the form given here. In literature (e.g. [497], [498]), modifications of these methods, taking into account special situations mentioned, are shown.

For a full symmetric matrix, we can recommend the reduction to tridiagonal form by the Householder method (§ 30.14), which requires less operations than the Givens method, followed by the computation of eigenvalues (and possibly also eigenvectors) by the QR (or QL) method (§ 30.13) with shifts, if needed. The QR method is, in general, numerically more stable than the LR method as it is based on orthogonal transformations. The LR method modified for symmetric matrices can be a good tool if the matrix given is positive definite. The calculation of eigenvectors demands a great number of arithmetic operations. If we need only several eigenvectors, we can, after reduction to tridiagonal form, find roots of the characteristic polynomial by methods of Chap. 31 and compute the corresponding eigenvectors by the inverse iteration (§ 30.15).

The procedure is similar for a full nonsymmetric matrix, but the first step should be balancing the matrix (§ 30.14) and the result of reduction is in Hessenberg form.

The power method (§ 30.11) is advantageous for calculating the dominant eigenvalue (and the eigenvector belonging to it). In this method as well as in various its modifications, the sparsity of the matrix given can be exploited.

We have not been concerned with the computation of eigenvalues and eigenvectors of complex matrices. Proper methods are presented, e.g., in [433], [498], together with much more detailed recommendations for the choice of the method for solving a particular real as well as complex eigenproblem.

The computation of eigenvalues and eigenvectors of matrices is a branch of numerical analysis, where efficient methods are known for most kinds of principal problems. The book [498] has become the foundation of methods and algorithms. It contains programs in Algol 60 that were translated into FORTRAN and thoroughly tested. The result of this work is the EISPACK package [433] that should be satisfactory for solving all current eigenproblems.

Standard numerical software of any computer usually includes also programs or subprograms for computing eigenvalues and eigenvectors of matrices. Very good universal libraries (in FORTRAN) available on the commercial basis are IMSL Library [231] and NAG Library [344] for mainframes, and personal computer programs from Numerical Recipes, the book [376] which contains all source programs in FORTRAN as well as in Pascal (C is also available).

In general, we can say that this part of Chap. 30 is to serve as a guide for the choice of a method, when we calculate eigenvalues and eigenvectors of a particular matrix. Next step should be to employ a suitable program found in the software available, preferably in the EISPACK package. In rather exceptional cases, it is necessary to code one's own program to implement a method which is not part of EISPACK.

31. NUMERICAL SOLUTION OF ALGEBRAIC AND TRANSCENDENTAL EQUATIONS

By MIROSLAV FIEDLER

References: [62], [209], [215], [224], [311], [360], [361], [378], [386], [468], [482].

31.1. Basic Properties of Algebraic Equations

An algebraic equation of degree n,

$$f(x) \equiv a_0 x^n + a_1 x^{n-1} + \ldots + a_n = 0 \quad (a_0 \neq 0), \tag{1}$$

with real or complex coefficients $a_0, \ldots, a_n$, has exactly n roots (in general complex) if each root is considered with its appropriate multiplicity (cf. § 1.14).

In the following text we consider only the case where all the coefficients in (1) are real. If some coefficients are not real, then we construct the polynomial

$$g(x) \equiv \left(a_0 x^n + a_1 x^{n-1} + \ldots + a_n\right)\left(\overline{a}_0 x^n + \overline{a}_1 x^{n-1} + \ldots + \overline{a}_n\right),$$

where the $\overline{a}_i$ denote the complex conjugate numbers to a_i; this polynomial has only real coefficients; the equation $g(x) = 0$ contains all the roots of the equation $f(x) = 0$.

REMARK 1. In some considerations, it is required that all the roots of equation (1) should be simple. Theoretically it is not difficult to find an equation with the same roots as those of (1) but all *simple*. If $d(x)$ is the greatest common divisor of $f(x)$ and its derivative $f'(x)$ (this can be found by using the Euclidean algorithm, cf. Theorem 1.14.7), then the quotient $g(x) = f(x)/d(x)$ is a polynomial with the above-mentioned property.

Theorem 1. *Let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be the roots of* (1). *Then,*

(a) *$1/\alpha_1, 1/\alpha_2, \ldots, 1/\alpha_n$ are the roots of the equation $g(x) \equiv x^n f(1/x) \equiv a_n x^n + a_{n-1} x^{n-1} + \ldots + a_0 = 0$ (for $a_n \neq 0$);*

(b) *$\alpha_1/c, \alpha_2/c, \ldots, \alpha_n/c$ are the roots of the equation $g(x) \equiv f(cx) \equiv a_0 c^n x^n + a_1 c^{n-1} x^{n-1} + \ldots + a_n = 0$ (for $c \neq 0$);*

(c) $-\alpha_1, -\alpha_2, \ldots, -\alpha_n$ *are the roots of the equation* $g(x) \equiv (-1)^n f(-x) \equiv$ $\equiv a_0 x^n - a_1 x^{n-1} + a_2 x^{n-2} + \ldots + (-1)^n a_n = 0$;

(d) $\alpha_1 - a, \alpha_2 - a, \ldots, \alpha_n - a$ *are the roots of the equation*

$$g(x) \equiv f(x+a) \equiv f(a) + \frac{x}{1!} f'(a) + \frac{x^2}{2!} f''(a) + \ldots + \frac{x^n}{n!} f^{(n)}(a) = 0.$$

In (d), the coefficients $f(a)$, $f'(a)$, $\frac{1}{2!} f''(a)$ etc. can be found, for example, by using Horner's scheme (cf. Remark 1.14.1).

31.2. Estimates for the Roots of Algebraic Equations

Theorem 1. *Let in the equation*

$$f(x) \equiv a_0 x^n + a_1 x^{n-1} + \ldots + a_n = 0 \tag{1}$$

a_0 *be positive. Then every real root* α *of* (1) *satisfies the following inequalities*:

(a) $$\alpha < 1 + \frac{|a_i|}{a_0} \qquad (\textit{Maclaurin}),$$

where a_i *is the negative coefficient in* (1) *with greatest modulus* (*if none of the coefficients are negative then* $\alpha \leqq 0$);

(b) $$\alpha < 1 + \left(\frac{|a_i|}{a_0}\right)^{1/r} \qquad (\textit{Lagrange}),$$

where a_i *is defined as in* (a) *and* a_r *is the first negative coefficient in* (1);

(c) $$\alpha < 1 + \left(\frac{|a_i|}{a_s}\right)^{1/(r-s)} \qquad (\textit{Tillot}),$$

where a_i *and* r *are the same as in* (a) *and* (b) *and* a_s *is the greatest of the first* r *positive coefficients in* (1).

Theorem 2. *Every real or complex root* α *of equation* (1) (*with real or complex coefficients*) *satisfies the inequality*

$$|\alpha| < 1 + \left|\frac{a_i}{a_0}\right|$$

where a_i *is the coefficient of* (1) *with the greatest modulus.*

REMARK 1. Similar estimates of the real roots of (1) from below can be obtained by the application of Theorem 1 to (c) of Theorem 31.1.1; an analogous estimate for the moduli of complex roots of (1) from below can be obtained by application of Theorem 2 to (a) of Theorem 31.1.1.

Example 1. Theorem 1 yields the following estimates for the real roots α_k of the equation $x^3 + 4x^2 + x - 6 = 0$ (here, $i = 3$, $r = 3$, $s = 1$):

(a) $\alpha_k < 7$,

(b) $\alpha_k < 1 + 6^{1/3} \doteq 2{\cdot}8171$,

(c) $\alpha_k < 1 + \left(\frac{6}{4}\right)^{1/2} \doteq 2{\cdot}2247$.

According to Theorem 2, all roots (real or complex) satisfy the inequality $|\alpha_k| < 7$; if we apply Theorem 2 to the equation $6x^3 - x^2 - 4x - 1 = 0$ with roots $1/\alpha_k$, we obtain $|1/\alpha_k| < 1 + \left|\frac{4}{6}\right| = \frac{5}{3}$ so that $|\alpha_k| > \frac{3}{5}$. (The roots of the equation $x^3 + 4x^2 + x - 6 = 0$ are 1, -2 and -3.)

Theorem 3. *Let x_1 be a (complex) number for which $f'(x_1) \neq 0$. Then the circle $|x - \xi| \leqq \rho$, where*

$$\xi = x_1 - \frac{n}{2}\frac{f(x_1)}{f'(x_1)}, \qquad \rho = \left|\frac{n}{2}\frac{f(x_1)}{f'(x_1)}\right|,$$

contains at least one root of the equation (1).

Example 2. Let us apply this Theorem to the equation $x^3 + 4x^2 + x - 6 = 0$ from Example 1. Choose $x_1 = 0{\cdot}9$, then $f(x_1) = -1{\cdot}141$, $f'(x_1) = 10{\cdot}63$. It follows that in the circle with centre $\xi = 0{\cdot}9 - \frac{3}{2} \cdot (-1{\cdot}141)/10{\cdot}63 \doteq 1{\cdot}061$ and radius $\rho = 0{\cdot}161$ there lies at least one root of the equation.

Theorem 4. (*Descartes' Theorem*). *The number of positive roots of equation* (1) *is either equal to the number of changes of sign in the sequence*

$$a_0, a_1, \ldots, a_n,$$

or it is smaller by an even number.

REMARK 2. The number of changes of sign is obtained by ignoring all the zero entries and by determining the number of pairs of consecutive numbers with different signs in the sequence.

Example 3. The corresponding sequence to the equation

$$x^4 + 3x^2 - 1 = 0 \tag{2}$$

is

$$1, 0, 3, 0, -1$$

with a single change of sign $(3, -1)$. It follows that equation (2) possesses exactly one positive root α (since the number of its roots cannot be smaller than one by an even number). According to the estimate (a) in Theorem 1, $\alpha < 2$, according to (b) we also have $\alpha < 2$, according to (c) $\alpha < 1 + \sqrt{\frac{1}{3}} \doteq 1{\cdot}58$.

Theorem 5. (*Budan-Fourier Theorem*). *Let $f(x)$ be a real polynomial, $a_0 > 0$. Let $\alpha < \beta$, $f(\alpha) \cdot f(\beta) \neq 0$ and let $\omega(x)$ denote the number of changes of sign in the sequence*

$$f(x), f'(x), f''(x), \ldots, f^{(n)}(x).$$

Then, the number of real roots of (1) *in the interval $[\alpha, \beta]$ is either equal to $\omega(\alpha) - \omega(\beta)$, or it is smaller by an even number.*

REMARK 3. In Theorems 4 and 5 each root is considered with its corresponding multiplicity.

Theorem 6. (*Sturm Theorem*). *Let all the roots of* (1) *be simple* (cf. Remark 31.1.1). *If neither of the real numbers α and β $(\alpha < \beta)$ is a root of* (1), *then there are exactly $V(\alpha) - V(\beta)$ real roots of* (1) *in the interval $[\alpha, \beta]$. Here, $V(x)$ denotes the number of changes of sign in the so-called Sturm sequence*

$$f(x), f_1(x), f_2(x), \ldots, f_m(x).$$

For this sequence the following special sequence can be chosen:

$$f(x), f'(x), r_1(x), \ldots, r_s(x),$$

where $-r_1(x)$ is the remainder obtained after the division $f(x)/f'(x)$, $-r_2(x)$ the remainder after the division $f'(x)/r_1(x)$ etc., and $-r_s(x)$ is the remainder after the division $r_{s-2}(x)/r_{s-1}(x)$, where $r_s(x)$ is a non-zero constant.

Example 4. Let us apply the Sturm theorem to the problem of finding the number of roots of the equation $x^3 + 4x^2 + x - 6 = 0$ in the interval $[0, 2]$. Here, $f(x) = x^3 + 4x^2 + x - 6 = 0$, $f_1(x) = f'(x) = 3x^2 + 8x + 1$, $f_2(x) = \frac{2}{9}(13x + 29)$ $(-f_2(x)$ is the remainder after the division $f(x)/f_1(x))$, $f_3(x) = \frac{324}{169}$. Hence $f(0) = -6$, $f_1(0) = 1$, $f_2(0) = \frac{58}{9}$, $f_3(0) = \frac{324}{169}$, so that $V(0) = 1$; further, $f(2) = 20$, $f_1(2) = 29$, $f_2(2) = \frac{110}{9}$, $f_3(2) = \frac{324}{169}$ so that $V(2) = 0$. Since $V(0) - V(2) = 1$, there is exactly one root of the given equation in the interval $[0, 2]$.

31.3. Connection of Roots with Eigenvalues of Matrices

It is easily seen that if $a_0 = 1$, equation (31.2.1) can be written (with λ instead of x) in the following determinant form:

$$\begin{vmatrix} \lambda, & -1, & 0, & \ldots, & 0, & 0 \\ 0, & \lambda, & -1, & \ldots, & 0, & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ 0, & 0, & 0, & \ldots, & \lambda, & -1 \\ a_n, & a_{n-1}, & a_{n-2}, & \ldots, & a_2, & \lambda + a_1 \end{vmatrix} = 0. \tag{1}$$

However, this is the characteristic equation of the matrix

$$\boldsymbol{A} = \begin{bmatrix} 0, & 1, & 0, & \ldots, & 0, & 0 \\ 0, & 0, & 1, & \ldots, & 0, & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ 0, & 0, & 0, & \ldots, & 0, & 1 \\ -a_n, & -a_{n-1}, & -a_{n-2}, & \ldots, & -a_2, & -a_1 \end{bmatrix}. \tag{2}$$

Hence estimates for eigenvalues of matrices yield immediate estimates for the roots of equation (31.2.1). This is of particular use in numerical solution of algebraic equations.

A well-known estimate for eigenvalues of square matrices has been given by *Gershgorin*:

Theorem 1. *Let* $\boldsymbol{A} = (a_{ij})$ *be a square matrix of order* n. *Then all eigenvalues of* $\boldsymbol{A}$ *are contained in the union of the following* n *circles in the complex plane*:

$$|a_{ii} - z| \leqq \sum_{j \neq i} |a_{ij}|, \qquad i = 1, \ldots, n.$$

Moreover, if $|a_{kk} - a_{11}| > \sum_{j \neq k} |a_{kj}| + \sum_{j \neq 1} |a_{1j}|$ *for* $k = 2, \ldots, n$, *the first circle contains exactly one eigenvalue of* $\boldsymbol{A}$.

31.4. Some Methods for Solving Algebraic and Transcendental Equations

Algebraic equations (with one unknown) of degree four at most, binomial equations and some other special types of equations can be solved directly (see §§ 1.20, 1.21, 1.22). To solve algebraic equations of higher degree or transcendental equations, numerical methods are mostly used. Some of these methods will now be described.

(a) Method of Bernoulli and Whittaker. The given algebraic equation of degree n is written in the form

$$x^n = a_1 x^{n-1} + a_2 x^{n-2} + \ldots + a_n. \tag{1}$$

Choose $u_0 = 1$, $u_{-1} = u_{-2} = \ldots = u_{-(n-1)} = 0$ and compute the numbers $u_1, u_2, \ldots$ according to the recurrence formula ($m = 1, 2, \ldots$)

$$u_m = a_1 u_{m-1} + a_2 u_{m-2} + \ldots + a_n u_{m-n}.$$

Let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be the roots (in general complex and not necessarily distinct) of equation (1), such that

$$|\alpha_1| = |\alpha_2| = \ldots = |\alpha_k| > |\alpha_{k+1}| \geqq \ldots \geqq |\alpha_n|$$

(where thus exactly k roots have the greatest modulus). Then, for integral r sufficiently large, $\alpha_1, \alpha_2, \ldots, \alpha_k$ are approximately equal to the roots of the equation

$$\begin{vmatrix} x^k, & x^{k-1}, & \ldots, & 1 \\ u_{r+k}, & u_{r+k-1}, & \ldots, & u_r \\ u_{r+k+1}, & u_{r+k}, & \ldots, & u_{r+1} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ u_{r+2k-1}, & u_{r+2k-2}, & \ldots, & u_{r+k-1} \end{vmatrix} = 0. \tag{2}$$

In general, the larger the value of r is, the more accurate is the approximation to the roots. When all the coefficients of (1) are real, the most frequent cases are $k = 1$ (for α_1 real), and $k = 2$ (for α_1, α_2 complex conjugate).

A convenient procedure for practical computation is the following: Compute the first, say, 20 (or more, according to the required accuracy and the magnitude of n) numbers $u_1, \ldots, u_{20}$ and then the ratios $u_{16}/u_{15}, u_{17}/u_{16}, \ldots, u_{20}/u_{19}$. If none of these ratios differs in sign or by more than $5 - 10\%$ in magnitude, u_{20}/u_{19} can be considered as an approximation to the root α_1 of equation (1), i.e. the case $k = 1$ in (2) has occurred. If the computed ratios differ by more than 10% (or even in sign), the determinants $\Delta_r = u_r^2 - u_{r-1}u_{r+1}$ for $r = 15, \ldots, 19$ and their ratios $Q_r = \Delta_r/\Delta_{r-1}$ for $r = 16, \ldots, 19$ should be computed. If these ratios Q_r are approximately equal and of the same sign, then the roots α_1 and α_2 of equation (1) are approximately equal to the roots of equation (2) for $k = 2$ and $r = 17$. If neither the first, nor the second case occurs, then the first three or more roots of (1) have almost equal moduli. It is then possible to proceed in a similar way for $k = 3$ or to solve another equation in y which has been obtained from the given equation by the substitution $x = y + u$. Here, u can be chosen for example as an approximation to $\sqrt[n]{|a_n|}$.

(b) The Graeffe Method and its Modifications. We have to solve the equation

$$a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0 = 0, \qquad a_n \neq 0. \tag{3}$$

Let us compute the numbers

$$\begin{aligned}
a_0^{(1)} &= a_0^2, \\
a_1^{(1)} &= -a_1^2 + 2a_0a_2, \\
a_2^{(1)} &= a_2^2 - 2a_1a_3 + 2a_0a_4, \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
a_{n-1}^{(1)} &= (-1)^{n-1}a_{n-1}^2 + (-1)^n \cdot 2a_{n-2}a_n, \\
a_n^{(1)} &= (-1)^n a_n^2
\end{aligned} \tag{4}$$

and repeat the computation to obtain

$$a_0^{(2)} = \left(a_0^{(1)}\right)^2, \quad a_1^{(2)} = -\left(a_1^{(1)}\right)^2 + 2a_0^{(1)}a_2^{(1)}, \quad \text{etc.},$$

and $a_0^{(3)} = \left(a_0^{(2)}\right)^2$ etc., up to $a_0^{(m)}, a_1^{(m)}, \ldots, a_n^{(m)}$, where m is a chosen integer (say $m = 10$).

The polynomial

$$a_n^{(m)}x^n + a_{n-1}^{(m)}x^{n-1} + \ldots + a_0^{(m)}$$

has for its roots the 2^m-th powers of the roots of equation (3). If m is sufficiently large (say $m = 10$), some of the coefficients (in any case $a_n^{(m)}$ and $a_0^{(m)}$) satisfy the approximate equalities

$$a_k^{(m)} \approx (-1)^k \left(a_k^{(m-1)}\right)^2$$

and, moreover, the remaining summands on the right-hand sides of (4) are, for these coefficients, sufficiently small. Suppose that these "well-behaved" coefficients are

$$a_n^{(m)}, a_{k_1}^{(m)}, a_{k_2}^{(m)}, \ldots, a_{k_s}^{(m)}, a_0^{(m)}, \quad \text{where} \quad n > k_1 > k_2 > \ldots > k_s > 0.$$

Then, the roots of equations

$$\begin{aligned}
L_1(x) &\equiv a_n^{(m)}x^{n-k_1} + a_{n-1}^{(m)}x^{n-k_1-1} + \ldots + a_{k_1}^{(m)} = 0, \\
L_2(x) &\equiv a_{k_1}^{(m)}x^{k_1-k_2} + a_{k_1-1}^{(m)}x^{k_1-k_2-1} + \ldots + a_{k_2}^{(m)} = 0, \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
L_{s+1}(x) &\equiv a_{k_s}^{(m)}x^{k_s} + a_{k_s-1}^{(m)}x^{k_s-1} + \ldots + a_0^{(m)} = 0
\end{aligned} \tag{5}$$

are approximately equal to the 2^m-th powers of the roots of equation (3) in such a sense that one of equations (5) corresponds to a group of roots of (3) with (approximately) equal moduli: the first with moduli r_1, the second with r_2, etc., the last with r_{s+1}. Here, $r_1 > r_2 > \ldots > r_{s+1}$.

In this manner, we obtain the moduli of the roots of equation (3) since for $i = 1, \ldots, s+1$ (if we put $k_0 = n$, $k_{s+1} = 0$)

$$r_i^{k_{i-1}-k_i} \approx \sqrt[2^m]{\left|\frac{a_{k_i}^{(m)}}{a_{k_i-1}^{(m)}}\right|}. \tag{6}$$

To compute the roots themselves, it is necessary to determine which of the 2^m-th roots of the corresponding root of equation (5) satisfies (3). This can easily be done when all the coefficients of (3) are real (we shall assume this from now on) and when the corresponding equation (5) is of degree one. Then, the corresponding root of (3) is real and it is sufficient to find out by substitution into (3) which of the two real 2^m-th roots

$$\sqrt[2^m]{\gamma} \qquad \text{and} \qquad -\sqrt[2^m]{\gamma}$$

of the root γ of (5) satisfies equation (3).

If all the equations (5) are linear, we obtain in this manner all roots. If one of equations (5) is quadratic, all the remaining being linear, we can compute the roots which correspond to the linear equations, then use relations for the sum of all roots (equal to $-a_{n-1}/a_n$) and for the product of all roots (equal to $(-1)^n a_0/a_n$), to compute, if $a_0 \neq 0$, the sum and product of the remaining two roots. Other cases lead to complications which can be avoided by the following *modified Lehmer's process*:

We compute, besides the sequence $a_n^{(k)}, a_{n-1}^{(k)}, \ldots, a_0^{(k)}$, $k = 1, \ldots, m$, another set of sequences $b_n^{(k)}, b_{n-1}^{(k)}, \ldots, b_0^{(k)}$, $k = 0, \ldots, m$, as follows:

$$b_n^{(0)} = a_{n-1},\ b_{n-1}^{(0)} = 2a_{n-2},\ b_{n-2}^{(0)} = 3a_{n-3}, \ldots, b_1^{(0)} = na_0,\ b_0^{(0)} = 0$$

(where $a_i^{(0)} = a_i$, $i = 0, \ldots, n$) and

$$\begin{aligned}
b_0^{(k+1)} &= a_0^{(k)} b_0^{(k)} (= 0),\\
b_1^{(k+1)} &= a_0^{(k)} b_2^{(k)} - a_1^{(k)} b_1^{(k)} + a_2^{(k)} b_0^{(k)},\\
b_2^{(k+1)} &= a_0^{(k)} b_4^{(k)} - a_1^{(k)} b_3^{(k)} + a_2^{(k)} b_2^{(k)} - a_3^{(k)} b_1^{(k)} + a_4^{(k)} b_0^{(k)},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
b_{n-1}^{(k+1)} &= (-1)^{n-2} a_{n-2}^{(k)} b_n^{(k)} + (-1)^{n-1} a_{n-1}^{(k)} b_{n-1}^{(k)} + (-1)^n a_n^{(k)} b_{n-2}^{(k)},\\
b_n^{(k+1)} &= (-1)^n a_n^{(k)} b_n^{(k)}
\end{aligned} \tag{7}$$

for $k = 0, \ldots, m-1$. If we denote the following polynomials (with k_i defined as in (5)) by $M_i(x)$, $i = 1, \ldots, s+1$, where

$$M_i(x) \equiv b_{k_{i-1}}^{(m)} x^{k_{i-1}-k_i} + b_{k_{i-1}-1}^{(m)} x^{k_{i-1}-k_i-1} + \ldots + b_{k_i}^{(m)},$$

then the following assertion holds:

If β is a simple root of $L_i(x) = 0$, then the corresponding root α of equation (3) satisfies the approximate equality

$$\alpha \approx -\frac{M_i(\beta)}{\beta L_i'(\beta)}. \tag{8}$$

If all the polynomials $L_j(x) = 0$ in (5) are of degree at most 2, it is sufficient to use the fact that the sum of those roots of (3) which correspond to the roots of $L_j(x) = 0$ is approximately equal to

$$\frac{b_{k_j}^{(m)}}{a_{k_j}^{(m)}} - \frac{b_{k_{j-1}}^{(m)}}{a_{k_{j-1}}^{(m)}}. \tag{9}$$

If the degree of $L_j(x) = 0$ is equal to one, the number (9) is approximately equal to the corresponding root of (3). If the degree is two, the number (9) is an approximation of the sum, and the right-hand side of (6) an approximation of the product of the two roots of (3). A slight complication can occur in the case where the number (9) is very close to zero. Then two cases are possible: either both roots of (3) are real and approximately equal to r_j and $-r_j$, or they are complex conjugates and approximately equal to $r_j\mathrm{i}$ and $-r_j\mathrm{i}$. Substitution of r_j into (3) determines the roots.

Example 1. Let us use the Graeffe method to solve the equation

$$4{\cdot}08x^4 - 6{\cdot}03x^3 + 6{\cdot}99x^2 - 9{\cdot}81x + 9{\cdot}72 = 0.$$

The numbers $a_4^{(k)}, a_3^{(k)}, a_2^{(k)}, a_1^{(k)}, a_0^{(k)}$ computed according to formula (4) for $k = 1, \ldots, 10$ ($m = 10$) are listed in Tab. 31.1.

TABLE 31.1

k	$a_4^{(k)}$	$a_3^{(k)}$	$a_2^{(k)}$	$a_1^{(k)}$	$a_0^{(k)}$
–	4·08	−6·03	6·99	−9·81	9·72
1	$1{\cdot}665 \cdot 10^{1}$	$2{\cdot}068 \cdot 10^{1}$	9·867	$3{\cdot}965 \cdot 10^{1}$	$9{\cdot}448 \cdot 10^{1}$
2	$2{\cdot}772 \cdot 10^{2}$	$-9{\cdot}909 \cdot 10^{1}$	$1{\cdot}604 \cdot 10^{3}$	$2{\cdot}923 \cdot 10^{2}$	$8{\cdot}926 \cdot 10^{3}$
3	$7{\cdot}684 \cdot 10^{4}$	$8{\cdot}794 \cdot 10^{5}$	$7{\cdot}579 \cdot 10^{6}$	$2{\cdot}855 \cdot 10^{7}$	$7{\cdot}967 \cdot 10^{7}$
4	$5{\cdot}904 \cdot 10^{9}$	$3{\cdot}914 \cdot 10^{11}$	$1{\cdot}947 \cdot 10^{13}$	$3{\cdot}925 \cdot 10^{14}$	$6{\cdot}347 \cdot 10^{15}$
5	$3{\cdot}486 \cdot 10^{19}$	$7{\cdot}671 \cdot 10^{22}$	$1{\cdot}468 \cdot 10^{26}$	$9{\cdot}310 \cdot 10^{28}$	$4{\cdot}028 \cdot 10^{31}$
6	$1{\cdot}215 \cdot 10^{39}$	$4{\cdot}350 \cdot 10^{45}$	$1{\cdot}008 \cdot 10^{52}$	$3{\cdot}159 \cdot 10^{57}$	$1{\cdot}622 \cdot 10^{63}$
7	$1{\cdot}476 \cdot 10^{78}$	$5{\cdot}572 \cdot 10^{90}$	$7{\cdot}806 \cdot 10^{103}$	$2{\cdot}272 \cdot 10^{115}$	$2{\cdot}631 \cdot 10^{126}$
8	$2{\cdot}179 \cdot 10^{156}$	$1{\cdot}994 \cdot 10^{182}$	$5{\cdot}848 \cdot 10^{207}$	$-1{\cdot}054 \cdot 10^{230}$	$6{\cdot}922 \cdot 10^{252}$
9	$4{\cdot}748 \cdot 10^{312}$	$-1{\cdot}427 \cdot 10^{364}$	$3{\cdot}424 \cdot 10^{415}$	$6{\cdot}985 \cdot 10^{460}$	$4{\cdot}791 \cdot 10^{505}$
10	$2{\cdot}254 \cdot 10^{625}$	$1{\cdot}215 \cdot 10^{728}$	$1{\cdot}172 \cdot 10^{831}$	$-1{\cdot}598 \cdot 10^{921}$	$2{\cdot}295 \cdot 10^{1011}$

We see that the coefficients a_4, a_2 and a_0 behave regularly. Thus $s = 1$, $k_0 = 4$, $k_1 = 2$, $k_2 = 0$ and (5) consists of two equations of degree 2. We thus compute the coefficients $b_4^{(k)}$, $b_3^{(k)}$, $b_2^{(k)}$, $b_1^{(k)}$, $b_0^{(k)}$ for $k = 0, 1, \ldots, 10$ using formula (7) (see Tab. 31.2).

TABLE 31.2

k	$b_4^{(k)}$	$b_3^{(k)}$	$b_2^{(k)}$	$b_1^{(k)}$	$b_0^{(k)}$
0	$-6{\cdot}03$	$1{\cdot}398 \,.\, 10^{1}$	$-2{\cdot}943 \,.\, 10^{1}$	$3{\cdot}888 \,.\, 10^{1}$	0
1	$-2{\cdot}460 \,.\, 10^{1}$	$-7{\cdot}792 \,.\, 10^{1}$	$1{\cdot}073 \,.\, 10^{2}$	$9{\cdot}535 \,.\, 10^{1}$	0
2	$-4{\cdot}096 \,.\, 10^{2}$	$3{\cdot}155 \,.\, 10^{3}$	$-1{\cdot}478 \,.\, 10^{2}$	$6{\cdot}357 \,.\, 10^{3}$	0
3	$-1{\cdot}135 \,.\, 10^{5}$	$-3{\cdot}853 \,.\, 10^{5}$	$-4{\cdot}185 \,.\, 10^{6}$	$-3{\cdot}177 \,.\, 10^{6}$	0
4	$-8{\cdot}721 \,.\, 10^{9}$	$-8{\cdot}430 \,.\, 10^{11}$	$-2{\cdot}697 \,.\, 10^{13}$	$-2{\cdot}427 \,.\, 10^{14}$	0
5	$-5{\cdot}149 \,.\, 10^{19}$	$9{\cdot}214 \,.\, 10^{20}$	$-1{\cdot}546 \,.\, 10^{26}$	$-7{\cdot}592 \,.\, 10^{28}$	0
6	$-1{\cdot}795 \,.\, 10^{39}$	$-1{\cdot}302 \,.\, 10^{46}$	$-1{\cdot}903 \,.\, 10^{52}$	$8{\cdot}409 \,.\, 10^{56}$	0
7	$-2{\cdot}181 \,.\, 10^{78}$	$1{\cdot}542 \,.\, 10^{91}$	$-1{\cdot}573 \,.\, 10^{104}$	$-3{\cdot}352 \,.\, 10^{115}$	0
8	$-3{\cdot}219 \,.\, 10^{156}$	$-4{\cdot}883 \,.\, 10^{182}$	$-1{\cdot}245 \,.\, 10^{208}$	$3{\cdot}477 \,.\, 10^{230}$	0
9	$-7{\cdot}014 \,.\, 10^{312}$	$5{\cdot}141 \,.\, 10^{364}$	$-7{\cdot}293 \,.\, 10^{415}$	$-4{\cdot}953 \,.\, 10^{460}$	0
10	$-3{\cdot}330 \,.\, 10^{625}$	$1{\cdot}472 \,.\, 10^{728}$	$-2{\cdot}497 \,.\, 10^{831}$	$-3{\cdot}441 \,.\, 10^{919}$	0

According to (6) and (9), two roots α_1 and α_2 of the given equation have modulus r_1, where

$$r_1^2 \approx \sqrt[1024]{\frac{1{\cdot}172 \,.\, 10^{831}}{2{\cdot}254 \,.\, 10^{625}}} \doteq 1{\cdot}588$$

and the sum

$$\alpha_1 + \alpha_2 \approx \frac{-2{\cdot}497 \,.\, 10^{831}}{1{\cdot}172 \,.\, 10^{831}} - \frac{-3{\cdot}330 \,.\, 10^{625}}{2{\cdot}254 \,.\, 10^{625}} \doteq -0{\cdot}653.$$

The remaining two roots α_3 and α_4 have modulus r_2, where

$$r_2^2 \approx \sqrt[1024]{\frac{2{\cdot}295 \,.\, 10^{1011}}{1{\cdot}172 \,.\, 10^{831}}} \doteq 1{\cdot}500,^{*)}$$

and the sum

$$\alpha_3 + \alpha_4 \approx \frac{2{\cdot}497 \,.\, 10^{831}}{1{\cdot}172 \,.\, 10^{831}} \doteq 2{\cdot}131.$$

It follows that the given equation has two pairs of complex conjugate roots satisfying the quadratic equations

$$\begin{aligned} x^2 + 0{\cdot}653x + 1{\cdot}588 &= 0, \\ x^2 - 2{\cdot}131x + 1{\cdot}500 &= 0. \end{aligned} \tag{10}$$

*) The absolute values of r_1 and r_2 are very close one to another, which is not convenient for computation. The Bernoulli-Whittaker method — without using the transformation $x = y + u$ — would fail in this case.

Hence

$$\alpha_{1,2} \doteq -0{\cdot}327 \pm 1{\cdot}217\mathrm{i},$$
$$\alpha_{3,4} \doteq \ 1{\cdot}065 \pm 0{\cdot}605\mathrm{i}.$$

REMARK 1. As seen from the preceding example, we meet with very large numbers (in absolute value), as usual, if we apply the Graeffe method. Thus necessary provosions are to be made when realizing this method on a computer.

(c) Newton's Method. This is a method of obtaining the roots of the equation $f(x) = 0$ where $f(x)$ is a polynomial or, more generally, a function of one variable which has a first derivative in the whole interval (or the whole complex region) containing the roots to be determined. Choose x_0 as an approximation to a particular root and construct a sequence $x_1, x_2, \ldots$ according to the formula

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

If the approximation x_0 is close enough to the value ξ of a simple root of $f(x) = 0$ the sequence x_k converges quadratically to ξ. This means that for large k

$$|x_{k+1} - \xi| \leqq C\,|x_k - \xi|^2$$

for some positive constant C. The numbers x_k can of course be complex (if $f(x)$ is a real function, no real sequence can converge to a non-real root of $f(x) = 0$).

In the real case, Newton's method has a simple geometric meaning: x_{k+1} is the abscissa of the point at which the tangent to the curve $y = f(x)$, at the point

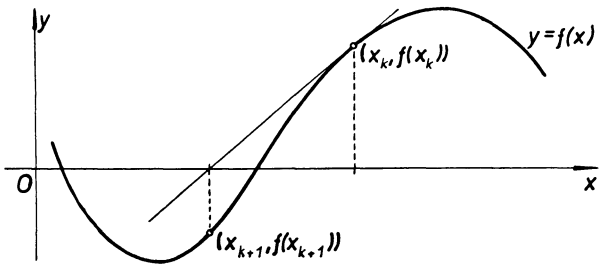


Fig. 31.1.

$(x_k, f(x_k))$, intersects the x-axis (Fig. 31.1). If in the whole interval $[a, b]$ the two derivatives $f'(x)$ and $f''(x)$ have unchanged signs and $f(a)f(b) < 0$, then Newton's process converges to some root of $f(x) = 0$ in $[a, b]$ if we choose $x_0 = a$ or $x_0 = b$ according to whether $f(a)$ or $f(b)$ has the same sign as $f''(x)$.

(d) The Regula Falsi Method. This method enables us to solve the equation $f(x) = 0$ ($f(x)$ being a real continuous function in an interval I), if two numbers x_0

and x_1 in the interval I are known such that $f(x_0)$ and $f(x_1)$ have opposite signs, i.e. $f(x_0)f(x_1) < 0$. In this case we construct a sequence $x_1, x_2, \ldots$ in the following manner:

We put

$$x_2 = \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)}. \tag{11}$$

If $f(x_2) = 0$, we have found a root. If $f(x_2) \neq 0$, then either $f(x_0)f(x_2) < 0$ or $f(x_1)f(x_2) < 0$. In the first case x_0 and x_2, or in the second case x_1 and x_2, satisfy

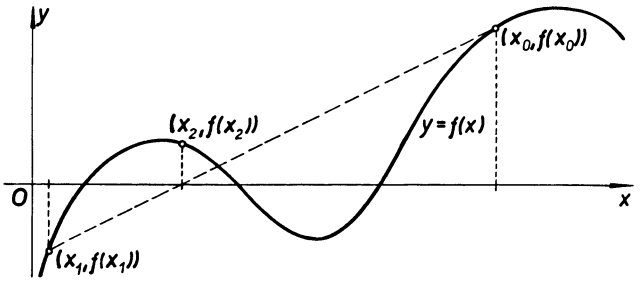


Fig. 31.2

the preceding condition and we compute x_3 from (11). Then, from x_3 and either x_2 or one of the previously chosen x_0 or x_1 we compute the number x_4, and repeat the computation to obtain $x_5, x_6, \ldots, x_k$. The sequence $\{x_k\}$ converges, but the convergence is usually slow.

The geometric meaning of this method is based upon the fact that x_2 in (11) is the abscissa of the point where the straight line connecting points $(x_0, f(x_0))$ and $(x_1, f(x_1))$ of the curve $y = f(x)$ meets the x-axis.

A similar process can be obtained if, under the same assumptions, the formula

$$x_2 = \frac{1}{2}(x_0 + x_1)$$

is used instead of formula (11).

(e) **Bairstow's method.** This iterative method is used for computing a factor of the given polynomial, mostly a real quadratic factor of a real polynomial. An initial approximation of the quadratic factor can be obtained by the Bernoulli-Whittaker method.

The idea is to perform the division of the given polynomial by the approximate factor and use the remainder to obtain a better approximation of the factor by solving a linearized system of equations. We describe the algorithm for a quadratic factor.

Let

$$f(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n, \qquad n \geqq 3,$$

be the given polynomial, let

$$\varphi_0(x) = x^2 - u_0x - v_0$$

be an approximate quadratic factor of $f(x)$. Define recursively

$$\begin{aligned} b_0 &= a_0, \\ b_1 &= a_1 + u_0b_0, \\ &\vdots \\ b_k &= a_k + u_0b_{k-1} + v_0b_{k-2}, \qquad k = 2, \ldots, n, \end{aligned}$$

and further

$$\begin{aligned} c_0 &= b_0, \\ c_1 &= b_1 + u_0c_0, \\ &\vdots \\ c_k &= b_k + u_0c_{k-1} + v_0c_{k-2}, \qquad k = 2, \ldots, n-1. \end{aligned}$$

If $c_{n-2}^2 - c_{n-1}c_{n-3} \neq 0$, set

$$u_1 = u_0 + \frac{b_nc_{n-3} - b_{n-1}c_{n-2}}{c_{n-2}^2 - c_{n-1}c_{n-3}}, \tag{12}$$

$$v_1 = v_0 + \frac{b_{n-1}c_{n-1} - b_nc_{n-2}}{c_{n-2}^2 - c_{n-1}c_{n-3}}, \tag{13}$$

and define the new quadratic factor

$$\varphi_1(x) = x^2 - u_1x - v_1.$$

If for the repeated process both the sequences $u_0, u_1, u_2, \ldots$ and $v_0, v_1, v_2, \ldots$ converge, say $u_k \to U$, $v_k \to V$, then the limit polynomial

$$\varphi(x) = x^2 - Ux - V$$

is a quadratic factor of $f(x)$, from which the roots of the polynomial can easily be computed.

Example 2. The polynomial on the left-hand side of the equation from Example 1,

$$4{\cdot}08x^4 - 6{\cdot}03x^3 + 6{\cdot}99x^2 - 9{\cdot}81x + 9{\cdot}72 = 0, \tag{14}$$

has an approximate quadratic factor $x^2 + 0{\cdot}6x + 1{\cdot}6$ (cf. (10)). Thus

$$u_0 = -0{\cdot}6, \qquad v_0 = -1{\cdot}6.$$

Individual steps of the algorithm are given in Tab. 31.3.

TABLE 31.3

k	a_k	b_k	c_k
0	4·08	4·08	4·08
1	−6·03	−8·478	−10·926
2	6·99	5·549	5·377
3	−9·81	0·425	14·560
4	9·72	0·587	–

Thus, by (12), (13)

$$u_1 = -0{\cdot}646, \qquad v_1 = -1{\cdot}585.$$

Continuing the procedure, we obtain

$$u_2 = -0{\cdot}647, \qquad v_2 = -1{\cdot}588.$$

The polynomial

$$\varphi_2(x) = x^2 + 0{\cdot}647x + 1{\cdot}588$$

is already a quadratic factor of the given polynomial (correct to three decimals, as can be shown). Thus two of the roots of equation (14) are

$$\alpha_{1,2} \doteq -0{\cdot}324 \pm 1{\cdot}218\mathrm{i}.$$

(f) The General Iterative Method. Let us write the given equation in an equivalent form

$$f_1(x) = f_2(x); \tag{15}$$

here, we choose the function $f_1(x)$ so that the equation $f_1(x) = c$ can easily be solved (for example $f_1(x)$ linear, of the form x^m, etc.). Then we take an initial approximation x_0 and construct the recurrent sequence $x_0, x_1, x_2, \ldots$ in such a manner that x_{k+1} is computed from the equation

$$f_1(x_{k+1}) = f_2(x_k). \tag{16}$$

If the sequence $x_0, x_1, x_2, \ldots$ tends to a limit z and if the two functions $f_1(x)$ and $f_2(x)$ are continuous at the point z, then z is a root of the given equation.

If the two function $f_1(x)$ and $f_2(x)$ have first derivatives in some neigbourhood of the root z for which

$$|f_1'(x)| > |f_2'(x)|,$$

then the sequence $x_0, x_1, x_2, \ldots$ is convergent whenever x_0 is close enough to z.

Example 3. We wish to determine by the iterative method that root of the transcendental equation $x^2 - x\tan x + 1 = 0$ which is near to the number 1. Write the given equation in the form $\tan x = x + 1/x$ and compute the first four terms of the sequence x_k, where $x_0 = 1$ and $\tan x_{k+1} = x_k + 1/x_k$. We obtain $x_1 \doteq 1{\cdot}1071$, $x_2 \doteq 1{\cdot}1092$, $x_3 \doteq 1{\cdot}1093$, $x_4 \doteq x_3$ so that the required root is equal to $1{\cdot}1093$ with error less than 10^{-4}.

31.5. Numerical Solution of (Nonlinear) Systems

We shall consider only the most important case of n equations for n unknowns $x_1, x_2, \ldots, x_n$ $(n \geqq 2)$:

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0, \\ f_2(x_1, x_2, \ldots, x_n) &= 0, \\ &\cdots\cdots\cdots\cdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0. \end{aligned} \tag{1}$$

Here we assume that all the functions f_i have continuous first partial derivatives with respect to $x_1, \ldots, x_n$ in some region containing the roots to be found and that the determinant of the *Jacobi-matrix*

$$\boldsymbol{J}(x_1, x_2, \ldots, x_n) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1}, & \dfrac{\partial f_1}{\partial x_2}, & \ldots, & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1}, & \dfrac{\partial f_2}{\partial x_2}, & \ldots, & \dfrac{\partial f_2}{\partial x_n} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \dfrac{\partial f_n}{\partial x_1}, & \dfrac{\partial f_n}{\partial x_2}, & \ldots, & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix} \tag{2}$$

(which we shall briefly denote by $\boldsymbol{J}(\boldsymbol{x})$) is not identically zero in this region.

The solution will be obtained by the (generalized) iterative method. It is a local method, i.e. it yields a solution as a limit of a convergent sequence of n-tuples $({}^k x_1, {}^k x_2, \ldots, {}^k x_n)$, $k = 0, 1, 2, \ldots$, under the assumption that the initial n-tuple $(k = 0)$ is already close enough to the solution and that for this solution the determinant of the matrix (2) is non-zero. Usually, the initial n-tuple is chosen either from a knowledge of the problem (physical, technical etc.) from which (1) has arisen, by trial, or, for small n, graphically.

To solve (1) by the iterative method, write the system (1) in some equivalent form

$$\begin{aligned} g_1(x_1, \ldots, x_n) &= h_1(x_1, \ldots, x_n), \\ g_2(x_1, \ldots, x_n) &= h_2(x_1, \ldots, x_n), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ g_n(x_1, \ldots, x_n) &= h_n(x_1, \ldots, x_n) \end{aligned} \tag{3}$$

satisfying two conditions: (a) the functions g_i should also have continuous first partial derivatives in the region mentioned, (b) the system

$$g_i(x_1, \ldots, x_n) = c_i \quad (i = 1, 2, \ldots, n)$$

should be easily solvable for any choice of c_i (for example, the g_i are linear functions or linear in x_k^m, etc.). We choose an initial approximation ${}^0\boldsymbol{x} = ({}^0x_1, {}^0x_2, \ldots, {}^0x_n)$ and construct a sequence ${}^1\boldsymbol{x}, {}^2\boldsymbol{x}, \ldots$ of n-tuples from the recurrence formula

$$g_i({}^{k+1}\boldsymbol{x}) = h_i({}^k\boldsymbol{x}) \quad (i = 1, 2, \ldots, n). \tag{4}$$

If the sequence converges (i.e. if all coordinates converge separately), the limit $\boldsymbol{z} = (z_1, \ldots, z_n)$ is a solution of the system (1).

The following theorem holds: Let a solution $\boldsymbol{z} = (z_1, \ldots, z_n)$ of (3) have the following properties: (i) $\det \boldsymbol{J}(\boldsymbol{z}) \neq 0$ and (ii) all (real or complex) roots λ of the equation

$$\det\left[\lambda \frac{\partial g_i}{\partial x_j}(\boldsymbol{z}) - \frac{\partial h_i}{\partial x_j}(\boldsymbol{z})\right] = 0$$

are smaller then 1 in modulus. Then there exists an n-dimensional region Ω such that whenever $\boldsymbol{x}$ is chosen in Ω, the sequence $\boldsymbol{x}$ converges to $\boldsymbol{z}$.

A special case of this method is Newton's method; here, equations (4) are of the form

$${}^{k+1}\boldsymbol{x} = {}^k\boldsymbol{x} - \left[\boldsymbol{J}({}^k\boldsymbol{x})\right]^{-1} \boldsymbol{f}({}^k\boldsymbol{x}),$$

where

$${}^k\boldsymbol{x} = \begin{bmatrix} {}^kx_1 \\ {}^kx_2 \\ \vdots \\ {}^kx_n \end{bmatrix}, \quad \boldsymbol{f}(x) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ f_2(\boldsymbol{x}) \\ \vdots \\ f_n(\boldsymbol{x}) \end{bmatrix}$$

and $\boldsymbol{J}(\boldsymbol{x})$ is the matrix in (2). This method converges rapidly in the neighbourhood of the solution.

Example 1. We have to find the minimum of the function

$$f(x, y) = 3x^3 + 2y^2 + xy^2 - 10x - 5y - 1 = 0$$

which lies near the point (1, 1). According to Theorem 12.12.1, this means solving the system $\partial f/\partial x = 0$, $\partial f/\partial y = 0$, i.e. the system

$$9x^2 + y^2 - 10 = 0, \qquad 4y + 2xy - 5 = 0$$

near the point (1, 1). Let us write this last system in the form

$$x = \frac{1}{3}\sqrt{(10 - y^2)},$$
$$y = \frac{1}{4}(5 - 2xy).$$

Let us choose $x_0 = 1$, $y_0 = 1$ and find the sequence of pairs (x_k, y_k) for which $x_{k+1} = \frac{1}{3}\sqrt{(10 - y_k^2)}$, $y_{k+1} = \frac{1}{4}(5 - 2x_k y_k)$ (see Tab. 31.3).

TABLE 31.4

k	x_k	y_k	$10 - y_k^2$	$2x_k y_k$
0	1·0000	1·0000	9·00000	2·0000
1	1·0000	0·7500	9·43750	1·5000
2	1·0240	0·8750	9·23437	1·7920
3	1·0129	0·8020	9·35680	1·6355
4	1·0196	0·8411	9·29251	1·7152
5	1·0161	0·8212	9·32563	1·6688
6	1·0179	0·8328	9·30644	1·6954
7	1·0169	0·8261	9·31748	1·6801
8	1·0171	0·8300	9·31110	1·6884
9	1·0171	0·8279	9·31458	1·6841
10	1·0173	0·8290	9·31276	1·6867
11	1·0172	0·8283	9·31387	1·6852
12	1·0173	0·8287	9·31324	1·6861
	1·0173	0·8285		

We see that the solution of the system is $x \doteq 1{\cdot}0173$, $y \doteq 0{\cdot}8285$. Since

$$\frac{\partial^2 f}{\partial x^2}\frac{\partial^2 f}{\partial y^2} - \left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 > 0 \quad \text{and} \quad \frac{\partial^2 f}{\partial x^2} > 0,$$

this is a local minimum. The corresponding value of the given function is $-10{\cdot}086$.

32. APPROXIMATION. INTERPOLATION, SPLINES

By EMIL VITÁSEK

References: [5], [38], [50], [52], [55], [104], [147], [150], [196], [215], [234], [292], [295], [359], [378], [412], [443].

The problem of approximating a given or a sought function by a function which is in some sense simpler is very frequent in numerical analysis. Usually, the approximating function is a linear combination of a finite number of functions which are given in advance. Thus, if in this case g_k, $k = 0, \ldots, n$, is a finite sequence of given functions, then under the approximation of a function f we understand a linear combination $c_0 g_0 + \cdots + c_n g_n$, where $c_0, \ldots, c_n$ are constants which are to be determined according to some criterion. The choice of this criterion is, hence, the central point of the problem of approximation. In this chapter, we will investigate two possibilities:

(i) The approximated function f as well as the functions $g_0, \ldots, g_n$ are supposed to be elements of a linear normed space (about normed spaces see § 22.4). Then the coefficients are chosen in such a way that the error has minimal norm.

(ii) We choose in advance a finite set of points in the domain of definition of f and demand that the values of the approximation, and possibly also of some of its derivatives, agree with the corresponding values of f, or of its derivatives, at those points. In this case, we speak about the *interpolation approximation* or, briefly, *interpolation.*

A further important possibility is the *least squares approximation.* In such an approximation, the coefficients are chosen from the condition that the sum of squares of the differences between f and its approximation taken over a fixed finite set of points be minimal. This approach is important especially for fitting curves to empirical data which are influenced by random errors. For this reason this case is investigated separately in Chap. 35.

32.1. The Best Approximation in a Linear Normed Space

Let X be a linear normed space (not necessarily complete) over the field of real or complex numbers, f any element of X and let $g_0, \ldots, g_n$ be $n+1$ linearly independent elements of X. The linear span of $g_0, \ldots, g_n$ (i.e., the space of all linear combinations of $g_0, \ldots, g_n$) is an $(n+1)$-dimensional subspace of X; denote it by K.

Definition 1. An element $g^* \in K$ is called the *best approximation of* $f \in X$ *in the subspace* K if

$$\|f - g^*\| = \inf_{g \in K} \|f - g\| . \tag{1}$$

Theorem 1. *The best approximation in a linear normed space always exists.*

Example 1. Let X be the space $C([-1, 1])$ of functions of one real variable which are continuous on $[-1, 1]$ and in which the norm is defined by $\|f\| = \max_{x \in [-1,1]} |f(x)|$ (cf. § 22.4). Approximate the function $f(x) = x^3$, $x \in [-1, 1]$, by an element of the linear subspace spanned by the functions $g_0 = 1$ and $g_1 = x^2$. Thus, the approximating function has the form $c_0 + c_1 x^2$.

First, we assert that

$$\|x^3 - (c_0 + c_1 x^2)\| = \max_{x \in [-1,1]} |x^3 - (c_0 + c_1 x^2)| \geqq 1 . \tag{2}$$

Really, if (2) did not hold it would be

$$|x^3 - (c_0 + c_1 x^2)| < 1 \tag{3}$$

for any $x \in [-1, 1]$. If we substitute in (3) -1 and $+1$ for x, we obtain

$$\begin{aligned} -1 < -1 - (c_0 + c_1) < 1 , \\ -1 < 1 - (c_0 + c_1) < 1 \end{aligned} \tag{4}$$

and these inequalities have to be satisfied simultaneously. However, from the first inequality in (4), $c_0 + c_1 < 0$ follows while the second one implies $c_0 + c_1 > 0$, and this is a contradiction proving the validity of (2). The error of the approximation of the function x^3 by a function of the form $c_0 + c_1 x^2$ thus cannot be smaller than 1.

Further, let ε be any real number satisfying $|\varepsilon| \leqq 1$. For this ε and for $x \in [-1, 1]$, we have

$$\begin{aligned} |x^3 - \varepsilon(1 - x^2)| \leqq |x^3| + |\varepsilon|(1 - x^2) \leqq x^2 + |\varepsilon|(1 - x^2) = \\ = 1 - (1 - x^2) + |\varepsilon|(1 - x^2) = \\ = 1 - (1 - x^2)(1 - |\varepsilon|) \leqq 1 , \end{aligned}$$

since the number $(1 - x^2)(1 - |\varepsilon|)$ is non-negative. Thus,

$$\|x^3 - \varepsilon(1 - x^2)\| \leqq 1$$

holds for any $|\varepsilon| \leqq 1$. This inequality and the inequality (2) imply that

$$\|x^3 - \varepsilon(1 - x^2)\| = 1 .$$

The last equation indicates that the function of the form $\varepsilon(1 - x^2)$, where $|\varepsilon| \leqq 1$, is the best approximation of x^3 in the space of functions of the form $c_0 + c_1 x^2$.

This simple example thus shows that the element of best approximation need not be determined uniquely.

Definition 2. A linear normed space is called *sharply normed* if the equality sign in the triangle inequality

$$\|f_1 + f_2\| \leqq \|f_1\| + \|f_2\| \tag{5}$$

occurs if and only if $f_2 = \alpha f_1$, where $\alpha \geqq 0$ or $f_1 = 0$.

REMARK 1. Any normed space in which the norm is defined by inner (scalar) product (see § 22.4) is sharply normed. Also all Banach spaces L_p of functions the p-th power of which is integrable are sharply normed for $1 < p < \infty$. On the other hand, the space $C([-1, 1])$ is an example of a space which is not sharply normed.

Theorem 2. *In a linear normed space which is sharply normed the best approximation is determined uniquely.*

32.2. The Best Approximation in a Hilbert Space

Since any Hilbert space is a sharply normed linear space, existence and uniqueness of the element of the best approximation follows from Theorems 32.1.1 and 32.1.2. The next theorem describes another important property of this element.

Theorem 1. *Let K be a finite-dimensional subspace of a Hilbert space H and let $f \in H$. An element $g^* \in K$ is the element of the best approximation of f if and only if*

$$(f - g^*, g) = 0 \quad \textit{for any } g \in K . \tag{1}$$

REMARK 1. The geometrical meaning of Theorem 1 is that the error of the best approximation is orthogonal to the subspace K. The element of the best approximation is thus the orthogonal projection of the approximated element into the subspace K.

REMARK 2. The uniquely determined element of the best approximation exists and is characterized by (1) even in the case that K is any (not necessarily finite-dimensional) closed subspace of H.

If $g_0, \ldots, g_n$ is the basis in K and if we take $g^* = c_0 g_0 + \cdots + c_n g_n$ for the best approximation of f, then the coefficients satisfy the system of linear algebraic equations

$$\sum_{j=0}^{n} (g_j, g_i) c_j = (f, g_i), \quad i = 0, \ldots, n, \tag{2}$$

the matrix of which is called the *Gram matrix* and is Hermitian (i.e. $(g_j, g_i) = \overline{(g_i, g_j)}$) and positive definite (see § 1.29).

REMARK 3. The element of the best approximation can be expressed in a substantially simpler form if the basis $g_0, \ldots, g_n$ is orthonormal, i.e., if it satisfies $(g_i, g_j) = 1$ for $i = j$ and $(g_i, g_j) = 0$ for $i \neq j$. In this case, the Gram matrix is the identity matrix and the element of best approximation is given by

$$g^* = \sum_{i=0}^{n} (f, g_i) g_i . \tag{3}$$

Let us now investigate the problem of convergence, i.e., the problem of the behaviour of elements of the best approximation in the case when the dimension of subspaces in which the approximation is sought tends to infinity.

Thus, let $g_0, \ldots, g_n, \ldots$ be a sequence of normed and pairwise orthonormal elements of H. Denote by H_n the space spanned by $g_0, \ldots, g_n$. Hence, the element of the best approximation is given by (3).

Theorem 2. *Let f be any element of H. Then the series*

$$\sum_{i=0}^{\infty} (f, g_i) g_i \tag{4}$$

converges (in the sense of convergence in the space H).

The series (4) is called the *generalized Fourier series* (cf. §§ 22.4 and 16.2).

Theorem 3. *The equality $f = \sum_{i=0}^{\infty} (f, g_i) g_i$ holds if and only if the sequence g_i, $i = 0, 1, \ldots$, is complete in H.*

For the definition of completeness of a sequence, see Definition 22.4.11. The criteria of completeness are described in the following theorem (cf. also Theorems 22.4.7 and 22.4.8).

Theorem 4. *The conditions*

(i) *the sequence* g_i, $i = 0, 1, \ldots$, *is complete in* H;

(ii) *the only element which is orthogonal to all* g_i*'s is the zero element*;

(iii) *for any* $f \in H$ *we have*

$$\|f\|^2 = \sum_{i=0}^{\infty} |(f, g_i)|^2 \tag{5}$$

are equivalent.

The identity (5) is called *Parceval's equality.*

32.3. The Best Approximation of Continuous Functions by Polynomials

Let C be the space of real continuous functions defined on $[a, b]$ with the usual norm $\|f\| = \max_{x\in[-1,1]} |f(x)|$. Since the polynomials of degree at most n obviously form a finite-dimensional subspace of C, there exists, for any $f \in C$, a polynomial Q_n^* of degree at most n such that

$$E_n(f) \equiv \|f - Q_n^*\| = \inf \|f - Q_n\|, \tag{1}$$

where inf is taken over all polynomials Q_n of degree n or less (see § 32.1). This polynomial is called the *polynomial of the best uniform approximation* or the *polynomial of the Chebyshev approximation.* It is also called the *minimax approximation of* f since its maximal error is minimal in $[a, b]$.

Theorem 1 (*Vallé-Poussin*). *Let* Q_n *be a polynomial of degree at most* n. *Let there exist* $n + 2$ *points* $x_0 < x_1 < \cdots < x_{n+1}$ *from* $[a, b]$ *such that*

$$\operatorname{sign}\{(-1)^i[f(x_i) - Q_n(x_i)]\} = \text{const.} \quad \text{for } i = 0, \ldots, n + 1 \tag{2}$$

(thus, the function $f - Q_n$ changes its sign when passing from any of the points x_i to the following one). *Then*

$$E_n(f) \geqq \min_{i=0,\ldots,n+1} |f(x_i) - Q_n(x_i)|. \tag{3}$$

Theorem 2 (*Chebyshev*). *A polynomial* Q_n *is the polynomial of best uniform approximation if and only if there exist at least* $n + 2$ *points* $x_0 < x_1 < \ldots < x_m$ $(m \geqq n + 1)$ *in* $[a, b]$ *such that*

$$f(x_i) - Q_n(x_i) = \alpha(-1)^i \|f - Q_n\|, \quad i = 0, \ldots, m, \tag{4}$$

and the number α, *common for all* i*'s, is equal to* 1 *or* -1.

REMARK 1. The property expressed in Theorem 2 means that the function $f - Q_n$ attains alternately its absolute maximum and minimum over the interval $[a, b]$ at the points $x_0, x_1, \ldots, x_m$. This property is called the *Chebyshev alternating property* and the corresponding set the *Chebyshev alternating set*. Hence, Theorem 2 says that the Chebyshev alternating property completely characterizes the polynomial of best uniform approximation.

Theorem 3. *The polynomial of best uniform approximation is uniquely determined.*

REMARK 2. Since the space C is not sharply normed, Theorem 3 implies that the condition on the space to be sharply normed is not necessary for uniqueness.

REMARK 3. Theorems 1 to 3 can be generalized to the case when given continuous function is approximated by a "generalized" polynomial, i.e., by a function of the form $c_0 g_0(x) + \cdots + c_n g_n(x)$ where $g_0, \ldots, g_n$ are linearly independent functions which satisfy the Haar condition. By the *Haar condition*, we mean here the fact that any generalized polynomial $c_0 g_0(x) + \cdots + c_n g_n(x)$, which has more than n different zeros in $[a, b]$, is identically equal to zero.

REMARK 4. Theorems similar to Theorems 1 to 3 can be formulated also for the case when the approximating function is a rational function.

Example 1. We have to determine a polynomial of the n-th degree with the coefficient at x^n equal to 1 whose deviation from zero (in the sense of the norm in C) is minimal on the interval $[-1, 1]$.

This problem can obviously be reformulated as a problem of finding a polynomial of degree at most $n - 1$ which is, in $[-1, 1]$, the polynomial of best uniform approximation of the function x^n. Thus, the polynomial sought (let us denote it by T_n) is completely characterized by the property that the number of consecutive points from $[-1, 1]$ at which its value is, with alternating signs, equal to $\|T_n\|$, is not less than $n + 1$ (this fact immediately follows from Theorem 2).

We assert that the polynomial T_n can be written in the form

$$T_n(x) = \frac{1}{2^{n-1}} \cos(n \arccos x). \tag{5}$$

Really, from formulae 2.5.4 it immediately follows that T_n is a polynomial. Further, we have $\|T_n\| = 1/2^{n-1}$ as follows from (5) at the first glance. Finally, if we put

$$x_k = -\cos\frac{k\pi}{n} = \cos\frac{(n-k)\pi}{n}, \quad k = 0, \ldots, n,$$

then

$$-1 = x_0 < x_1 < \cdots < x_n = 1$$

and,

$$T_n(x_k) = \frac{1}{2^{n-1}} \cos \frac{(n-k)\pi}{n} = \frac{(-1)^{n-k}}{2^{n-1}} = (-1)^{n-k} \|T_n\| .$$

The polynomial which we have constructed in Example 1 is called the *Chebyshev polynomial.*

32.4. Jackson's Theorems

In this paragraph, we introduce some important theorems concerning the behaviour of the quantity $E_n(f)$ defined by (32.3.1).

Definition 1. We say that f satisfies the *Hölder condition* in $[a, b]$ if there exist constants $K > 0$ and α, $0 < \alpha \leqq 1$, such that

$$|f(x) - f(y)| \leqq K|x - y|^{\alpha}, \quad x, y \in [a, b] . \tag{1}$$

REMARK 1. For $\alpha = 1$, we apparently obtain the well-known Lipschitz condition.

Definition 2. We say that the function f *belongs to the class* $C^{n,\alpha}$ on $[a, b]$ ($n \geqq 0$ an integer, $0 < \alpha \leqq 1$), if its n-th derivative exists and satisfies the Hölder condition with exponent α in $[a, b]$.

Definition 3. Let f be a positive function defined on $[a, b]$. The function $\omega_f(t)$ defined for positive t's by the formula

$$\omega_f(t) = \sup_{\substack{x, y \in [a, b] \\ |x-y| \leqq t}} |f(x) - f(y)| , \tag{2}$$

where also ∞ is admitted as a function value, is called the *modulus of continuity* (or, in more detail, the *uniform modulus of continuity*) of f.

Theorem 1. *A function f is uniformly continuous on $[a, b]$ if and only if*

$$\lim_{t \to 0+} \omega_f(t) = 0 . \tag{3}$$

A function f is an element of $C^{0,\alpha}$ if and only if

$$\omega_f(t) \leqq Mt^{\alpha} . \tag{4}$$

REMARK 2. The modulus of continuity of a continuous function measures the rate of convergence of $f(x)$ to $f(y)$ for $x \to y$, where y is any point from $[a, b]$. In addition, there exists, to any sequence of numbers converging to zero, a continuous function whose modulus of continuity converges to zero more slowly than the given sequence.

Theorem 2 (*1st Jackson's Theorem*). *If $E_n(f)$ is the error of best uniform approximation of a function $f \in C$ by a polynomial, then*

$$E_n(f) \leqq 12\,\omega_f\left(\frac{b-a}{2n}\right). \tag{5}$$

Theorem 3 (*2nd Jackson's Theorem*). *If f has p continuous derivatives in $[a, b]$ and if we denote the modulus of continuity of $f^{(p)}$ by $\omega_p(t)$, then we have, for $n > p$, that*

$$E_n(f) \leqq \frac{C_p(b-a)^p}{n^p}\,\omega_p\left(\frac{b-a}{2(n-p)}\right), \tag{6}$$

where C_p is a constant depending only on p.

Corollary 1. *Let the assumption of Theorem 2 be satisfied and let, moreover, $f^{(p)} \in C^{0,\alpha}$. Then*

$$E_n(f) \leqq \left(\frac{p+1}{2}\right)^{\alpha} C_p M \frac{(b-a)^{p+\alpha}}{n^{p+\alpha}}, \tag{7}$$

where M is the constant from (4).

Corollary 2. *If $|f^{(p+1)}(x)| \leqq M_{p+1}$ for $x \in [a, b]$, then*

$$E_n(f) \leqq \frac{p+1}{2} C_p M_{p+1} \frac{(b-a)^{p+1}}{n^{p+1}}. \tag{8}$$

Corollary 3. *If f has derivatives of all orders, then*

$$\lim_{n\to\infty}(n^p E_n(f)) = 0 \tag{9}$$

holds for any p.

32.5. The Remes Algorithm

In those situations, when we have to approximate a function by a polynomial at a point about which we know a priori only that it lies in an interval but we

do not know exactly its position, the best uniform approximation can be useful since it minimizes the maximal error in the considered interval. In what follows, we therefore present an iterative procedure for its computation. It is based on the Vallé-Poussin Theorem 32.3.1 and consists in improving the Chebyshev alternating set.

Assume that we have found the k-th approximation of the Chebyshev alternating set, i.e., we have at our disposal the points

$$a \leqq x_0^{(k)} < \cdots < x_{n+1}^{(k)} \leqq b. \tag{1}$$

First, solve the system of $n+2$ linear algebraic equations

$$\sum_{j=0}^{n} a_j^{(k)} (x_i^{(k)})^j + (-1)^i E^{(k)} = f(x_i^{(k)}), \quad i = 0, \ldots, n+1, \tag{2}$$

for $n+2$ unknowns $a_0^{(k)}, \ldots, a_n^{(k)}$ and $E^{(k)}$. This is possible since the determinant of (2) is always different from zero. Put

$$Q^{(k)}(x) = \sum_{j=0}^{n} a_j^{(k)} x^j. \tag{3}$$

Further, determine the number $y^{(k+1)}$ in $[a, b]$ in such a way that the function

$$R^{(k)} = f - Q^{(k)} \tag{4}$$

attains its absolute extreme at $y^{(k+1)}$.

The next, $(k+1)$-st approximation of the alternating set is then defined as a set containing always $y^{(k+1)}$ and, further, some suitable points from the preceding approximation.

If $E^{(k)} = 0$, then the $(k+1)$-st approximation is the set containing $y^{(k+1)}$ and any $n+1$ points from the k-th approximation. The new approximation thus arises in such a way that an arbitrary point of the k-th approximation is replaced by $y^{(k+1)}$ and the resulting set is then ordered according to the magnitude of its elements.

If $E^{(k)} \neq 0$, the following three cases take place:

(i) $y^{(k+1)} \in [a, x_0^{(k)}]$,

(ii) $y^{(k+1)} \in [x_{n+1}^{(k)}, b]$,

(iii) there exists an index i_0, $0 \leqq i_0 \leqq n$, such that $y^{(k+1)} \in [x_{i_0}^{(k)}, x_{i_0+1}^{(k)}]$.

In case (i) we set $x_0^{(k+1)} = y^{(k+1)}$ and, for $i = 1, \ldots, n+1$, we set $x_i^{(k+1)} = x_i^{(k)}$ if $R^{(k)}(x_0^{(k)}) R^{(k)}(y^{(k+1)}) > 0$, or $x_i^{(k+1)} = x_{i-1}^{(k)}$ if $R^{(k)}(x_0^{(k)}) R^{(k)}(y^{(k+1)}) < 0$. Thus, the point $x_0^{(k)}$ or $x_{n+1}^{(k)}$ is omitted from the k-th approximation.

In case (ii) we set $x_{n+1}^{(k+1)} = y^{(k+1)}$ and, for $i = 0, \ldots, n$, we set $x_i^{(k+1)} = x_i^{(k)}$ if $R^{(k)}(x_{n+1}^{(k)})R^{(k)}(y^{(k+1)}) > 0$, or $x_i^{(k+1)} = x_{i+1}^{(k)}$ if $R^{(k)}(x_{n+1}^{(k)})R^{(k)}(y^{(k+1)}) < 0$. Hence, here the point $x_{n+1}^{(k)}$ or $x_0^{(k)}$ is omitted.

In case (iii) we set $x_{i_0}^{(k+1)} = y^{(k+1)}$ and $x_i^{(k+1)} = x_i^{(k)}$ for $i = 0, \ldots, n$, $i \neq i_0$, if $R^{(k)}(x_{i_0}^{(k)})R^{(k)}(y^{(k+1)}) > 0$, or $x_{i_0+1}^{(k+1)} = y^{(k+1)}$ and $x_i^{(k+1)} = x_i^{(k)}$ for $i = 0, \ldots, n$, $i \neq i_0 + 1$, if $R^{(k)}(x_{i_0}^{(k)})R^{(k)}(y^{(k+1)}) < 0$. Thus, the point $x_{i_0}^{(k)}$ or $x_{i_0+1}^{(k)}$ is omitted from the preceding approximation.

If f is sufficiently smooth, the polynomials $Q^{(k)}$ converge to the polynomial of best uniform approximation for an arbitrary initial approximation. The convergence, however, is the faster the closer the starting set is to the Chebyshev alternating set. Thus, if we succeed in finding a polynomial R of degree n having the property that the difference $f - R$ attains its local extrema, the signs of which alternate at $n + 2$ successive points, then such set will be a suitable starting approximation for the described algorithm.

We now show two relatively simple ways to construct such polynomial approximations of the given function that their error has the above property.

(a) Chebyshev's Expansions

Chebyshev's polynomials (32.3.5) are orthogonal in $[-1, 1]$ with weight $1/(1-x^2)^{1/2}$ (see also § 16.6) and a function f possessing suitable properties can be expanded in the generalized Fourier series

$$f(x) = \frac{1}{2}c_0 + \sum_{j=1}^{\infty} c_j T_j(x), \tag{5}$$

where

$$c_j = \frac{2^{2j-1}}{\pi} \int_{-1}^{1} \frac{f(x)T_j(x)}{(1-x^2)^{1/2}}\,dx. \tag{6}$$

If we truncate the series in (5) after $n + 1$ terms, we obtain a polynomial which is usually a very good initial approximation for Remes' algorithm.

(b) Economized Power Series

The approximation of a given (sufficiently smooth) function f by the polynomial

$$C_n(x) = \sum_{j=0}^{n+1} \frac{1}{j!} f^{(j)}(0)x^j - \frac{1}{(n+1)!} f^{(n+1)}(0)\alpha^{n+1} T_{n+1}\left(\frac{x}{\alpha}\right), \tag{7}$$

where T_{n+1} is the Chebyshev polynomial of degree $n+1$, has very similar properties on the interval $[-\alpha, \alpha]$ as the approximation gained from the truncated Chebyshev expansion. At the same time, the construction of C_n is substantially easier. Note that C_n is really a polynomial of degree n since the terms with x^{n+1} cancel.

32.6. Polynomial Interpolation. Lagrange's Interpolation Formula. Hermite's Interpolation Formula

Let a function f of a real variable, $n+1$ mutually different points $a_0, \ldots, a_n$ and $n+1$ positive integers $r_0, \ldots, r_n$ be given. The basic problem of the polynomial interpolation is the problem of finding the polynomial P of least possible degree satisfying, for $i = 0, \ldots, n$, the conditions

$$f^{(j)}(a_i) = P^{(j)}(a_i), \quad j = 0, \ldots, r_i - 1. \tag{1}$$

Definition 1. The polynomial P of above properties is called the *interpolation polynomial* and the points $a_0, \ldots, a_n$ at which we ask for the agreement of the values of f and, possibly, also of the values of its derivatives up to some definite order with the corresponding values of P, or of its derivatives, are called the *nodes of intrepolation* or *tabular points.*

Theorem 1. *There exists exactly one polynomial of degree at most $m-1$, where*

$$m = r_0 + r_1 + \cdots + r_n, \tag{2}$$

which satisfies (1).

Theorem 2. *Let f have m continuous derivatives in an interval $[a, b]$ which contains the nodes $a_0, \ldots, a_n$. Further, let x be any point from $[a, b]$. Then there exists $\xi \in [a, b]$ such that*

$$f(x) - P(x) = \frac{1}{n!}\Omega(x) f^{(m)}(\xi), \tag{3}$$

where

$$\Omega(x) = (x - a_0)^{r_0} \ldots (x - a_n)^{r_n}.$$

REMARK 1. The formula (3) gives the error which is made when replacing f by its interpolation polynomial. It is not possible to compute the error directly from it since the number ξ depends on x in a manner which is not known, in general. Its importance consists mainly in the fact that it allows to bound the error of interpolation if we are able to bound the m-th derivative of the interpolated function.

REMARK 2. The error of an interpolation formula is also called its *remainder*.

The equations (1) form a system of linear algebraic equations for determining coefficients of a general interpolation polynomial. In special cases, when $r_i = 1$ and $r_i = 2$ for $i = 0, \ldots, n$, we speak about *Lagrange's* and *Hermite's interpolation*, respectively. In those situations the interpolation polynomials can be expressed by simple formulae.

Theorem 3. *Let*

$$\omega(x) = (x - a_0) \ldots (x - a_n) \tag{4}$$

and

$$l_i(x) = \frac{\omega(x)}{(x - a_i)\omega'(a_i)}, \quad i = 0, \ldots, n. \tag{5}$$

Then the polynomial $L_n(x)$ defined by

$$L_n(x) = \sum_{i=0}^{n} f(a_i) l_i(x) \tag{6}$$

has the degree at most n and satisfies

$$L_n(a_i) = f(a_i) \tag{7}$$

for $i = 0, \ldots, n$.

Definition 2. The polynomial L_n given by (6) is called *Lagrange's interpolation polynomial* and the polynomials l_i are called *elementary polynomials of Lagrange's interpolation.*

REMARK 3. The elementary polynomial $l_i(x)$ of Lagrange's interpolation is a polynomial of degree n which attains the value 1 at a_i and the values 0 at the other tabular points. Alternately, it is possible to write

$$l_i(x) = \frac{(x - a_0) \ldots (x - a_{i-1})(x - a_{i+1}) \ldots (x - a_n)}{(a_i - a_0) \ldots (a_i - a_{i-1})(a_i - a_{i+1}) \ldots (a_i - a_n)}. \tag{8}$$

REMARK 4. Supposing that the function f has $m = n + 1$ continuous derivatives, the error of the Lagrange interpolation polynomial is given by (3), where the function Ω is replaced by ω defined in (4).

Theorem 4. *The polynomial H_{2n+1}, defined by*

$$H_{2n+1}(x) = \sum_{i=0}^{n} [f(a_i) h_i(x) + f'(a_i) \tilde{h}_i(x)], \tag{9}$$

where

$$h_i(x) = [1 - 2(x - a_i)l_i'(a_i)]l_i^2(x), \qquad \tilde{h}_i(x) = (x - a_i)l_i^2(x) \tag{10}$$

and l_i are elementary polynomials of the Lagrange interpolation, is of degree at most $2n + 1$ and satisfies

$$H_{2n+1}(a_i) = f(a_i), \qquad H_{2n+1}'(a_i) = f'(a_i), \quad i = 0, \ldots, n. \tag{11}$$

Definition 3. The polynomial (9) is called the *Hermite interpolation polynomial.*

REMARK 5. The error of the Hermite interpolation polynomial is given again by (3). Now we must put $m = 2n + 1$ and $\Omega = \omega^2$ in it.

REMARK 6. The general interpolation polynomial from Definition 1 is often called the *Hermite interpolation polynomial*, too.

The Lagrange interpolation polynomial can alternatively be written in the Newton form using the concept of a divided difference.

Definition 4. Let $n + 1$ mutually different points x_i, $i = 0, \ldots, n$, be given. The *first divided difference* $f[x_0, x_1]$ is defined by

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}. \tag{12}$$

Generally, the *n-th divided difference* is defined recurrently by

$$f[x_0, \ldots, x_n] = \frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0}. \tag{13}$$

Theorem 5. *The n-th divided difference $f[x_0, \ldots, x_n]$ can be expressed in the form*

$$f[x_0, \ldots, x_n] = \sum_{i=0}^{n} \frac{f(x_i)}{\prod_{\substack{k=0 \\ k \neq i}}^{n} (x_i - x_k)}. \tag{14}$$

Theorem 6. *The Lagrange interpolation polynomial from formula* (6) *can be written in the form*

$$L_n(x) = f(a_0) + (x - a_0)f[a_0, a_1] + (x - a_0)(x - a_1)f[a_0, a_1, a_2] + \ldots \\ \cdots + (x - a_0)(x - a_1)\ldots(x - a_{n-1})f[a_0, a_1, \ldots, a_n] \tag{15}$$

and its remainder in the form

$$f(x) - L_n(x) = (x - a_0) \dots (x - a_n) f[x, a_0, \dots, a_n].$$

Definition 5. The polynomial on the right-hand side of (15) is called the *general Newton interpolation polynomial.*

REMARK 7. Notice that it was not necessary to assume that the points $a_0, \dots, a_n$ are ordered according to their magnitude.

32.7. Differences. Interpolation Polynomial for Equidistant Arguments

If the tabular points are *equidistant*, i.e., if

$$a_k = a_0 + kh, \quad k \text{ an integer}, \tag{1}$$

then the Lagrange polynomial can be written in terms of differences.

Definition 1. The *k-th forward difference* (k is a non-negative integer) *of a function f at the point x* is the number $\Delta^k f(x)$ defined by the recurrence

$$\begin{aligned} \Delta^0 f(x) &= f(x), \\ \Delta^k f(x) &= \Delta^{k-1} f(x + h) - \Delta^{k-1} f(x). \end{aligned} \tag{2}$$

Similarly, the *k-th backward difference of a function f at the point x* is the number $\nabla^k f(x)$ defined by

$$\begin{aligned} \nabla^0 f(x) &= f(x), \\ \nabla^k f(x) &= \nabla^{k-1} f(x) - \nabla^{k-1} f(x - h). \end{aligned} \tag{3}$$

Theorem 1. *The k-th forward difference of a function f can be expressed as a linear combination of the values of f in the form*

$$\Delta^k f(x) = \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i} f(x + ih) \tag{4}$$

and the k-th backward difference in the form

$$\nabla^k f(x) = \sum_{i=0}^{k} (-1)^i \binom{k}{i} f(x - ih). \tag{5}$$

Theorem 2. *For any non-negative integer k we have*

$$\nabla^k f(x) = \Delta^k f(x - kh) \tag{6}$$

and

$$\Delta^k f(x) = \nabla^k f(x + kh). \tag{7}$$

Theorem 3. *The value of a function f at the point $x + kh$ and $x - kh$ (k is a non-negative integer) can be expressed as a linear combination of the 0-th up to the k-th forward and backward differences of f at x, respectively, according to the formulae*

$$f(x + kh) = \sum_{i=0}^{k} \binom{k}{i} \Delta^i f(x) \tag{8}$$

and

$$f(x - kh) = \sum_{i=0}^{k} (-1)^i \binom{k}{i} \nabla^i f(x), \tag{9}$$

respectively.

Theorem 4. *Let a point x, equidistant points $a_0, \ldots, a_n$ or $a_0, \ldots, a_{-n}$ and a function f defined on the interval I spanned by the points $a_0, \ldots, a_n, x$ or $a_0, \ldots, a_{-n}, x$, respectively, be given. Further, let f have $n + 1$ continuous derivatives on I. Then there exist points $\xi_1, \xi_2 \in I$ such that*

$$f(x) = \sum_{i=0}^{n} \binom{s}{i} \Delta^i f(a_0) + \binom{s}{n+1} h^{n+1} f^{(n+1)}(\xi_1), \tag{10}$$

or

$$f(x) = \sum_{i=0}^{n} (-1)^i \binom{-s}{i} \nabla^i f(a_0) + (-1)^{n+1} \binom{-s}{n+1} h^{n+1} f^{(n+1)}(\xi_2), \tag{11}$$

respectively, where

$$s = \frac{x - a_0}{h} \tag{12}$$

and

$$\binom{m}{i} = \frac{m(m-1)\ldots(m-i+1)}{i!}. \tag{13}$$

REMARK 1. The formula (13) must be used exactly in the form indicated and not in the form $m!/[i!(m-i)!]$ which holds only for an integer $m \geqq i$.

REMARK 2. Taking into account the definition of s in (12) we see that the sums in the right-hand terms of (10) and (11) are really polynomials in x of degree at most n.

Definition 2. The formula (10), or (11) is called *Newton's forward*, or *Newton's backward interpolation formula*, respectively.

REMARK 3. The polynomials in (10) and (11) are polynomials of degree at most n attaining the same values as the given function f at points $a_0, \ldots, a_n$, or $a_0, \ldots, a_{-n}$. They are, naturally, equal to the Lagrange interpolation polynomial for the same tabular points. To be able to construct such a polynomial it is not necessary to assume that f has derivatives up to the order $n+1$. This smoothness enables us only to write the interpolation polynomials with remainders.

Definition 3. The *k-th central difference of a function f at the point x* is the number $\delta^k f(x)$ defined recurrently by the formulae

$$\begin{aligned} \delta f(x) &= f(x + h/2) - f(x - h/2)\,, \\ \delta^k f(x) &= \delta^{k-1} f(x + h/2) - \delta^{k-1} f(x - h/2)\,. \end{aligned} \tag{14}$$

Theorem 5. *For any positive integer k, we have*

$$\delta^k f(x + kh/2) = \Delta^k f(x)\,. \tag{15}$$

Theorem 6. *Let a point x, equidistant nodes $a_{-n}, \ldots, a_0, \ldots, a_n$ and a function f defined on the interval I spanned by $a_{-n}, \ldots, a_n$ and x be given. Further, let f have $2n+1$ continuous derivatives in I. Then there exist points $\xi_1, \xi_2 \in I$ such that*

$$f(x) = f(a_0) + \sum_{i=1}^{n} \left[\binom{s+i-1}{2i-1} \delta^{2i-1} f\left(a_0 + \frac{h}{2}\right) + \binom{s+i-1}{2i} \delta^{2i} f(a_0) \right] + \\ + \binom{s+n}{2n+1} h^{2n+1} f^{(2n+1)}(\xi_1) \tag{16}$$

and

$$f(x) = f(a_0) + \sum_{i=1}^{n} \left[\binom{s+i-1}{2i-1} \delta^{2i-1} f\left(a_0 - \frac{h}{2}\right) + \binom{s+i}{2i} \delta^{2i} f(a_0) \right] + \\ + \binom{s+n}{2n+1} h^{2n+1} f^{(2n+1)}(\xi_2)\,, \tag{17}$$

where s is defined by (12) *again.*

REMARK 4. The odd central differences at the point $a_0 + h/2$ or $a_0 - h/2$, respectively, are given by the values of f at tabular points. Thus, the values of f only at tabular points occur in the right-hand terms of (16) or (17).

Definition 4. The formula (16) is called the *Gauss forward interpolation formula* and the formula (17) the *Gauss backward interpolation formula.*

REMARK 5. The sum in the right-hand term of (16) or (17) is a polynomial of degree at most $2n$ which attains the same values as f at the tabular points $a_{-n}, \ldots, a_n$. Thus, each of these polynomials is again the Lagrange interpolation polynomial expressed only in a different form, naturally for an odd number of equidistant tabular points.

REMARK 6. Omitting the term containing the n-th central difference in the sum on the right-hand side of (16), or (17), we obtain the interpolation polynomial for an even number of tabular points $a_{-(n-1)}, \ldots, a_n$, or $a_{-n}, \ldots, a_{n-1}$.

REMARK 7. There exist many other special interpolation formulae written by means of differences. Many of them have special names. Thus, for example, if we add the Gauss interpolation formula (16) to the Gauss interpolation formula (17) and divide by two, we obtain the *Stirling interpolation formula*; the *Bessel interpolation formula* is the arithmetic mean of the Gauss forward interpolation formula for tabular points $a_{-(n-1)}, \ldots, a_n$ and the Gauss backward interpolation formula written for the same points.

REMARK 8 (*Fraser's Diagram*). If we order forward differences and the binomical coefficients in a scheme shown in Tab.32.1, we obtain the Fraser diagram. With the help of it we can generate most of the interesting interpolation formulae written by means of differences. To generate such an interpolation formula, we procceed as follows:

(i) Start at an entry in the first (functional value) column and proceed along *any* path in the Fraser diagram (i.e., if a segment terminates on a difference, the path may be continued along any of the other three paths leading from the difference). End the path at any difference.

(ii) Then construct the interpolation formula by

(1) writing down the functional value at which the path started and then

(2a) for every left to right segment in the path *add* a term consisting of the difference on which the segment *terminates* multiplied by the binomial coefficient directly *below* this difference, if the slope of the segment is positive (i.e., if the segment goes upward and to the right), and directly *above*, if the slope of the segment is negative, and

(2b) for every right to left segment *subtract* a term consisting of the difference at which the segment *originates* multiplied by the binomial coefficient directly *below*

TABLE 32.1

$f(a_{-3})$	$\binom{s+3}{1}$	$\Delta^2 f(a_{-4})$	$\binom{s+4}{3}$	$\Delta^4 f(a_{-5})$	$\binom{s+5}{5}$	$\Delta^6 f(a_{-6})$
1	$\Delta f(a_{-3})$	$\binom{s+3}{2}$	$\Delta^3 f(a_{-4})$	$\binom{s+4}{4}$	$\Delta^5 f(a_{-5})$	$\binom{s+5}{6}$	$\Delta^7 f(a_{-6})$
$f(a_{-2})$	$\binom{s+2}{1}$	$\Delta^2 f(a_{-3})$	$\binom{s+3}{3}$	$\Delta^4 f(a_{-4})$	$\binom{s+4}{5}$	$\Delta^6 f(a_{-5})$
1	$\Delta f(a_{-2})$	$\binom{s+2}{2}$	$\Delta^3 f(a_{-3})$	$\binom{s+3}{4}$	$\Delta^5 f(a_{-4})$	$\binom{s+4}{6}$	$\Delta^7 f(a_{-5})$
$f(a_{-1})$	$\binom{s+1}{1}$	$\Delta^2 f(a_{-2})$	$\binom{s+2}{3}$	$\Delta^4 f(a_{-3})$	$\binom{s+3}{5}$	$\Delta^6 f(a_{-4})$
1	$\Delta f(a_{-1})$	$\binom{s+1}{2}$	$\Delta^3 f(a_{-2})$	$\binom{s+2}{4}$	$\Delta^5 f(a_{-3})$	$\binom{s+3}{6}$	$\Delta^7 f(a_{-4})$
$f(a_0)$	$\binom{s}{1}$	$\Delta^2 f(a_{-1})$	$\binom{s+1}{3}$	$\Delta^4 f(a_{-2})$	$\binom{s+2}{5}$	$\Delta^6 f(a_{-3})$
1	$\Delta f(a_0)$	$\binom{s}{2}$	$\Delta^3 f(a_{-1})$	$\binom{s+1}{4}$	$\Delta^5 f(a_{-2})$	$\binom{s+2}{6}$	$\Delta^7 f(a_{-3})$
$f(a_1)$	$\binom{s-1}{1}$	$\Delta^2 f(a_0)$	$\binom{s}{3}$	$\Delta^4 f(a_{-1})$	$\binom{s+1}{5}$	$\Delta^6 f(a_{-2})$
1	$\Delta f(a_1)$	$\binom{s-1}{2}$	$\Delta^3 f(a_0)$	$\binom{s}{4}$	$\Delta^5 f(a_{-1})$	$\binom{s+1}{6}$	$\Delta^7 f(a_{-2})$
$f(a_2)$	$\binom{s-2}{1}$	$\Delta^2 f(a_1)$	$\binom{s-1}{3}$	$\Delta^4 f(a_0)$	$\binom{s}{5}$	$\Delta^6 f(a_{-1})$
1	$\Delta f(a_2)$	$\binom{s-2}{2}$	$\Delta^3 f(a_1)$	$\binom{s-1}{4}$	$\Delta^5 f(a_0)$	$\binom{s}{6}$	$\Delta^7 f(a_{-1})$

this difference, if the slope of the segment is positive (i.e., if the segment goes downward and to the left), and directly *above*, if the slope is negative.

Any interpolation polynomial constructed in the way indicated which terminates with the n-th difference, no matter by which path it reaches that difference, is equal to the Lagrange interpolation polynomial using the same tabular points which are included in the difference with which we have terminated.

Thus, for example, starting at $f(a_0)$, proceeding along lines sloping downward to the right and terminating with the n-th difference, we get the Newton forward interpolation polynomial. If we proceed in a zigzag way, downward and to the right, then upward and to the right, then downward and to the right, etc., if we stop at the n-th difference and replace the forward differences by central differences according to Theorem 5, we obtain the Gauss forward interpolation polynomial, etc.

32.8. Trigonometric Interpolation

The basic problem of *trigonometric interpolation* is similar to that of polynomial interpolation. Again, we look for a polynomial Q_n, now a trigonometric one, i.e., a function of the form

$$Q_n(x) = \frac{1}{2}A_0 + \sum_{j=0}^{n}(A_j \cos jx + B_j \sin jx), \tag{1}$$

which attains the same values (eventually, has the same derivatives up to some order) as a given function at given tabular points.

Since the trigonometric polynomial (1) is 2π-periodic, we will assume that the interpolated function is also 2π-periodic and restrict ourselves to the interpolation of Lagrange type with equidistant tabular points.

Theorem 1. *Let the number of equidistant tabular points be odd, i.e., let they be given by the formula*

$$a_k = \frac{2\pi k}{2n+1}, \quad k = 0, \ldots, 2n. \tag{2}$$

Then the coefficients of the trigonometric polynomial Q_n which satisfies the conditions $Q_n(a_k) = f(a_k)$, $k = 0, \ldots, 2n$, are given by the formulae

$$\begin{aligned} A_j &= \frac{2}{2n+1}\sum_{k=0}^{2n} f(a_k)\cos ja_k, \quad j = 0, \ldots, n, \\ B_j &= \frac{2}{2n+1}\sum_{k=0}^{2n} f(a_k)\sin ja_k, \quad j = 1, \ldots, n. \end{aligned} \tag{3}$$

Theorem 2. *The polynomial Q_n, having the properties described in Theorem* 1, *can be written in the form*

$$Q_n(x) = \frac{1}{2n+1} \sum_{k=0}^{2n} \frac{\sin[(n+1/2)(a_k - x)]}{\sin[(a_k - x)/2]} f(a_k) . \tag{4}$$

REMARK 1. The formula (4) is an analogue of the Lagrange interpolation formula; the role of the elementary polynomials l_k is played here by the functions

$$t_k(x) = \frac{1}{2n+1} \frac{\sin[(n+1/2)(a_k - x)]}{\sin[(a_k - x)/2]} . \tag{5}$$

Theorem 3. *Let the number of equidistant tabular points be even, i.e., let they be given by the formula*

$$a_k = \frac{2\pi k}{2n} , \quad k = 0, \ldots, 2n-1 . \tag{6}$$

Then the trigonometric polynomial R_{n-1} satisfying $R_{n-1}(a_k) = f(a_k)$, $k = 0, \ldots, 2n-1$, is given by

$$R_{n-1}(x) = \frac{1}{2} A_0 + \sum_{j=1}^{n-1} (A_j \cos jx + B_j \sin jx) + \frac{1}{2} A_n \cos nx \tag{7}$$

with the coefficients A_j and B_j given by

$$\begin{aligned} A_j &= \frac{1}{n} \sum_{k=0}^{2n-1} f(a_k) \cos j a_k , \quad j = 0, \ldots, n , \\ B_j &= \frac{1}{n} \sum_{k=0}^{2n-1} f(a_k) \sin j a_k , \quad j = 1, \ldots, n-1 . \end{aligned} \tag{8}$$

REMARK 2. The computation involved in determining A_j and B_j, can be advantageously performed using the fast Fourier transform (cf. Remark 16.3.13).

32.9. Interpolation by Splines

The basic idea underlying the interpolation of this type is similar to that of the Lagrange or Hermite interpolation. The only difference consists in the circumstance that, instead of a polynomial, we take, for interpolating the given function, a *spline*, i.e., a piecewise polynomial function.

(a) Interpolation of the Lagrange Type

Definition 1. The *classical spline of degree* k for the nodes $a_0 < a_1 < \cdots < a_n$ is a function which is a polynomial of degree at most k in any interval $[a_i, a_{i+1}]$, $i = 0, \ldots, n-1$ (generally different in different intervals) and which has, in the whole interval $[a_0, a_n]$, continuous derivatives up to the order $k-1$.

Definition 2. By the *interpolation of a given function* f *by a classical spline of degree* k for nodes $a_0 < a_1 < \cdots < a_n$ we understand the spline f_s of degree k which satisfies the conditions $f_s(a_i) = f(a_i)$ for $i = 0, \ldots, n$.

Among classical splines, the most popular one is the spline of degree 3 called the *cubic spline*. The reason is that the interpolation by the cubic spline has, among all functions which interpolate the given function and which are sufficiently smooth, the least flexion.

Theorem 1. *Let* f *be a function continuous in* $[a, b]$, *let* $a = a_0 < a_1 < \ldots$ $\ldots < a_n = b$ *be any partition of this interval and let* σ_0 *and* σ_n *be arbitrary real numbers. Then there exists one and only one* f_s *which satisfies* $f_s''(a) = \sigma_0$, $f_s''(b) = \sigma_n$ *and which is the interpolation of* f *by the classical cubic spline. Moreover, if we put* $f_s(x) = s_i(x)$ *for* $x \in [a_i, a_{i+1}]$, *we have*

$$s_i(x) = w_i(x) f(a_{i+1}) + \tilde{w}_i(x) f(a_i) + \\ + \tfrac{1}{6} h_i^2 [(w_i^3(x) - w_i(x))\sigma_{i+1} + (\tilde{w}_i^3(x) - \tilde{w}_i(x))\sigma_i], \tag{1}$$

where

$$h_i = a_{i+1} - a_i, \quad w_i(x) = \frac{x - a_i}{h_i}, \quad \tilde{w}_i(x) = \frac{a_{i+1} - x}{h_i}, \tag{2}$$

and the numbers σ_i, $i = 1, \ldots, n-1$, *are obtained as the solution of the system of* $n-1$ *linear equations*

$$h_{i-1}\sigma_{i-1} + 2(h_{i-1} + h_i)\sigma_i + h_i\sigma_{i+1} = 6(\Delta_i - \Delta_{i-1}), \quad i = 1, \ldots, n-1, \tag{3}$$

where

$$\Delta_i = \frac{f(a_{i+1}) - f(a_i)}{h_i}. \tag{4}$$

REMARK 1. The numbers σ_i are the values of the second derivatives of f_s at a_i's.

REMARK 2. If we want to construct the interpolation of a function by the classical cubic spline, we must solve the system of linear algebraic equations (3). The matrix of this system is tridiagonal and diagonally dominant so that the Gaussian

elimination method can be used without pivoting (see § 30.1). Moreover, since the matrix is also well-conditioned for any reasonable choice of nodes, the solution of (3) is not accompanied with any computational problems.

REMARK 3. If the nodes of interpolation are equidistant, then the matrix of (3) is symmetric.

REMARK 4. The classical cubic spline which interpolates a given function depends on two parameters σ_0 and σ_n. Their choice, i.e., the choice of $f_s''(a_0)$ and $f_s''(a_n)$, guarantees uniqueness. If we put $\sigma_0 = \sigma_n = 0$, the corresponding spline f_s satisfies $f_s''(a_0) = f_s''(a_n) = 0$. Such a spline is called the *natural spline*.

REMARK 5. Alternatively, the function s_i from Theorem 1 can be written in the form

$$\begin{aligned} s_i(x) = f(a_i)[2(x-a_i)+h_i](a_{i+1}-x)^2\frac{1}{h_i^3} + m_i(x-a_i)(a_{i+1}-x)^2\frac{1}{h_i^2} + \\ + f(a_{i+1})[2(a_{i+1}-x)+h_i](x-a_i)^2\frac{1}{h_i^3} - m_{i+1}(a_{i+1}-x)(x-a_i)^2\frac{1}{h_i^2}, \end{aligned} \tag{5}$$

where m_0 and m_n are any numbers and the numbers $m_1, \ldots, m_{n-1}$ satisfy the system

$$\begin{aligned} h_i m_{i-1} + 2(h_{i-1}+h_i)m_i + h_{i-1}m_{i+1} = \\ = 3\frac{h_{i-1}}{h_i}[f(a_{i+1}) - f(a_i)] + 3\frac{h_i}{h_{i-1}}[f(a_i) - f(a_{i-1})]. \end{aligned} \tag{6}$$

The constants m_k are now the values of the first derivatives of f_s at a_k's. Thus, the choice of $f_s'(a_0)$ and $f_s'(a_n)$ leads also to uniqueness.

The most practicable error bound for the interpolation by the classical cubic spline is obtained supposing that f belongs to a convenient Sobolev space $H^k = W_2^k(a, b)$ (see § 22.4).

Theorem 2. *Let $f \in H^r$, where $r = 2$, or 3, or 4 and let f_s be such an interpolation of f by the classical cubic spline for the nodes $a = a_0 < a_1 < \cdots < a_n = b$ that $f_s'(a) = f'(a)$ and $f_s'(b) = f'(b)$. Then there exists an absolute constant M such that*

$$\|f - f_s\|_{H^p} \leqq M h^{r-p}\|f\|_{H^r}, \quad p = 0, \ldots, r-1, \tag{7}$$

where

$$h = \max_{i=0,\ldots,n-1}(a_{i+1} - a_i). \tag{8}$$

REMARK 6. If the functions of one variable are concerned, the elements of the Sobolev space H^k are exactly those functions which have, in the given interval, absolutely continuous derivatives up to the order $k-1$ and the k-th derivative of which is square integrable.

(b) Interpolation of the Hermite Type

Definition 3. The *Hermite spline of degree* $2k-1$ ($k \geqq 1$ integer) for the nodes $a = a_0 < a_1 < \cdots < a_n = b$ is a function which is a polynomial of degree at most $2k-1$ in any interval $[a_i, a_{i+1}]$, $i = 0, \ldots, n-1$, and which has continuous derivatives up to the order $k-1$ in $[a, b]$.

Theorem 3. *Let f be a function having $k-1$ continuous derivatives in $[a, b]$ and let $a = a_0 < a_1 < \cdots < a_n = b$ be an arbitrary partition of $[a, b]$. Then there exists one and only one Hermite spline f_s of degree $2k-1$ which satisfies $f_s^{(j)}(a_i) = f^{(j)}(a_i)$, $i = 0, \ldots, n$, $j = 0, \ldots, k-1$.*

Definition 4. The function f_s from Theorem 3 is called the *interpolation of a given function f by the Hermite spline of degree* $2k-1$ for the nodes $a = a_0 < a_1 < \ldots$ $\ldots < a_n = b$.

Theorem 4. *Let $f \in H^r$, where r is any number from the set $\{k, k+1, \ldots, 2k\}$ and let f_s be the interpolation of f by the Hermite spline of degree $2k-1$ for the partition $a = a_0 < a_1 < \cdots < a_n = b$. Then there exists an absolute constant M such that*

$$\|f - f_s\|_{H^p} \leqq M h^{r-p} \|f\|_{H^r}, \quad p = 0, \ldots, \min(r-1, k).$$

33. PROBABILITY THEORY

By TOMÁŠ CIPRA

References: [31], [43], [81], [82], [101], [137], [141], [194], [202], [205], [243], [244], [245], [249], [269], [278], [294], [302], [330], [331], [362], [364], [366], [367], [368], [391], [402], [440], [491], [499].

33.1. Random Event and Probability

Random Experiment. Probability theory is concerned with mathematical models of *random experiments*. They are such experiments the outcomes of which – in contrast to deterministic experiments – are not uniquely determined by the prescribed conditions of the experiment and can be, moreover, repeated many times (in principle without limitation) under these conditions. The term "random experiment" is used not only in classical examples of throwing a die or tossing a coin but also in such situations when one draws items from a lot to control quality in mass production, observes the time to failure of technical devices in tests of reliability or considers the sex of new–born children in demography, etc.

Random Event. A *random event* is such a statement concerning the outcome of a random experiment that after performing the experiment one can uniquely conclude whether this statement is valid or not.

Operations similar to those used in logic (§ 1.1) can be applied to random events. The *complementary event* to an event A is such an event denoted by $\bar{A}$ that occurs if and only if the event A does not. The *union of events* A and B is the event denoted by $A \cup B$ that occurs if and only if at least one of the events A and B occurs. The *intersection of events* A and B is the event denoted by $A \cap B$ that occurs if and only if both events A and B occur simultaneously. In a quite natural way one can generalize the operations of union and intersection to an arbitrary family of events. The *difference of events* A and B is the event denoted by $A \setminus B$ that occurs if and only if the event A occurs and the event B does not. An event A *implies* an event B (in this case one writes $A \subset B$) if the event B occurs whenever the event A does. The events A and B are *equivalent* if and only if the implications $A \subset B$ and $B \subset A$ hold simultaneously. The *certain event* denoted by Ω occurs

in each possible realization of the random experiment while the *impossible event* denoted by $\emptyset$ does not occur in any of its realizations. The events A and B are *disjoint* if $A \cap B = \emptyset$. More generally, if a family of events is considered, these events are called disjoint if any two different events of this family are disjoint. The *total system of events* is a family of events that are disjoint and whose union is equal to the certain event. The *elementary event* is the event that cannot be expressed as a union of two events distinct from the event considered.

REMARK 1. Elementary events coincide with individual outcomes of the experiment that can no longer be decomposed so that the random events can be identified with subsets of the set Ω of all possible experiment outcomes (Ω is also called the *space of elementary events*). Then the individual operations with random events correspond to the operations with sets from § 1.23 (the complementary event corresponds to the complement of a set, the implication of events to the inclusion of sets, disjoint events to disjoint sets, the certain event to the whole space Ω and the impossible event to the empty set).

Example 1. When throwing a fair die, the random experiment has six possible outcomes. Let ω_i denote the outcome (the elementary event) that the number i turns up ($i = 1, \ldots, 6$). The space of elementary events is obviously $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$. Let us consider the random events A: "a number greater than 4 turns up ", B: "an even number turns up", C: " an odd number turns up", and D: "the number 6 turns up", i.e.

$$A = \{\omega_5, \omega_6\},\ B = \{\omega_2, \omega_4, \omega_6\},\ C = \{\omega_1, \omega_3, \omega_5\},\ D = \{\omega_6\}.$$

Then it holds, e.g.,

$$\bar{A} = \{\omega_1, \omega_2, \omega_3, \omega_4\},\ A \cup B = \{\omega_2, \omega_4, \omega_5, \omega_6\},\ A \cap B = D,\ A \setminus B = \{\omega_5\},\ D \subset A.$$

The events B and C are disjoint and, in addition, they form the total system of events.

Example 2. If the random experiment consists in throwing two fair dice, then it has 36 possible outcomes and one can write

$$\Omega = \{(\omega_1, \omega_1), (\omega_1, \omega_2), (\omega_1, \omega_3), \ldots, (\omega_6, \omega_6)\}.$$

The random event A: "the sum 5 turns up "can be written as

$$A = \{(\omega_1, \omega_4), (\omega_2, \omega_3), (\omega_3, \omega_2), (\omega_4, \omega_1)\}.$$

Probability. Despite the fact that one cannot forecast the outcomes of the individual performances of a random experiment it is possible to obtain a certain

general information if one repeats the experiment many times. If an event A has occurred precisely k times in a series of n performances of the experiment, then k is called the *frequency of the event A* and k/n is the *relative frequency of the event A* in the performance considered. When n increases, then due to the so-called *stability of relative frequencies* which is the empirical foundation of probability theory, the relative frequencies converge to a value called the *probability of the event A*.

In the modern probability theory, the existence of probabilities of random events is introduced axiomatically. Since the probabilities have their counterparts in the relative frequencies, their properties have to be analogous to those of the relative frequencies. These properties are postulated in the form of the so-called *axioms of probability*:

A1. To every event A, a real value $P(A)$ is associated fulfilling

$$0 \leqq P(A) \leqq 1 . \tag{1}$$

The value $P(A)$ is called the probability of the event A.

A2. The probability of the certain event is

$$P(\Omega) = 1 . \tag{2}$$

A3. If A and B are disjoint events, then

$$P(A \cup B) = P(A) + P(B) . \tag{3}$$

Theorem 1 (*Properties of Probability*). *Let A, B, $A_1, \ldots, A_n$ be arbitrary events. Then the following statements hold:*

(i) $P(\emptyset) = 0 .$ (4)

(ii) $P(A \cup B) = P(A) + P(B) - (A \cap B) .$ (5)

(iii) $P(A \setminus B) = P(A) - P(A \cap B) .$ (6)

(iv) *If $A \subset B$, then $P(A) \leqq P(B)$.*

(v) $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} (-1)^{i-1} S_i^{(n)} ,$ (7)

where

$$S_i^{(n)} = \sum_{1 \leqq k_1 < k_2 < \cdots < k_i \leqq n} P(A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_i}).$$

If the events $A_1, \ldots, A_n$ are disjoint, then

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) . \tag{8}$$

(vi) $P(\bigcap_{i=1}^{n} A_i) \geqq 1 - \sum_{i=1}^{n} [1 - P(A_i)]$ (*Bonferroni Inequality*). (9)

Classical Definition of Probability. If the number of all possible outcomes (the number of all elementary events) of a random experiment is finite and the individual outcomes can be considered equally probable, then the probability of an arbitrary random event A can be found according to the so-called *classical definition of probability* by means of the formula

$$P(A) = \frac{K}{N}, \tag{10}$$

where N is the number of all outcomes of the experiment and K is the number of such outcomes for which the event A occurs. One frequently looks for the values K and N by means of combinatorial methods (see § 1.12). This approach is known as the *combinatorial calculation of probability.*

Example 3. A lot of 50 products contains 6 defective ones. One draws 5 products from the lot randomly. Then the probability of the random event A that there are precisely 2 defective products in the selected sample is

$$P(A) = \frac{\binom{6}{2}\binom{50-6}{5-2}}{\binom{50}{5}} = \frac{15 \,.\, 13244}{2118760} = 0{\cdot}094\,.$$

Example 4. 11 letters A, A, C, E, H, I, M, M, S, T, T are ordered randomly on the magnetic blackboard. Then the probability of the random event A that the word MATHEMATICS arises is

$$P(A) = \frac{2!\,2!\,2!}{11!} = \frac{8}{39916800} = 0{\cdot}0000002\,.$$

REMARK 2. Modern probability theory usually introduces the *probability space* $(\Omega, \mathcal{A}, P)$. In this triplet, Ω is a space of elementary events, $\mathcal{A}$ is a *Borel field* (σ–*algebra*) of sets in Ω (i.e. the non–empty family of sets in Ω containing with each set also its complement in Ω and with each countable sequence of sets its union) and P is a *probability measure* on $\mathcal{A}$ which fulfils axioms (1), (2) given above and the following generalization of the axiom (3): if $A_1, A_2, \ldots$ is an at most countable sequence of mutually disjoint sets in $\mathcal{A}$, then $P(\bigcup A_i) = \sum P(A_i)$.

33.2. Conditional Probability and Independent Events

Conditional Probability. In conditioning, the set of all possible outcomes of a random experiment is reduced to such outcomes which satisfy the considered condition.

Definition 1. The *conditional probability* of an event A given an event B (or briefly the conditional probability of A given B) is the value

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{1}$$

The conditional probability makes sense only for $P(B) > 0$.

If the conditioning event is fixed, then the conditional probability has the same properties as the probability without conditioning. In the following assertions we shall suppose that all conditional probabilities make sense:

Theorem 1 (*Probability of Intersection of Events*). *For arbitrary random events* $A_1, \ldots, A_n$ *we have*

$$P(\bigcap_{i=1}^{n} A_i) = P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1 \cap A_2) \ldots P(A_n \mid \bigcap_{i=1}^{n-1} A_i). \tag{2}$$

In particular,

$$P(A \cap B) = P(A)\, P(B \mid A) = P(B)\, P(A \mid B). \tag{3}$$

Example 1. In a telephone exchange, 7 calls wait for acceptance; precisely one of them is an international call and one is a trunk–call. The telephone operator randomly puts through one of the calls, and after it is over, another call again randomly. Find the probability that the international call was put through as the first one and the trunk–call as the second one.

Let A be the random event that the international call was put through as the first one and B be the random event that the trunk–call was put through as the second one. Then the probability of interest is

$$P(A \cap B) = P(A)\, P(B \mid A) = \tfrac{1}{7} \cdot \tfrac{1}{6} = 0{\cdot}024.$$

Theorem 2 (*Total Probability Rule*). *Let* $A_1, \ldots, A_n$ *be a total system of events* (see § 33.1). *For an arbitrary event* B *we then have*

$$P(B) = \sum_{i=1}^{n} P(A_i)\, P(B \mid A_i). \tag{4}$$

Example 2. A warehouse is supplied by bulbs from three producers. The first producer delivered 1000 bulbs with 2% defective products, the second one 1500 bulbs with 1% defective products and the third one 3000 bulbs with 0·5% defective products. Find the probability that a randomly selected bulb is defective.

Let A_i be the random event that a randomly selected bulb comes from the i–th producer ($i = 1, 2, 3$) and B be the random event that a randomly selected bulb is defective. Then the probability of interest is

$$P(B) = \sum_{i=1}^{3} P(A_i)\, P(B \mid A_i) = \frac{1000}{5500} \cdot \frac{2}{100} + \frac{1500}{5500} \cdot \frac{1}{100} + \frac{3000}{5500} \cdot \frac{0{\cdot}5}{100} = 0{\cdot}009\,.$$

Theorem 3 (*Bayes's Theorem*). *Let* $A_1, \ldots, A_n$ *be a total system of events* (see § 33.1) *and* B *an arbitrary event. Then*

$$P(A_k \mid B) = \frac{P(A_k)\, P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)\, P(B \mid A_i)} \qquad (k = 1, \ldots, n)\,. \tag{5}$$

REMARK 1. Bayes's theorem is sometimes called the "theorem on probabilities of causes". If the events $A_1, \ldots, A_n$ are possible causes of the event B, then formula (5) gives the probability that the event B that has occurred is the consequence of the cause A_k. In this context $P(A_k)$ are called *a priori probabilities* (i.e. the probabilities before experiment) and $P(A_k \mid B)$ are called *a posteriori probabilities* (i.e. the probabilities after the experiment in which the event B occurred).

Example 3. A special technology has been used for production of 30 % devices while the other devices have been produced using standard technology. The probability that a device will work without failure during a period t is 0·97 for the devices produced with the special technology and 0·82 for the devices produced with the standard technology. Find the probability that a randomly chosen device that worked without failure during the period t was produced by means of the special technology.

Let A_1 and A_2 be the events that the device was produced by means of the special and standard technology, respectively. If B is the event that the device worked without failure during the period t, then the probability of interest is

$$P(A_1 \mid B) = \frac{P(A_1)\, P(B \mid A_1)}{P(A_1)\, P(B \mid A_1) + P(A_2)\, P(B \mid A_2)} = \\ = \frac{0{\cdot}30\,.\,0{\cdot}97}{0{\cdot}30\,.\,0{\cdot}97 + 0{\cdot}70\,.\,0{\cdot}82} = 0{\cdot}336\,.$$

Independent Events.

Definition 2. Events A and B are *independent* if

$$P(A \cap B) = P(A)\, P(B)\,. \tag{6}$$

Events $A_1, \ldots, A_n$ are *independent* if for arbitrary indices $1 \leqq i_1 < i_2 < \ldots < i_r \leqq n$,

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_r}) = P(A_{i_1})\, P(A_{i_2}) \ldots P(A_{i_r}) \tag{7}$$

holds.

REMARK 2. If the events A and B are independent and $P(B) > 0$, then

$$P(A \mid B) = P(A)\,,$$

i.e., in this case the probability of the event A does not depend on the occurrence of the event B.

Example 4. A type of the independence of events weaker than that introduced in (7) is represented by the so-called *pairwise independent* events for which the validity of (7) is sufficient only for $r = 2$. Let three faces of a regular tetrahedron be successively labelled by numbers 1, 2, 3 and let the fourth face contain the whole group of numbers 1, 2, 3. Let A_i be the event that after throwing the tetrahedron the face lying on the ground contains the number i ($i = 1, 2, 3$). Then

$$P(A_1) = P(A_2) = P(A_3) = \tfrac{1}{2}\,,$$
$$P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = \tfrac{1}{4}\,,$$
$$P(A_1 \cap A_2 \cap A_3) = \tfrac{1}{4}\,.$$

The events A_1, A_2, A_3 are thus pairwise independent but they are not independent in the sense of (7).

33.3. Random Variables and Probability Distributions

Random Variable. Outcomes of random experiments can be often expressed numerically. For example, the outcomes of the random experiment from Example 33.1.1. can be described by the numerical values 1, 2, ... , 6. The outcomes of a random experiment expressed as real numbers are values of a variable which is called the *random variable.*

Probability Distribution. The *probability distribution* of a random variable X is a rule that enables us to determine probabilities of all random events which can be described by this random variable. Examples of such events are $X = 4$, $X > 2$, $1 < X \leqq 3$, etc.

REMARK 1. The random variable X is rigorously defined as a mapping that assigns a real number to each elementary event ω (i.e., it is a real function $X(\omega)$ of elementary events) and which, in addition, has to have the property that the set $\{\omega : X(\omega) \leqq x\}$ is a random event for each real x. Then the probability distribution of the random variable X is a mapping that assigns a probability to each set A of importance by means of X; this probability is formally denoted by the symbol $P(X \in A)$. In particular, one can write e.g. $P(X \leqq x)$.

Definition 1. The *distribution function* of a random variable X is a real function defined by

$$F(x) = P(X \leqq x) \qquad (-\infty < x < \infty). \tag{1}$$

REMARK 2. There exists a one–to–one correspondence between the distribution function and the probability distribution of a given random variable. Therefore one also speaks of the distribution function of a probability distribution.

Theorem 1 (*Properties of Distribution Function*). *The distribution function of an arbitrary random variable is non–decreasing* (*in particular, the set of its points of discontinuity is at most countable*) *and continuous from the right. Furthemore,*

$$\lim_{x \to -\infty} F(x) = 0\,, \quad \lim_{x \to \infty} F(x) = 1\,. \tag{2}$$

REMARK 3. Conversely, each function with the properties from Theorem 1 is the distribution function of a random variable. Further important formulae are

$$P(a < X \leqq b) = F(b) - F(a) \qquad (-\infty < a < b < \infty) \tag{3}$$

and

$$P(X = a) = F(a) - F(a - 0)\,, \tag{4}$$

where the symbol $F(a - 0)$ denotes the left–hand limit of the function F at the point a.

REMARK 4. Sometimes the distribution function is defined by the relation $F(x) = P(X < x)$. Then one has to replace continuity from the right by continuity from the left in Theorem 1.

Two types of random variables and their corresponding probability distributions play the most important role from practical point of view.

The *discrete random variable* X attains, with positive probabilities p_j, only values x_j from an at most countable system:

$$p_j = P(X = x_j) > 0\,, \quad \sum_{x_j} p_j = 1 \tag{5}$$

(the system extends over all the values x_j). The corresponding *discrete probability distribution* is completely defined by the system of values x_j with their assigned probabilities p_j (the so-called *probability function*). The distribution function of the discrete random variable is

$$F(x) = \sum_{x_j \leqq x} p_j \,. \tag{6}$$

It is thus a step function with jumps at the points x_j.

Example 1. An example of discrete probability distribution is the *alternative distribution*. The random variable X with this distribution fulfils the relations

$$P(X = 0) = 1 - p\,, \quad P(X = 1) = p\,, \tag{7}$$

where $0 < p < 1$. If "0"denotes, that the tail appeared and "1"denotes, that the head did, then the random variable X with $p = \frac{1}{2}$ describes the random experiment consisting in tossing a fair coin. The probability function of this random variable is shown in Fig. 33.1 and the distribution function in Fig. 33.2.

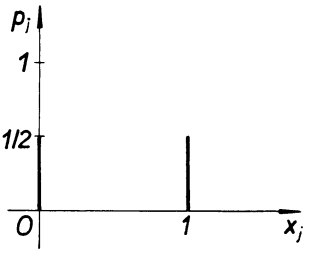


Fig. 33.1. Probability function of the alternative distribution ($p = \frac{1}{2}$).

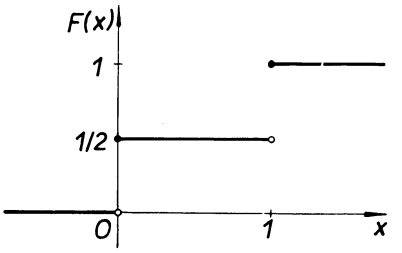


Fig. 33.2. Distribution function of the alternative distribution ($p = \frac{1}{2}$).

The *continuous random variable* X is not confined to discrete values. Its distribution function $F(x)$ can be expressed in the form

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t\,, \tag{8}$$

where $f(x)$ is a non-negative function. The function $f(x)$ is called the *probability density* (or briefly *density*) of the random variable X and it completely defines the corresponding *continuous probability distribution*. The distribution function (8) is continuous (it possesses even the property of the so-called absolute continuity). In particular,

$$P(a < X \leqq b) = \int_{a}^{b} f(x)\,\mathrm{d}x \tag{9}$$

holds for arbitrary real a and b, $a < b$, and

$$P(X = a) = 0 \tag{10}$$

for an arbitrary real a.

Example 2. An example of continuous probability distribution is the *exponential distribution* with the density

$$f(x) = \begin{cases} \frac{1}{\delta} e^{-x/\delta} & \text{for } x > 0, \\ 0 & \text{for } x \leqq 0 \end{cases} \tag{11}$$

and with the distribution function

$$F(x) = \begin{cases} \int_0^x \frac{1}{\delta} e^{-t/\delta} \, dt = 1 - e^{-x/\delta} & \text{for } x > 0, \\ 0 & \text{for } x \leqq 0 \end{cases} \tag{12}$$

(δ is a positive constant). The random variable with this distribution describes e.g. the time to failure of an electronic component which, under the assumption that it survived through the time period x, breaks down in an interval $(x, x + h)$ of a small length h with the probability h/δ. The density (11) and the distribution function (12) for $\delta = 1$ are shown in Fig. 33.3 and Fig. 33.4, respectively.

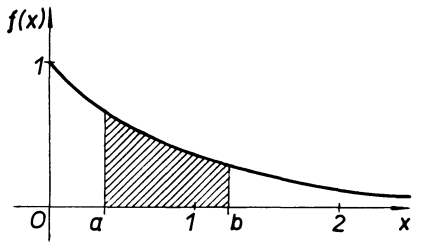


Fig. 33.3. Probability density of the exponential distribution ($\delta = 1$).

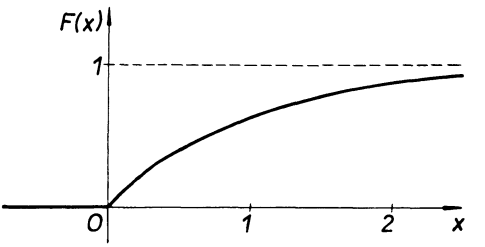


Fig. 33.4. Distribution function of the exponential distribution ($\delta = 1$).

The shaded area under the curve of the density $f(x)$ between $x = a$ and $x = b$ in Fig. 33.3 corresponds to the probability (9).

33.4. Basic Characteristics of Random Variables

Characteristics of random variable. *Characteristics of a random variable* enable us, in contrast to the distribution function, probability function or density, to summarize the whole information on the random variable or on the probability distribution into several numerical values.

Definition 1. The *mean* (or *mean value* or *expectation*) of a random variable X is defined in the discrete case by the formula

$$\mathrm{E}(X) = \sum_{x_j} x_j p_j \tag{1}$$

and in the continuous case by the formula

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x\,, \tag{2}$$

where one assumes the absolute convergence of the sum (1) or the integral (2).

REMARK 1. If the sum (1) or the integral (2) do not converge absolutely, then we say that the random variable X does not have the mean.

REMARK 2. The mean can be expressed for all types of random variables in a uniform way by means of the Lebesgue–Stieltjes integral $\int_{-\infty}^{\infty} x\,\mathrm{d}F(x)$, where $F(x)$ is the distribution function.

Theorem 1. *Let $g(X)$ be a random variable which arises from a function $g(x)$ of a random variable X. Then*

$$\mathrm{E}[g(X)] = \sum_{x_j} g(x_j) p_j \tag{3}$$

holds in the case of the discrete random variable X and

$$\mathrm{E}[g(x)] = \int_{-\infty}^{\infty} g(x) f(x)\,\mathrm{d}x \tag{4}$$

in the case of the continuous random variable X, provided that the sum (3) *or the integral* (4) *converge absolutely.*

Definition 2. Let k be a non–negative integer. Then the *k–th moment* of a random variable X is defined by

$$\mu_k' = \mathrm{E}(X^k) \tag{5}$$

and the *k–th central moment* of a random variable X by

$$\mu_k = \mathrm{E}\{[X - \mathrm{E}(X)]^k\} \tag{6}$$

provided that the corresponding means exist.

REMARK 3. In particular,

$$\mu_0' = \mu_0\,,\ \mu_1' = \mathrm{E}(X)\,,\ \mu_1 = 0\,.$$

Definition 3. The second central moment μ_2 is called the *variance* (it is usually denoted by the symbol $\mathrm{var}(X)$ or σ^2).

REMARK 4. Moments can be calculated according to Theorem 1. For example, in the case of a discrete random variable X one obtains

$$\mathrm{var}(X) = \sum_{x_j}[x_j - \mathrm{E}(X)]^2 p_j = \sum_{x_j} x_j^2 p_j - [\mathrm{E}(X)]^2\,, \tag{7}$$

and in the case of a continuous random variable X,

$$\mathrm{var}(X) = \int_{-\infty}^{\infty}[x - \mathrm{E}(X)]^2 f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x - [\mathrm{E}(X)]^2\,. \tag{8}$$

Example 1. The mean and the variance of the random variable with the alternative distribution from Example 33.3.1 are

$$\mathrm{E}(X) = 0\,.\,(1-p) + 1\,.\,p = p\,, \tag{9}$$

$$\mathrm{var}(X) = 0^2\,.\,(1-p) + 1^2\,.\,p - p^2 = p(1-p)\,. \tag{10}$$

Example 2. The mean and variance of the random variable with the exponential distribution from Example 33.3.2 are

$$\mathrm{E}(X) = \int_0^{\infty} x\frac{1}{\delta}\mathrm{e}^{-x/\delta}\,\mathrm{d}x = \delta\,, \tag{11}$$

$$\mathrm{var}(X) = \int_0^{\infty} x^2\frac{1}{\delta}\mathrm{e}^{-x/\delta}\,\mathrm{d}x - \delta^2 = \delta^2\,. \tag{12}$$

Theorem 2. *The following relations hold between the moments and central moments*:

$$\mu_k = \sum_{j=0}^{k}\binom{k}{j}(-\mu_1')^j\mu_{k-j}', \quad \mu_k' = \sum_{j=0}^{k}\binom{k}{j}(\mu_1')^j\mu_{k-j} \qquad (k = 0,\,1,\,2,\dots)\,. \tag{13}$$

In particular, it is (compare with (7) and (8))

$$\mathrm{var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2\,. \tag{14}$$

Definition 4. Let P be a real number, $0 < P < 1$. Then the *P–quantile* of a random variable with the distribution function $F(x)$ is defined as the value x_P for that

$$F(x_P - 0) \leqq P\,,\ F(x_P) \geqq P\,. \tag{15}$$

In particular, the quantile $x_{0\cdot5}$ is called the *median*, $x_{0\cdot25}$ the *lower quartile*, $x_{0\cdot75}$ the *upper quartile*, $x_{k/10}$ the *k–th decile* ($k = 1, \ldots, 9$) and $x_{k/100}$ the *k–th percentile* ($k = 1, \ldots, 99$).

REMARK 5. In general, the P–quantile is not determined uniquely. In the case of a continuous random variable the relationship (15) is reduced to

$$F(x_P) = P\,. \tag{16}$$

Example 3. The P–quantile of the random variable with the exponential distribution from Example 33.3.2 fulfils the relation

$$1 - \mathrm{e}^{-x_P/\delta} = P$$

so that one obtains

$$x_P = -\delta \ln(1 - P)\,.$$

In particular, the median is

$$x_{0\cdot5} = \delta \ln 2 = 0{\cdot}69315\,\delta\,.$$

Characteristics of Location. *Characteristics of location* describe, in certain ways, the location of a random variable. These characteristics can be regarded as a "centre of gravity" about which the values of the random variable considered are spread. The usual characteristics of location are the mean (see Definition 1), median (see Definition 4) and mode.

Definition 5. The *mode* of a random variable X is a value $\hat{x}$ that fulfils, in the discrete case, the relation

$$P(X = \hat{x}) \geqq P(X = x_j) \tag{17}$$

for all values x_j of the random variable and, in the continuous case, the relation

$$f(\hat{x}) \geqq f(x) \tag{18}$$

for all real x.

REMARK 6. As in the case of the median, the mode need not be determined uniquely. The probability distribution with precisely one mode is called *unimodal*.

REMARK 7. The sizes of the mean, median and mode can be in various mutual relations as shown in Fig. 33.5.

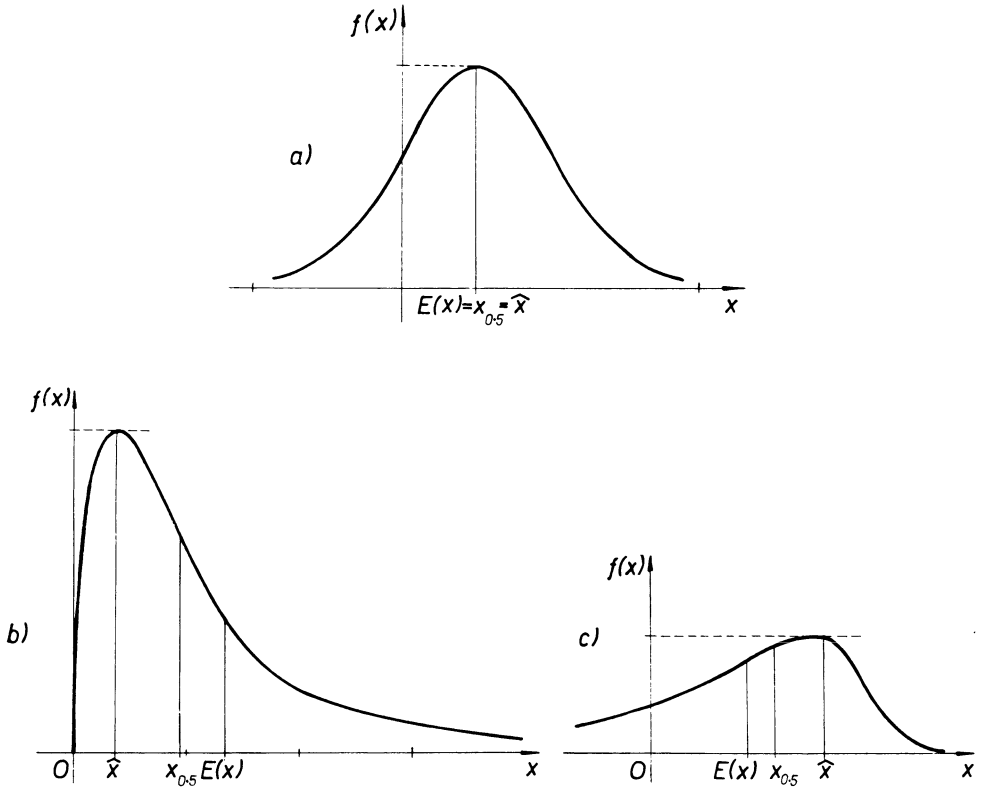


Fig. 33.5 a, b, c. Mean $\mathrm{E}(x)$, median $x_{0\cdot5}$ and mode $\hat{x}$ of various probability distributions.

Characteristics of Variability. *Characteristics of variability* measure, in certain ways, the degree of dispersion (or spread) of the considered random variable about some of the location characteristics. The usual characteristics of variability are the variance (see Definition 3), standard deviation, mean deviation, coefficient of variation and interquartile range.

Definition 6. The *standard deviation* is

$$\sigma = [\mathrm{var}(X)]^{1/2}.$$

Definition 7. The *mean deviation* is $\mathrm{E}[|X - \mathrm{E}(X)|]$.

Definition 8. The *coefficient of variation* is $[\mathrm{var}(X)]^{1/2}/|\mathrm{E}(X)|$ for $\mathrm{E}(X) \neq 0$.

Definition 9. The *interquartile range* is $x_{0\cdot75} - x_{0\cdot25}$, the *interdecile range* $x_{0\cdot9} - x_{0\cdot1}$, and the *interpercentile range* $x_{0\cdot99} - x_{0\cdot01}$.

Characteristics of skewness and kurtosis. *Characteristics of skewness and kurtosis* concern the shape of the curve of the probability density of the considered random variable.

Definition 10. The *coefficient of skewness* is

$$\gamma_1 = \frac{\mu_3}{\sigma^3}, \tag{19}$$

where μ_3 is the third central moment (see Definition 2) and σ is the standard deviation (see Definition 6).

REMARK 8. The coefficient of skewness is zero for the probability distribution with symmetric curves of density (see Fig. 33.5a), it is positive for the distributions with such curves of density that have larger right–hand tails (see Fig 33.5b) and negative for the distributions with such curves of density that have larger left–hand tails (see Fig. 33.5c). Obviously, γ_1 is a measure of asymmetry of distributions.

REMARK 9. Some random variables have the property that

$$P(X = a - x) = P(X = a + x)$$

holds for an arbitrary real x in the discrete case and

$$f(a - x) = f(a + x)$$

in the continous case. Then we say that the distribution of these random variables is *symmetric about the point a*. In this case one has $\mathrm{E}(X) = a$, $\mu_k = 0$ for k odd, and $\gamma_1 = 0$ provided that these values exist.

Definition 11. The *coefficient of kurtosis* (or *coefficient of excess*) is

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3, \tag{20}$$

where μ_4 is the fourth central moment (see Definition 2) and σ is the standard deviation (see Definition 6).

REMARK 10. The distributions with tails of curves of density thicker or thinner than those of the normal distribution with the same mean and variance (see § 33.7) have $\gamma_2 > 0$ or $\gamma_2 < 0$, respectively. The normal distribution has always $\gamma_2 = 0$. Obviously, γ_2 is a measure of peakedness of distributions.

Example 4. The coefficient of skewness and kurtosis of the random variable with the exponential distribution from Example 33.3.2 is

$$\gamma_1 = \frac{2\delta^3}{\delta^3} = 2\,,\quad \gamma_2 = \frac{9\delta^4}{\delta^4} - 3 = 6\,.$$

Characteristic Function. The characteristic function is an important tool for the calculation of moments and for various theoretical considerations.

Definition 12. The *characteristic function* of a random variable X is a complex function defined by the formula

$$\varphi(t) = \mathrm{E}(\mathrm{e}^{\mathrm{i}tX}) \qquad (-\infty < t < \infty)\,. \tag{21}$$

REMARK 11. The characteristic function exists for every random variable. One can rewrite (21) as

$$\varphi(t) = \sum_{x_j} \mathrm{e}^{\mathrm{i}tx_j} p_j \tag{22}$$

in the discrete case and as

$$\varphi(t) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}tx} f(x)\,\mathrm{d}x \tag{23}$$

in the continuous case. Conversely, using the theory of Fourier transform, one obtains

$$p_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}tx_j} \varphi(t)\,\mathrm{d}t \tag{24}$$

in the discrete case and

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}tx} \varphi(t)\,\mathrm{d}t \tag{25}$$

in the continuous case. It means that there exists a one–to–one correspondence between the probability distribution and the characteristic function.

REMARK 12. The characteristic function enables us to calculate the k–th moment by the formula

$$\mu'_k = \frac{1}{\mathrm{i}^k} \varphi^{(k)}(0)\,. \tag{26}$$

Example 5. The random variable with the alternative distribution from Example 33.3.1 has the characteristic function

$$\varphi(t) = \mathrm{e}^{\mathrm{i}t\,.\,0}(1-p) + \mathrm{e}^{\mathrm{i}t\,.\,1}p = 1 + p(\mathrm{e}^{\mathrm{i}t} - 1)\,. \tag{27}$$

Example 6. The random variable with the exponential distribution from Example 33.3.2 has the characteristic function

$$\varphi(t) = \int_0^\infty \frac{1}{\delta} e^{x(it-1/\delta)}\, dx = \frac{1}{1 - i\delta t}. \tag{28}$$

Using (26) one obtains

$$E(X) = \frac{1}{i}\varphi'(0) = \frac{1}{i} i\delta = \delta,$$

in accordance with the result (11).

REMARK 13. Sometimes it is advantageous to replace the moments by the so-called cumulants. If we set $\psi(t) = \ln \varphi(t)$, then the *k–th cumulant* κ_k is defined by the formula

$$\kappa_k = \frac{1}{i^k} \psi^{(k)}(0) \qquad (k = 0, 1, 2, \dots). \tag{29}$$

In particular,

$$\kappa_1 = \mu_1', \quad \kappa_2 = \mu_2, \quad \kappa_3 = \mu_3, \quad \kappa_4 = \mu_4 - 3\mu_2^2. \tag{30}$$

33.5. Random Vectors

REMARK 1. In what follows, $\boldsymbol{x} = (x_1, \dots, x_n)'$ means the n-component *column vector* with components $x_1, \dots, x_n$, thus an $n \times 1$ matrix with the element x_i in its i–th row ($i = 1, \dots, n$). The symbol $\boldsymbol{x}' = (x_1, \dots, x_n)$ means the n-component *row vector* with components $x_1, \dots, x_n$.

Random Vector. Frequently, an outcome of a random experiment has to be described by several numbers. The column vector $\boldsymbol{X}$, whose components are random variables $X_1, \dots, X_n$ defined on the same space of elementary events, is called the n–component *random vector*. One writes formally $\boldsymbol{X} = (X_1, \dots, X_n)'$. The *probability distribution of the random vector* $\boldsymbol{X}$ (or the *joint probability distribution* of the random variables $X_1, \dots, X_n$) is a rule that enables us to determine the probabilities of all random events which can be described by the random variables $X_1, \dots, X_n$. An example is the event $X_1 \leqq 1, \dots, X_n \leqq 1$. In general, one also speaks of the *multivariate probability distribution.*

Definition 1. The *distribution function of a random vector* $\boldsymbol{X}$ (or the *joint distribution function* of random variables $X_1, \dots, X_n$) is a real function $F(x_1, \dots, x_n)$ defined by

$$F(x_1, \dots, x_n) = P(X_1 \leqq x_1, \dots, X_n \leqq x_n) \qquad (-\infty < x_i < \infty, i = 1, \dots, n). \tag{1}$$

Theorem 1 (*Properties of Joint Distribution Function*). *The distribution function* $F(x_1, \ldots, x_n)$ *of an arbitrary random vector is non-decreasing and continuous from the right in each of its arguments. Furthermore,*

$$\lim_{x_j \to -\infty} F(x_1, \ldots, x_n) = 0 \quad (j = 1, \ldots, n), \qquad \lim_{x_1 \to \infty, \ldots, x_n \to \infty} F(x_1, \ldots, x_n) = 1. \tag{2}$$

REMARK 2. The relation

$$P(a_1 < X_1 \leqq b_1, \ldots, a_n < X_n \leqq b_n) = \sum_{\delta_1, \ldots, \delta_n} (-1)^{\sum_{i=1}^{n} \delta_i} F(c_1, \ldots, c_n) \tag{3}$$

holds for arbitrary real a_i, b_i, $a_i < b_i$ $(i = 1, \ldots, n)$, where $c_i = \delta_i a_i + (1 - \delta_i) b_i$ and the quantities $\delta_1, \ldots, \delta_n$ acquire the values 0 and 1 mutually independently. In particular,

$$P(a_1 < X_1 \leqq b_1, a_2 < X_2 \leqq b_2) = F(b_1, b_2) - F(a_1, b_2) - F(a_2, b_1) + F(a_1, a_2). \tag{4}$$

The *discrete random vector* $\boldsymbol{X} = (X_1, \ldots, X_n)'$ attains, with positive probabilities $P(X_1 = x_1, \ldots, X_n = x_n)$, values $(x_1, \ldots, x_n)'$ from an at most countable system of n-component vectors only. The system of all these probabilities determines the *probability function of the random vector* $\boldsymbol{X}$.

The *continuous random vector* $\boldsymbol{X} = (X_1, \ldots, X_n)'$ is not confined to discrete values. Its distribution function $F(x_1, \ldots, x_n)$ can be expressed in the form

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \ldots, t_n)\, \mathrm{d}t_1 \ldots \mathrm{d}t_n, \tag{5}$$

where $f(x_1, \ldots, x_n)$ is a non-negative function. The function $f(x_1, \ldots, x_n)$ is called the *probability density of the random vector* $\boldsymbol{X}$ (or the *joint probability density* of the random variables $X_1, \ldots, X_n$). In particular,

$$P(a_1 < X_1 \leqq b_1, \ldots, a_n < X_n \leqq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \ldots, x_n)\, \mathrm{d}x_1 \ldots \mathrm{d}x_n \tag{6}$$

holds for arbitrary real a_i and b_i, $a_i < b_i$ $(i = 1, \ldots, n)$.

Marginal Distribution. If $\boldsymbol{X} = (X_1, \ldots, X_n)'$ is a random vector, then the probability distribution of the random vector $(X_{i_1}, \ldots, X_{i_k})'$ for an arbitrary subset $I = \{i_1, \ldots, i_k\}$ of the index set $\{1, \ldots, n\}$ $(1 \leqq k < n)$ is called the *marginal probability distribution*. The *marginal distribution function*

$F_{i_1,\ldots,i_k}(x_{i_1}, \ldots, x_{i_k})$ of the random vector $(X_{i_1}, \ldots, X_{i_k})'$ is obtained by letting $x_i \to \infty$ in $F(x_1, \ldots, x_n)$ for all $i \notin I$. The *marginal probability function* $P(X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k})$ is obtained by summing up over x_i in $P(X_1 = x_1, \ldots, X_n = x_n)$ for all $i \notin I$. The *marginal probability density* $f_{i_1,\ldots,i_k}(x_{i_1}, \ldots, x_{i_k})$ is obtained by integrating over x_i in $f(x_1, \ldots, x_n)$ for all $i \notin I$.

REMARK 3. If $\boldsymbol{X} = (X_1, X_2)'$ is a two-variate random vector, then we have

$$F_1(x_1) = \lim_{x_2 \to \infty} F(x_1, x_2), \quad F_2(x_2) = \lim_{x_1 \to \infty} F(x_1, x_2). \tag{7}$$

Furthemore,

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2), \; P(X_2 = x_2) = \sum_{x_1} P(X_1 = x_1, X_2 = x_2) \tag{8}$$

holds in the discrete case and

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2, \quad f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1 \tag{9}$$

in the continuous case.

Conditional Distribution. Let $\boldsymbol{X} = (X_1, \ldots, X_n)'$ be a random vector. Let $I = \{i_1, \ldots, i_k\}$ and $J = \{j_1, \ldots, j_m\}$ $(1 \leqq k < n, k + m = n)$ form a decomposition of the index set $\{1, \ldots, n\}$ into two disjoint subsets. Given $X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}$ for fixed values $x_{i_1}, \ldots, x_{i_k}$, the *conditional probability distribution* of the random vector $(X_{j_1}, \ldots, X_{j_m})'$ is defined in the discrete case by the *conditional probability function*

$$\begin{aligned} &P(X_{j_1} = x_{j_1}, \ldots, X_{j_m} = x_{j_m} \mid X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}) = \\ &= \begin{cases} P(X_1 = x_1, \ldots, X_n = x_n)/P(X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}) \\ \qquad\qquad \text{for } P(X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}) \neq 0, \\ 0 \qquad\qquad \text{for } P(X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}) = 0 \end{cases} \end{aligned} \tag{10}$$

and in the continuous case by the *conditional probability density*

$$\begin{aligned} &f(x_{j_1}, \ldots, x_{j_m} \mid x_{i_1}, \ldots, x_{i_k}) = \\ &= \begin{cases} f(x_1, \ldots, x_n)/f_{i_1,\ldots,i_k}(x_{i_1}, \ldots, x_{i_k}) & \text{for } f_{i_1,\ldots,i_k}(x_{i_1}, \ldots, x_{i_k}) \neq 0, \\ 0 & \text{for } f_{i_1,\ldots,i_k}(x_{i_1}, \ldots, x_{i_k}) = 0. \end{cases} \end{aligned} \tag{11}$$

Definition 2. Let $S(X_1, \ldots, X_n)$ be a random variable defined as a function $S(x_1, \ldots, x_n)$ of a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$. Using the above notation, we define the *conditional mean* of the random variable $S(X_1, \ldots, X_n)$, given

$X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}$, in the discrete case by

$$\mathrm{E}[S(X_1, \ldots, X_n) \mid X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}] =$$
$$= \sum_{x_{j_1}} \ldots \sum_{x_{j_m}} S(x_1, \ldots, x_n) P(X_{j_1} = x_{j_1}, \ldots, X_{j_m} = x_{j_m} \mid X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}) \tag{12}$$

and in the continuous case by

$$\mathrm{E}[S(X_1, \ldots, X_n) \mid X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}] =$$
$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} S(x_1, \ldots, x_n) f(x_{j_1}, \ldots, x_{j_m} \mid x_{i_1}, \ldots, x_{i_k}) \, \mathrm{d}x_{j_1} \ldots \mathrm{d}x_{j_m} . \tag{13}$$

REMARK 4. The conditional mean from Definition 2 is frequently considered as a random variable of the form $\mathrm{E}[S(X_1, \ldots, X_n) \mid X_{i_1}, \ldots, X_{i_k}]$.

The formula (12), or (13) holds also for $k = 0$, when it represents a generalization of Theorem 33.4.1 for (unconditional) mean

$$\mathrm{E}[S(X_1, \ldots, X_n)] = \sum_{x_1} \cdots \sum_{x_n} S(x_1, \ldots, x_n) P(X_1 = x_1, \ldots, X_n = x_n) ,$$

or

$$\mathrm{E}[S(X_1, \ldots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} S(x_1, \ldots, x_n) f(x_1, \ldots, x_n) \, \mathrm{d}x_1 \ldots \mathrm{d}x_n ,$$

respectively.

REMARK 5. If $\boldsymbol{X} = (X_1, X_2)'$ is a two-variate continuous random vector, then we have

$$f(x_1 \mid x_2) = \begin{cases} f(x_1, x_2)/f_2(x_2) & \text{for } f_2(x_2) \neq 0 , \\ 0 & \text{for } f_2(x_2) = 0 \end{cases} \tag{14}$$

and

$$\mathrm{E}(X_1 \mid X_2 = x_2) = \int_{-\infty}^{\infty} x_1 f(x_1 \mid x_2) \, \mathrm{d}x_1 . \tag{15}$$

The formulae for $f(x_2 \mid x_1)$ and $\mathrm{E}(X_2 \mid X_1 = x_1)$ can be obtained by interchanging the indices 1 and 2.

Independent Random Variables. The independence of random variables corresponds to the independence of random events described by the individual random variables.

Definition 3. Random variables $X_1, \ldots, X_n$ are called *independent* if

$$F(x_1, \ldots, x_n) = F(x_1) \ldots F(x_n) \tag{16}$$

holds for any real $x_1, \ldots, x_n$.

REMARK 6. For independent random variables $X_1, \ldots, X_n$, the relation (3) can be rewritten in the form

$$P(a_1 < X_1 \leqq b_1, \ldots, a_n < X_n \leqq b_n) = [F(b_1) - F(a_1)] \ldots [F(b_n) - F(a_n)]. \quad (17)$$

Moreover, if the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$ is discrete, then

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \ldots P(X_n = x_n) \quad (18)$$

and if it is continuous, then

$$f(x_1, \ldots, x_n) = f(x_1) \ldots f(x_n). \quad (19)$$

Theorem 2. *Let $X_1, \ldots, X_n$ be independent random variables such that $\mathrm{E}(|X_i|^{k_i}) < \infty$ $(i = 1, \ldots, n)$, where $k_1, \ldots, k_n$ are given non-negative integers. Then*

$$\mathrm{E}(X_1^{k_1} \ldots X_n^{k_n}) = \mathrm{E}(X_1^{k_1}) \ldots \mathrm{E}(X_n^{k_n}). \quad (20)$$

REMARK 7. The expression on the left-hand side of (20) is called the *mixed moment* of the random variables $X_1, \ldots, X_n$ and is denoted by $\mu'_{k_1,\ldots,k_n}$. In general, it is calculated according to the formulae from Remark 4. Theorem 2 remains valid even for general $k_1, \ldots, k_n$ provided that the powers $X_1^{k_1}, \ldots, X_n^{k_n}$ have sense.

Example 1. Let a random experiment consist in throwing four fair dice and let the random variable X_i denote the number that turns up on the i-th die $(i = 1, \ldots, 4)$. The random variables $X_1, \ldots, X_4$ are independent. In particular, the probability of the random event that the number 6 turns up on all the dice is

$$P(X_1 = 6, \ldots, X_4 = 6) = P(X_1 = 6) \ldots P(X_4 = 6) = \left(\tfrac{1}{6}\right)^4 = 0{\cdot}0008$$

according to (18).

Characteristics of Random Vector.

Definition 4. The *mean of a random vector* $\boldsymbol{X} = (X_1, \ldots, X_n)'$ is an n-component vector of the form

$$\mathrm{E}(\boldsymbol{X}) = (\mathrm{E}(X_1), \ldots, \mathrm{E}(X_n))'. \quad (21)$$

Definition 5. The *covariance* of random variables X and Y is defined by

$$\mathrm{cov}(X, Y) = \mathrm{E}\{[X - \mathrm{E}(X)][Y - \mathrm{E}(Y)]\}. \quad (22)$$

The *correlation coefficient* of random variables X and Y is defined, under the assumptions $\operatorname{var}(X) \neq 0$ and $\operatorname{var}(Y) \neq 0$, by

$$\varrho(X, Y) = \frac{\operatorname{cov}(X, Y)}{[\operatorname{var}(X) \operatorname{var}(Y)]^{1/2}} . \tag{23}$$

REMARK 8. The expression (22) can be rewritten in the form

$$\operatorname{cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) . \tag{24}$$

The mean on the right-hand side of (22) and $\mathrm{E}(XY)$ in (24) are calculated as special cases of the mixed moment from Remark 7 for $n = 2$.

Theorem 3 (*Properties of Correlation Coefficient*). *The correlation coefficient has the following properties:*

(i) $-1 \leqq \varrho(X, Y) \leqq 1$ (*Schwarz Inequality*). (25)

(ii) *If the random variables X and Y are independent, then $\varrho(X, Y) = 0$.*

(iii) $|\varrho(X, Y)| = 1$ *if and only if*

$$Y = aX + b \tag{26}$$

holds with probability 1, *where a and b are real constants ($a \neq 0$). Moreover, $\varrho(X, Y) = 1$ or $\varrho(X, Y) = -1$ if $a > 0$ or $a < 0$, respectively.*

Definition 6. If $\operatorname{cov}(X, Y) = 0$, then we say that the random variables X and Y are *uncorrelated*.

REMARK 9. Due to Theorem 3, the correlation coefficient is used as a measure of linear dependence between two random variables. In particular, $\varrho(X, X) = 1$. Uncorrelated random variables need not be independent. Besides the correlation coefficient, one also uses the *multiple correlation coefficient* that is a (scalar) measure of linear dependence between a random variable and a random vector, and the *partial correlation coefficient* which is a measure of linear dependence between two random variables provided that the values of a given random vector are fixed (in this way one eliminates a possible dependence caused by the influence of this random vector on both random variables of interest). Details can be found e.g. in [101] and [499].

Definition 7. The *covariance matrix* of a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$ is an $n \times n$ matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ with elements $\sigma_{ij} = \operatorname{cov}(X_i, X_j)$ $(i, j = 1, \ldots, n)$. The *correlation matrix* of this random vector is an $n \times n$ matrix with elements $\varrho_{ij} = \varrho(X_i, X_j)$ $(i, j = 1, \ldots, n)$.

REMARK 10. The correlation matrix has the values 1 on its main diagonal.

Definition 8. The *characteristic function of a random vector* $\boldsymbol{X} = (X_1, \ldots, X_n)'$ is a complex function defined by

$$\varphi(t_1, \ldots, t_n) = \mathrm{E}\left[\exp\left(\mathrm{i}\sum_{j=1}^{n} t_j X_j\right)\right] \qquad (-\infty < t_i < \infty,\ i = 1, \ldots, n). \tag{27}$$

Theorem 4. *Random variables* $X_1, \ldots, X_n$ *are independent if and only if*

$$\varphi(t_1, \ldots, t_n) = \varphi_1(t_1) \ldots \varphi_n(t_n) \tag{28}$$

holds for any real $t_1, \ldots, t_n$, *where* $\varphi_i(t_i)$ *is the characteristic function of the random variable* X_i $(i = 1, \ldots, n)$.

REMARK 11. One can calculate mixed moments by means of the characteristic function according to the formula

$$\mathrm{E}(X_1^{k_1} \ldots X_n^{k_n}) = \frac{1}{\mathrm{i}^{k_1+\ldots+k_n}} \frac{\partial^{k_1+\ldots+k_n}}{\partial t_1^{k_1} \ldots \partial t_n^{k_n}} \varphi(0, \ldots, 0). \tag{29}$$

33.6. Important Discrete Distributions

Among discrete random variables, the so-called *integer* (or *integral-valued*) *random variables* (*integer probability distributions*) which can attain only the values $x = 0, 1, 2, \ldots$, are most important.

1. *Binomial distribution* with parameters n, p (n is a positive integer, $0 < p < 1$):

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad (x = 0, 1, \ldots, n),$$

$$\mathrm{E}(X) = np, \quad \mathrm{var}(X) = np(1-p), \quad (n+1)p - 1 \leqq \hat{x} \leqq (n+1)p \quad (\hat{x} \text{ integer}),$$

$$\gamma_1 = \frac{1-2p}{[np(1-p)]^{1/2}}, \quad \gamma_2 = \frac{1-6p(1-p)}{np(1-p)},$$

$$\varphi(t) = [1 + p(\mathrm{e}^{\mathrm{i}t} - 1)]^n.$$

The binomial distribution describes the probability behaviour of the number x of occurrences of an event A in n independent performances of a random experiment, assuming that A has probability p in a single experiment (trial). This model is known as n *Bernoulli trials*. The occurrence of A is called the *success* and the non-occurrence of A is called the *failure* so that the binomial distribution gives the

probability of obtaining exactly x successes in n Bernoulli trials for $x = 0, 1, \ldots, n$. The special case of the binomial distribution for $n = 1$ is the alternative distribution (see Example 33.3.1). Sometimes one uses the *generalized binomial distribution* whose probability of success is not constant but has a value p_i in the i–th trial $(i = 1, \ldots, n)$.

Example 1. A random experiment which consists in throwing four fair dice can be interpreted equivalently as four independent trials each of which consists in throwing one fair die. Let a random variable X denote the number of sixes which turn up in this experiment. Then X has the binomial distribution with parameters $n = 4$ and $p = \frac{1}{6}$ and, e.g.,

$$P(X = 2) = \binom{4}{2}\left(\frac{1}{6}\right)^2\left(1 - \frac{1}{6}\right)^2 = 0{\cdot}116\,,$$

$$\mathrm{E}(X) = 4\,.\,\frac{1}{6} = 0{\cdot}667\,,\ \mathrm{var}(X) = 4\,.\,\frac{1}{6}\left(1 - \frac{1}{6}\right) = 0{\cdot}556\,.$$

2. *Negative binomial distribution* with parameters r, p $(r > 0,\ 0 < p < 1)$:

$$P(X = x) = \begin{cases} p^r & \text{for } x = 0, \\ \dfrac{(r + x - 1)(r + x - 2)\ldots r}{x!}\,p^r(1 - p)^x & \text{for } x = 1, 2, \ldots, \end{cases}$$

$$\mathrm{E}(X) = \frac{r(1 - p)}{p}\,,\ \mathrm{var}(X) = \frac{r(1 - p)}{p^2}\,,$$

$$\frac{(r - 1)(1 - p)}{p} - 1 \leqq \hat{x} \leqq \frac{(r - 1)(1 - p)}{p} \quad (\hat{x} \text{ integer})\,,$$

$$\gamma_1 = \frac{2 - p}{[r(1 - p)]^{1/2}}\,,\ \gamma_2 = \frac{p^2 - 6p + 6}{r(1 - p)}$$

$$\varphi(t) = \left[\frac{p}{1 - (1 - p)\mathrm{e}^{\mathrm{i}t}}\right]^r.$$

In particular, if r is a positive integer, then one speaks of the *Pascal distribution*, where we can write

$$P(X = x) = \binom{r + x - 1}{x}p^r(1 - p)^x \qquad (x = 0, 1, \ldots)\,.$$

Finally, if $r = 1$, then the distribution is called the *geometric distribution* and

$$P(X = x) = p(1 - p)^x \quad (x = 0, 1, \ldots)\,.$$

The Pascal distribution describes the probability behaviour of the number x of failures before the r–th success in Bernoulli trials with the probability p of success

(therefore one also speaks of the *binomial waiting-time distribution*). The probabilities of the geometrical distribution that describes the waiting time before the first success have the form of a geometrical sequence.

3. *Poisson distribution* with parameter λ ($\lambda > 0$):

$$P(X = x) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!} \qquad (x = 0, 1, \dots),$$
$$\mathrm{E}(X) = \lambda, \ \mathrm{var}(X) = \lambda, \ \lambda - 1 \leqq \hat{x} \leqq \lambda \quad (\hat{x} \text{ integer}),$$
$$\gamma_1 = 1/\sqrt{\lambda}\ , \ \gamma_2 = 1/\lambda,$$
$$\varphi(t) = \exp[\lambda(\mathrm{e}^{it} - 1)].$$

Theorem 1 (*Approximation of the Binomial Distribution by the Poisson Distribution*). *Let a sequence of probabilities p_n be given with*

$$\lim_{n\to\infty} np_n = \lambda, \tag{1}$$

where λ is a positive number. Then

$$\lim_{n\to\infty} \binom{n}{x} p_n^x (1 - p_n)^{n-x} = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!} \qquad (x = 0, 1, \dots). \tag{2}$$

REMARK 1. The assumption (1) means, especially, that the sequence of probabilities p_n converges to zero with the rate n^{-1}.

The Poisson distribution is sometimes called the "law of rare events" since it describes the probability behaviour of the number of successes in a long series of Bernoulli trials when the probability of success in each individual trial is very small (e.g. the number of the red blood-corpuscles in the field of a microscope, the number of defective items in a lot of products, the number of calls in a telephone exchange, etc.). It can approximate some more complicated distributions. The approximation of the binomial distribution with parameters n, p by the Poisson distribution with parameter $\lambda = np$ is recommended for $p < 0{\cdot}1$ and $n > 30$.

Example 2. A transmitted signal is damaged in one of 100 cases in the communication channel. Find the probability that none of the 200 signals transmitted by this channel is damaged.

The probability that a signal is damaged is $p = 0{\cdot}01$. According to Theorem 1 one can assume that the number X of the damaged signals among the 200 transmitted signals has the Poisson distribution with parameter $\lambda = 200 \,.\, 0{\cdot}01 = 2$. Hence the probability of interest is

$$P(X = 0) = \mathrm{e}^{-\lambda}\frac{\lambda^0}{0!} = \mathrm{e}^{-2} = 0{\cdot}135.$$

For comparison, the result obtained when using the binomial distribution with parameters $n = 200$ and $p = 0{\cdot}01$, is

$$P(X = 0) = \binom{n}{0} p^0 (1-p)^{n-0} = 0{\cdot}99^{200} = 0{\cdot}134\,.$$

Theorem 2 (*Distribution of the Sum of Independent Random Variables*). *Let independent random variables X_i $(i = 1, \ldots, k)$ have the binomial distribution with parameters n_i, p or the negative binomial distribution with parameters r_i, p or the Poisson distribution with parameters λ_i. Then the random variable $X_1 + \cdots + X_k$ has the binomial distribution with parameters $n_1 + \cdots + n_k$, p or the negative binomial distribution with parameters $r_1 + \cdots + r_k$, p or the Poisson distribution with parameter $\lambda_1 + \cdots + \lambda_k$, respectively.*

4. *Hypergeometric distribution* with parameters N, M, n (N, M, n are positive integers, $M < N$, $n < N$):

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \qquad (x = \max(0,\, M + n - N), \ldots, \min(M, n))\,,$$

$$\mathrm{E}(X) = n\frac{M}{N}\,, \quad \mathrm{var}(X) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1}\,,$$

$$\frac{(M+1)(n+1)}{N+2} - 1 \leqq \hat{x} \leqq \frac{(M+1)(n+1)}{N+2} \qquad (\hat{x} \text{ integer})\,.$$

Let a lot have N items and let M of these items have a certain property. Then the hypergeometric distribution describes the probability behaviour of the number of items with this property among n items that are randomly selected from the lot. If $n/N < 0{\cdot}1$, $M/N < 0{\cdot}1$, then the hypergeometric distribution can be approximated by the binomial distribution with parameters n, $p = M/N$. If $n/N < 0{\cdot}1$, $M/N < 0{\cdot}1$ and $n > 30$, then it can be approximated by the Poisson distribution with parameter $\lambda = nM/N$.

Example 3. Obviously $N = 50$, $M = 6$, $n = 5$ and $x = 2$ in Example 33.1.3.

REMARK 2. Distribution functions or probability functions of important discrete distributions are tabulated in various statistical tables (see e.g. [202], [205], [362], [368], [491]). In addition, they are included as an appendix in majority of textbooks on probability and statistics or they can be calculated by commonly available statistical software. One can also take advantage of their approximations by the normal distribution (see Remark 33.13.2).

33.7. Important Continuous Distributions

1. *Uniform* (or *rectangular*) *distribution* on interval (a, b) $(-\infty < a < b < \infty)$:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a < x < b, \\ 0 & \text{otherwise}, \end{cases}$$

$$\mu'_k = \frac{b^{k+1} - a^{k+1}}{(b-a)(k+1)} \quad (k = 1, 2, \dots), \quad \operatorname{var}(X) = \frac{(b-a)^2}{12},$$

$$\gamma_1 = 0, \quad \gamma_2 = -\tfrac{6}{5},$$

$$\varphi(t) = \frac{e^{itb} - e^{ita}}{it(b-a)}.$$

Theorem 1. *Let U be a random variable with the uniform distribution on the interval $(0, 1)$. Let F be an increasing distribution function. Then the random variable $X = F^{-1}(U)$, where F^{-1} is the inverse function to the function F, has the distribution function F.*

Theorem 1 which can be generalized for an arbitrary distribution function F shows the importance of the uniform distribution for simulations of various probability distributions (in the so-called random number generators).

2. *Normal distribution* with parameters μ, σ^2 $(-\infty < \mu < \infty, \sigma > 0)$:

$$f(x) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \qquad (-\infty < x < \infty),$$

$$\mathrm{E}(X) = \mu, \quad \operatorname{var}(X) = \sigma^2,$$

$$\mu_{2k-1} = 0, \quad \mu_{2k} = \frac{(2k)!}{k!\,2^k}\sigma^{2k} \quad (k \text{ positive integer}),$$

$$\hat{x} = x_{0\cdot 5} = \mu,$$

$$\gamma_1 = 0, \quad \gamma_2 = 0,$$

$$\varphi(t) = \exp(i\mu t - \tfrac{1}{2}\sigma^2 t^2).$$

Due to the Central Limit Theorem (see § 33.13), the normal distribution (sometimes also called the *Gaussian distribution* or the *Gauss–Laplace distribution*) plays a key role in theory of probability and mathematical statistics. It is mostly denoted by the symbol $N(\mu, \sigma^2)$. In particular, $N(0, 1)$ is the so-called *standard normal distribution* with the probability density $\varphi(x)$ and the distribution function $\Phi(x)$ of the form

$$\varphi(x) = \frac{1}{\sqrt{(2\pi)}} e^{-x^2/2}, \quad \Phi(x) = \int_{-\infty}^{x} \varphi(t)\,dt \qquad (-\infty < x < \infty) \tag{1}$$

(see Figs. 33.6 and 33.7).

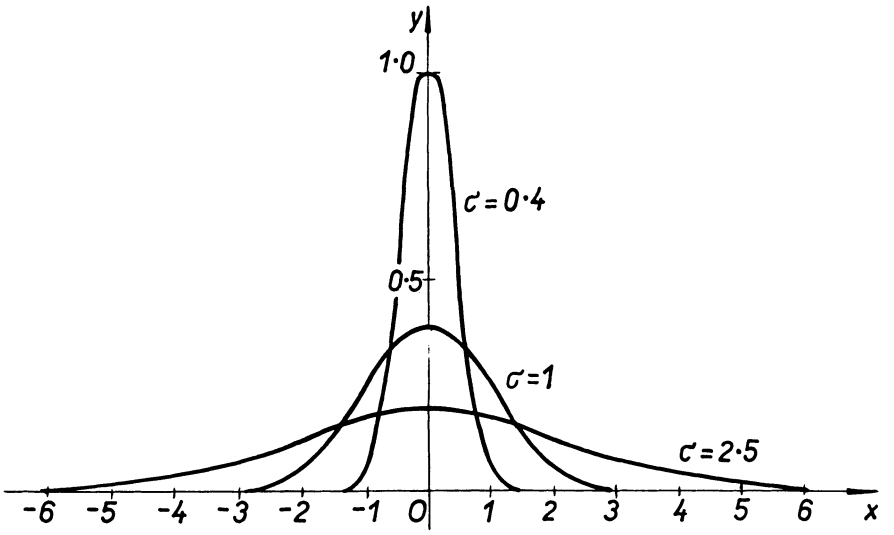


Fig. 33.6. Probability densities of the normal distribution $N(0, \sigma^2)$ for $\sigma = 0{\cdot}4, 1, 2{\cdot}5$. The case $\sigma = 1$ represents the density $\varphi(x)$ in (1).

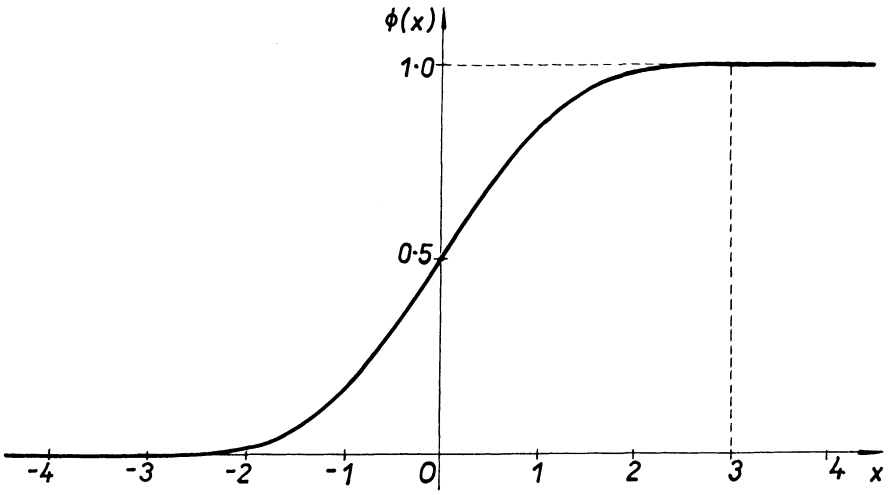


Fig. 33.7. Distribution function of the standard normal distribution $N(0, 1)$.

Theorem 2. *Let the random variable X have the distribution $N(\mu, \sigma^2)$. Then*

$$P(a \leqq X \leqq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right), \tag{2}$$

$$P(|X-\mu| \leqq a) = 2\Phi\left(\frac{a}{\sigma}\right) - 1, \quad P(|X-\mu| > a) = 2\left[1 - \Phi\left(\frac{a}{\sigma}\right)\right] \tag{3}$$

for any real a and b, $a < b$.

REMARK 1. For an arbitrary real x,

$$\Phi(x) = \frac{1}{2} + \frac{1}{\sqrt{(2\pi)}} e^{-x^2/2} \left(\frac{x}{1} + \frac{x^3}{1.3} + \frac{x^5}{1.3.5} + \dots \right) = \tag{4}$$

$$= \frac{1}{2} + \frac{1}{\sqrt{(2\pi)}} \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{k!(2k+1)2^k} . \tag{5}$$

REMARK 2. The values $\varphi(x)$ and $\Phi(x)$ (or $2\Phi(x) - 1$) are tabulated for $x \geqq 0$ (see e.g. [362] and further references in Remark 33.6.2); for $x < 0$, one can make use of the relations $\varphi(x) = \varphi(-x)$ and $\Phi(x) = 1 - \Phi(-x)$. The P–quantiles of the distribution $N(0, 1)$ are usually denoted by u_P and are tabulated for $0{\cdot}5 \leqq P < 1$ (see e.g. [362], [368] and Remark 33.6.2); for $0 < P < 0{\cdot}5$, one can make use of the relation $u_P = -u_{1-P}$ (in some tables, the values u_{1-P} are tabulated for $0 < P \leqq 0{\cdot}5$). Important quantiles u_P used frequently in mathematical statistics are given in Tab. 33.1 below.

REMARK 3 (*Sigma Limits*). For a random variable X with the distribution $N(\mu, \sigma^2)$ one has

$$P(|X - \mu| > k\sigma) = \begin{cases} 0{\cdot}3173 & \text{for } k = 1, \\ 0{\cdot}0455 & \text{for } k = 2, \\ 0{\cdot}0027 & \text{for } k = 3. \end{cases} \tag{6}$$

Most frequently one uses the *three-sigma limits*, i.e. the fact that the random variable X lies outside the interval $(\mu - 3\sigma, \mu + 3\sigma)$ with probability $0{\cdot}0027$. The value $0{\cdot}6745\sigma$ is called the *probable error* since

$$P(|X - \mu| > 0{\cdot}6745\sigma) = 0{\cdot}5 . \tag{7}$$

3. *Logarithmic normal* (or *lognormal*) *distribution* with parameters μ, σ^2 $(-\infty < \mu < \infty, \sigma > 0)$:

$$f(x) = \begin{cases} \dfrac{1}{\sqrt{(2\pi)}\sigma x} \exp\left[-\dfrac{(\ln x - \mu)^2}{2\sigma^2} \right] & \text{for } x > 0, \\ 0 & \text{for } x \leqq 0, \end{cases}$$

$$\mu'_k = \exp(k\mu + \tfrac{1}{2}k^2\sigma^2) \quad (k = 1, 2, \dots), \quad \operatorname{var}(X) = (\mu'_1)^2[\exp(\sigma^2) - 1],$$

$$\hat{x} = \exp(\mu - \sigma^2), \quad x_{0{\cdot}5} = \exp(\mu),$$

$$\gamma_1 = [\exp(\sigma^2) + 2][\exp(\sigma^2) - 1]^{1/2},$$

$$\gamma_2 = \exp(4\sigma^2) + 2\exp(3\sigma^2) + 3\exp(2\sigma^2) - 6 .$$

Theorem 3. *A random variable X has the logarithmic normal distribution with parameters μ, σ^2 if and only if the random variable* $\ln X$ *has the distribution* $N(\mu, \sigma^2)$.

The logarithmic normal distribution is suitable for the description of magnitudes of grains in loose materials, in theory of reliability (e.g. for modelling the time to failure of semiconductors or of materials with possible fatigue defects), in statistical physics and elsewhere.

4. *Exponential distribution* with parameter δ ($\delta > 0$):

$$f(x) = \begin{cases} \frac{1}{\delta} \exp\left(-\frac{x}{\delta}\right) & \text{for } x > 0, \\ 0 & \text{for } x \leqq 0, \end{cases}$$

$$\mathrm{E}(X) = \delta, \ \mathrm{var}(X) = \delta^2,$$

$$x_{0\cdot 5} = \delta \ln 2,$$

$$\gamma_1 = 2, \ \gamma_2 = 6,$$

$$\varphi(t) = 1/(1 - \mathrm{i}\delta t).$$

Theorem 4 (*Properties of the Exponential Distribution*). *Let a random variable X have the exponential distribution. Then*

$$P(X > t + s \mid X > s) = P(X > t) \tag{8}$$

holds for any positive s and t.

The exponential distribution is applied in theory of reliability to model the time to failure of some products or in queueing theory to model the waiting time to service. Since it is the "distribution without memory" it suits for models that are independent of the previous development (e.g. for modelling the time to failure that occurs due to random causes and not due to a wear so that, in accordance with (8), the probability of the future failureless work does not depend on the length of the past failureless one).

5. *Double exponential distribution* with parameters a, δ ($-\infty < a < \infty$, $\delta > 0$):

$$f(x) = \frac{1}{2\delta} \exp\left(-\frac{|x - a|}{\delta}\right) \qquad (-\infty < x < \infty),$$

$$\mathrm{E}(X) = a, \ \mathrm{var}(X) = 2\delta^2,$$

$$\hat{x} = x_{0\cdot 5} = a,$$

$$\gamma_1 = 0, \ \gamma_2 = 3,$$

$$\varphi(t) = \frac{\mathrm{e}^{\mathrm{i}at}}{1 + \delta^2 t^2}.$$

6. *Weibull distribution* with parameters p, δ ($p > 0$, $\delta > 0$):

$$f(x) = \begin{cases} \dfrac{px^{p-1}}{\delta^p} \exp\left[-\left(\dfrac{x}{\delta}\right)^p\right] & \text{for } x > 0, \\ 0 & \text{for } x \leqq 0, \end{cases}$$

$$\mathrm{E}(X) = \Gamma\left(\frac{p+1}{p}\right)\delta, \quad \mathrm{var}(X) = \left[\Gamma\left(\frac{p+2}{p}\right) - \Gamma^2\left(\frac{p+1}{p}\right)\right]\delta^2,$$

$$\hat{x} = \delta\left(\frac{p-1}{p}\right)^{1/p}, \quad x_{0\cdot5} = \delta(\ln 2)^{1/p}$$

(the function Γ is defined in § 13.11). The Weibull distribution has applications in theory of reliability to modelling the time to failure due to wear (e.g. bearings have the value of parameter δ approximately 1·5). It is suitable for the description of physical properties of materials as well. The special case of the Weibull distribution for $p = 1$ is the exponential distribution and for $p = 2$, $\delta = \sigma\sqrt{2}$ ($\sigma > 0$) the *Rayleigh distribution* with the probability density

$$f(x) = \begin{cases} \dfrac{x}{\sigma^2} \exp\left(-\dfrac{x^2}{2\sigma^2}\right) & \text{for } x > 0, \\ 0 & \text{for } x \leqq 0, \end{cases}$$

7. *Gamma distribution* with parameters p, δ ($p > 0$, $\delta > 0$):

$$f(x) = \begin{cases} \dfrac{x^{p-1}}{\delta^p \Gamma(p)} \exp\left(-\dfrac{x}{\delta}\right) & \text{for } x > 0, \\ 0 & \text{for } x \leqq 0, \end{cases}$$

$$\mathrm{E}(X) = p\delta, \ \mathrm{var}(X) = p\delta^2,$$

$$\hat{x} = (p-1)\delta \qquad (p \geqq 1),$$

$$\gamma_1 = 2/\sqrt{p}, \quad \gamma_2 = 6/p,$$

$$\varphi(t) = (1 - \mathrm{i}\delta t)^{-p}.$$

If p is a positive integer, then the gamma distribution is called the *Erlang distribution* that is useful in reliability theory for modelling the time to failure of a system with redundant components or in queueing theory if the service has several phases. The special case for $p = 1$ is the exponential distribution.

Theorem 5. *Let $X_1, \ldots, X_n$ be independent random variables such that X_i has the gamma distribution with parameters p_i, δ ($i = 1, \ldots, n$). Then the random variable $X_1 + \cdots + X_n$ has the gamma distribution with parameters $p_1 + \cdots + p_n$, δ. In particular, if X_i has the exponential distribution with parameter δ ($i = 1, \ldots, n$), then $X_1 + \cdots + X_n$ has the Erlang distribution with parameters n, δ.*

8. χ^2 (*chi-square*) *distribution* with n degrees of freedom (n is a positive integer):

$$f(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \exp(-x/2) & \text{for } x > 0, \\ 0 & \text{for } x \leqq 0. \end{cases}$$

Theorem 6. *Let $X_1, \ldots, X_n$ be independent random variables with the distribution $N(0, 1)$. Then the random variable $X_1^2 + \cdots + X_n^2$ has the distribution $\chi^2(n)$* (i.e. *the χ^2 distribution with n degrees of freedom*).

The χ^2 distribution is a special case of the gamma distribution for $p = n/2$, $\delta = 2$ so that, e.g.,

$$\mathrm{E}(X) = n, \quad \mathrm{var}(X) = 2n.$$

9. t (or *Student*) *distribution* with n degrees of freedom (n is a positive integer):

$$f(x) = \frac{\Gamma\left(\dfrac{n+1}{2}\right)}{(n/2)\sqrt{(\pi n)}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \qquad (-\infty < x < \infty),$$

$$\mathrm{E}(X) = 0 \quad (n > 1), \quad \mathrm{var}(X) = n/(n-2) \quad (n > 2).$$

Theorem 7. *Let independent random variables X and Y be such that X has the distribution $N(0, 1)$ and Y has the distribution $\chi^2(n)$. Then the random variable $X/\sqrt{(Y/n)}$ has the distribution $t(n)$* (i.e. *the t distribution with n degrees of freedom*).

10. F (or *Fisher–Snedecor*) *distribution* with n_1 and n_2 degrees of freedom (n_1, n_2 are positive integers):

$$f(x) = \begin{cases} \dfrac{1}{\mathrm{B}(n_1/2, n_2/2)} \left(\dfrac{n_1}{n_2}\right)^{n_1/2} x^{n_1/2-1} \left(1 + \dfrac{n_1}{n_2} x\right)^{-(n_1+n_2)/2} & \text{for } x > 0, \\ 0 & \text{for } x \leqq 0, \end{cases}$$

$$\mathrm{E}(X) = \frac{n_2}{n_2 - 2} \quad (n_2 > 2), \quad \mathrm{var}(X) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)} \quad (n_2 > 4)$$

(the function B is defined in § 13.11).

Theorem 8. *Let independent random variables X and Y be such that X has the distribution $\chi^2(n_1)$ and Y has the distribution $\chi^2(n_2)$. Then the random variable $(X/n_1)/(Y/n_2)$ has the distribution $F(n_1, n_2)$* (i. e. *the F-distribution with n_1 and n_2 degrees of freedom*).

REMARK 4. The distribution $\chi^2(n), t(n)$ and $F(n_1, n_2)$ play an important role in mathematical statistics. Therefore their quantiles $\chi^2_P(n), t_P(n)$ and $F_P(n_1, n_2)$ are tabulated (see e.g. [362] and further references in Remark 33.6.2). Figs. 33.8, 33.9 and 33.10 show the probability densities of the distributions $\chi^2(4), t(4)$ and $F(4, 8)$ including the corresponding 0·95–quantiles (i.e. the 95–th percentiles). Tab. 33.1

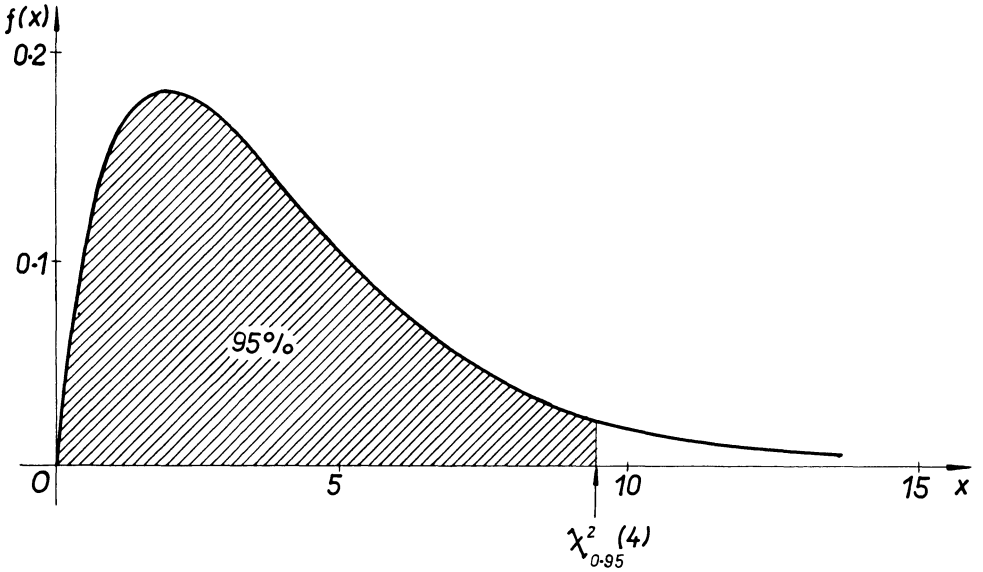


Fig. 33.8. Probability density and 0·95 quantile of the distribution $\chi^2(4)$.

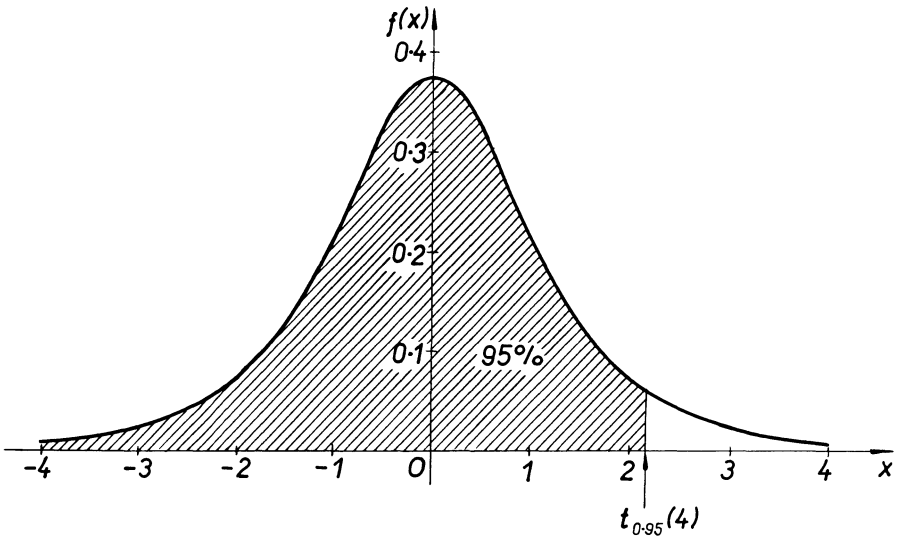


Fig. 33.9. Probability density and 0·95 quantile of the distribution $t(4)$.

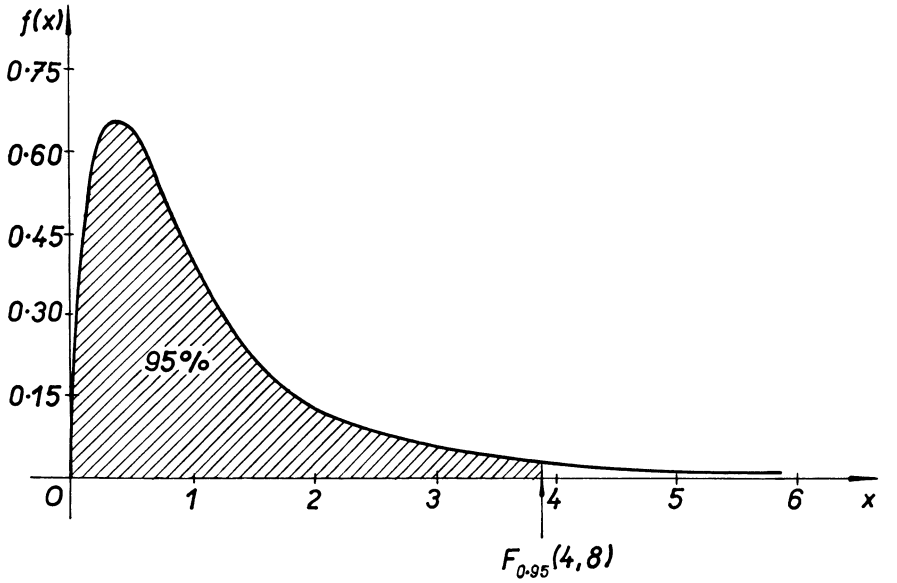


Fig. 33.10. Probability density and 0·95 quantile of the distribution $F(4, 8)$.

contains some quantiles of these distributions. Tabs. 33.2 and 33.3 contain important quantiles $\chi^2_P(n)$ and $t_P(n)$ for various degrees of freedom n that are frequently used in statistical problems (see § 34.10). Sometimes one uses the following approximations by the quantiles u_P of the distribution $N(0, 1)$ (see Remark 2):

$$\chi^2_P(n) \approx n\left(1 - \frac{2}{9n} + u_P\sqrt{\frac{2}{9n}}\right)^3, \tag{9}$$

$$t_P(n) \approx u_P\left[1 + \frac{1}{4n}(1 + u_P^2) + \frac{1}{96n^2}(3 + 16u_P^2 + 5u_P^4)\right]. \tag{10}$$

In particular,

$$\chi^2_P(1) \approx u^2_{(1+P)/2}, \quad \chi^2_P(2) \approx -2\ln(1 - P).$$

TABLE 33.1

P	0·01	0·025	0·05	0·1	0·9	0·95	0·975	0·99
u_P	−2·3263	−1·9600	−1·6449	−1·2816	1·2816	1·6449	1·9600	2·3263
$\chi^2_P(4)$	0·29711	0·48442	0·71072	1·0636	7·7794	9·4877	11·143	13·277
$t_P(4)$	−3·7469	−2·7764	−2·1318	−1·5332	1·5332	2·1318	2·7764	3·7469
$F_P(4; 8)$	0·06757	0·11136	0·16554	0·25285	3·4579	5·3177	7·5709	11·259

TABLE 33.2

n	$\chi^2_{0\cdot025}(n)$	$\chi^2_{0\cdot05}(n)$	$\chi^2_{0\cdot95}(n)$	$\chi^2_{0\cdot975}(n)$
1	0·00098	0·00393	3·8415	5·0239
2	0·05064	0·10259	5·9915	7·3778
3	0·21580	0·35185	7·8147	9·3484
4	0·48442	0·71072	9·4877	11·143
5	0·83121	1·1455	11·070	12·833
6	1·2373	1·6354	12·592	14·449
7	1·6899	2·1673	14·067	16·013
8	2·1797	2·7326	15·507	17·535
9	2·7004	3·3251	16·919	19·023
10	3·2470	3·9403	18·307	20·483
11	3·8157	4·5748	19·675	21·920
12	4·4038	5·2260	21·026	23·337
13	5·0088	5·8919	22·362	24·736
14	5·6287	6·5706	23·685	26·119
15	6·2621	7·2609	24·996	27·488
16	6·9077	7·9616	26·296	28·845
17	7·5642	8·6718	27·587	30·191
18	8·2307	9·3905	28·869	31·526
19	8·9065	10·117	30·144	32·852
20	9·5908	10·851	31·410	34·170
21	10·283	11·591	32·671	35·479
22	10·982	12·338	33·924	36·781
23	11·689	13·091	35·172	38·076
24	12·401	13·848	36·415	39·364
25	13·120	14·611	37·652	40·646
30	16·791	18·493	43·773	46·979
35	20·569	22·465	49·802	53·203
40	24·433	26·509	55·758	59·342
45	28·366	30·612	61·656	65·410
50	32·357	34·764	67·505	71·420
60	40·482	43·188	79·082	83·298
70	48·758	51·739	90·531	95·023
80	57·153	60·391	101·88	106·63
90	65·647	69·126	113·15	118·14
100	74·222	77·929	124·34	129·56

TABLE 33.3

n	$t_{0\cdot95}(n)$ $(= -t_{0\cdot05}(n))$	$t_{0\cdot975}(n)$ $(= -t_{0\cdot025}(n))$
1	6·3138	12·706
2	2·9200	4·3027
3	2·3534	3·1824
4	2·1318	2·7764
5	2·0150	2·5706
6	1·9432	2·4469
7	1·8946	2·3646
8	1·8595	2·3060
9	1·8331	2·2622
10	1·8125	2·2281
11	1·7959	2·2010
12	1·7823	2·1788
13	1·7709	2·1604
14	1·7613	2·1448
15	1·7531	2·1314
16	1·7459	2·1199
17	1·7396	2·1098
18	1·7341	2·1009
19	1·7291	2·0930
20	1·7247	2·0860
21	1·7207	2·0796
22	1·7171	2·0739
23	1·7139	2·0687
24	1·7109	2·0639
25	1·7081	2·0595
30	1·6973	2·0423
35	1·6896	2·0301
40	1·6839	2·0211
45	1·6749	2·0141
50	1·6759	2·0086
60	1·6706	2·0003
70	1·6669	1·9944
80	1·6641	1·9901
90	1·6620	1·9867
100	1·6602	1·9840

REMARK 5. For the quantiles x_P of continuous distributions, the following relations are valid:

(i) logarithmic normal distribution:

$$x_P = \exp(\mu + \sigma u_P); \tag{11}$$

(ii) exponential distribution:

$$x_P = \frac{\delta}{2}\chi^2_P(2); \tag{12}$$

(iii) Weibull distribution:

$$x_P = \delta[\chi^2_P(2)/2]^{1/p}; \tag{13}$$

(iv) Erlang distribution:

$$x_P = \frac{\delta}{2}\chi^2_P(2p). \tag{14}$$

11. *Beta distribution* with parameters p, q ($p > 0$, $q > 0$):

$$f(x) = \begin{cases} \dfrac{1}{\mathrm{B}(p, q)} x^{p-1}(1-x)^{q-1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise}, \end{cases}$$

$$\mathrm{E}(X) = \frac{p}{p+q}, \quad \mathrm{var}(X) = \frac{pq}{(p+q)^2(p+q+1)}$$

(the function B is defined in § 13.11). A special case of the beta distribution for $p = q = 1$ is the uniform distribution on the interval (0, 1).

12. *Cauchy distribution* with parameters a, λ ($-\infty < a < \infty$, $\lambda > 0$):

$$f(x) = \frac{1}{\pi}\frac{\lambda}{\lambda^2 + (x-a)^2} \qquad (-\infty < x < \infty),$$

$$\hat{x} = x_{0.5} = a,$$

$$\varphi(t) = \exp(\mathrm{i}at - \lambda|t|).$$

The Cauchy distribution has no mean and no variance.

REMARK 6. Further continuous distributions with applications to various fields are *triangular*, *logistic*, χ, *Maxwell*, *Pareto* and others (see [244], [249], [269]).

33.8. Important Multivariate Distributions

1. k–variate *multinomial distribution* with parameters $n, p_1, \ldots, p_k$ (n is a positive integer, $0 < p_j < 1$, $p_1 + \cdots + p_k = 1$):

$$P(X_1 = x_1, \ldots, X_k = x_k) = \begin{cases} \dfrac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k} & \text{for } x_j = 0, 1, \ldots n, \ \sum_{j=1}^{k} x_j = n, \\ 0 & \text{otherwise}, \end{cases}$$

$$\mathrm{E}(X_j) = np_j, \quad \mathrm{var}(X_j) = np_j(1 - p_j) \qquad (j = 1, \ldots, k),$$

$$\mathrm{cov}(X_i, X_j) = -np_ip_j \qquad (i, j = 1, \ldots, k;\ i \neq j),$$

$$\varphi(t_1, \ldots, t_k) = \left[\sum_{j=1}^{k} p_j \exp(\mathrm{i}t_j)\right]^n.$$

The multinomial distribution generalizes the binomial distribution in such a way that each of n independent trials has k possible outcomes with probabilities p_j now and the random variable X_j describes the number of occurrences of the j–th outcome ($j = 1, \ldots, k$) in these n trials. The marginal distribution (see § 33.5) of the random variable X_j is the binomial distribution with parameters n, p_j ($j = 1, \ldots, k$).

Example 1. Let a random experiment consist in throwing four fair dice and let a random variable X_j denote the number of j's that turn up in this experiment ($j = 1, \ldots, 6$). Then the random vector $\boldsymbol{X} = (X_1, \ldots, X_6)'$ has the 6–variate multinomial distribution with parameters $n = 4$, $p_j = \frac{1}{6}$ ($j = 1, \ldots, 6$). The random variable X_6 coincides with the random variable X from Example 33.6.1. In particular,

$$\mathrm{cov}(X_i, X_j) = -4 \cdot \tfrac{1}{6} \cdot \tfrac{1}{6} = -0{\cdot}111 \qquad (i, j = 1, \ldots, 6;\ i \neq j).$$

2. *n–variate normal distribution* with parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ ($\boldsymbol{\mu}$ is an n–component column vector with real components μ_i, $\boldsymbol{\Sigma}$ is an $n \times n$ symmetric positive definite matrix with real elements σ_{ij}):

$$f(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp[-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})]$$
$$(-\infty < x_i < \infty,\ i = 1, \ldots, n),$$

$$\mathrm{E}(X_i) = \mu_i, \quad \mathrm{cov}(X_i, X_j) = \sigma_{ij} \qquad (i, j = 1, \ldots, n),$$
$$\varphi(t_1, \ldots, t_n) = \exp(\mathrm{i}\boldsymbol{\mu}'\boldsymbol{t} - \tfrac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t})$$

($|\Sigma|$ is the determinant of Σ, Σ^{-1} is the inverse of Σ, $\boldsymbol{x} = (x_1, \ldots, x_n)'$, $\boldsymbol{t} = (t_1, \ldots, t_n)'$). This distribution is usually denoted by $N_n(\boldsymbol{\mu}, \Sigma)$. If a random vector has the distribution $N_n(\boldsymbol{\mu}, \Sigma)$, then its mean is $\boldsymbol{\mu}$ and its covariance matrix is Σ.

REMARK 1. The positive definite matrix Σ is nonsingular but the multivariate normal distribution can be introduced also for a singular matrix Σ. Then it is defined by its characteristic function rather than by the probability density.

Theorem 1 (*Probability Distribution of a Linear Transformation*). *Let a random vector have the distribution $N_n(\boldsymbol{\mu}, \Sigma)$. Let $\boldsymbol{a}$ be a real m-component column vector and $\boldsymbol{A}$ be a real $m \times n$ matrix. Then the random vector $\boldsymbol{AX} + \boldsymbol{a}$ has the distribution $N_m(\boldsymbol{A\mu} + \boldsymbol{a}, \boldsymbol{A\Sigma A'})$. In particular, the random variable X_i has the marginal distribution $N(\mu_i, \sigma_{ii})$ $(i = 1, \ldots, n)$.*

Example 2. If a random variable X has the distribution $N(\mu, \sigma^2)$, then the random variable

$$Y = \frac{X - \mu}{\sigma} \tag{1}$$

has the distribution $N(0, 1)$.

Theorem 2. *A random vector $\boldsymbol{X}$ has the n-variate normal distribution if and only if the random variable $\boldsymbol{c'X}$ has the (one-variate) normal distribution for any real n-component column vector $\boldsymbol{c}$.*

REMARK 2. The probability density of the two-variate (or bi-variate) normal distribution can be written in the form

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\} \quad (-\infty < x_i < \infty,\ i = 1, 2), \tag{2}$$

where $\sigma_i^2 = \sigma_{ii}$ is the variance of the random variable X_i with the marginal distribution $N(\mu_i, \sigma_i^2)$ $(i = 1, 2)$ and ρ is the correlation coefficient of the random variables X_1 and X_2,

$$\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}. \tag{3}$$

REMARK 3. Further multivariate distributions with applications to various fields are *multivariate t distribution, Wishart distribution* (a multivariate version of the χ^2 distribution), the *Dirichlet distribution* (a multivariate version of the beta distribution) and others (see e.g. [245], [269], [499]).

33.9. Transformations of Random Variables

Theorem 1 (*Probability Density of Transformed Random Variable*). *Let X be a continuous random variable with the probability density $f(x)$. Let $h(x)$ be a strictly monotonic differentiable function on such interval I that $P(X \in I) = 1$. Then the random variable $Z = h(X)$ has the probability density*

$$g(z) = f(h^{-1}(z)) \left| \frac{\mathrm{d}h^{-1}(z)}{\mathrm{d}z} \right| , \tag{1}$$

where $h^{-1}(z)$ is the inverse function to the function $z = h(x)$.

REMARK 1. Theorem 1 can be generalized to hold for functions h that are not strictly monotonic and for transformations of random vectors (see e.g. [499]).

REMARK 2. Let a continuous random variable X have the probability density $f(x)$. Then the following formulae hold for the probability densities $q(z)$ of random variables Z:

(i) $Z = aX + b$: $q(z) = \frac{1}{|a|} f\left(\frac{z-b}{a}\right)$ $(a \neq 0)$;

(ii) $Z = X^2$: $q(z) = \begin{cases} \frac{1}{2\sqrt{z}}[f(\sqrt{z}) + f(-\sqrt{z})] & \text{for } z > 0, \\ 0 & \text{for } z \leqq 0; \end{cases}$

(iii) $Z = |X|$: $q(z) = \begin{cases} f(z) + f(-z) & \text{for } z > 0, \\ 0 & \text{for } z \leqq 0; \end{cases}$

(iv) $Z = \mathrm{e}^X$: $q(z) = \begin{cases} \frac{1}{|z|} f(\ln z) & \text{for } z > 0, \\ 0 & \text{for } z \leqq 0; \end{cases}$

(v) $Z = 1/X$: $q(z) = \frac{1}{z^2} f\left(\frac{1}{z}\right)$ $(z \neq 0)$.

If, moreover, $f(x) = 0$ for $x \leqq 0$, then

(vi) $Z = \sqrt{X}$: $q(z) = \begin{cases} 2zf(z^2) & \text{for } z > 0, \\ 0 & \text{for } z \leqq 0; \end{cases}$

(vii) $Z = \ln X$: $q(z) = \mathrm{e}^z f(\mathrm{e}^z)$.

REMARK 3. Let independent continuous random variables X and Y have probability densities $f(x)$ and $g(y)$. Then the following formulae hold for the probability

densities $q(z)$ of random variables Z:

$$\text{(i)} \quad Z = X + Y: \quad q(z) = \int_{-\infty}^{\infty} f(x)g(z-x)\,dx = \int_{-\infty}^{\infty} f(z-y)g(y)\,dy;$$

$$\text{(ii)} \quad Z = XY: \quad q(z) = \int_{-\infty}^{\infty} \frac{1}{|x|} f(x) g\left(\frac{z}{x}\right) dx = \int_{-\infty}^{\infty} \frac{1}{|y|} f\left(\frac{z}{y}\right) g(y)\,dy;$$

$$\text{(iii)} \quad Z = X/Y: \quad q(z) = \int_{-\infty}^{\infty} |y| f(yz) g(y)\,dy.$$

Example 1. Let X and Y be independent random variables that have the exponential distribution with parameter δ. Then the random variable $Z = X + Y$ has the probability density

$$q(z) = \int_{-\infty}^{\infty} f(x)g(z-x)\,dx = \int_0^z \frac{1}{\delta} \exp\left(-\frac{x}{\delta}\right) \frac{1}{\delta} \exp\left(-\frac{z-x}{\delta}\right) dx =$$
$$= \frac{z}{\delta^2} \exp\left(-\frac{z}{\delta}\right) \text{ for } z > 0, \qquad q(z) = 0 \text{ for } z \leqq 0.$$

The probability density $q(z)$ corresponds to the Erlang distribution with parameters $p = 2$, δ. It is possible to show by induction, that the sum of n independent random variables possessing the exponential distribution with parameter δ has the Erlang distribution with parameters $p = n$, δ (see Theorem 33.7.5).

Theorem 2 (*Mean and Variance of a Linear Transformation*). *Let $X_1, \ldots, X_n$ be random variables and $a_1, \ldots, a_n, b$ be arbitrary real numbers. If $\mathrm{E}(|X_i|) < \infty$ $(i = 1, \ldots, n)$, then*

$$\mathrm{E}\left(\sum_{i=1}^{n} a_i X_i + b\right) = \sum_{i=1}^{n} a_i \mathrm{E}(X_i) + b. \tag{2}$$

If $\mathrm{E}(X_i^2) < \infty$ $(i = 1, \ldots, n)$, then

$$\mathrm{var}\left(\sum_{i=1}^{n} a_i X_i + b\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \,\mathrm{cov}(X_i, X_j). \tag{3}$$

If $\mathrm{E}(X_i^2) < \infty$ $(i = 1, \ldots, n)$ and the random variables $X_1, \ldots, X_n$ are independent, then (3) takes a simpler form

$$\mathrm{var}\left(\sum_{i=1}^{n} a_i X_i + b\right) = \sum_{i=1}^{n} a_i^2 \,\mathrm{var}(X_i). \tag{4}$$

Example 2. Let independent random variables $X_1, \ldots, X_n$ have the same mean μ and the same variance σ^2. Then

$$\mathrm{E}(\overline{X}) = \mu, \ \mathrm{var}(\overline{X}) = \sigma^2/n \tag{5}$$

holds for their arithmetic mean $\overline{X} = \sum_{i=1}^{n} X_i/n$.

REMARK 4. If X is a random variable with a small coefficient of variation (smaller than 0·2), then it is possible to approximate

$$\mathrm{E}(X/Y) \approx \mathrm{E}(X)/\mathrm{E}(Y). \tag{6}$$

33.10. Some Inequalities

Theorem 1 (*Markov Inequality*). *Let X be a positive random variable* (i.e. $P(X > 0) = 1$) *such that* $\mathrm{E}(X) < \infty$. *Then*

$$P(X > \lambda \mathrm{E}(X)) \leqq 1/\lambda \tag{1}$$

for every $\lambda > 1$.

Theorem 2 (*Chebyshev Inequality*). *Let X be a random variable such that* $\mathrm{E}(X^2) < \infty$. *Then*

$$P(|X - \mathrm{E}(X)| \geqq \varepsilon) \leqq \mathrm{var}(X)/\varepsilon^2, \tag{2}$$

$$P(X - \mathrm{E}(X) \geqq \varepsilon) \leqq \frac{\mathrm{var}(X)}{\mathrm{var}(X) + \varepsilon^2} \tag{3}$$

for every $\varepsilon > 0$.

REMARK 1. The inequality (3) is sometimes called the *Cantelli inequality*.

Theorem 3 (*Camp–Meidell Inequality*). *Let X be a random variable with a continuous unimodal probability distribution such that* $\mathrm{E}(X^2) < \infty$. *Let* $|\hat{x} - \mathrm{E}(X)| \leqq \leqq [\mathrm{var}(X)]^{1/2}$. *Then*

$$P(|X - \mathrm{E}(X)| \geqq \varepsilon) \leqq \frac{\mathrm{var}(X)}{2{\cdot}25\varepsilon^2} \tag{4}$$

for every $\varepsilon > 0$.

Theorem 4 (*Kolmogorov Inequality*). *Let $X_1, \ldots, X_n$ be independent random variables such that $\mathrm{E}(X_i^2) < \infty$ $(i = 1, \ldots, n)$. Then*

$$P\Big(\max_{1 \leqq k \leqq n} \Big|\sum_{i=1}^{k} [X_i - \mathrm{E}(X_i)]\Big| \geqq \varepsilon\Big) \leqq \sum_{i=1}^{n} \mathrm{var}(X_i)/\varepsilon^2 \tag{5}$$

for every $\varepsilon > 0$.

Theorem 5 (*Jensen Inequality*). *Let X be a random variable such that $\mathrm{E}(|X|) < \infty$. Let $g(x)$ be a convex function on an interval I such that $P(X \in I) = 1$. Then*

$$\mathrm{E}[g(X)] \geqq g[\mathrm{E}(X)]\,. \tag{6}$$

If $g(x)$ is concave then the converse inequality holds.

Theorem 6 (*Berry-Essén Inequality*). *Let $X_1, \ldots, X_n$ be independent identically distributed random variables with the mean μ and the variance σ^2 and such that $\mathrm{E}(|X_1|^3) < \infty$. Let $F_n(x)$ be the distribution function of the random variable $\sum_{i=1}^{n} (X_i - \mu)/(\sigma\sqrt{n})$. Then*

$$|F_n(x) - \Phi(x)| \leqq A \frac{\mathrm{E}(|X_1 - \mu|^3)}{\sigma^3 \sqrt{n}} \qquad (-\infty < x < \infty)\,, \tag{7}$$

where $\Phi(x)$ is the distribution function of $N(0, 1)$ and A is a constant independent of x.

33.11. Limit Theorems in Probability Theory

Limit Theorems in Probability Theory. These theorems are concerned with the behaviour of sequences of random variables or probability distributions. An example of such assertions is Theorem 33.6.1 on convergence of binomial distributions to the Poisson distribution. Probability laws called the law of large numbers and the central limit theorem play the key role here.

The *law of large numbers* claims that the variability of a large number of independent (or weakly dependent) random variables is mutually compensated so that their arithmetic mean is nearly constant. In particular, the law of large numbers justifies the stability of relative frequencies mentioned in § 33.1.

The *central limit theorem* expresses the fact that, under very general conditions, the sum of a large number of independent (or weakly dependent) random variables

with general distributions has approximately the normal distribution. In particular, the central limit theorem justifies the assumption on the normality of random variables that result from a large number of random factors such that each of them has a negligible effect by itself.

Convergence of a Sequence of Random Variables. The limit theorems concern various types of convergence of sequences of random variables.

Definition 1. A sequence of random variables $X_1, X_2, \ldots$ *converges in probability* to a random variable X if

$$\lim_{n\to\infty} P(|X_n - X| \geqq \varepsilon) = 0 \tag{1}$$

for every $\varepsilon > 0$.

Definition 2. A sequence of random variables $X_1, X_2, \ldots$ *converges almost surely* (or *with probability* 1) to a random variable X if

$$P(\lim_{n\to\infty} X_n = X) = 1. \tag{2}$$

REMARK 1. The almost sure convergence implies the convergence in probability. The converse is not true.

Definition 3. A sequence of random variables $X_1, X_2, \ldots$ with distribution functions $F_1, F_2, \ldots$ *converges in distribution* (or *weakly*) to a random variable X with a distribution function F if

$$\lim_{n\to\infty} F_n(x) = F(x) \tag{3}$$

holds at every point x of continuity of the function F.

33.12. Law of Large Numbers

Weak Law of Large Numbers. This law concerns the convergence in probability.

Theorem 1 (*Bernoulli Theorem*). *Let $X_1, X_2, \ldots$ be a sequence of independent random variables which have the alternative distribution with parameter p* (see Example 33.3.1 or § 33.6). *Then*

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| \geqq \varepsilon\right) = 0. \tag{1}$$

for every $\varepsilon > 0$.

REMARK 1. The random variable $(X_1 + \cdots + X_n)/n$ in Theorem 1 represents the relative frequency of successes in Bernoulli trials. Theorem 1 thus shows the stability of these relative frequencies that converge in probability to the probability of success when the number of trials increases.

Theorem 2 (*Khintchine Theorem*). *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with the mean μ. Then*

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqq \varepsilon\right) = 0 \tag{2}$$

for every $\varepsilon > 0$.

REMARK 2. According to Theorem 2, the mean of a random variable can be found approximately as the arithmetic mean of independent observations of this random variable. Theorem 1 is a special case of Theorem 2.

Theorem 3 (*Markov Theorem*). *Let $X_1, X_2, \ldots$ be a sequence of independent random variables such that $\mathrm{E}(X_i^2) < \infty$ $(i = 1, 2, \ldots)$. Let the condition*

$$\lim_{n\to\infty} \frac{1}{n}\left[\sum_{i=1}^{n} \mathrm{var}(X_i)\right]^{1/2} = 0 \tag{3}$$

be fulfilled. Then

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(X_i)\right| \geqq \varepsilon\right) = 0 \tag{4}$$

for every $\varepsilon > 0$.

REMARK 3. In particular, the condition (3) is fulfilled if the variances of the random variables $X_1, X_2, \ldots$ are identical or at least bounded by the same constant.

Strong Law of Large Numbers. This law concerns the almost sure convergence.

Theorem 4 (*Kolmogorov Theorem*). *Let the assumptions of Theorem 2 hold. Then*

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1\,. \tag{5}$$

If the mean of the random variables X_i does not exist, then the sequence of the random variables $(X_1 + \cdots + X_n)/n$ is unbounded with probability 1.

33.13. Central Limit Theorems

Theorem 1 (*Moivre-Laplace Theorem*). *Let $X_1, X_2, \ldots$ be a sequence of independent random variables that have the alternative distribution with parameter p* (see Example 33.3.1 or § 33.6). *Then the sequence of the random variables*

$$Y_n = \frac{1}{[np(1-p)]^{1/2}} \left(\sum_{i=1}^{n} X_i - np \right) \tag{1}$$

converges in distribution to a random variable with the distribution $N(0, 1)$.

REMARK 1. Preserving the notation from Theorem 1 we can use the following approximation for every $\varepsilon > 0$ and sufficiently large n:

$$P\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - p \right| \geqq \varepsilon \right) \approx 2\left\{ 1 - \Phi\left(\left[\frac{1}{p(1-p)} \right]^{1/2} \varepsilon \right) \right\}, \tag{2}$$

where Φ is the distribution function of $N(0, 1)$.

Theorem 2 (*Lévy–Lindeberg Theorem*). *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with the mean μ and variance σ^2. Then the sequence of the random variables*

$$Y_n = \frac{1}{\sigma \sqrt{n}} \left(\sum_{i=1}^{n} X_i - n\mu \right) \tag{3}$$

converges in distribution to a random variable with the distribution $N(0, 1)$.

Theorem 3 (*Lyapunov Theorem*). *Let k be a real number, $k > 2$. Let X_1, $X_2, \ldots$ be a sequence of independent random variables such that $\mathrm{E}[|X_i - \mathrm{E}(X_i)|^k] < \infty$ $(i = 1, 2, \ldots)$. Let the condition*

$$\lim_{n \to \infty} \frac{\sum_{i=1}^{n} \mathrm{E}[|X_i - \mathrm{E}(X_i)|^k]}{\left[\sum_{i=1}^{n} \mathrm{var}(X_i) \right]^{k/2}} = 0 \tag{4}$$

be fulfilled. Then the sequence of the random variables

$$Y_n = \frac{1}{\left[\sum_{i=1}^{n} \mathrm{var}(X_i) \right]^{1/2}} \left[\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathrm{E}(X_i) \right] \tag{5}$$

converges in distribution to a random variable with the distribution $N(0, 1)$.

REMARK 2. Central limit theorems justify the approximation of some discrete distributions by the normal distribution: Let a random variable X have the binomial, negative binomial, Poisson or hypergeometric distribution (see § 33.6) and let $\mathrm{var}(X) > 9$ (in the case of the hypergeometric distribution we require, moreover, $n/N < 0{\cdot}1$). Then for any non-negative integers a and b, $a < b$, one can use the approximation

$$P(a \leqq X \leqq b) \approx \Phi\left(\frac{b + \frac{1}{2} - \mathrm{E}(X)}{[\mathrm{var}(X)]^{1/2}}\right) - \Phi\left(\frac{a - \frac{1}{2} - \mathrm{E}(X)}{[\mathrm{var}(X)]^{1/2}}\right), \tag{6}$$

where Φ is the distribution function of $N(0, 1)$.

REMARK 3. Let a random variable X_n have the distribution $\chi^2(n)$. Then the sequence of random variables $(X_n - n)/\sqrt{(2n)}$ converges in distribution to $N(0, 1)$.

REMARK 4. Let a random variable X_n have the distribution $t(n)$. Then the sequence of the random variables X_n converges in distribution to $N(0, 1)$.

34. MATHEMATICAL STATISTICS

By TOMÁŠ CIPRA

References: [57], [64], [73], [74], [92], [100], [101], [106], [117], [123], [127], [137], [194], [198], [200], [201], [202], [205], [217], [218], [219], [225], [228], [230], [249], [258], [266], [269], [278], [294], [297], [298], [306], [330], [331], [345], [362], [368], [380], [382], [396], [407], [411], [416], [424], [427], [448], [453], [473], [484], [491], [499], [504], [505].

34.1. Basic Concepts

This chapter is devoted both to theoretical and practical aspects of mathematical statistics. Necessary foundations for elementary statistical analysis are given in § 34.1, § 34.2 and § 34.5. Furthermore, the chapter is concerned with problems of statistical estimation and hypothesis testing; § 34.6 and § 34.9 are concentrated on theoretical analysis of these problems while practical computational advice for estimation can be found in § 34.7 and § 34.8 and for testing in § 34.10 – § 34.12.

Mathematical Statistics. Mathematical statistics is concerned with analysis of numerical data obtained under circumstances affected by chance. These data are collected in order to draw certain conclusions concerning lots (populations) with a large number of items. The substantial feature of mathematical statistics, in contrast to descriptive statistics, consists in the fact that it regards the investigated data as realizations of random variables and that it aims at obtaining certain information concerning probability distributions of these random variables. Therefore procedures of mathematical statistics exploit substantially probability theory discussed in Chap. 33.

Statistical Model. The working premises on probability distributions that generate numerical data for statistical analysis are usually called the *statistical model.* The statistical analysis must include the description which model is suitable for the data investigated.

Random Sample. The *random sample* (or briefly *sample*) from a given probability distribution is a sequence of independent random variables $X_1, \ldots, X_n$ that have this distribution. It can be written as a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$.

The number n is called the *size of the random sample.* The random sample of size n corresponds to the situation when a random experiment is repeated n times under the same conditions and with mutually independent repetitions (e.g., the weighing of n randomly selected casts of the same type can be regarded as a random sample from the normal distribution, the measuring of time to failure for n bearings as a random sample from the Weibull distribution). The random sample can also be formed from a multivariate probability distribution (e.g. the simultaneous weighing of casts and measuring of their tensile strength) or mutually independent random samples from several probability distributions may be available (e.g. the weighing of two batches of casts produced by different technologies). The random sample defined by means of random variables is only a hypothetical concept that conforms to the probability formulation of the statistical model. For practical treatment, a statistician obtains only n realized values $x_1, \ldots, x_n$ which are called *observations* of the random sample $X_1, \ldots, X_n$. The family of all possible observations $x_1, \ldots, x_n$ of a given random sample is called the *sample space.*

Statistics and Estimators. The values constructed on the basis of a random sample in order to draw certain conclusions are called *statistics.* From the mathematical point of view, the statistic is a function $S(X_1, \ldots, X_n)$ of the random variables $X_1, \ldots, X_n$ that is itself a random variable and that can be constructed without knowledge of the probability distribution of these random variables. The statistics whose objective is an approximate determination of characteristics of a probability distribution are called *sample* (or *empirical*) *characteristics* since they are the sample counterparts of actual *theoretical characteristics* (see § 34.2). The statistics whose objective is an approximate determination of parameters of a statistical model are called *estimators.*

34.2. Sample Characteristics

1. *Sample mean*:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i . \tag{1}$$

2. *Sample variance*:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} X_i^2 - \frac{1}{n} \Big(\sum_{i=1}^{n} X_i \Big)^2 \right] \qquad (n \geqq 2). \tag{2}$$

Theorem 1 (*Properties of Sample Mean and Sample Variance*). *Let* $X_1, \ldots, X_n$ *be a random sample from a distribution with the mean* μ *and variance* σ^2. *Then*

$$\mathrm{E}(\overline{X}) = \mu, \quad \mathrm{var}(\overline{X}) = \sigma^2/n , \tag{3}$$

$$\mathrm{E}(S^2) = \sigma^2 . \tag{4}$$

If n increases, then $\overline{X}$ converges almost surely (§ 33.11) *to μ and S^2 to σ^2. If μ_4 exists, then*

$$\operatorname{var}(S^2) = \frac{\mu_4}{n} - \frac{n-3}{n(n-1)}\sigma^4 \qquad (n \geqq 4). \tag{5}$$

3. *Sample standard deviation*:

$$S = \sqrt{S^2}. \tag{6}$$

4. *Sample coefficient of variation*:

$$C = S/|\overline{X}|. \tag{7}$$

5. *Sample k-th moment*:

$$M'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \qquad (k = 1, 2, \dots). \tag{8}$$

6. *Sample k-th central moment*:

$$M_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k \qquad (k = 1, 2, \dots). \tag{9}$$

REMARK 1. For the sample second central moment we have $\mathrm{E}(M_2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$. Therefore one mostly uses the statistic S^2 instead of M_2. For this statistic $\mathrm{E}(S^2) = \sigma^2$ holds and the term "sample variance" usually refers to it.

7. *Sample coefficient of skewness and kurtosis*:

$$G_1 = \frac{M_3}{M_2^{3/2}}, \qquad G_2 = \frac{M_4}{M_2^2} - 3. \tag{10}$$

8. *Sample correlation coefficient* (for a two-variate random sample $(X_1, Y_1)', \dots, (X_n, Y_n)'$):

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\left\{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]\left[\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right]\right\}^{1/2}} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{\left\{\left[\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right]\left[\sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2\right]\right\}^{1/2}}. \tag{11}$$

A *sample covariance* equals the numerators of the expressions in (11) divided by n.

9. *Sample correlation matrix* (for a k-variate random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$) is a $k \times k$ matrix $\boldsymbol{R}_{\boldsymbol{X}}$ with elements r_{ij}, where r_{ij} is the sample correlation coefficient of the i-th and j-th component of the considered random sample. A *sample covariance matrix* $\boldsymbol{S}_{\boldsymbol{X}}$ is defined analogously by means of the sample covariances.

34.3. Random Sample from Normal Distribution

Theorem 1 (*Random Sample from Normal Distribution*). *Let $X_1, \ldots, X_n$ be a random sample from the normal distribution $N(\mu, \sigma^2)$. Then the sample mean $\overline{X}$ has the distribution $N(\mu, \sigma^2/n)$ and the random variable $(n-1)S^2/\sigma^2$ has the distribution $\chi^2(n-1)$, the random variables $\overline{X}$ and S^2 being independent. The random variable*

$$T = \frac{\overline{X} - \mu}{S} \sqrt{n} \tag{1}$$

has the distribution $t(n-1)$ (§ 33.7).

REMARK 1. Under the assumptions of Theorem 1,

$$\operatorname{var}(S^2) = \frac{2}{n-1} \sigma^4 \qquad (n \geqq 2). \tag{2}$$

Theorem 2 (*Two Independent Random Samples from Normal Distributions*). *Let $X_1, \ldots, X_{n_1}$ be a random sample from the distribution $N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_{n_2}$ a random sample from the distribution $N(\mu_2, \sigma_2^2)$, the samples being independent. Let $\overline{X}$ and S_1^2 denote the sample mean and sample variance of the random sample $X_1, \ldots, X_{n_1}$ and, analogously, $\overline{Y}$ and S_2^2 these quantities of the random sample $Y_1, \ldots, Y_{n_2}$. Then the random variable*

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \tag{3}$$

has the distribution $F(n_1 - 1, n_2 - 1)$ (§ 33.7). *If, moreover, $\sigma_1^2 = \sigma_2^2 = \sigma^2$ then the random variable*

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\left[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\right]^{1/2}} \left[\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}\right]^{1/2} \tag{4}$$

has the distribution $t(n_1 + n_2 - 2)$ (§ 33.7).

Theorem 3 (*Random Sample from Two-Variate Normal Distribution*). *Let* $(X_1, Y_1)', \ldots, (X_n, Y_n)'$ *be a random sample from the two-variate normal distribution with the probability density* (33.8.2). *Let* r *be the sample correlation coefficient* (see (34.2.11)). *If* $\varrho = 0$ (see (33.8.3)), *then the random variable*

$$T = \frac{r}{[1 - r^2]^{1/2}} [n - 2]^{1/2} \qquad (n \geqq 3) \tag{5}$$

has the distribution $t(n - 2)$ (§ 33.7).

REMARK 2. For a general ϱ one frequently uses the so-called *Z-transformation* of the form

$$Z = \frac{1}{2} \ln \frac{1 + r}{1 - r}. \tag{6}$$

If ϱ is not too close to the number 1 or -1, then for $n \geqq 10$ it is possible to approximate the probability distribution of the random variable Z by the distribution

$$N\left(\frac{1}{2} \ln \frac{1 + \varrho}{1 - \varrho} + \frac{\varrho}{2(n - 1)}, \frac{1}{n - 3}\right). \tag{7}$$

REMARK 3. Theorems 1 – 3 are usually applied in order to draw various conclusions in statistical models based on the normal distribution (see e.g. § 34.10). Derivation of properties of sample statistics for other probability distributions is mostly much more difficult. For example, if $X_1, \ldots, X_n$ is a random sample from the gamma distribution with parameters p, δ, then the sample mean $\overline{X}$ has the gamma distribution with parameters np, δ/n and, for larger n, its distribution can be approximated by the distribution $N(p, \delta^2/n)$.

34.4. Ordered Random Sample

Frequently it is necessary to analyze data ordered from the smallest to the largest item or to concentrate one's attention to the behaviour of the smallest or largest data items, etc. In such a case one can exploit the following theory.

Definition 1. Let $X_1, \ldots, X_n$ be a random sample. Let random variables $X_{(1)}, \ldots, X_{(n)}$ be obtained from $X_1, \ldots, X_n$ by rearranging from the smallest to the largest item so that

$$X_{(1)} \leqq X_{(2)} \leqq \cdots \leqq X_{(n)}. \tag{1}$$

Then $X_{(1)}, \ldots, X_{(n)}$ is called the *ordered random sample.* The i-th random variable $X_{(i)}$ is called the *i-th order statistic* ($i = 1, \ldots, n$).

Definition 2. Let $X_1, \ldots, X_n$ be a random sample. Then the statistic $\tilde{X}$ of the form

$$\tilde{X} = \begin{cases} X_{((n+1)/2)} & \text{for } n \text{ odd}, \\ \frac{1}{2}[X_{(n/2)} + X_{(n/2+1)}] & \text{for } n \text{ even} \end{cases} \tag{2}$$

is called the *sample median.* The statistic $R = X_{(n)} - X_{(1)}$ is called the *sample range.*

Theorem 1 (*Probability Distribution of Order Statistics*). *Let $X_1, \ldots, X_n$ be a random sample from a probability distribution with distribution function $F(x)$. Then the distribution function $F_{(i)}(x)$ of the i-th order statistic is given by*

$$F_{(i)}(x) = \sum_{r=i}^{n} \binom{n}{r} [F(x)]^r [1 - F(x)]^{n-r} \qquad (-\infty < x < \infty,\ i = 1, \ldots, n). \tag{3}$$

If the probability distribution considered is continuous with probability density $f(x)$, then the probability density $f_{(i)}(x)$ of the i-th order statistic is

$$f_{(i)}(x) = n \binom{n-1}{i-1} f(x)[F(x)]^{i-1}[1 - F(x)]^{n-i} \qquad (-\infty < x < \infty,\ i = 1, \ldots, n). \tag{4}$$

REMARK 1. In particular, for the statistic $X_{(1)} = \min(X_1, \ldots, X_n)$ one has

$$F_{(1)}(x) = 1 - [1 - F(x)]^n, \qquad f_{(1)}(x) = nf(x)[1 - F(x)]^{n-1} \tag{5}$$

and for the statistic $X_{(n)} = \max(X_1, \ldots, X_n)$

$$F_{(n)}(x) = [F(x)]^n, \qquad f_{(n)}(x) = nf(x)[F(x)]^{n-1}. \tag{6}$$

Example 1. Let n bulbs be connected in series in a circuit. The time to failure of the i-th bulb can be modelled by the Weibull distribution with parameters p, δ. This distribution has the probability density $f(x)$ given in § 33.7 and the distribution function

$$F(x) = \begin{cases} 1 - \exp[-(x/\delta)^p] & \text{for } x > 0, \\ 0 & \text{for } x \leqq 0. \end{cases}$$

The time to failure for the whole system is described by the statistic $X_{(1)}$ since the first failure of any bulb destroys the whole system. The probability density of this statistic is, according to (5)

$$f_{(1)}(x) = \begin{cases} nf(x)[1 - F(x)]^{n-1} = \dfrac{px^{p-1}}{(\delta/n^{1/p})^p} \exp\left[-\left(\dfrac{x}{\delta/n^{1/p}}\right)^p\right] & \text{for } x > 0, \\ 0 & \text{for } x \leqq 0. \end{cases}$$

Hence the time to failure of the whole system has the Weibull distribution as well, but with parameters p, $\delta/n^{1/p}$. In particular, we obtain for $p = 1$ that the statistic $X_{(1)}$ in a random sample from the exponential distribution with parameter δ has the exponential distribution with parameter δ/n.

34.5. Elementary Statistical Treatment

Frequencies of Observations. Some values among the observations that are obtained as realization of a random sample of size n can repeat several times. The number n_i of occurrences of a value x_i is called the *frequency of the observation* x_i and the quotient n_i/n is called the *relative frequency of the observation* x_i. The sum of the frequencies of all observations is equal to n and the sum of the relative frequencies to one. The *cumulative frequency of the observation* x_i is the sum of frequencies of all the observations that do not exceed the value x_i and analogously for the *cumulative relative frequency of the observation* x_i.

Frequency and Correlation Tables. The *frequency table* contains the values x_i (usually ordered from the smallest to the largest) that appeared at least once among the observations in the first column and their frequencies n_i in the second one. The *correlation table* used for two-variate random samples contains the values x_i that appeared at least once among the observations of the first component of the random sample on the left-hand side and the values y_j that appeared at least once among the observations of the second component on the top edge. At the points of intersections of the individual rows and columns frequencies n_{ij} corresponding to the pairs (x_i, y_j) are recorded.

Frequency and correlation tables can be used for calculations of various sample characteristics. One frequently calculates the sample mean, sample standard deviation and sample correlation coefficient according to the formulae

$$\overline{x} = \frac{1}{n} \sum x_i n_i , \tag{1}$$

$$s = \left\{ \frac{1}{n-1} \sum (x_i - \overline{x})^2 n_i \right\}^{1/2} =$$

$$= \left\{ \frac{1}{n-1} \left[\sum x_i^2 n_i - \frac{1}{n} \left(\sum x_i n_i \right)^2 \right] \right\}^{1/2} , \tag{2}$$

$$r = \frac{\sum\sum x_i y_j n_{ij} - \frac{1}{n} (\sum x_i n_i)(\sum y_j n_j)}{\left\{ [\sum x_i^2 n_i - \frac{1}{n} (\sum x_i n_i)^2][\sum y_j^2 n_j - \frac{1}{n} (\sum y_j n_j)^2] \right\}^{1/2}} , \tag{3}$$

respectively (the summation runs always over all possible values x_i or y_j). It is also possible to find any *sample quantile* in such a simple way that the observations that do not exceed this value form the prescribed part of the sample, e.g. 25 % for the sample lower quartile, 50 % for the sample median, etc.

Example 1. Twenty laboratory tests have given the following data representing the percentage of carbon content in coal:

85, 82, 81, 83, 74, 84, 83, 84, 83, 84,
87, 79, 87, 81, 83, 90, 77, 84, 85, 85.

The corresponding frequency table is given as Tab. 34.1. The sample mean is $\bar{x} = 83{\cdot}050$ and the sample standard deviation $s = 3{\cdot}546$. The sample lower quartile is 81·5 (the arithmetic mean of the values 81 and 82), the sample median 83·5 (the arithmetic mean of the values 83 and 84) and the sample upper quartile 85.

TABLE 34.1

x_i	Frequency n_i	Relative Frequency n_i/n	Cumulative Frequency	Cumulative Relative Frequency
74	1	0·05	1	0·05
77	1	0·05	2	0·10
79	1	0·05	3	0·15
81	2	0·10	5	0·25
82	1	0·05	6	0·30
83	4	0·20	10	0·50
84	4	0·20	14	0·70
85	3	0·15	17	0·85
87	2	0·10	19	0·95
90	1	0·05	20	1·00
$\sum$	20	1·00		

Grouping. If a random sample of size n consists of too many numerically different observed values, then the range of all observations is usually divided into k *class intervals* (or *cells*). The recommended value of k ranges from 5 to 20, one sometimes takes $k \approx \sqrt{n}$ or $k \approx 1 + 3{\cdot}3 \ln n$ (the so-called *Sturges rule*). The observations lying in the same class interval are said to form a *class* and they are usually approximated by the midpoint of this interval. The number of values in a class is called the *class frequency* and, after dividing by n, the *relative class frequency*. If an observation coincides with the common endpoint of two class intervals, then it is either assigned systematically to the lower interval or the class frequency of both the intervals is increased by $\frac{1}{2}$. The grouped observations are presented again in the form of a frequency table with the individual class intervals and their midpoints. The sample characteristics are calculated according to the formulas (1) – (3) (and others) but now the values x_i or y_j are the class midpoints and n_i or n_j are the corresponding class frequencies.

The calculation of $\overline{x}$ and s can be simplified if all the class intervals have the same width d. In this case, one can replace the original class midpoints by the values $\ldots, -2, -1, 0, 1, 2, \ldots$ of an auxiliary variable u, where the value 0 corresponds to a suitably chosen class interval (e.g. to the middle one or to that with the largest class frequency), the value 1 corresponds to the right-hand adjacent interval, etc. Then

$$\overline{x} = x_0 + d\overline{u}, \quad s = ds_u, \tag{4}$$

where $\overline{u}$ and s_u are the sample mean and the sample standard deviation of the values u, respectively, and x_0 is the midpoint of the interval with $u = 0$.

Example 2. Tab. 34.2 is the frequency table that describes the results of weighing a produced chemical (in grams) in 300 tests. The data are grouped into 12 class intervals of the same length 100 grams. If we set $x_0 = 1500$, then the auxiliary variable u attains the values given in Tab. 34.2. For these values of u it is easy to calculate

$$\overline{u} = 0{\cdot}027, \quad s_u = 2{\cdot}041.$$

Finally, one obtains

$$\overline{x} = 1500 + 100 \,.\, 0{\cdot}027 = 1502{\cdot}7, \quad s = 100 \,.\, 2{\cdot}041 = 204{\cdot}1$$

according to (4).

TABLE 34.2

i	Class Interval	Class Midpoint x_i	Auxiliary Variable u	Class Frequency n_i	Relative Class Freq. n_i/n	Cumulative Relative Class Freq.
1	950 – 1050	1000	−5	4	0·013	0·013
2	1050 – 1150	1100	−4	9	0·030	0·043
3	1150 – 1250	1200	−3	19	0·063	0·106
4	1250 – 1350	1300	−2	36	0·120	0·226
5	1350 – 1450	1400	−1	51	0·170	0·396
6	1450 – 1550	1500	0	58	0·193	0·589
7	1550 – 1650	1600	1	53	0·177	0·766
8	1650 – 1750	1700	2	37	0·123	0·889
9	1750 – 1850	1800	3	20	0·067	0·956
10	1850 – 1950	1900	4	9	0·030	0·986
11	1950 – 2050	2000	5	3	0·010	0·996
12	2050 – 2150	2100	6	1	0·003	0·999
Σ				300	0·999	

REMARK 1. Errors of sample moments due to grouping data into class intervals of the same length d can be reduced by means of the so-called *Sheppard correction*. The corrected standard deviation s^* has the form

$$s^* = [s^2 - d^2/12]^{1/2} \tag{5}$$

Histogram and Empirical Distribution Function. The histogram and empirical distribution function are constructed in order to obtain an idea on the form of the probability density and distribution function of the probability distribution from which the given random sample has arisen. If the observations are grouped into class intervals, then the *histogram* is a system of rectangles whose horizontal edges located on the horizontal axis coincide with the individual class intervals and whose areas correspond to the relative class frequencies. The *empirical distribution function* denoted often by $F_n(x)$ can be constructed e.g. in such a way that one assigns the cumulative relative class frequencies to the upper endpoints of the corresponding class intervals. The histogram and empirical distribution function for the data from Example 2 are shown in Fig. 34.1 and 34.2.

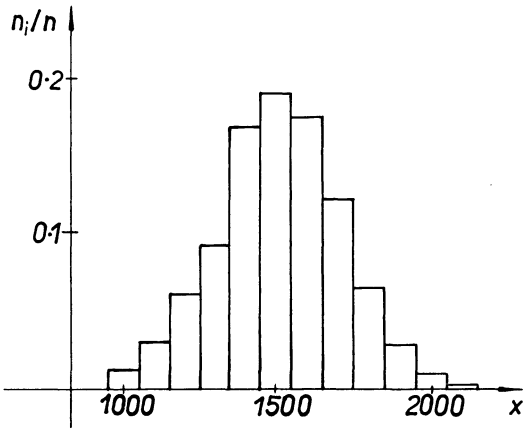


Fig. 34.1. Histogram of the data from Example 34.5.2.

REMARK 2. In general, the empirical distribution function of a random sample $X_1, \ldots, X_n$ is defined by the formula

$$F_n(x) = \begin{cases} 0 & \text{for } x < X_{(1)}\,, \\ i/n & \text{for } X_{(i)} \leqq x < X_{(i+1)} \quad (i = 1, \ldots, n-1)\,, \\ 1 & \text{for } x \geqq X_{(n)}\,, \end{cases} \tag{6}$$

where $X_{(1)}, \ldots, X_{(n)}$ is the corresponding ordered random sample (see § 34.4). The empirical distribution function defined in this way has a random character.

However, if n increases, then $F_n(x)$ converges with probability 1 to the distribution function of the probability distribution from which the observed random sample has been obtained, the convergence being even uniform for real x (the so-called *Glivenko theorem*).

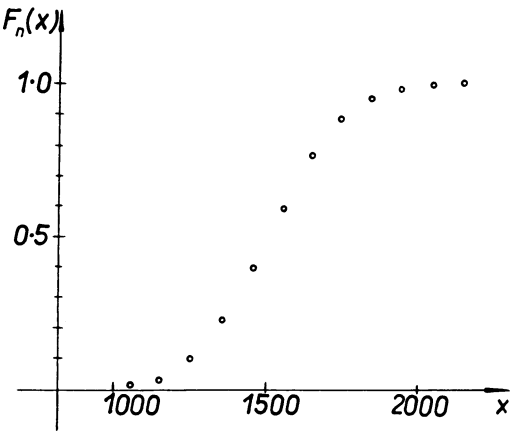


Fig. 34.2. Empirical distribution function of the data from Example 34.5.2.

Probability Paper. The probability paper is a graphical instrument that facilitates the decision on the type of the probability distribution from which the observed random sample has been obtained. In order to verify the normality one uses the *normal probability paper.* It is a graph paper whose horizontal axis has the usual linear scale. However, the vertical axis is scaled nonlinearly so that the actual distance u_P is denoted by P (or $100P$ in per cent), where u_P is the P-quantile of $N(0, 1)$. We plot the empirical distribution function $F_n(x)$ of a random sample on this probability paper. If the random sample has been obtained from the distribution $N(\mu, \sigma^2)$, then the graph on the paper can be fitted approximately with a straight line as in Fig. 34.3 for the data from Example 2. Moreover, this line achieves the value $P = 0{\cdot}5$ (or $100P = 50$ in per cent) at the point with abscissa $x \approx \mu$ so that it can be taken for a rough estimate of the corresponding mean. There exist also other types of probability papers, e.g. the exponential or Weibull probability paper.

34.6. Estimation Theory

The task of estimation theory can be formulated as follows: Let $X_1, \ldots, X_n$ be a random sample from a probability distribution which depends on an unknown parameter ϑ. The values to be considered for the parameter ϑ can be taken only

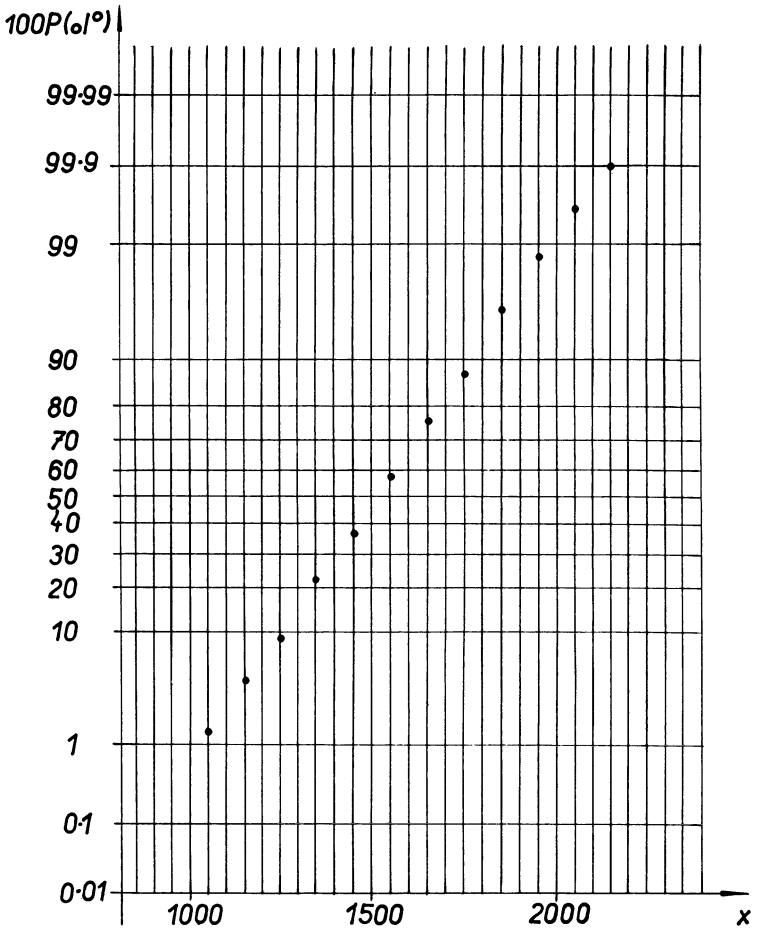


Fig. 34.3. Empirical distribution function of the data from Example 34.5.2 plotted on the normal probability paper.

from a known *parameter space* Ω. In the simplest case, Ω is a subset of the real line. The estimation of the parameter ϑ consists in the construction of a statistic $T(X_1, \ldots, X_n)$ whose distribution is concentrated (maximally, in a specified sense) about such a value of the parameter $\vartheta \in \Omega$ for which the random sample was observed.

REMARK 1. The above formulation concerns a *parametric estimation.* However, it is also possible to use a *nonparametric estimation.* An example is the sample median (see Definition 34.4.2) that estimates the median without specifying the type of probability distribution parametrically.

REMARK 2. Instead of a scalar parameter ϑ one can, in the same manner, estimate a value $\tau(\vartheta)$, where $\vartheta = (\vartheta_1, \ldots, \vartheta_m)'$ is an m-component (vector) parameter (i.e., Ω is a subset of the m-dimensional Euclidean space E_m) and τ is a known function defined on Ω. In particular, the functions $\tau_i(\vartheta) = \vartheta_i$ make possible the estimation of the individual components ϑ_i of the vector parameter ϑ $(i = 1, \ldots, m)$.

Point and Interval Estimation. The statistic $T(X_1, \ldots, X_n)$ in the above formulation of the estimation problem represents the so-called *point estimator* of the parameter ϑ since, in practice, after substituting the observations $x_1, \ldots, x_n$, we estimate the parameter ϑ by a single number $T(x_1, \ldots, x_n)$ called the *point estimate*. For simplicity, one sometimes denotes the point estimator of ϑ by $\hat{\vartheta}$.

However, we can also construct the *interval estimator* where, using two statistics $T_l(X_1, \ldots, X_n)$ and $T_u(X_1, \ldots, X_n)$, we specify an interval that covers the actual value of the parameter ϑ with a prescribed (sufficiently large) probability, e.g. $P(T_l(X_1, \ldots, X_n) < \vartheta < T_u(X_1, \ldots, X_n)) = 0{\cdot}95$. In an individual problem, after substituting observations $x_1, \ldots, x_n$, we obtain a certain interval $(T_l(x_1, \ldots, x_n), T_u(x_1, \ldots, x_n))$ called an *interval estimate*.

Example 1. Let the time to failure of a device have the exponential distribution with distribution function $F(x)$ given in (33.3.12). Then it is often important to estimate the probability that the device will work without failure at least during a period x_0:

$$P(X > x_0) = 1 - F(x_0) = \exp(-x_0/\delta) \tag{1}$$

($x_0 > 0$ being a given constant). If $X_1, \ldots, X_n$ is the corresponding random sample from the exponential distribution with parameter δ then, due to certain properties of optimality, the statistic

$$T(X_1, \ldots, X_n) = \begin{cases} \left(1 - x_0 / \sum_{i=1}^{n} X_i\right)^{n-1} & \text{for } \sum_{i=1}^{n} X_i > x_0\,, \\ 0 & \text{for } \sum_{i=1}^{n} X_i \leqq x_0 \end{cases} \tag{2}$$

can be taken for the point estimator of the value $\tau(\delta) = \exp(-x_0/\delta)$, while the recommended estimator covering the actual value $\tau(\delta)$ with probability $1 - \alpha$ has the form

$$\left(\exp\left(-\frac{x_0 \chi^2_{1-\alpha/2}(2n)}{2\sum_{i=1}^{n} X_i}\right), \exp\left(-\frac{x_0 \chi^2_{\alpha/2}(2n)}{2\sum_{i=1}^{n} X_i}\right)\right), \tag{3}$$

where $\chi^2_P(2n)$ is the P-quantile of the distribution $\chi^2(2n)$ (see Remark 33.7.4).

Properties of Estimators. In applications, one prefers estimators with certain properties. Let $X_1, \ldots, X_n$ be a random sample from a probability distribution with an unknown parameter $\vartheta \in \Omega$.

Definition 1. An estimator $T(X_1, \ldots, X_n)$ of the parameter ϑ is called *unbiased* if the relation

$$\mathrm{E}[T(X_1, \ldots, X_n)] = \vartheta \tag{4}$$

holds for all $\vartheta \in \Omega$. The value

$$b(\vartheta) = \mathrm{E}[T(X_1, \ldots, X_n)] - \vartheta \tag{5}$$

is called the *bias of the estimator* $T(X_1, \ldots, X_n)$.

Definition 2. An estimator $T(X_1, \ldots, X_n)$ of the parameter ϑ is called *best unbiased* if it is unbiased and if, for any unbiased estimator $S(X_1, \ldots, X_n)$ of the parameter ϑ,

$$\mathrm{var}[T(X_1, \ldots, X_n)] \leqq \mathrm{var}[S(X_1, \ldots, X_n)] \tag{6}$$

holds for all $\vartheta \in \Omega$.

Sometimes an estimator $T(X_1, \ldots, X_n)$ is acceptable only for large values of n. The estimator properties that become apparent only when $n \to \infty$ are called *asymptotic.*

Definition 3. An estimator $T(X_1, \ldots, X_n)$ of the parameter ϑ is called *asymptotically unbiased* if, for all $\vartheta \in \Omega$,

$$\lim_{n\to\infty} \mathrm{E}[T(X_1, \ldots, X_n)] = \vartheta\,. \tag{7}$$

Definition 4. An estimator $T(X_1, \ldots, X_n)$ of the parameter ϑ is called *consistent* if, for every $\varepsilon > 0$ and all $\vartheta \in \Omega$,

$$\lim_{n\to\infty} P(|T(X_1, \ldots, X_n) - \vartheta| \geqq \varepsilon) = 0\,. \tag{8}$$

REMARK 3. According to Definition 33.11.1 the consistency of an estimator means the convergence in probability to the corresponding value of the parameter.

Theorem 1. *Let* $T(X_1, \ldots, X_n)$ *be an asymptotically unbiased estimator of the parameter* ϑ *such that*

$$\lim_{n\to\infty} \mathrm{var}[T(X_1, \ldots, X_n)] = 0\,. \tag{9}$$

Then $T(X_1, \ldots, X_n)$ *is a consistent estimator of* ϑ.

Example 2. Let $X_1, \ldots, X_n$ be a random sample from a probability distribution with the mean μ and variance σ^2. According to (34.2.3),

$$\mathrm{E}(\overline{X}) = \mu\,, \quad \lim_{n\to\infty} \mathrm{var}(\overline{X}) = \lim_{n\to\infty} \sigma^2/n = 0$$

so that the sample mean $\overline{X}$ is an unbiased and consistent estimator of the parameter μ according to Theorem 1. Furthermore, the sample variance S^2 is the unbiased estimator of the parameter σ^2 according to (34.2.4) (Theorem 34.2.1 even guarantees that if n increases, then $\overline{X}$ converges to μ and S^2 converges to σ^2 almost surely).

Efficiency of Estimator. The quality of an estimator $T(X_1, \ldots, X_n)$ of the parameter ϑ is frequently appreciated according to its *mean square error* (or briefly *mean error*)

$$\mathrm{E}[T(X_1, \ldots, X_n) - \vartheta]^2 \tag{10}$$

that expresses the degree of concentration of the estimator about the estimated parameter. The mean square error of an unbiased estimator coincides with the variance of this estimator. The natural aim of estimation is the construction of such estimators whose mean square error is as small as possible. Under certain conditions on a given random sample $X_1, \ldots, X_n$ with parameter ϑ one can find a lower bound of mean square errors of all possible estimators $T(X_1, \ldots, X_n)$ of the parameter ϑ. In particular, one uses mainly the lower bound of variances of the unbiased estimators (the so-called *Cramer-Rao lower bound*). An unbiased estimator whose variance achieves this bound is called the *efficient* (or *minimum variance*) *estimator*. If this bound is achieved only in limit for increasing size of the sample, then the corresponding estimator is called *asymptotically efficient*.

34.7. Point Estimators

In this Section two methods for systematic construction of point estimators are presented. Their description is given in terms of continuous probability distributions even though it is valid for discrete distributions as well.

Method of Maximum Likelihood. Let $X_1, \ldots, X_n$ be a random sample from a probability distribution with density $f(x, \vartheta)$, where ϑ is an m-component parameter from an m-dimensional open interval Ω.

Definition 1. A real function $L(\vartheta)$ of the parameter ϑ defined for the observed values $x_1, \ldots, x_n$ of the random sample $X_1, \ldots, X_n$ by

$$L(\vartheta) = \prod_{i=1}^{n} f(x_i, \vartheta) \tag{1}$$

is called the *likelihood function*. A value $\hat{\vartheta}$ of the parameter ϑ fulfilling the condition

$$L(\hat{\vartheta}) \geqq L(\vartheta) \qquad \text{for all } \vartheta \in \Omega \tag{2}$$

is called the *maximum likelihood estimator* of the parameter ϑ.

REMARK 1. By (2), the maximum likelihood estimator is such a value of the parameter ϑ for which the likelihood function (1) attains its maximum when the values $x_1, \ldots, x_n$ have been observed. However, investigating its theoretical properties, one regards the maximum likelihood estimator as a random vector $\hat{\vartheta}(X_1, \ldots, X_n)$.

REMARK 2. In some cases (usually if the exponential function occurs in the density $f(x, \vartheta)$) it is suitable to maximize $\ln L(\vartheta)$ (the so-called *log likelihood function*) instead of $L(\vartheta)$. If one sets $\ln 0 = -\infty$, then the logarithm does not influence the determination of $\hat{\vartheta}$.

There is a connection between the value $\hat{\vartheta}$ and the solution of the *likelihood equations*

$$\frac{\partial \ln L(\vartheta)}{\partial \vartheta_j} = 0 \quad (j = 1, \ldots, m) \tag{3}$$

provided that the corresponding partial derivatives exist. The method of maximum likelihood is usually applied just by solving the equations (3).

Example 1. Let $X_1, \ldots, X_n$ be a random sample from the Weibull distribution with parameters $p > 0$ and $\delta > 0$. Then we have

$$\ln L(p, \delta) = n \ln p - np \ln \delta + (p-1) \sum_{i=1}^{n} \ln x_i - \sum_{i=1}^{n} \left(\frac{x_i}{\delta}\right)^p$$

for the observed values $x_1, \ldots, x_n$ so that the likelihood equations (3) have the form

$$\frac{n}{p} - n \ln \delta + \sum_{i=1}^{n} \ln x_i - \frac{1}{\delta^p} \sum_{i=1}^{n} x_i^p (\ln x_i - \ln \delta) = 0,$$

$$-\frac{np}{\delta} + \frac{p}{\delta^{p+1}} \sum_{i=1}^{n} x_i^p = 0.$$

After some rearrangement, one obtains for the estimator $\hat{p}$ the equation

$$\hat{p} = \left(\sum_{i=1}^{n} x_i^{\hat{p}} \ln x_i \Big/ \sum_{i=1}^{n} x_i^{\hat{p}} - \frac{1}{n} \sum_{i=1}^{n} \ln x_i \right)^{-1}$$

that has to be solved numerically. Afterwards the estimator $\hat{\delta}$ can be calculated as

$$\hat{\delta} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^{\hat{p}} \right)^{1/\hat{p}}.$$

Method of Moments. Although this method need not provide estimators of optimal properties, its advantage lies in the fact that it is usually simple from the numerical point of view. The estimators obtained by this method are frequently used as initial values for more complicated estimation procedures.

Let $X_1, \ldots, X_n$ be a random sample from a probability distribution depending on an m-component parameter $\vartheta \in \Omega$. Let the moments $\mu_k'(\vartheta)$ $(k = 1, \ldots, m)$

of this distribution exist for all $\vartheta \in \Omega$ (see Definition 33.4.2), the symbol $\mu'_k(\vartheta)$ expressing explicitly the dependence of the moment μ'_k on the parameter ϑ. Finally, let M'_k denote the sample k-th moment corresponding to the random sample $X_1, \ldots, X_n$ (see (34.2.8)). Then the estimator $\hat{\vartheta}$ of ϑ is constructed by the method of moments as the solution of the equations

$$\mu'_k(\vartheta) = M'_k \quad (k = 1, \ldots, m). \tag{4}$$

REMARK 3. Sometimes one must enlarge the number of equations (4) in order to achieve uniqueness of the solution. If the solution of (4) is unique and its components are continuous functions of the sample moments M'_k, then the estimator obtained by the method of moments is consistent.

Example 2. Let $X_1, \ldots, X_n$ be a random sample from the logarithmic normal distribution with parameters $-\infty < \mu < \infty$ and $\sigma^2 > 0$. Then equations (4) have the form

$$\exp(\mu + \sigma^2/2) = \overline{X},$$
$$\exp(2\mu + 2\sigma^2) = M'_2.$$

Solving these equations, we obtain the estimators

$$\hat{\mu} = \ln(\overline{X}^2 / \sqrt{M'_2}), \quad \hat{\sigma}^2 = \ln(M'_2/\overline{X}^2).$$

Point Estimators for Some Important Distributions. In what follows, we give a survey of point estimators constructed by means of a random sample $X_1, \ldots, X_n$ for the parameters of some important probability distributions.

1. *Binomial distribution* (N a positive integer, $0 < p < 1$):

$$\hat{p} = \frac{1}{N}\overline{X} \tag{5}$$

(an efficient estimator of p if N is known).

2. *Poisson distribution* ($\lambda > 0$):

$$\hat{\lambda} = \overline{X} \tag{6}$$

(an efficient estimator).

3. *Uniform distribution* ($a < b$):

$$\begin{aligned} \hat{a} &= \frac{n}{n-1}X_{(1)} - \frac{1}{n-1}X_{(n)}, \\ \hat{b} &= \frac{n}{n-1}X_{(n)} - \frac{1}{n-1}X_{(1)} \end{aligned} \tag{7}$$

(the best unbiased estimator).

4. *Normal distribution* $(-\infty < \mu < \infty, \sigma^2 > 0)$:

$$\hat{\mu} = \overline{X}, \quad \hat{\sigma}^2 = S^2 \tag{8}$$

(the best unbiased estimator);

$$\hat{\mu} = \overline{X}, \quad \hat{\sigma}^2 = \frac{n-1}{n} S^2 \tag{9}$$

(the maximum likelihood estimator).

5. *Logarithmic normal distribution* $(-\infty < \mu < \infty, \sigma^2 > 0)$:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \ln X_i, \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\ln X_i - \hat{\mu})^2 \tag{10}$$

(the best unbiased estimator).

6. *Exponential distribution* $(\delta > 0)$:

$$\hat{\delta} = \overline{X} \tag{11}$$

(the best unbiased estimator).

7. *Weibull distribution* $(p > 0, \delta > 0)$:
see the maximum likelihood method in Example 1.

8. *Gamma distribution* $(p > 0, \delta > 0)$:

$$\hat{\delta} = \frac{1}{p} \overline{X} \tag{12}$$

(an efficient estimator of δ if p is known).

34.8. Interval Estimators

Let $X_1, \ldots, X_n$ be a random sample from a probability distribution which depends on an m-component parameter $\vartheta \in \Omega$ and let $\tau(\vartheta)$ be a known real function on Ω.

Definition 1. Let $0 < \alpha < 1$ be a given number. Let $T_l(X_1, \ldots, X_n)$ and $T_u(X_1, \ldots, X_n)$ be such statistics that

$$P(T_l(X_1, \ldots, X_n) < \tau(\vartheta) < T_u(X_1, \ldots, X_n)) = 1 - \alpha \tag{1}$$

holds for all $\vartheta \in \Omega$. Then $(T_l(X_1, \ldots, X_n), T_u(X_1, \ldots, X_n))$ is called the *two-sided confidence interval* for $\tau(\vartheta)$ with the *confidence level* $1 - \alpha$ while $T_l(X_1, \ldots, X_n)$ and $T_u(X_1, \ldots, X_n)$ are called the *lower* and *upper confidence limits* for $\tau(\vartheta)$. If

$$P(\tau(\vartheta) < T_u(X_1, \ldots, X_n)) = 1-\alpha \quad \text{or} \quad P(\tau(\vartheta) > T_l(X_1, \ldots, X_n)) = 1-\alpha, \tag{2}$$

then $(-\infty, T_u(X_1, \ldots, X_n))$ and $(T_l(X_1, \ldots, X_n), \infty)$ are called the *one-sided confidence intervals* with the confidence level $1 - \alpha$.

REMARK 1. It is usual to give the confidence levels in per cent (e.g., a 95% confidence interval covers the actual value $\tau(\vartheta)$ with probability 0·95, i.e. $\alpha = 0{\cdot}05$). The most frequent confidence intervals in statistical practice are 95% and 99% ones. The one-sided confidence intervals are used in such situations when we are interested in the lowest or highest parameter value that can be expected on the given confidence level. Multivariate generalization of confidence intervals provides the so-called *confidence regions.*

Interval Estimators for Some Important Distributions. There exists a general theory of interval estimator construction that is related to the testing of statistical hypotheses (see e.g. [297]). We shall give a survey of two-sided confidence intervals with confidence level $1 - \alpha$ constructed by means of a random sample $X_1, \ldots, X_n$ for parameters of some important probability distributions. Let us put

$$T = X_1 + \cdots + X_n .$$

1. *Alternative distribution* $(0 < p < 1)$:

$$\frac{T}{T + (n - T + 1)F_{1-\alpha/2}(2(n - T + 1),\, 2T)} < p < \\ < \frac{(T+1)F_{1-\alpha/2}(2(T+1),\, 2(n-T))}{n - T + (T+1)F_{1-\alpha/2}(2(T+1),\, 2(n-T))} \tag{3}$$

(the lower confidence limit is equal to 0 for $T = 0$ and the upper confidence limit is equal to 1 for $T = n$). An approximate confidence interval for large n is

$$\overline{X} - u_{1-\alpha/2}[\overline{X}(1 - \overline{X})/n]^{1/2} < p < \overline{X} + u_{1-\alpha/2}[\overline{X}(1 - \overline{X})/n]^{1/2} . \tag{4}$$

2. *Binomial distribution* (N a positive integer, $0 < p < 1$):
If N is known, approximate lower and upper confidence limits for p can be obtained for large n as the roots of the quadratic equation

$$(nN + u_{1-\alpha/2}^2)x^2 - (2T + u_{1-\alpha/2}^2)x + T^2/(nN) = 0 . \tag{5}$$

3. *Poisson distribution* $(\lambda > 0)$:

$$\frac{1}{2n}\chi_{\alpha/2}^2(2T) < \lambda < \frac{1}{2n}\chi_{1-\alpha/2}^2(2T + 2) \tag{6}$$

(the lower confidence limit is equal to 0 for $T = 0$). An approximate confidence interval for large n is

$$\frac{1}{n}(T - u_{1-\alpha/2}\sqrt{T}) < \lambda < \frac{1}{n}(T + 1 + u_{1-\alpha/2}\sqrt{(T+1)}) . \tag{7}$$

4. *Normal distribution* $(-\infty < \mu < \infty,\ \sigma^2 > 0)$:

$$\overline{X} - t_{1-\alpha/2}(n-1)S/\sqrt{n} < \mu < \overline{X} + t_{1-\alpha/2}(n-1)S/\sqrt{n}\ , \tag{8}$$

$$(n-1)S^2/\chi^2_{1-\alpha/2}(n-1) < \sigma^2 < (n-1)S^2/\chi^2_{\alpha/2}(n-1)\,, \tag{9}$$

$$S[(n-1)/\chi^2_{1-\alpha/2}(n-1)]^{1/2} < \sigma < S[(n-1)/\chi^2_{\alpha/2}(n-1)]^{1/2}\,. \tag{10}$$

5. *Exponential distribution* $(\delta > 0)$:

$$2T/\chi^2_{1-\alpha/2}(2n) < \delta < 2T/\chi^2_{\alpha/2}(2n)\,. \tag{11}$$

6. *Two-variate normal distribution* $(-\infty < \mu_1 < \infty,\ -\infty < \mu_2 < \infty,\ \sigma_1^2 > 0,\ \sigma_2^2 > 0,\ -1 < \varrho < 1)$:
Approximate lower and upper confidence limits for ϱ can be obtained for large n as the roots of the nonlinear equation

$$(n-3)^{1/2}\left(\frac{1}{2}\ln\frac{1+r}{1-r} - \frac{1}{2}\ln\frac{1+x}{1-x} - \frac{x}{2(n-1)}\right) = \pm u_{1-\alpha/2}\,, \tag{12}$$

where r is the sample correlation coefficient (34.2.11).

REMARK 2. Sometimes one uses the following approximate confidence intervals with confidence level $1-\alpha$ for the basic characteristics $\mathrm{E}(X)$, $\mathrm{var}(X)$, $[\mathrm{var}(X)]^{1/2}$, $\varrho(X, Y)$ corresponding to random samples of large size n:

$$\overline{X} - u_{1-\alpha/2}S/\sqrt{n} < \mathrm{E}(X) < \overline{X} + u_{1-\alpha/2}S/\sqrt{n}\ , \tag{13}$$

$$S^2 - u_{1-\alpha/2}S^2\sqrt{(2/n)} < \mathrm{var}(X) < S^2 + u_{1-\alpha/2}S^2\sqrt{(2/n)}\,, \tag{14}$$

$$S - u_{1-\alpha/2}S/\sqrt{(2n)} < [\mathrm{var}(X)]^{1/2} < S + u_{1-\alpha/2}S/\sqrt{(2n)}\,, \tag{15}$$

$$r - u_{1-\alpha/2}(1-r^2)/\sqrt{n} < \varrho(X,Y) < r + u_{1-\alpha/2}(1-r^2)/\sqrt{n}\ . \tag{16}$$

Example 1. In Example 34.5.1, we have $n = 20$, $\overline{x} = 83{\cdot}050$, $s = 3{\cdot}546$. If we assume that it is a random sample from the normal distribution $N(\mu, \sigma^2)$, then the 95 % confidence intervals for μ and σ are

$$81{\cdot}390 < \mu < 84{\cdot}710\,, \quad 2{\cdot}697 < \sigma < 5{\cdot}179$$

according to (8) and (10). For example, the lower confidence limit for μ has been calculated by the substitution of the observed values $\overline{x}$ and s for $\overline{X}$ and S:

$$\overline{x} - t_{1-\alpha/2}(n-1)s/\sqrt{n} = 83{\cdot}050 - 2{\cdot}0930\,.\,3{\cdot}546/\sqrt{20} = 81{\cdot}390$$

(we have used $t_{0{\cdot}975}(19) = 2{\cdot}0930$). One can compare these intervals with the approximate 95 % confidence intervals (13) and (15) which are

$$81{\cdot}496 < \mu < 84{\cdot}604\,, \quad 2{\cdot}447 < \sigma < 4{\cdot}645\,.$$

34.9. Hypothesis Testing

Testing of statistical hypotheses is a statistical procedure in which we have to choose, on base of an observed random sample, exactly one of two mutually exclusive decisions (e.g., using experimental results, one must choose the better of the two production technologies or to decide on the efficiency of a new therapy). This section is devoted to the so-called *parametric tests* in which the possible decisions are formulated by means of suitable parameters. Various *nonparametric tests* can be found e.g. in [92], [219], [427].

Statistical Hypothesis. Let a probability distribution from which a random sample $X_1, \ldots, X_n$ has been chosen depends on an unknown m-component parameter ϑ from a known parameter space Ω. Let ω be a given subset of Ω. Then the assertion of the form $\vartheta \in \omega$ is called the *statistical hypothesis* and the aim is to decide on its validity using observations $x_1, \ldots, x_n$ of the considered random sample. We write H_0: $\vartheta \in \omega$ and H_1: $\vartheta \in \Omega \setminus \omega$, where H_0 is called the *null hypothesis* (or briefly the *hypothesis*) and H_1 the *alternative hypothesis* (or briefly the *alternative*), and say that we test the null hypothesis H_0 against the alternative hypothesis H_1 (or briefly the hypothesis H_0 against the alternative H_1). Construction of a test consists in finding out a suitable subset W of the sample space that is called the *critical region*. If $(x_1, \ldots, x_n)' \in W$ holds for the observations $x_1, \ldots, x_n$, then we reject H_0 and accept H_1. In the opposite case we cannot reject H_0 and it mostly means in practice that we prefer H_0 to H_1.

Example 1. A lucid example of a critical region can be formulated in the field of quality control (see § 35.12 and § 35.13). Let x be the number of defective items found among n items that were drawn randomly from a lot of products. The consignee refuses to accept the lot if the fraction ϑ of defective items in the whole lot exceeds 0·01. However, the fraction ϑ is unknown so that one must statistically test the null hypothesis H_0: $\vartheta \leqq 0{\cdot}01$ against the alternative hypothesis H_1: $\vartheta > 0{\cdot}01$ (i.e. $\Omega = (-\infty, \infty)$, $\omega = (-\infty, 0{\cdot}01]$). It is shown in § 35.12 that the hypothesis H_0 on the acceptable quality is rejected (and the lot is returned to the producer in this case) if $x > c$, where c is a suitable number constructed for this test. Obviously, the inequality $x > c$ represents a special form of the general relation $(x_1, \ldots, x_n)' \in W$ here.

Definition 1. The function $\beta(\vartheta)$ defined on Ω by

$$\beta(\vartheta) = P_\vartheta(\boldsymbol{X} \in W) \tag{1}$$

is called the *power function* (or briefly the *power*) of the corresponding test (the symbol P_ϑ means that the probability of the event $\boldsymbol{X} \in W$ is calculated using the value ϑ of the parameter).

Every test of statistical hypothesis takes a risk of a wrong decision. There are two possibilities of making an error. *Type-one error* is committed if H_0 is rejected when it is true. On the contrary, *type-two error* is committed if H_0 is not rejected when it is false. The objective is to find such a critical region that the probability of type-one error does not exceed a prescribed number α (i.e. $\beta(\vartheta) \leqq \alpha$ for $\vartheta \in \omega$) and, at the same time, the probability of type-two error is as small as possible (i.e. the values of the power function $\beta(\vartheta)$ for $\vartheta \in \Omega \setminus \omega$ are as large as possible).

Definition 2. The number α for which

$$\alpha = \sup_{\vartheta \in \omega} \beta(\vartheta) \tag{2}$$

is called the *significance level* of the corresponding test. Sometimes this test is called the *test of size* α.

REMARK 1. If $(x_1, \ldots, x_n)' \in W$ holds for observations $x_1, \ldots, x_n$, then we say that the hypothesis H_0 is rejected on the significance level α. In statistical practice one usually takes $\alpha = 0{\cdot}05$ or $\alpha = 0{\cdot}01$ (the significance levels larger than $0{\cdot}1$ are used exceptionally). It is common to give the significance level in per cent (e.g., for $\alpha = 0{\cdot}05$ one speaks of the 5 % significance level). In some special cases the *uniformly most powerful test* for H_0 against H_1 with a given significance level can be constructed whose power function attains the largest value for each $\vartheta \in \Omega \setminus \omega$ among all the tests for H_0 against H_1 with the same significance level.

In practice one most frequently tests H_0: $\tau(\vartheta) = \tau_0$ against H_1: $\tau(\vartheta) > \tau_0$ or H_0: $\tau(\vartheta) = \tau_0$ against H_1: $\tau(\vartheta) < \tau_0$ (the so-called *one-sided tests*) and H_0: $\tau(\vartheta) = \tau_0$ against H_1: $\tau(\vartheta) \neq \tau_0$ (the so-called *two-sided tests*) where $\tau(\vartheta)$ is a known function defined on Ω and τ_0 is a given constant.

34.10. Tests of Hypotheses on Parameters of Normal Distributions

This section gives a survey of tests with a significance level α that concern parameters of normal distributions. The assumption of normality is frequently acceptable and therefore these tests are very common in statistical practice. For simplicity, only the inequalities that determine the corresponding critical region will be given (e.g. (1) instead of the full form $W = \{(x_1, \ldots, x_n)'$: $\sqrt{n}\,(\bar{x} - \mu_0)/s \geqq \geqq t_{1-\alpha}(n-1)\}$), and only one of the two one-sided tests will be described (the form of the other test should be clear, e.g., $\sqrt{n}\,(\bar{x} - \mu_0)/s \leqq t_\alpha(n-1)$ is the other one-sided test to the test (1)). The following tests on parameters of normal distributions make use of the quantiles of the distributions χ^2, t, F (see Remark 33.7.4).

1. *One-sample tests.* Let $X_1, \ldots, X_n$ $(n > 1)$ be a random sample from the distribution $N(\mu, \sigma^2)$ where μ, σ^2 are unknown parameters and μ_0, σ_0^2 are given constants.

(i) H_0: $\mu = \mu_0$ against H_1: $\mu > \mu_0$:

$$\sqrt{n}\,(\overline{x} - \mu_0)/s \geqq t_{1-\alpha}(n-1)\,; \tag{1}$$

(i′) H_0: $\mu = \mu_0$ against H_1: $\mu \neq \mu_0$:

$$\sqrt{n}\,|\overline{x} - \mu_0|\,/s \geqq t_{1-\alpha/2}(n-1) \tag{2}$$

(the so-called *one-sample t tests*). For example, if the inequality (1) is fulfilled for $\alpha = 0{\cdot}05$ after substituting the observed numerical values, then we can reject the null hypothesis $\mu = \mu_0$ and accept the alternative hypothesis $\mu > \mu_0$ with the 95 % certainty.

(ii) H_0: $\sigma^2 = \sigma_0^2$ against H_1: $\sigma^2 > \sigma_0^2$:

$$(n-1)s^2/\sigma_0^2 \geqq \chi^2_{1-\alpha}(n-1)\,; \tag{3}$$

(ii′) H_0: $\sigma^2 = \sigma_0^2$ against H_1: $\sigma^2 \neq \sigma_0^2$:

$$(n-1)s^2/\sigma_0^2 \leqq \chi^2_{\alpha/2}(n-1) \quad \text{or} \quad (n-1)s^2/\sigma_0^2 \geqq \chi^2_{1-\alpha/2}(n-1)\,. \tag{4}$$

Example 1. A machine fills packages of a certain weight, the maximum tolerated standard deviation being $\sigma_0 = 0{\cdot}4$ grams. Using a sample of 10 packages, the sample standard deviation $s = 0{\cdot}5$ grams was calculated. The objective is to decide whether the machine works properly, i.e., to test H_0: $\sigma^2 = \sigma_0^2$ against H_1: $\sigma^2 > \sigma_0^2$ $(\alpha = 0{\cdot}05)$.

According to (3) one obtains

$$(n-1)s^2/\sigma_0^2 = 9\,.\,0{\cdot}5^2/0{\cdot}4^2 = 14{\cdot}063 < 16{\cdot}919 = \chi^2_{0{\cdot}95}(9)$$

so that the hypothesis that the machine works properly cannot be rejected on the 5 % level significance.

2. *Two-sample tests.* Let $X_1, \ldots, X_{n_1}$ $(n_1 > 1)$ be a random sample from the distribution $N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_{n_2}$ $(n_2 > 1)$ a random sample from the distribution $N(\mu_2, \sigma_2^2)$ where μ_1, μ_2, σ_1^2, σ_2^2 are unknown parameters. Let these random samples are mutually independent.

(i) H_0: $\mu_1 = \mu_2$ against $\mu_1 > \mu_2$:
if $\sigma_1^2 = \sigma_2^2$, then

$$\left[\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}\right]^{1/2} \frac{\overline{x} - \overline{y}}{[(n_1-1)s_1^2 + (n_2-1)s_2^2]^{1/2}} \geqq t_{1-\alpha}(n_1 + n_2 - 2)\,; \tag{5}$$

if $\sigma_1^2 \neq \sigma_2^2$, then

$$\frac{\overline{x} - \overline{y}}{(s_1^2/n_1 + s_2^2/n_2)^{1/2}} \geqq \frac{t_{1-\alpha}(n_1 - 1)s_1^2/n_1 + t_{1-\alpha}(n_2 - 1)s_2^2/n_2}{s_1^2/n_1 + s_2^2/n_2}; \tag{6}$$

(i′) H_0: $\mu_1 = \mu_2$ against $\mu_1 \neq \mu_2$:
if $\sigma_1^2 = \sigma_2^2$, then

$$\left[\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}\right]^{1/2} \frac{|\overline{x} - \overline{y}|}{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]^{1/2}} \geqq t_{1-\alpha/2}(n_1 + n_2 - 2); \tag{7}$$

if $\sigma_1^2 \neq \sigma_2^2$, then

$$\frac{|\overline{x} - \overline{y}|}{(s_1^2/n_1 + s_2^2/n_2)^{1/2}} \geqq \frac{t_{1-\alpha/2}(n_1 - 1)s_1^2/n_1 + t_{1-\alpha/2}(n_2 - 1)s_2^2/n_2}{s_1^2/n_1 + s_2^2/n_2} \tag{8}$$

(the so-called *two-sample t tests*).

(ii) H_0: $\sigma_1^2 = \sigma_2^2$ against H_1: $\sigma_1^2 > \sigma_2^2$:

$$s_1^2/s_2^2 \geqq F_{1-\alpha}(n_1 - 1, n_2 - 1); \tag{9}$$

(ii′) H_0: $\sigma_1^2 = \sigma_2^2$ against H_1: $\sigma_1^2 \neq \sigma_2^2$:

$$s_1^2/s_2^2 \leqq F_{\alpha/2}(n_1 - 1, n_2 - 1) \quad \text{or} \quad s_1^2/s_2^2 \geqq F_{1-\alpha/2}(n_1 - 1, n_2 - 1). \tag{10}$$

REMARK 1. The significance level of the tests (6) and (8) is equal to α only approximately. In order to decide whether to apply the test (5) or (6) (analogously (7) or (8)) one can first perform the test (10).

Example 2. A sample of 20 bulbs from the first producer has the mean lifetime $\overline{x} = 1230$ hours with the sample standard deviation $s_1 = 75$ hours while a sample of 30 bulbs from the second producer has the mean lifetime $\overline{y} = 1180$ hours with the sample standard deviation $s_2 = 80$ hours. The objective is to decide whether the bulbs from both producers have the same lifetime. Since the first producer is known to have better conditions for production of bulbs with long lifetime than the second producer we shall test H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 > \mu_2$ ($\alpha = 0{\cdot}05$).

First it is necessary to verify whether we can assume $\sigma_1^2 = \sigma_2^2$. According to (10) we obtain

$$F_{0{\cdot}025}(19, 29) = 0{\cdot}416 < s_1^2/s_2^2 = 75^5/80^2 = 0{\cdot}8789 < 2{\cdot}2313 = F_{0{\cdot}975}(19, 29).$$

Hence the assumption $\sigma_1^2 = \sigma_2^2$ is not rejected so that the test (5) can be applied:

$$\left[\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}\right]^{1/2} \frac{\overline{x} - \overline{y}}{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]^{1/2}} =$$

$$= \left[\frac{20 \,.\, 30(20 + 30 - 2)}{20 + 30}\right]^{1/2} \frac{1230 - 1180}{[19 \,.\, 75^2 + 29 \,.\, 80^2]^{1/2}} =$$

$$= 2{\cdot}2189 > 1{\cdot}6772 = t_{0{\cdot}95}(48)\,.$$

Therefore the hypothesis that both types of bulbs have the same lifetime must be rejected on the 5 % significance level and we accept the alternative that the lifetime of bulbs from the first producer is longer.

3. *Paired tests.* Let $(X_1, Y_1)', \ldots, (X_n, Y_n)'$ $(n > 2)$ be a random sample from the two-variate normal distribution with unknown parameters μ_1, μ_2, σ_1^2, σ_2^2, ϱ.

(i) H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 > \mu_2$:

$$\sqrt{n}\, \frac{\overline{x} - \overline{y}}{\left\{\frac{1}{n-1} \sum_{i=1}^{n} [(x_i - y_i) - (\overline{x} - \overline{y})]^2\right\}^{1/2}} \geqq t_{1-\alpha}(n-1)\,; \tag{11}$$

(i′) H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$:

$$\sqrt{n}\, \frac{|\overline{x} - \overline{y}|}{\left\{\frac{1}{n-1} \sum_{i=1}^{n} [(x_i - y_i) - (\overline{x} - \overline{y})]^2\right\}^{1/2}} \geqq t_{1-\alpha/2}(n-1) \tag{12}$$

(so called *paired t tests*).

(ii) H_0: $\varrho = 0$ against H_1: $\varrho > 0$:

$$(n-2)^{1/2} r / (1 - r^2)^{1/2} \geqq t_{1-\alpha}(n-2)\,; \tag{13}$$

(ii′) H_0: $\varrho = 0$ against H_1: $\varrho \neq 0$:

$$(n-2)^{1/2} |r| / (1 - r^2)^{1/2} \geqq t_{1-\alpha/2}(n-2)\,. \tag{14}$$

(iii) H_0: $\sigma_1^2 = \sigma_2^2$ against H_1: $\sigma_1^2 > \sigma_2^2$:

see the critical region (13); however, here r is the sample correlation coefficient of the random samples $X_1 + Y_1, \ldots, X_n + Y_n$ and $X_1 - Y_1, \ldots, X_n - Y_n$;

(iii′) H_0: $\sigma_1^2 = \sigma_2^2$ against H_1: $\sigma_1^2 \neq \sigma_2^2$:
see the critical region (14) with the same r as before.

REMARK 2. The tests (13) and (14) enables us to decide whether the components of the observed two-variate normal distribution are uncorrelated (see Definition 33.5.6).

Example 3. The following 10 pairs of values show what is the wear (thickness in millimeters) of the left-hand front tire (the first value in the pair) and of the right-hand front tire (the second value in the pair) in a sample of 10 cars of the same make during some time period:

$$(2{\cdot}1,\ 1{\cdot}9),\ (0{\cdot}9,\ 0{\cdot}8),\ (0{\cdot}5,\ 0{\cdot}7),\ (1{\cdot}8,\ 1{\cdot}9),\ (1{\cdot}7,\ 1{\cdot}3),$$
$$(2{\cdot}0,\ 2{\cdot}1),\ (0{\cdot}9,\ 0{\cdot}7),\ (1{\cdot}4,\ 1{\cdot}1),\ (1{\cdot}7,\ 1{\cdot}6),\ (0{\cdot}8,\ 0{\cdot}9).$$

The objective is to decide whether the left-hand and right-hand tire wear off uniformly, i.e. to test H_0: $\mu_1 = \mu_2$ against H_0: $\mu_1 \neq \mu_2$ ($\alpha = 0{\cdot}05$). The rejection of H_0 and acceptance of H_1 should mean that there is a systematic defect in the symmetry of the front axle.

According to (12),

$$\sqrt{n}\,\frac{|\overline{x} - \overline{y}|}{\left\{\frac{1}{n-1}\sum_{i=1}^{n}[(x_i - y_i) - (\overline{x} - \overline{y})]^2\right\}^{1/2}} =$$
$$= \sqrt{10}\,\frac{0{\cdot}08}{0{\cdot}1989} = 1{\cdot}2719 < 2{\cdot}2622 = t_{0{\cdot}975}(9)$$

so that the hypothesis that the tires wear off uniformly cannot be rejected on the 5 % significance level.

REMARK 3. One can see that there is a close connection between the critical regions and the corresponding confidence intervals (compare e.g. (2) and (34.8.8)).

34.11. Goodness of Fit Tests

The *goodness of fit tests* form a category of statistical tests that enable us to test, on a prescribed significance level α, the null hypothesis H_0 that a given random sample $X_1, \ldots, X_n$ was chosen from probability distribution of a given type, possibly with unknown parameters (in the case of unknown parameters one must assume their independence without mutual functional relations). They are also called the *tests for distribution functions.* For example, one can test the null hypothesis that the corresponding probability distribution is $N(\mu, \sigma^2)$ with known or unknown parameters μ, σ^2.

Chi-Square Test. The chi-square test is generally carried out in the following steps:

1. The range of values of the tested probability distribution (e.g. the interval $(0, \infty)$ for the exponential distribution) is divided into k class intervals $I_1, \ldots, I_k$ (see § 34.5); the end intervals I_1 and I_k are frequently unbounded. The class frequencies n_i corresponding to the intervals I_i $(i = 1, \ldots, k)$ are called the *empirical frequencies* in this test.

2. One finds the probability p_i of the event that the random variable with the tested probability distribution lies in interval I_i $(i = 1, \ldots, k)$. If this distribution depends on an unknown m-component parameter $\vartheta = (\vartheta_1, \ldots, \vartheta_m)'$ $(m < k - 1)$, then the probabilities p_i depend on this parameter as well and it is necessary to write $p_i(\vartheta)$. The terms np_i or $np_i(\vartheta)$ are called the *theoretical frequencies* in this test.

3. If the tested probability distribution depends on an unknown parameter ϑ, then one constructs its maximum likelihood estimate $\hat{\vartheta}$. The corresponding likelihood equations (34.7.3) have the form

$$\sum_{i=1}^{k} \frac{n_i}{p_i(\vartheta)} \frac{\partial p_i(\vartheta)}{\partial \vartheta_j} = 0 \qquad (j = 1, \ldots, m). \tag{1}$$

If parameters of the tested distribution are known, then this step is omitted.

4. One performs the test with the significance level α that has the critical region

$$\sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n \geqq \chi^2_{1-\alpha}(k-1) \tag{2}$$

provided that the parameters are known, and

$$\sum_{i=1}^{k} \frac{[n_i - np_i(\hat{\vartheta})]^2}{np_i(\hat{\vartheta})} = \sum_{i=1}^{k} \frac{n_i^2}{np_i(\hat{\vartheta})} - n \geqq \chi^2_{1-\alpha}(k-m-1) \tag{3}$$

provided that the parameters are unknown.

REMARK 1. The class intervals should be chosen in such a way that the theoretical frequencies are not too small ($np_i \geqq 5$ or $np_i(\hat{\vartheta}) \geqq 5$ is recommended). The test (2) or (3) is only approximate since it has been derived asymptotically for $n \to \infty$.

Example 1 (*Goodness of Fit Test for the Exponential Distribution*). The objective is to test the null hypothesis that a random sample $X_1, \ldots, X_n$ was chosen from the exponential distribution with an unknown parameter $\delta > 0$. Let all the

class intervals $(0, b_1]$, $(b_1, b_2]$, ..., $(b_{k-2}, b_{k-1}]$, (b_{k-1}, ∞) except for the last one have the same length h, i.e. $b_i = ih$ $(i = 0, 1, \ldots, k-1)$. Then the solution of the likelihood equation (1) has the form

$$\hat{\delta} = h \left[\ln \frac{\sum_{i=1}^{k} i n_i - n_k}{\sum_{i=1}^{k} i n_i - n} \right]^{-1} . \tag{4}$$

Finally, $\chi^2_{1-\alpha}(k-2)$ is on the right-hand side of the inequality (3) and one substitutes

$$p_i(\hat{\delta}) = \begin{cases} \exp(-ih/\hat{\delta})[\exp(h/\hat{\delta}) - 1] & \text{for } i = 1, \ldots, k-1, \\ \exp[-(k-1)h/\hat{\delta}] & \text{for } i = k \end{cases} \tag{5}$$

on the left-hand side.

Example 2 (*Goodness of Fit Test for the Normal Distribution*). The objective is to test the null hypothesis that a random sample $X_1, \ldots, X_n$ was chosen from the distribution $N(\mu, \sigma^2)$ with unknown parameters $-\infty < \mu < \infty$ and $\sigma^2 > 0$. Let the class intervals be $(-\infty, b_1]$, $(b_1, b_2]$, ..., $(b_{k-2}, b_{k-1}]$, (b_{k-1}, ∞) $(k \geqq 4)$. Let c_i be the midpoints of the intervals $(b_{i-1}, b_i]$ $(i = 2, \ldots, k-1)$ and, moreover, $c_1 = b_1 - (b_2 - b_1)/2$, $c_k = b_{k-1} + (b_{k-1} - b_{k-2})/2$. Then the likelihood equations (1) have the form

$$\begin{aligned} \mu &= \frac{1}{n} \sum_{i=1}^{k} \frac{n_i}{p_i} \left\{ \mu p_i + \frac{\sigma^2}{\sqrt{(2\pi)}} [\exp(-A_{i-1}^2/2) - \exp(-A_i^2/2)] \right\} , \\ \sigma^2 &= \frac{1}{n} \sum_{i=1}^{k} \frac{n_i}{p_i} \left\{ \sigma^2 p_i + \frac{\sigma^2}{\sqrt{(2\pi)}} [A_{i-1} \exp(-A_{i-1}^2/2) - A_i \exp(-A_i^2/2)] \right\} , \end{aligned} \tag{6}$$

where

$$\begin{aligned} &p_i = \Phi(A_i) - \Phi(A_{i-1}), \quad A_i = (b_i - \mu)/\sigma \quad (i = 1, \ldots, k-1), \\ &A_0 = -\infty, \quad A_k = \infty, \quad A_0 \exp(-A_0^2/2) = A_k \exp(-A_k^2/2) = 0. \end{aligned}$$

One frequently uses the approximate solution

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{k} c_i n_i , \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{k} (c_i - \hat{\mu})^2 n_i \tag{7}$$

or solves equations (6) iteratively with the initial approximation (7). Finally, $\chi^2_{1-\alpha}(k-3)$ is on the right-hand side of the inequality (3) and one substitutes

$$p_i(\hat{\mu}, \hat{\sigma}^2) = \Phi(\hat{A}_i) - \Phi(\hat{A}_{i-1}) \quad (i = 1, \ldots, k) \tag{8}$$

on the left-hand side, where

$$\hat{A}_i = (b_i - \hat{\mu})/\hat{\sigma} \quad (i = 1, \ldots, k-1), \quad \hat{A}_0 = -\infty, \quad \hat{A}_k = \infty.$$

The values of the distribution function $\Phi(x)$ of the distribution $N(0, 1)$ can be found in statistical tables (see Remark 33.7.2) or can be calculated by means of (33.7.4) or (33.7.5) (see also Fig. 33.7).

Let us test the goodness of fit for the normal distribution in Example 34.5.2 (see also the graphical verification in Fig. 34.3). For this purpose we shall use the class intervals from Tab. 34.2. According to Example 34.5.2, the approximate solution (7) is $\hat{\mu} = \bar{x} = 1502{\cdot}7$ and $\hat{\sigma} = s = 204{\cdot}1$. There is a considerably good coincidence between the empirical and theoretical frequencies here (e.g., $n_1 = 4$ and $np_1(\hat{\mu}, \hat{\sigma}^2) = 3{\cdot}9627$, $n_2 = 9$ and $np_2(\hat{\mu}, \hat{\sigma}^2) = 8{\cdot}5818$, etc.). The test (3) gives

$$\sum_{i=1}^{12} \frac{[n_i - 300p_i(\hat{\mu}, \hat{\sigma}^2)]^2}{300p_i(\hat{\mu}, \hat{\sigma}^2)} = 0{\cdot}106 < 16{\cdot}919 = \chi^2_{0{\cdot}95}(9)$$

so that the hypothesis on the normality cannot be rejected on the 5 % significance level.

Kolmogorov-Smirnov Test. Let a random sample $X_1, \ldots, X_n$ have the empirical distribution function $F_n(x)$ (see Remark 34.5.2). The *Kolmogorov-Smirnov test* enables us to test the null hypothesis that the probability distribution from which the random sample was chosen has a known continuous distribution function $F_0(x)$. The corresponding critical region with the significance level α has the form

$$\sup_x |F_n(x) - F_0(x)| = \max_{1 \leqq i \leqq n} \left\{ \max\left[\left| F_0(x_{(i)}) - \frac{i-1}{n} \right|, \left| F_0(x_{(i)}) - \frac{i}{n} \right| \right] \right\} \geqq k_{1-\alpha/2}(n), \tag{9}$$

where $k_P(n)$ are tabulated values. For a sufficiently large n, one can use the approximation

$$k_{1-\alpha/2}(n) = (1/\sqrt{n})[-(1/2)\ln(\alpha/2)]^{1/2}. \tag{10}$$

34.12. Contingency Tables

An $r \times s$ *contingency table* is used in the situation when we classify a sample of size n by two criteria A and B and distinguish among r possible categories according to the criterion A and among s possible categories according to the criterion B. The intersection of the i-th row and j-th column (these intersections are called *cells* of the contingency table) contains the observed frequency n_{ij} among n observations

that corresponds to the i-th category of the criterion A simultaneously with the j-th category of the criterion B ($i = 1, \ldots, r$, $j = 1, \ldots, s$). The row sums $n_{i.}$ and the column sums $n_{.j}$ defined by

$$n_{i.} = \sum_{j=1}^{s} n_{ij} \quad (i = 1, \ldots, r), \quad n_{.j} = \sum_{i=1}^{r} n_{ij} \quad (j = 1, \ldots, s) \tag{1}$$

are called *marginal frequencies* (see Tab. 34.3).

From the point of view of mathematical statistics, the described contingency table gives the frequencies of observed values of a two-component discrete vector $(X, Y)'$ where $X = i$ corresponds to the i-th category of the criterion A and $Y = j$ to the j-th category of the criterion B. Obviously, the contingency tables form a special discrete case of the correlation tables from § 34.5. If we put

$$P(X = i, Y = j) = p_{ij} \quad (i = 1, \ldots, r, \ j = 1, \ldots, s) \tag{2}$$

then

$$\begin{aligned} P(X = i) = p_{i\cdot} &= \sum_{j=1}^{s} p_{ij} \quad (i = 1, \ldots, r), \\ P(Y = j) = p_{\cdot j} &= \sum_{i=1}^{r} p_{ij} \quad (j = 1, \ldots, s) \end{aligned} \tag{3}$$

holds for the marginal probability distributions of the random variables X and Y.

After recording frequencies in a contingency table we can test various hypotheses. One most frequently tests the *hypothesis on the independence* of the criteria A and B, i.e. H_0: $p_{ij} = p_{i\cdot}p_{\cdot j}$ ($i = 1, \ldots, r$, $j = 1, \ldots, s$). The critical region corresponding to the test on the significance level α has the form

$$\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 1\right) \geqq \chi^2_{1-\alpha}((r-1)(s-1)). \tag{4}$$

TABLE 34.3

A	B				$\sum$
	1	2	...	s	
1	n_{11}	n_{12}	...	n_{1s}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2s}	$n_{2.}$
...	...	...	...	...	...
r	n_{r1}	n_{r2}	...	n_{rs}	$n_{r.}$
$\sum$	$n_{.1}$	$n_{.2}$	...	$n_{.s}$	

TABLE 34.4

Failure	Failure B		$\sum$
A	no	yes	
no	99	11	110
yes	14	6	20
$\sum$	113	17	130

REMARK 1. The test (4) is approximate since it has been derived asymptotically for $n \to \infty$. Furthermore, we should have $n_{i.}n_{.j}/n \geqq 5$ $(i = 1, \ldots, r,\ j = 1, \ldots, s)$.

REMARK 2. If $r = s = 2$ (i.e., the contingency table has 4 cells), then the critical region (4) can be expressed as

$$\frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}} \geqq \chi^2_{1-\alpha}(1) \approx u_{1-\alpha/2}\,. \tag{5}$$

REMARK 3. The described contingency table for two criteria is called the *two-way contingency table.* However, one can also have the *three-way contingency table* for three criteria, etc.

Example 1. For 130 devices of the same type one has observed the frequencies of failures of two types A and B given in Tab.34.4. The objective is to decide on the independence of the failures A and B.

The test (5) gives

$$\frac{130(99\,.\,6 - 11\,.\,14)^2}{110\,.\,20\,.\,113\,.\,17} = 5{\cdot}9552 > 3{\cdot}8415 = \chi^2_{0{\cdot}95}(1)$$

so that the hypothesis on the independence of the failures A and B is rejected on the 5 % significance level.

35. TOPICS IN STATISTICAL INFERENCE

By TOMÁŠ CIPRA

References: [1], [10], [15], [40], [57], [63], [64], [71], [73], [86], [95], [100], [101], [114], [120], [124], [126], [127], [138], [175], [178], [179], [188], [208], [275], [278], [306], [308], [310], [336], [338], [382], [407], [411], [414], [416], [417], [421], [422], [439], [442], [448], [466], [481], [487], [499], [509].

A. REGRESSION ANALYSIS. FITTING CURVES TO EMPIRICAL DATA. CALCULUS OF OBSERVATIONS

35.1. Regression in Statistics

Regression Function . In many practical situations one investigates the dependence of a quantity on an other quantity, or quantities. An example of it is the dependence of the petrol consumption on the driving speed of a car. However, the main interest of mathematical statistics *does not lie* in the study of functional dependences of the form

$$y = f(x_1, \ldots, x_r)\,, \tag{1}$$

where y is a dependent variable and $x_1, \ldots, x_r$ are independent variables. Due to random influences (e.g., due to measurement errors or negligence of important aspects that are not included in $x_1, \ldots, x_r$), it is suitable to consider the dependent variable as a random variable Y and to rewrite (1) in the form

$$Y = f(x_1, \ldots, x_r) + e\,, \tag{2}$$

where e is an *error variable.* The random variable e should fluctuate about zero level, i.e. it should have zero mean. The objective of statistical analysis is to find (to estimate) the function f using observations $y_1, \ldots, y_n$ of the random variable Y that correspond to values $(x_{11}, \ldots, x_{1r}), \ldots, (x_{n1}, \ldots, x_{nr})$ of the variables $x_1, \ldots, x_r$. The specification of the function f mostly consists in the estimation of parameters that determine this function uniquely. For example, if one supposes f to

be of the form $f(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2$, we have to estimate the parameters β_1 and β_2. The estimated function f enables us to estimate (predict), for arbitrary values of variables $x_1, \ldots, x_r$, the corresponding value of the variable Y. In mathematical statistics, $f(x_1, \ldots, x_r)$ is called the *regression function*, Y is called the *response variable* and $x_1, \ldots, x_r$ are called the *explanatory variables* (or *regressors*).

Method of Least Squares . In order to estimate parameters of a regression function, one most frequently uses the *method of least squares* that consists in minimization of the expression

$$\sum_{i=1}^{n} [y_i - f(x_{i1}, \ldots, x_{ir})]^2 \tag{3}$$

over all possible values of the parameters of the regression function. The method of least squares is often interpreted in a deterministic way as a method that enables us to fit a suitable curve $y = f(x_1, \ldots, x_r)$ optimally to n observed points $[x_{11}, \ldots, x_{1r}, y_1], \ldots, [x_{n1}, \ldots, x_{nr}, y_n]$ (one speaks on *fitting curves* to empirical data). For example, Fig. 35.1 shows the fitting of a straight line $y = \beta_0 + \beta_1 x$ to points $[x_1, y_1], \ldots, [x_5, y_5]$. The estimates b_0 and b_1 of the parameters β_0 and β_1 are found by minimizing the expression

$$\sum_{i=1}^{5} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

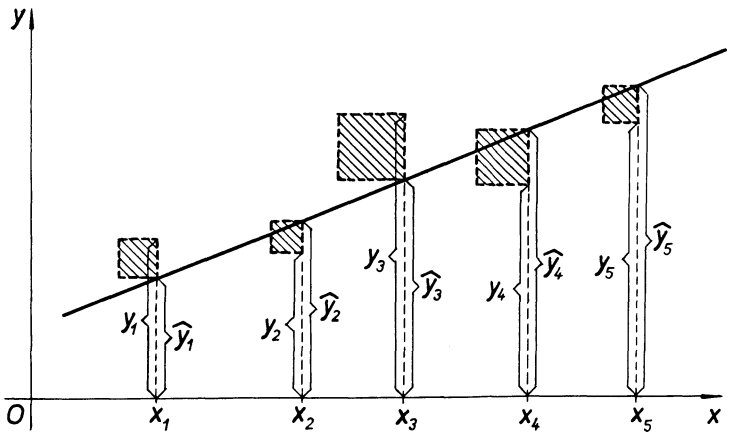


Fig. 35.1. Fitting a straight line to given points by means of the method of least squares.

over all real β_0 and β_1, i.e., by minimizing the sum of areas of the squares shown in Fig. 35.1. As concerns the determination of a suitable type of regression curve

to fit given points, its choice in statistical practice is often obvious due to the configuration of the points being fitted or can be carried out with help of an objective method (see e.g. Remarks 35.3.1 and 35.5.1).

35.2. Linear Regression Model

The simplest type of dependence among variables is the linear dependence. This type of dependence with the linear regression function

$$f(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{1}$$

is treated in statistics by means of the *linear regression model*

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + e_i \quad (i = 1, \ldots, n), \tag{2}$$

where Y_i denotes the response variable Y when the values of the explanatory variables $x_1, \ldots, x_k$ are $x_{i1}, \ldots, x_{ik}$. The error variables e_i are assumed to fulfil the relations

$$\mathrm{E}(e_i) = 0 \quad (i = 1, \ldots, n), \tag{3}$$

$$\mathrm{cov}(e_i, e_j) = \begin{cases} \sigma^2 & \text{for } i = j \quad (i, j = 1, \ldots, n), \\ 0 & \text{for } i \neq j \quad (i, j = 1, \ldots, n). \end{cases} \tag{4}$$

The model (2) has $k+1$ unknown real *regression parameters* $\beta_0, \beta_1, \ldots, \beta_k$ and an unknown parameter $\sigma^2 > 0$.

REMARK 1. The assumptions (3) and (4) mean that the error variables e_i fluctuate about zero level, have a constant unknown variance σ^2 and are mutually uncorrelated.

The model (2) is often written in the matrix form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{5}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ is an n-component random vector, $\boldsymbol{X}$ is an $n \times (k+1)$ matrix with elements x_{ij} $(i = 1, \ldots, n, j = 0, 1, \ldots, k)$ and $x_{i0} = 1$ $(i = 1, \ldots, n)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ is a $(k+1)$-component vector and $\boldsymbol{e} = (e_1, \ldots, e_n)'$ is an n-component random vector. The first column of the matrix $\boldsymbol{X}$ is formed only from numbers 1. Moreover, one assumes that the rank of the matrix $\boldsymbol{X}$ is equal to $k+1$ where $k+1 < n$. Hence the columns of this matrix are linearly independent and one speaks of a *full rank model*. If the rank of $\boldsymbol{X}$ is less than $k+1$, then one has a *model not possessing full rank* that demands special procedures based on pseudoinverse matrices (see e.g. [382]).

REMARK 2. Some important cases of the linear regression model have their individual names. We shall give some of them including the corresponding regression function:

(i) *linear regression*: $\beta_0 + \beta_1 x$;

(ii) *quadratic regression*: $\beta_0 + \beta_1 x + \beta_2 x^2$

(it takes the form (1) if we put $k = 2$, $x_1 = x$, $x_2 = x^2$);

(iii) *polynomial regression*: $\beta_0 + \beta_1 x + \cdots + \beta_k x^k$

(it takes the form (1) if we put $x_1 = x$, $x_2 = x^2, \ldots, x_k = x^k$);

(iv) *hyperbolic regression*: $\beta_0 + \beta_1 / x$

(it takes the form (1) if we put $k = 1$, $x_1 = 1/x$).

Estimation in Linear Regression Model.

Theorem 1. *The estimator* $\boldsymbol{b} = (b_0, b_1, \ldots, b_k)'$ *of the regression parameters* $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ *constructed in the linear regression model by the method of least squares* (see § 35.1) *has the form*

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \tag{6}$$

Its covariance matrix is

$$\boldsymbol{\Sigma}_{\boldsymbol{b}} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{7}$$

REMARK 3. In practical applications we substitute the vector of observations $\boldsymbol{y} = (y_1, \ldots, y_n)'$ for the random vector $\boldsymbol{Y}$ in the formula (6) (see Example 1). Then for any given values $x_1, \ldots, x_k$ of explanatory variables one can estimate (predict) the corresponding response variable Y as

$$\hat{Y} = b_0 + b_1 x_1 + \cdots + b_k x_k. \tag{8}$$

In particular, we have

$$\hat{Y}_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}. \tag{9}$$

REMARK 4. Instead of direct substitution to the formula (6) it is common that one solves numerically the system

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{y} \tag{10}$$

of $k+1$ linear equations with $k+1$ unknowns $b_0, b_1, \ldots, b_k$. The equations (10) are called the *normal equations.* They can be rewritten as

$$\begin{aligned} b_0\sum_{i=1}^n x_{i0}^2 + b_1\sum_{i=1}^n x_{i0}x_{i1} + \ldots + b_k\sum_{i=1}^n x_{i0}x_{ik} &= \sum_{i=1}^n x_{i0}y_i \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots&\cdots\cdots \\ b_0\sum_{i=1}^n x_{ik}x_{i0} + b_1\sum_{i=1}^n x_{ik}x_{i1} + \ldots + b_k\sum_{i=1}^n x_{ik}^2 &= \sum_{i=1}^n x_{ik}y_i\,. \end{aligned} \tag{11}$$

REMARK 5. The estimator $\boldsymbol{b}$ defined by (6) is called *linear* since each of its components can be expressed as a linear function of the random variables $Y_1, \ldots, Y_n$. Furthermore, it can be called *unbiased* since it is not difficult to show that each of its components is an unbiased estimator of the corresponding component of the vector $\boldsymbol{\beta}$ (see Definition 34.6.1). According to the so-called *Gauss-Markov theorem* $\boldsymbol{b}$ is even the *best linear unbiased estimator* (BLUE) since for any linear unbiased estimator of the vector $\boldsymbol{\beta}$ with covariance matrix $\boldsymbol{\Sigma}$ the matrix difference $\boldsymbol{\Sigma} - \boldsymbol{\Sigma_b}$ is positive semidefinite. In general, the estimator $\boldsymbol{b}$ is not the best unbiased estimator (see Definition 34.6.2). However, if the error variables e_i have the normal distribution, then $\boldsymbol{b}$ is not only the best unbiased estimator but even the efficient estimator (see § 34.6).

Definition 1. The values

$$\hat{e}_i = Y_i - \hat{Y}_i\,, \tag{12}$$

where $\hat{Y}_i$ are given in (9), are called the *residuals* of the linear regression model. Furthermore, one introduces the following sums:

$$S_t = \sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2 \quad (\textit{total sum of squares}), \tag{13}$$

$$S_r = \sum_{i=1}^n \left(\hat{Y}_i - \overline{Y}\right)^2 \quad (\textit{regression sum of squares}), \tag{14}$$

$$S_e = \sum_{i=1}^n \hat{e}_i^2 \quad (\textit{residual sum of squares}). \tag{15}$$

Theorem 2. *In the linear regression model*

$$S_t = S_r + S_e\,. \tag{16}$$

Definition 2. The *coefficient of determination* of the linear regression model is defined by

$$R^2 = \frac{S_r}{S_t} = 1 - \frac{S_e}{S_t}. \tag{17}$$

REMARK 6. The coefficient of determination expresses inasmuch the considered linear regression model is capable to explain the sample variance of the response variable, i.e., R^2 assesses the quality of the model for the given data. It is always $0 \leqq R^2 < 1$. One frequently takes $R^2 > 0{\cdot}95$ for the criterion of acceptance of a model. It is common to express R^2 in per cent.

Theorem 3. *The estimator*

$$s^2 = \frac{1}{n-k-1} S_e = \frac{1}{n-k-1} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \tag{18}$$

is an unbiased estimator of the parameter σ^2 *and the estimator*

$$\boldsymbol{S_b} = s^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{19}$$

is an unbiased estimator of the covariance matrix $\boldsymbol{\Sigma_b}$ *of the estimator* $\boldsymbol{b}$.

Example 1. One investigates the dependence of the monthly consumption of heating oil y (in litres) on the average monthly temperature x_1 (in °C) and on the floorage x_2 (in m^2). The observed values are given in Tab. 35.1.

TABLE 35.1

y_i	140	200	370	600	620	1300	1050	1280	1100	550
x_{i1}	17·8	16·6	12·2	7·1	2·8	0·1	−2·9	−3·1	−0·7	4·4
x_{i2}	170	210	150	190	110	250	140	155	180	130

From (6) we compute

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} 10, & 54{\cdot}3, & 1685 \\ 54{\cdot}3, & 837{\cdot}37, & 9583{\cdot}5 \\ 1685, & 9583{\cdot}5, & 299125 \end{bmatrix}^{-1} \begin{bmatrix} 7210 \\ 11089 \\ 1243400 \end{bmatrix} = \begin{bmatrix} 438{\cdot}73 \\ -54{\cdot}468 \\ 3{\cdot}4304 \end{bmatrix}.$$

From (17) we obtain

$$R^2 = 1 - \frac{S_e}{S_t} = 1 - \frac{59632{\cdot}9}{1685890} = 0{\cdot}965.$$

The value $R^2 = 96{\cdot}5\,\%$ seems to be sufficiently large. From (18) we compute

$$s^2 = \frac{1}{n-k-1} S_e = \frac{1}{7}\ 59632{\cdot}9 = 8518{\cdot}98$$

and from (19)

$$\boldsymbol{S_b} = s^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 16768{\cdot}2, & -9{\cdot}9669, & -94{\cdot}138 \\ -9{\cdot}9669, & 16{\cdot}070, & -0{\cdot}45870 \\ -94{\cdot}138, & -0{\cdot}45870, & 0{\cdot}57346 \end{bmatrix}.$$

Hence, in particular, the standard deviation of the estimates b_0, b_1, b_2 can be estimated as

$$s_{b_0} = \sqrt{16768{\cdot}2} = 129{\cdot}5, \quad s_{b_1} = \sqrt{16{\cdot}070} = 4{\cdot}009, \quad s_{b_2} = \sqrt{0{\cdot}57346} = 0{\cdot}7573.$$

If the objective is to predict the consumption $\hat{y}$ of heating oil for the average temperature $x_1 = -1{\cdot}5°\mathrm{C}$ and the floorage $x_2 = 200\ \mathrm{m}^2$, then we obtain from (8)

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 = 438{\cdot}73 + (-54{\cdot}468).(-1{\cdot}5) + 3{\cdot}4304\ .\ 200 = 1206{\cdot}5 \text{ (litres)}.$$

REMARK 7. In order to simplify calculations, one sometimes replaces the model (2) with $k+1$ regression parameters by the following one with only k parameters:

$$Y_i - \overline{Y} = \beta_1^*(x_{i1} - \bar{x}_1) + \cdots + \beta_k^*(x_{ik} - \bar{x}_k) + e_i^* \quad (i = 1, \ldots, n), \tag{20}$$

where

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad (j = 1, \ldots, k).$$

The estimators $\boldsymbol{b} = (b_0, b_1, \ldots, b_k)'$ and $\boldsymbol{b}^* = (b_1^*, \ldots, b_k^*)'$ constructed by the method of least squares in the models (2) and (20), respectively, are related by

$$b_0 = \overline{Y} - b_1^*\bar{x}_1 - \cdots - b_k^*\bar{x}_k, \quad b_j = b_j^* \quad (j = 1, \ldots, k). \tag{21}$$

REMARK 8. For routine computer calculations it is advantageous to employ the *recursive method of least squares.* The estimate $\boldsymbol{b}_t$ based on the observations y_i and $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})$ $(i = 1, \ldots, t)$ is calculated by means of the recursive formula

$$\boldsymbol{b}_t = \boldsymbol{b}_{t-1} + \boldsymbol{P}_t \boldsymbol{x}_t (y_t - \boldsymbol{x}_t' \boldsymbol{b}_{t-1}), \tag{22}$$

where

$$\boldsymbol{P}_t = \boldsymbol{P}_{t-1} - (\boldsymbol{x}_t' \boldsymbol{P}_{t-1} \boldsymbol{x}_t + 1)^{-1} \boldsymbol{P}_{t-1} \boldsymbol{x}_t \boldsymbol{x}_t' \boldsymbol{P}_{t-1} \tag{23}$$

is an auxiliary matrix calculated recursively, too. The formulae (22) and (23) require a choice of suitable initial values $\boldsymbol{b}_0$ and $\boldsymbol{P}_0$ (see e.g. [138] , [509]).

35.3. Normal Linear Regression Model

The *normal linear regression model* fulfills all the assumptions introduced for the linear regression model in § 35.2 and, in addition, its error variables e_i are independent random variables with the normal distribution, i.e.

$$e_i \sim N\left(0, \sigma^2\right) . \tag{1}$$

Theorem 1. *The following statements hold in the normal linear regression model:*

(i)

$$\boldsymbol{b} \sim N_{k+1}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right) . \tag{2}$$

(ii) *The random vector* $\boldsymbol{b}$ *and the random variable* s^2 *are independent.*

(iii) *Let* $s_{b_i}^2$ *be the* i*-th diagonal element* ($i = 0, 1, \ldots, k$) *of the matrix* $\boldsymbol{S}_{\boldsymbol{b}}$ *defined by* (35.2.19). *Then the random variable*

$$(b_i - \beta_i)/s_{b_i} \quad (i = 0, 1, \ldots, k) \tag{3}$$

has the distribution $t(n - k - 1)$.

Tests of Significance in the Normal Linear Regression Model. Some useful tests on a significance level α can be constructed by means of Theorem 1 in the normal linear regression model. For simplicity we shall give their critical regions only (i.e., if the given inequality is fulfilled, then the corresponding null hypothesis is rejected) :

(i) H_0: $\beta_i = 0$ against H_1: $\beta_i \neq 0$ (*test of significance for the parameter* β_i):

$$|b_i| / s_{b_i} \geqq t_{1-\alpha/2}(n - k - 1). \tag{4}$$

(ii) H_0: $(\beta_1, \ldots, \beta_k)' = (0, \ldots, 0)'$ against H_1: $(\beta_1, \ldots, \beta_k)' \neq (0, \ldots, 0)'$ (*test of significance for the model*):

$$\frac{n - k - 1}{k} \frac{R^2}{1 - R^2} = F_{1-\alpha}(k, n - k - 1). \tag{5}$$

REMARK 1. The tests (4) and (5) enable us to decide whether the presence of some explanatory variables in the model is statistically significant. They are used if we want to choose an optimal system of explanatory variables (in particular, an optimal degree of polynomial regression). The test (5) enables us to decide on the significance of a chosen system of explanatory variables if it is considered as a whole.

Theorem 1 also enables us to construct interval estimators for parameters of the model. Most frequently one constructs the so-called *prediction interval* that covers the predicted value of the response variable Y corresponding to given values of explanatory variables $x_1, \ldots, x_k$ on a given confidence level $1 - \alpha$. This interval has the form

$$\left(\hat{Y} - t_{1-\alpha/2}(n-k+1)s\left[1 + \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}\right]^{\frac{1}{2}},\right.$$
$$\left.\hat{Y} + t_{1-\alpha/2}(n-k-1)s\left[1 + \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}\right]^{\frac{1}{2}}\right), \quad (6)$$

where $\boldsymbol{x} = (1, x_1, \ldots, x_k)'$, $\hat{Y}$ is the prediction (35.2.8) and for t see § 33.7.

Example 1. The test of significance (4) for the individual parameters β_0, β_1, β_2 in the model from Example 35.2.1 gives

$$|b_0|/s_{b_0} = 438{\cdot}73/129{\cdot}5 = 3{\cdot}39 > 2{\cdot}3646 = t_{0{\cdot}975}(7),$$
$$|b_1|/s_{b_1} = 54{\cdot}468/4{\cdot}009 = 13{\cdot}59 > 2{\cdot}3646 = t_{0{\cdot}975}(7),$$
$$|b_2|/s_{b_2} = 3{\cdot}4304/0{\cdot}7573 = 4{\cdot}53 > 2{\cdot}3646 = t_{0{\cdot}975}(7).$$

Hence each of the parameters differs significantly from zero on the significance level 5% so that none of the terms of the regression function can be omitted. Furthermore, from (5) we have

$$\frac{n-k-1}{k}\frac{R^2}{1-R^2} = \frac{7}{2}\frac{0{\cdot}965}{1-0{\cdot}965} = 96{\cdot}5 > 4{\cdot}7374 = F_{0{\cdot}95}(2, 7)$$

so that the estimated model is distinctly significant as a whole. Finally, the prediction interval with the confidence level 95 % (the so-called 95 % prediction interval) for the prediction in Example 35.2.1 corresponding to the values $x_1 = -1{\cdot}5$ and $x_2 = 200$ is

$$(959{\cdot}5,\ 1453{\cdot}5)$$

since

$$\hat{y} \pm t_{0{\cdot}975}(7)s\left[1 + \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}\right]^{\frac{1}{2}} =$$
$$= 1206{\cdot}5 \pm 2{\cdot}3646.\sqrt{8518{\cdot}98}\,\sqrt{(1 + 0{\cdot}2809)} = \begin{cases} 959{\cdot}5, \\ 1453{\cdot}5. \end{cases}$$

35.4. Linear Regression

The *linear regression* (or *simple linear regression*) is a special case of the model (35.2.2) for $k = 1$ (i.e. $n > 2$) and can thus be written as

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad (i = 1, \ldots, n). \tag{1}$$

The regression function determines a straight line. The parameters can be estimated by the formulae

$$b_0 = \left(\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} x_i\right) \Big/ n = \overline{Y} - b_1 \bar{x}\,, \tag{2}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \tag{3}$$

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left[Y_i - \overline{Y} - b_1(x_i - \bar{x})\right]^2 = \frac{1}{n-2}\left(\sum_{i=1}^{n} Y_i^2 - b_0\sum_{i=1}^{n} Y_i - b_1\sum_{i=1}^{n} x_i Y_i\right). \tag{4}$$

The standard deviations of the estimators b_0 and b_1 can be estimated by

$$s_{b_0} = s\left[1/n + \bar{x}^2 \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{1}{2}},$$
$$s_{b_1} = s \Big/ \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{1}{2}}. \tag{5}$$

The test of significance (35.3.4) for the parameter β_1 (the *test of linearity*) has the critical region of the form

$$|b_1|\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{1}{2}} \Big/ s \geqq t_{1-\alpha/2}(n-2) \tag{6}$$

(for t see § 33.7). The prediction interval (35.3.6) corresponding to a value x of the explanatory variable has the endpoints of the form

$$b_0 + b_1 x \pm t_{1-\alpha/2}(n-2)s\left[1 + 1/n + (x - \bar{x})^2 \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{1}{2}}. \tag{7}$$

35.5. Polynomial Regression

The *polynomial regression* of degree k is a special case of the model (35.2.2) (see Remark 35.2.2) and is of the form

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + e_i \quad (i = 1, \ldots, n). \tag{1}$$

If we proceed in accordance with the general scheme from § 35.2, then the formula (35.2.6) for the estimation of the regression parameters by the method of least squares has the form

$$\begin{bmatrix} b_0 \\ b_1 \\ \cdots \\ b_k \end{bmatrix} = \begin{bmatrix} n, & \sum_{i=1}^n x_i, & \ldots, & \sum_{i=1}^n x_i^k \\ \sum_{i=1}^n x_i, & \sum_{i=1}^n x_i^2, & \ldots, & \sum_{i=1}^n x_i^{k+1} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^n x_i^k, & \sum_{i=1}^n x_i^{k+1}, & \ldots, & \sum_{i=1}^n x_i^{2k} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \\ \cdots \\ \sum_{i=1}^n x_i^k Y_i \end{bmatrix}. \tag{2}$$

Sometimes one carries out the so-called *orthogonalization of regressors.* It means that one treats the model

$$Y_i = \alpha_0 \varphi_0(x_i) + \alpha_1 \varphi_1(x_i) + \cdots + \alpha_k \varphi_k(x_i) + e_i \quad (i = 1, \ldots, n) \tag{3}$$

instead of the model (1), where $\varphi_j(x)$ $(j = 0, 1, \ldots, k)$ is the *Chebyshev polynomial* of degree j. These polynomials have the following orthogonality property:

$$\sum_{i=1}^n \varphi_j(x_i)\varphi_m(x_i) = 0 \quad (j \neq m; \quad j = 0, 1, \ldots, k; \quad m = 0, 1, \ldots, k).$$

The model (3) is again a polynomial regression of degree k so that it can be used equivalently instead of the model (1). However, the least squares estimators a_i of the regression parameters α_i can be found directly according to the formulae

$$a_j = \sum_{i=1}^n \varphi_j(x_i) Y_i \Big/ \left[\sum_{i=1}^n \varphi_j(x_i)\right]^2 \quad (j = 0, 1, \ldots, k). \tag{4}$$

Chebyshev polynomials are usually constructed recursively by the formula

$$\varphi_j(x) = x^j - \frac{\sum_{i=1}^n x_i^j \varphi_{j-1}(x_i)}{\sum_{i=1}^n \varphi_{j-1}^2(x_i)} \varphi_{j-1}(x) - \cdots - \frac{\sum_{i=1}^n x_i^j \varphi_0(x_i)}{\sum_{i=1}^n \varphi_0^2(x_i)} \varphi_0(x), \tag{5}$$

where $\varphi_0(x) = 1$.

Example 1. To orthogonalize the quadratic regression

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \quad (i = 1, \ldots, n)$$

it is sufficient to use the following Chebyshev polynomials up to the second degree:

$$\varphi_0(x) = 1,$$

$$\varphi_1(x) = x - \sum_{i=1}^{n} x_i/n = x - \bar{x},$$

$$\varphi_2(x) = x^2 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}(x - \bar{x}) - \sum_{i=1}^{n} x_i^2/n.$$

REMARK 1. There are various methods that enable us to find the unknown degree k of the polynomial regression. Either tests of significance of the type (35.3.4) are applied to polynomial terms that are added successively with increasing power (see e.g. [188]) or special criterial functions are used whose minimization provides directly the optimal degree of the polynomial regression.

35.6. Generalized Linear Regression Model. Calculus of Observations

The *generalized linear regression model* fulfils all the assumptions of the linear regression model from § 35.2 except for the assumption (35.2.4). Instead of it one supposes, more generally, that the covariance matrix of the vector of error variables $\boldsymbol{e} = (e_1, \ldots, e_n)'$ has the form

$$\boldsymbol{\Sigma}_{\boldsymbol{e}} = \sigma^2 \boldsymbol{\Omega}, \tag{1}$$

where $\sigma^2 > 0$ and $\boldsymbol{\Omega}$ is a positive definite matrix. It means that the error variables e_i need not have constant variance and need not be mutually uncorrelated. The linear regression model is a special case of the generalized linear regression model whose matrix $\boldsymbol{\Omega}$ is the identity matrix.

Theorem 1 (*Aitken Theorem*). *The best linear unbiased estimator of the regression parameters $\boldsymbol{\beta}$ in the generalized linear regression model has the form*

$$\tilde{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{Y} \quad (\textit{Aitken Estimator}). \tag{2}$$

Its covariance matrix is

$$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{b}}} = \sigma^2(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}. \tag{3}$$

Possibilities of the practical application of the Aitken estimator are restricted due to the fact that the matrix $\boldsymbol{\Omega}$ is only seldom a priori known in practice and it is impossible, in general, to estimate $n(n+1)/2$ unknown elements of this symmetric matrix using n observations. Therefore, from the practical point of view, only such special cases are important in which the matrix $\boldsymbol{\Omega}$ is given a priori or can be described using a small number of (unknown) parameters.

Weighted Method of Least Squares. Due to practical reasons, one sometimes assigns various weights to the individual terms in the sum (35.1.3) that is minimized in the method of least squares, so that, in the case of the linear regression function, one minimizes the expression of the form

$$\sum_{i=1}^{n} w_i \left[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})\right]^2 , \tag{4}$$

where $w_1, \ldots, w_n$ are given positive numbers (the *weights*). This procedure is equivalent to the construction of the Aitken estimator (2) whose matrix $\boldsymbol{\Omega}$ is diagonal with diagonal elements $1/w_1, \ldots, 1/w_n$. Another equivalent interpretation consists in the estimation of the regression parameters $\boldsymbol{\beta}$ using the classical method of least squares in the transformed model

$$\sqrt{w_i}\, y_i = \beta_0 \sqrt{w_i} + \beta_1 \sqrt{w_i}\, x_{i1} + \cdots + \beta_k \sqrt{w_i}\, x_{ik} + e_i \quad (i = 1, \ldots, n). \tag{5}$$

Calculus of Observations. Calculus of observations is used in geodesy, in the first place. Every measurement is subject to errors. If a quantity with the true value a is measured n times and the individual results are denoted by $y_1, \ldots, y_n$, then we can write

$$y_i = a + e_i \quad (i = 1, \ldots, n), \tag{6}$$

where e_i is the error of the i-th measurement. In this context, the probability distribution of the random variables $e_1, \ldots, e_n$ is called the *law of error*. Generally, all measurements of the same quantity cannot be regarded as equally precise. The quality of the individual measurements is then expressed by weights so that a larger weight is assigned to more precise measurements. Let us denote the given positive weight of the measurement y_i by w_i $(i = 1, \ldots, n)$. Then, using the weighted method of least squares, the expression (4) to be minimized has the simple form

$$\sum_{i=1}^{n} w_i (y_i - a)^2 . \tag{7}$$

The corresponding estimator $\tilde{a}$ of a is called the *adjusted value of a*. According to (2) or (5) one obtains

$$\tilde{a} = \sum_{i=1}^{n} w_i y_i \Big/ \sum_{i=1}^{n} w_i . \tag{8}$$

The standard deviation of the adjusted value $\tilde{a}$ can be estimated by (see (35.2.18) and (3))

$$s_{\tilde{a}} = \left(\frac{1}{n-1} \sum_{i=1}^{n} w_i \hat{e}_i^2 \Big/ \sum_{i=1}^{n} w_i \right)^{\frac{1}{2}}, \tag{9}$$

where the residual $\hat{e}_i = y_i - \tilde{a}$ is often called the *correction of the measurement* y_i. A generally adopted practice in the calculus of observations is to present the result of the adjustment in the form $\tilde{a} \pm s_{\tilde{a}}$.

REMARK 1. For $w_i = 1$ $(i = 1, \ldots, n)$ the formulas (8) and (9) reduce to

$$\tilde{a} = \sum_{i=1}^{n} y_i / n$$

(the arithmetic mean) and

$$s_{\tilde{a}} = \left\{ \sum_{i=1}^{n} \hat{e}_i^2 / [n(n-1)] \right\}^{\frac{1}{2}} .$$

35.7. Nonlinear Regression

If the regression function is not a linear function of parameters, then one speaks of the *nonlinear regression*. (If the regression function is a linear function of parameters then, according to Remark 35.2.2, the corresponding model can always be expressed in the form of a linear regression model without changing the parameters.) The *nonlinear regression model* can be written as

$$Y_i = f(\boldsymbol{x}_i, \boldsymbol{\vartheta}) + e_i \quad (i = 1, \ldots, n), \tag{1}$$

where Y_i denotes a response variable corresponding to the values $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ir})'$ of explanatory variables, $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_m)'$ is an m-component vector of regression parameters and e_i's are error variables satisfying the assumptions (35.2.3) and (35.2.4) in the simplest case.

The estimator $\hat{\boldsymbol{\vartheta}}$ of regression parameters $\boldsymbol{\vartheta}$ obtained by the method of least squares minimizing the expression (35.1.3) over all possible values $\boldsymbol{\vartheta}$ cannot usually be given explicitly but one employs various optimization procedures that construct the estimator $\hat{\boldsymbol{\vartheta}}$ iteratively. If the first partial derivatives $\partial f(\boldsymbol{x}_i, \boldsymbol{\vartheta})/\partial \vartheta_j$

($i = 1, \ldots, n$, $j = 1, \ldots, m$) exist for all values $\boldsymbol{\vartheta}$ from a given parameter space Ω, then the *Gauss-Newton method* is commonly used. The iterative formula of this method describing the calculation of the value $\hat{\boldsymbol{\vartheta}}_{k+1}$ in the $(k+1)$-st step by means of the value $\hat{\boldsymbol{\vartheta}}_k$ from the k-th step has the form

$$\hat{\boldsymbol{\vartheta}}_{k+1} = \hat{\boldsymbol{\vartheta}}_k + [\boldsymbol{F}'(\hat{\boldsymbol{\vartheta}}_k)\boldsymbol{F}(\hat{\boldsymbol{\vartheta}}_k)]^{-1}\boldsymbol{F}'(\hat{\boldsymbol{\vartheta}}_k)[\boldsymbol{y} - \boldsymbol{f}(\hat{\boldsymbol{\vartheta}}_k)] \quad (k = 0, 1, \ldots), \tag{2}$$

where $\boldsymbol{F}(\hat{\boldsymbol{\vartheta}}_k)$ is the $n \times m$ matrix with elements $\partial f(\boldsymbol{x}_i, \hat{\boldsymbol{\vartheta}}_k)/\partial \hat{\vartheta}_j$ $(i = 1, \ldots, n$, $j = 1, \ldots, m)$, $\boldsymbol{f}(\hat{\boldsymbol{\vartheta}}_k)$ is the n-component vector with components $f(\boldsymbol{x}_i, \hat{\boldsymbol{\vartheta}}_k)$ $(i = 1, \ldots, n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is the n-component vector of observed values of the response variable. Under general assumptions one takes $\hat{\boldsymbol{\vartheta}}_k$ for the estimate $\hat{\boldsymbol{\vartheta}}$ to be found provided that k is sufficiently large (there exist various stopping criteria indicating the end of the iterative procedure). The variance σ^2 of the error variables can be estimated by

$$s^2 = \frac{1}{n-m} S(\hat{\boldsymbol{\vartheta}}), \tag{3}$$

where $S(\hat{\boldsymbol{\vartheta}})$ denotes the value of the expression (35.1.3) for the value $\hat{\boldsymbol{\vartheta}}$ of regression parameters. Finally, the covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\vartheta}}}$ of the estimator $\hat{\boldsymbol{\vartheta}}$ can be estimated by

$$\boldsymbol{S}_{\hat{\boldsymbol{\vartheta}}} = s^2[\boldsymbol{F}'(\hat{\boldsymbol{\vartheta}})\boldsymbol{F}(\hat{\boldsymbol{\vartheta}})]^{-1}. \tag{4}$$

REMARK 1. In order to accelerate the rate of convergence of the iterative procedure, the formula (2) is sometimes modified to the form

$$\hat{\boldsymbol{\vartheta}}_{k+1} = \hat{\boldsymbol{\vartheta}}_k + [\boldsymbol{F}'(\hat{\boldsymbol{\vartheta}}_k)\boldsymbol{F}(\hat{\boldsymbol{\vartheta}}_k) + \lambda_k \boldsymbol{I}]^{-1}\boldsymbol{F}'(\hat{\boldsymbol{\vartheta}}_k)[\boldsymbol{y} - \boldsymbol{f}(\hat{\boldsymbol{\vartheta}}_k)] \quad (k = 0, 1, \ldots), \tag{5}$$

where $\boldsymbol{I}$ is the $m \times m$ identity matrix and λ_k's are suitably chosen numbers (the so-called *Levenberg-Marquardt method*, see e.g. [175]). Another possible modification has the form

$$\hat{\boldsymbol{\vartheta}}_{k+1} = \hat{\boldsymbol{\vartheta}}_k + \nu_k[\boldsymbol{F}'(\hat{\boldsymbol{\vartheta}}_k)\boldsymbol{F}(\hat{\boldsymbol{\vartheta}}_k)]^{-1}\boldsymbol{F}'(\hat{\boldsymbol{\vartheta}}_k)[\boldsymbol{y} - \boldsymbol{f}(\hat{\boldsymbol{\vartheta}}_k)] \quad (k = 0, 1, \ldots), \tag{6}$$

where the numbers ν_k are chosen from the interval $(0, 1)$ in such a way that the replacement of $\hat{\boldsymbol{\vartheta}}_k$ by $\hat{\boldsymbol{\vartheta}}_{k+1}$ reduces the value of the minimized expression (35.1.3) (the so-called *Hartley method*, see [208]).

REMARK 2. The calculation of the partial derivatives $\partial f(\boldsymbol{x}_i, \hat{\boldsymbol{\vartheta}}_k)/\partial \vartheta_j$ is performed either analytically or using the numerical approximation of the form

$$\frac{\partial f(\boldsymbol{x}_i, \hat{\boldsymbol{\vartheta}}_k)}{\partial \hat{\vartheta}_j} \approx \frac{f(\boldsymbol{x}_i, \hat{\vartheta}_{k1}, \ldots, \hat{\vartheta}_{kj} + \delta_j, \ldots, \hat{\vartheta}_{km}) - f(\boldsymbol{x}_i, \hat{\vartheta}_{k1}, \ldots, \hat{\vartheta}_{kj}, \ldots, \hat{\vartheta}_{km})}{\delta_j}, \tag{7}$$

where δ_j is a sufficiently small positive number.

Linearization Method . In some cases the regression function $f(\boldsymbol{x}, \boldsymbol{\vartheta})$ can be transformed by a simple transformation T into the function $T(f(\boldsymbol{x}, \boldsymbol{\vartheta}))$ that depends on its parameters linearly so that the transformed model with the transformed response variable $T(Y)$ can be treated as a linear regression model.

Example 1. The regression function

$$f(x, \gamma, \delta) = \delta [g(x)]^{\gamma}, \tag{8}$$

where $g(x)$ is a positive function of a real variable x and γ, δ are parameters ($\delta > 0$), can be logarithmically transformed into the linear form

$$\ln f(x, \gamma, \delta) = \ln \delta + \gamma \ln g(x) = \beta_0 + \beta_1 x_1, \tag{9}$$

where we put $x_1 = \ln g(x)$, $\beta_0 = \ln \delta$, $\beta_1 = \gamma$. Then the rough estimates $\hat{\gamma}, \hat{\delta}$ of parameters γ, δ are

$$\hat{\gamma} = b_1, \quad \hat{\delta} = \exp(b_0), \tag{10}$$

where b_0, b_1 are the estimates obtained by the method of least squares in the linearized model.

REMARK 3. The procedure shown in Example 1 can considerably distort the estimates of parameters of the original model. An improvement can be achieved by the so-called *linearization method with transformed weights* (see [439]). For example, in case of the logarithmic transformation that is the most frequent linearization transformation, one estimates the transformed model by the weighted method of least squares (see § 35.6) using the weights $w_i = y_i^2$, where y_i is the i-th observed value of the response variable in the original model.

Example 2. If we use the linearization method with transformed weights in Example 1, then the estimates b_0, b_1 of the parameters β_0, β_1 in the transformed regression function (9) are obtained by minimizing the expression

$$\sum_{i=1}^{n} y_i^2 [\ln y_i - (\beta_0 + \beta_1 x_{i1})]^2 \tag{11}$$

or, equivalently, by applying the method of least squares to the model (cf. (35.6.5))

$$|y_i| \ln y_i = \beta_0 |y_i| + \beta_1 |y_i| x_{i1} + e_i . \tag{12}$$

The estimates of the parameters γ, δ of the original model are obtained again from (10).

B. ANALYSIS OF VARIANCE

35.8. Principle of the Analysis of Variance

In practice it is often necessary to decide whether different values (*levels*) of a certain factor have significantly different effects (expressed usually by different means) on units of a sample. For example, a producer of cars is interested in the effect of different sorts of oil on the wear of piston rings. It is natural to subdivide the observed sample into groups corresponding to particular levels of the considered factor. If the grouping is carried out according to a single factor, then one speaks of the so-called *one-way classification* (see § 35.9). However, the observed sample of piston rings can be subdivided into groups corresponding not only to different sorts of oil but simultaneously also to different periods between oil exchanges. In such a case one speaks of the *two-way classification* according to two factors. In general, it is possible to classify according to k factors (see e.g. [414], [499]).

The term "analysis of variance" (sometimes abbreviated ANOVA) is justified by the fact that the sample variance of the observed sample can be decomposed into components that correspond to the influence of individual factors (i.e., they explain variability in data due to different values of the factors), and into a residual component that corresponds to random fluctuations. This decomposition is usually written by means of the so-called *table of analysis of variance* (see Tab. 35.2 and others). Suitable statistical tests compare, on a given significance level, the sizes of the components corresponding to individual factors with the size of the residual component and thus they decide on a possible significance of individual factors.

If the analysis of variance confirms the significance of a factor, then some of the *method of multiple comparison* is usually applied in order to find out such groups of the given classification that are significantly distinct (e.g. *Scheffé's method* described in §35.9, *Tukey's method, Duncan' s method* and others).

The analysis of variance exploits the following notation. If the sample values are distinguished by means of several indices, then dots in place of some indices denote sums over all possible values of these indices, and bars above such symbols denote the corresponding averages. For example, for a sample with values x_{ip} ($i = 1, \ldots, I$; $p = 1, \ldots, n_i$) we use the symbols

$$x_{i.} = \sum_{p=1}^{n_i} x_{ip}, \quad x_{..} = \sum_{i=1}^{I} \sum_{p=1}^{n_i} x_{ip},$$

$$\bar{x}_{i.} = x_{i.}/n_i, \quad \bar{x}_{..} = x_{..}/(n_1 + \cdots + n_I).$$

35.9. One-Way Classification

In the one-way classification according to a factor A, let a sample of size n be subdivided into I groups with values $x_{i1}, x_{i2}, \ldots, x_{in_i}$ in the i-th group ($i = 1, \ldots, I$). It means that $n_1 + \cdots + n_I = n$. In this situation, the analysis of variance uses the following model of one-way classification:

$$X_{ip} = \mu + \alpha_i + e_{ip} \quad (i = 1, \ldots, I;\ p = 1, \ldots, n_i), \tag{1}$$

where e_{ip}'s are independent random variables with the distribution $N(0, \sigma^2)$ and μ, α_i, σ^2 are unknown parameters (the parameters α_i can be interpreted as effects of the factor A). The objective is to test the hypothesis

$$H_0\colon\ \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0. \tag{2}$$

If H_0 is rejected on a prescribed significance level α, then the effects of the factor A are statistically significant.

REMARK 1. If $I = 2$, then the two-sample t test (see § 34.10) is also applicable. For $I > 2$, however, two-sample t test cannot be applied to each from the $I(I-1)/2$ pairs that can be selected from the I groups considered (what is connected with the problem how to achieve a prescribed significance level of the composite test).

The appropriate procedure is based on the decomposition of the form

$$S_T = S_A + S_e\,, \tag{3}$$

where S_T, S_A and S_e are the *total sum of squares*, *A-factor sum of squares* and *residual sum of squares*, respectively:

$$S_T = \sum_i \sum_p (x_{ip} - \bar{x}_{..})^2 = \sum_i \sum_p x_{ip}^2 - x_{..}^2/n, \tag{4}$$

$$S_A = \sum_i n_i(\bar{x}_{i.} - \bar{x}_{..})^2 = \sum_i x_{i.}^2/n_i - x_{..}^2/n, \tag{5}$$

$$S_e = \sum_i \sum_p (x_{ip} - \bar{x}_{i.})^2 = S_T - S_A\,. \tag{6}$$

The critical region of the hypothesis H_0 has the form

$$\frac{n-I}{I-1}\,\frac{S_A}{S_e} \geqq F_{1-\alpha}(I-1, n-I)\,. \tag{7}$$

The situation is displayed in the table of analysis of variance (see Tab. 35.2) that serves as a check, at the same time (the last row must contain the corresponding column sums).

Provided that the hypothesis H_0 is rejected we reject the equality of the r-th and s-th group ($r = 1, \ldots, I$; $s = 1, \ldots, I$; $r \neq s$) in the consequent analysis if

$$(\bar{x}_{r.} - \bar{x}_{s.})^2 \geqq \frac{n_r + n_s}{n_r n_s} \frac{I-1}{n-I} S_e F_{1-\alpha}(I-1, n-I) \tag{8}$$

(the so-called *Scheffé method*).

TABLE 35.2

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Test Statistic
Factor A	S_A	$I-1$	$S_A/(I-1)$	$\frac{S_A/(I-1)}{S_e/(n-I)}$
Residual	S_e	$n-I$	$S_e/(n-I)$	—
Total	S_T	$n-1$	—	—

Example 1. The objective is to compare the content of copper (in per cent) for three technologies of the bronze casting. One has performed 5, 3 and 4 laboratory experiments for the first, second and third technology, respectively. The observed values are summarized in Tab. 35.3. From (4) – (6) we have

$$S_T = 77347{\cdot}31 - 963{\cdot}1^2/12 = 50{\cdot}5092,$$
$$S_A = 392{\cdot}1^2/5 + 246{\cdot}9^2/3 + 324{\cdot}1^2/4 - 963{\cdot}1^2/12 = 31{\cdot}7537,$$
$$S_e = 50{\cdot}5092 - 31{\cdot}7537 = 18{\cdot}7555.$$

The corresponding table of analysis of variance is Tab. 35.4. The test (7) gives

$$7{\cdot}6188 > 4{\cdot}2565 = F_{0{\cdot}95}(2, 9).$$

TABLE 35.3

i	x_{ip}	n_i	$x_{i.}$	$\bar{x}_{i.}$	$\sum_p x_{ip}^2$
1	80·1 78·7 79·1 76·1 78·1	5	392·1	78·42	30757·33
2	82·8 80·5 83·6	3	246·9	82·3	20325·05
3	79·8 81·2 82·7 80·4	4	324·1	81·025	26264·93
$\sum$		12	963·1	—	77347·31

Hence, on the significance level 5 %, we reject the hypothesis that the mean content of copper does not depend on the technology used. Tab. 35.5 contains the left--hand and right-hand sides of the test (8) ($\alpha = 0{\cdot}05$) for particular r and s. We can conclude that only the first and second technologies differ from each other significantly on the significance level 5 %.

TABLE 35.4

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Test Statistic
Factor A	31·7537	2	15·8769	7·6188
Residual	18·7555	9	2·0839	—
Total	50·5092	11	—	—

TABLE 35.5

r	s	Left-hand side of (8)	Right-hand side of (8)
1	2	15·0544	9·4617
1	3	6·7860	7·9833
2	3	1·6256	10·3487

C. MULTIVARIATE ANALYSIS

Multivariate statistical analysis makes use of methods that are capable to treat various aspects of information contained in multivariate data, i.e., in observations of random vectors. This analysis includes *multivariate correlation analysis* (which investigates various aspects of dependence of random vectors), *multivariate analysis of variance* MANOVA (which generalizes methods of the analysis of variance in such a way that the data classified according to individual factors can be multivariate), *canonical correlation* (pairs of linear combinations of components of two random vectors are constructed that successively exhaust the maximal portions of correlation between both vectors), *factor analysis* (this method suggested originally for evaluating psychological tests enables us to draw certain conclusions on unobserved variables), *principal components* (that reduce a set of multivariate data in an optimal way with minimal loss of information), *discriminant analysis* (this method is employed in the case when multivariate data must be classified into se-

veral groups called *clusters*) and other methods (see e.g. [10], [15], [95], [114], [178], [275], [338], [499]).

D. RELIABILITY THEORY

35.10. Basic Reliability Concepts

Reliability theory is concerned with technological and mathematical aspects of reliability of products (in general, one speaks of elements or components) that are to perform required functions for a certain time period. As the mathematical aspects are concerned, the tools of probability theory and mathematical statistics enable us to investigate various reliability characteristics (like the mean time to failure, 100γ % life, mean costs on repairs and others) and, consequently, to design a suitable strategy that optimizes the work of a device without failures. Special types of reliability problems are *renewal theory* (which is concerned with processes of replacement of failed elements by new ones), *system reliability* (which investigates reliability of composite systems on the basis of reliability characteristics of individual components), *maintenance strategy*, etc.

Definition 1. Let a non-negative random variable X with the distribution function $F(x) = P(X \leqq x)$ describe the *time to failure* (or *lifetime*) of an element. Then the function

$$R(x) = 1 - F(x) = P(X > x) \quad (x \geqq 0) \tag{1}$$

is called the *reliability function*. For $0 < \gamma < 1$, the quantile $x_{1-\gamma}$ of the random variable X is called the 100γ % *life*.

REMARK 1. The quantity $R(x)$ is interpreted as the probability of the event that no failure occurs in the interval $[0, x]$ (sometimes one calls $R(x)$ the *probability of survival* and $F(x)$ the *probability of failure*). The interpretation of the 100γ % life is such that approximately 100γ per cent products will work without failure at least till the time $x_{1-\gamma}$.

Let us suppose for simplicity that the time to failure X has a continuous probability distribution with probability density $f(x)$.

Theorem 1. *Let k be a positive integer and let the time to failure X fulfil the assumption $\mathrm{E}(X^k) < \infty$. Then*

$$\mathrm{E}(X^k) = k \int_0^\infty x^{k-1} R(x) \mathrm{d}x \,. \tag{2}$$

REMARK 2. In particular

$$\mathrm{E}(X) = \int_0^\infty R(x)\mathrm{d}x \quad (\text{the } \textit{mean time to failure}), \tag{3}$$

$$\mathrm{var}(X) = 2\int_0^\infty xR(x)\mathrm{d}x - [\mathrm{E}(X)]^2 \tag{4}$$

according to Theorem 1.

Definition 2. The function

$$r(x) = \frac{f(x)}{1 - F(x)} = -\frac{R'(x)}{R(x)} \quad (x > 0), \tag{5}$$

which is defined for such x that $F(x) < 1$, is called the *hazard rate* (or *failure rate* or *force of mortality*).

REMARK 3. The reliability function can be expressed by means of the hazard rate as

$$R(x) = \exp\left[-\int_0^x r(t)\mathrm{d}t\right]. \tag{6}$$

The following approximation is admissible for small values of h:

$$P(x < X < x + h \mid X > x) \approx r(x)h. \tag{7}$$

Thus the quantity $r(x)$ approximately gives the probability of the event that an element which survived to time x will fail in the time interval $(x, x + 1)$. Some engineering devices feature a hazard rate that resembles the so-called bathtub curve shown in Fig. 35.2. Period I is referred to as the period of "infant mortality" ("running-in"), period II represents the useful life period and period III is called the "wear-out" period.

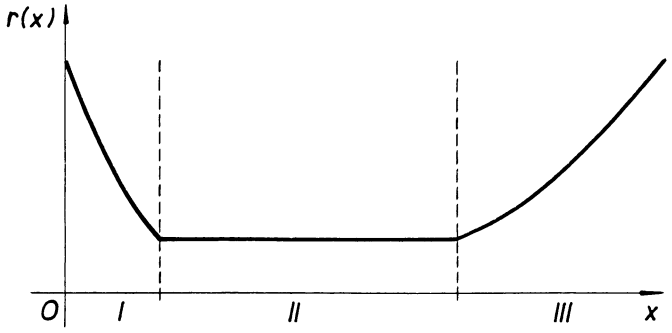


Fig. 35.2. Typical form of a hazard rate.

Theorem 2. *Let a system be composed of n independent components with their times to failure $X_1, \ldots, X_n$ and hazard rates $r_1(x), \ldots, r_n(x)$. Let the time to failure X of the system be*

$$X = \min(X_1, \ldots, X_n), \tag{8}$$

i.e., the failure of the system occurs if any of the n considered components fails. Then

$$r(x) = r_1(x) + \cdots + r_n(x)$$

holds for the hazard rate $r(x)$ of the system.

Probability Distribution of the Time to Failure . As probability distributions of the time to failure are concerned, one frequently uses the exponential, Weibull, Erlang and logarithmic normal distribution in practice (see notes in § 33.7 devoted to the application of these continuous distributions to reliability theory) but also other special distributions or mixtures of several distributions. We shall give a survey of hazard rates that correspond to the most frequent distributions of the time to failure.

1. *Exponential distribution* ($\delta > 0$):

$$r(x) = 1/\delta .$$

The exponential distribution, as the "distribution without memory", is suitable for such situations where a failure occurs due to random causes and not due to wear, thus it has the constant hazard rate.

2. *Weibull distribution* ($p > 0$, $\delta > 0$):

$$r(x) = \frac{p}{\delta^p} x^{p-1} .$$

The popularity of the Weibull distribution in reliability theory follows just from the flexibility of the corresponding hazard rate ($r(x)$ is decreasing for $0 < p < 1$, constant for $p = 1$, increasing and concave for $1 < p < 2$, increasing and linear for $p = 2$, and increasing and convex for $p > 2$).

3. *Erlang distribution* (p a positive integer, $\delta > 0$):

$$r(x) = x^{p-1} \exp(-x/\delta) \Big/ \int_x^\infty t^{p-1} \exp(-t/\delta)\mathrm{d}t .$$

4. *Logarithmic normal distribution* ($-\infty < \mu < \infty$, $\sigma^2 > 0$):

$$r(x) = \frac{1}{\sigma x} \varphi\left(\frac{\ln x - \mu}{\sigma}\right) \Big/ \Phi\left(\frac{\mu - \ln x}{\sigma}\right) ,$$

where the functions $\varphi(x)$ and $\Phi(x)$ are given in (33.7.1).

Redundancy. Redundancy means in reliability theory that given elements which carry the load of the system function are accompanied by components which are redundant for the proper function of the system until the load-carrying elements fail. The *active* (or *parallel*) *redundancy* means that the redundant elements operate and share the system load while the *standby redundancy* refers to the case of redundant elements that are inactive until the load-carrying elements fail.

Example 1. Let n independent elements have their times to failure $X_1, \ldots, X_n$ possessing the exponential distribution with parameter δ. If only one element can carry the load of the system (i.e., $n-1$ elements are redundant), then the mean time to failure (3) of this system with the standby redundancy is

$$\mathrm{E}(X_1 + \cdots + X_n) = n\delta$$

while for the case of the active redundancy we have (see (3) and (34.4.6))

$$\mathrm{E}[\max(X_1, \ldots, X_n)] = \int_0^\infty \left[1 - (1 - \mathrm{e}^{-x/\delta})^n\right] \mathrm{d}x = \left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right)\delta .$$

In particular, the mean time to failure with the standby redundancy is larger than that with the active redundancy.

35.11. Estimation of Reliability Characteristics

Censoring. In the classical estimation theory (see § 34.6) one uses a random sample $X_1, \ldots, X_n$. However, if performing reliability experiments, when one observes the times to failure $X_1, \ldots, X_n$ of n independent elements of the same type starting at time $t = 0$, the experiment must often be finished before all n elements fail. In this case one speaks of the *censored random sample.*

There are three basic types of censoring. In the following text, let $X_{(1)} \leqq X_{(2)} \leqq \cdots \leqq X_{(n)}$ denote the ordered random sample corresponding to the random sample $X_1, \ldots, X_n$ (see Definition 34.4.1).

1. *Type I censoring* (or *time censoring*). A positive number T (*time censor*) is prescribed and the experiment is finished as soon as its time length achieves T. The number r of observed failures is a random variable which can assume the values $0, 1, \ldots, n$. The result of the experiment consists of the observed values $x_{(1)} \leqq x_{(2)} \leqq \cdots \leqq x_{(r)}$ and the information that the value of the $(r+1)$-st order statistic is larger than T.

2. *Type II censoring* (or *failure censoring*). A positive integer r $(r \leqq n)$ is prescribed and the experiment is finished as soon as the r-th failure occurs. The

time length of the experiment is a random variable which coincides with the r-th order statistic $X_{(r)}$. The observed values $x_{(1)} \leqq x_{(2)} \leqq \cdots \leqq x_{(r)}$ are the result of the experiment.

3. *Random censoring.* This type of censoring is most frequent in practice. Here, in addition to the random variable X (time to failure), one has to consider a random variable T (*random time censor*) describing the time length of observation of an element. The time length of the experiment and the number of failures observed are random variables. The observed values $(w_1, i_1)', \ldots, (w_n, i_n)'$ of two-variate random vectors $(W_1, I_1)', \ldots, (W_n, I_n)'$ are the result of the experiment, where

$$W_i = \min(X_i, T_i), \tag{1}$$

$$I_i = \begin{cases} 1 & \text{if } W_i = X_i \quad (\text{the } i\text{-th observation is not censored}), \\ 0 & \text{if } W_i = T_i \quad (\text{the } i\text{-th observation is censored}). \end{cases} \tag{2}$$

The random variable I informs us whether the observation of an element is stopped due to failure or due to time censor.

Method of Maximum Likelihood for Censored Random Samples. The approach represents parametric estimation methods used in the case of censoring. Using this method, one must derive the form of the likelihood function corresponding to the considered censored random sample from a given probability distribution.

Example 1. Maximum likelihood estimators of the parameter δ of the exponential distribution for the individual types of censoring have the following form:

(1) type I censoring:

$$\hat{\delta} = \frac{1}{r}\left[\sum_{i=1}^{r} X_{(i)} + (n-r)T\right]; \tag{3}$$

(2) type II censoring:

$$\hat{\delta} = \frac{1}{r}\left[\sum_{i=1}^{r} X_{(i)} + (n-r)X_{(r)}\right]; \tag{4}$$

(3) random censoring:

$$\hat{\delta} = \sum_{i=1}^{n} W_i \Big/ \sum_{i=1}^{n} I_i . \tag{5}$$

Then the reliability function $R(x)$ can be estimated by

$$\hat{R}(x) = \exp(-x/\hat{\delta}). \tag{6}$$

Nonparametric Estimation for Censored Random Samples. As far as nonparametric estimation for censored random samples is concerned, one frequently uses the so-called *Kaplan-Meier estimator* (or *product-limit estimator*) of the reliability function $R(x)$. In random censoring, let no coincidences occur in the random sample $W_1, \ldots, W_n$ so that $W_{(1)} < W_{(2)} < \cdots < W_{(n)}$ for the corresponding ordered random sample. Let the random sample $I_1, \ldots, I_n$ be ordered in such a way that $I_{(i)}$ corresponds to the i-th order statistic $W_{(i)}$, i.e. let the random sample $(W_1, I_1)', \ldots, (W_n, I_n)'$ be ordered according to the first component. Then the Kaplan-Meier estimator of the reliability function has the form

$$\hat{R}(x) = \begin{cases} 1 & \text{for } x < W_{(1)}, \\ \prod_{\{i:\, W_{(i)} \leqq x\}} \left(\dfrac{n-i}{n-i+1}\right)^{I_{(i)}} & \text{for } W_{(1)} \leqq x < W_{(n)}, \\ 0 & \text{for } x \geqq W_{(n)}. \end{cases} \tag{7}$$

Example 2. Reliability tests of 8 products with random censoring have provided the following values of the random sample $(W_{(1)}, I_{(1)})', \ldots, (W_{(n)}, I_{(n)})'$ ordered according to the first component:

$$(4, 1),\ (7, 0),\ (12, 1),\ (15, 1),\ (27, 0),\ (31, 1),\ (35, 0),\ (47, 1).$$

The first component is the time length of the observation of the corresponding product. According to (2), zero in the second component means that the random time censoring was active so that the observation of the corresponding product was stopped before its failure, while one means that the failure occurred. The formula (7) gives

$$\hat{R}(4) = \left(\frac{n-1}{n-1+1}\right)^{I_{(1)}} = \left(\frac{7}{8}\right)^{1} = 0{\cdot}8750,$$

$$\hat{R}(7) = \hat{R}(4)\left(\frac{n-2}{n-2+1}\right)^{I_{(2)}} = \hat{R}(4)\left(\frac{6}{7}\right)^{0} = \hat{R}(4) = 0{\cdot}8750,$$

$$\hat{R}(12) = \hat{R}(7)\left(\frac{n-3}{n-3+1}\right)^{I_{(3)}} = \hat{R}(7)\left(\frac{5}{6}\right)^{1} = 0{\cdot}7292$$

and similarly $\hat{R}(15) = \hat{R}(27) = 0{\cdot}5834$, $\hat{R}(31) = \hat{R}(35) = 0{\cdot}3889$, $\hat{R}(47) = 0$. The estimate $\hat{R}(x)$ for $x \geqq 0$ is a jump function that is continuous from the right and has jumps at the above given points (i.e., $\hat{R}(x) = 1$ for $0 \leqq x < 4$, $\hat{R}(x) = 0{\cdot}8750$

for $4 \leqq x < 12$, $\hat{R}(x) = 0{\cdot}7292$ for $12 \leqq x < 15$, $\hat{R}(x) = 0{\cdot}5834$ for $15 \leqq x < 31$, $\hat{R}(x) = 0{\cdot}3889$ for $31 \leqq x < 47$, $\hat{R}(x) = 0$ for $x \geqq 47$). If one now wants e.g. to evaluate the probability that no failure occurs in the interval $[0, 20]$ for the tested type of products, then the estimate of this probability is obviously $\hat{R}(20) = 0{\cdot}5834$.

E. STATISTICAL PRINCIPLES OF QUALITY CONTROL

35.12. Acceptance Sampling

Statistical methods are successfully applied to quality control of mass production when the inspection of every piece in a lot is uneconomical (e.g. in screw production) or impossible due to the destructive character of control tests. The practical importance stimulates the efforts devoted to this problem (see e.g. [63], [86], [120]). Nowadays one exploits effective strategies not only in the quality control but also in the consequent regulation of production processes.

The objective of the *acceptance sampling procedures* (or *sampling inspections* or *sampling plans*) is to decide whether a consumer can accept a lot of products from a producer. Each acceptance sampling procedure is determined by two numbers. The first of them is the number n of products which must by randomly sampled from the lot and inspected. The second one is the so-called *acceptance number* which determines a rule that enables us to decide on the acceptance or rejection of the whole lot using the test sample of size n.

Operating Characteristic and Risks . Each acceptance sampling procedure has its *operating characteristic* $P(\vartheta)$ which presents the probability that a lot with the fraction ϑ of defective items will be accepted (ϑ is the so-called *fraction defective*). Let ϑ_0 and ϑ_1 $(0 < \vartheta_0 < \vartheta_1 < 1)$ be such values that the lots with the fraction defective $\vartheta \leqq \vartheta_0$ are accepted while those with $\vartheta \geqq \vartheta_1$ are rejected (the values ϑ in the interval $(\vartheta_0, \vartheta_1)$ are indifferent). Then the probability $\alpha = 1 - P(\vartheta_0)$ that a lot with the admissible fraction defective $\vartheta = \vartheta_0$ will be rejected is called the *producer's risk* while the probability $\beta = P(\vartheta_1)$ that a lot with the inadmissible fraction defective $\vartheta = \vartheta_1$ will be accepted is called the *consumer's risk*. There exist close relations to hypothesis testing (see § 34.9) where α and β correspond to the type-one and type-two errors, respectively, and $1 - P(\vartheta)$ corresponds to the power function for the null hypothesis H_0: $\vartheta \leqq \vartheta_0$ against the alternative H_1: $\vartheta \geqq \vartheta_1$. The graph of $P(\vartheta)$ is called the *operating characteristic curve*. Its typical form is sketched in Fig. 35.3.

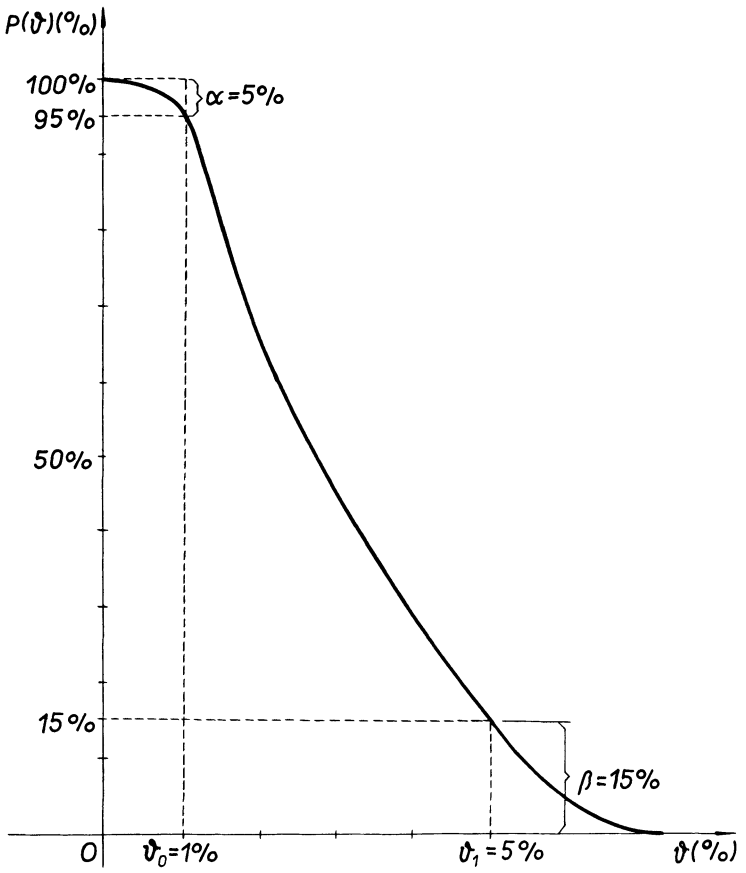


Fig. 35.3. Operating characteristic curve ($\vartheta_0 = 1\,\%$, $\vartheta_1 = 5\,\%$, $\alpha = 5\,\%$, $\beta = 15\,\%$).

Acceptance Sampling Procedures by Attributes. In *acceptance sampling procedures by attributes*, the only information about an inspected item is whether the item is defective or not. The decision on accepting or rejecting a lot of size N is based on the number x of defective items in a sample of size n from this lot. One compares x with the acceptance number c. If $x \leqq c$ or $x > c$, then the lot is accepted or rejected, respectively.

REMARK 1. If M is the number of defective items in the lot (i.e. $\vartheta = M/N$), then the number X of defective items in the sample has the hypergeometric distribution with parameters N, M, n (see § 33.6) so that the operating characteristic is

$$P(\vartheta) = P(X \leqq c) = \sum_{x=0}^{c} \binom{N\vartheta}{x} \binom{n - N\vartheta}{n - x} \Big/ \binom{N}{n}. \tag{1}$$

For $n/N < 0{\cdot}1$, $\vartheta < 0{\cdot}1$, $n > 30$ one can use its approximation by the Poisson distribution in the form

$$P(\vartheta) = \mathrm{e}^{-n\vartheta} \sum_{x=0}^{c} \frac{(n\vartheta)^x}{x!} . \tag{2}$$

Special tables (see, e.g. [120], [442]) provide the numbers n and c for prescribed values of ϑ_0, ϑ_1, α, β such that the corresponding acceptance sampling procedure by attributes has the producer's risk α for the admissible fraction defective ϑ_0 and the customer's risk β for the inadmissible fraction defective ϑ_1.

Example 1. The hypergeometric distribution (1) can be also approximated by the normal distribution (see Remark 33.13.2). Then the numbers n and c can be found by solving the system of equations

$$\frac{c + \frac{1}{2} - n\vartheta_0}{\sqrt{(n\vartheta_0)}} = u_{1-\alpha} , \qquad \frac{c + \frac{1}{2} - n\vartheta_1}{\sqrt{(n\vartheta_1)}} = u_{\beta} , \tag{3}$$

where u_P is the P-quantile of the distribution $N(0, 1)$ (see Remark 33.7.2). The explicit formulae for n and c are

$$\begin{aligned} n &= \frac{1}{\vartheta_0} \left[\frac{u_{1-\alpha}\vartheta_0/\vartheta_1 + u_{1-\beta}\sqrt{(\vartheta_0/\vartheta_1)}}{1 - \vartheta_0/\vartheta_1} \right]^2 , \\ c &= u_{1-\alpha}\sqrt{(n\vartheta_0)} + n\vartheta_0 - \tfrac{1}{2} . \end{aligned} \tag{4}$$

For the situation in Fig. 35.3 ($\vartheta_0 = 0{\cdot}01$, $\vartheta_1 = 0{\cdot}05$, $\alpha = 0{\cdot}05$, $\beta = 0{\cdot}15$) the formulae (4) give ($u_{1-\alpha} = u_{0{\cdot}95} = 1{\cdot}645$, $u_{1-\beta} = u_{0{\cdot}85} = 1{\cdot}036$)

$$n = 98{\cdot}09 \approx 98, \quad c = 2{\cdot}11 .$$

If a sample of size 98 contains at most 2 defective items, then the lot is accepted while in the opposite case it is rejected.

REMARK 2. Sometimes the consumer demands that all rejected lots be inspected completely and all defective items be replaced by acceptable ones. In such a situation one uses the so-called *rectifying acceptance sampling procedures.* After the rectification the original value ϑ of the fraction defective is replaced by the new value $\vartheta P(\vartheta)$ where $P(\vartheta)$ is the operating characteristic of the original acceptance sampling procedure before the rectification.

REMARK 3. If the fraction defective fluctuates largely in various lots, then it is more effective to use the so-called *multiple acceptance sampling procedures*, which accelerate the decision on acceptance or rejection provided that the fraction defective is low or high. For example, in the double acceptance sampling procedure one divides the original sample of size n into two subsamples of sizes n_1 and

n_2 $(n_1+n_2=n)$ with the corresponding acceptance numbers c_1 and c_2. If $c_1 < c_2$, then one first finds the number x_1 of defective items in the sample of size n_1. If $x_1 \leqq c_1$, then the whole lot is accepted while for $x_1 > c_2$ it is rejected. However, if $c_1 < x_1 \leqq c_2$, then one also has to find the number x_2 of defective items in the sample of size n_2. If $x_1 + x_2 \leqq c_2$, then the whole lot is accepted while for $x_1 + x_2 > c_2$ it is rejected.

Acceptance Sampling Procedures by Variables . If an average value μ of some quality factor is prescribed for products, then one can use *acceptance sampling procedures by variables* that compare an acceptance number c with the value of a suitable sample characteristic. This sample characteristic has frequently the form

$$u = \frac{|\bar{x} - \mu|}{s}, \tag{5}$$

where $\bar{x}$ is the sample mean and s is the sample standard deviation calculated from the values of the considered quality factor that are observed for items of a sample of size n. The corresponding numbers n and c are used in the way analogous to acceptance sampling procedures by attributes (the lot is accepted for $u \leqq c$ while it is rejected for $u > c$) and they are again tabulated (see [120], [442]). In order to simplify the calculations, the sample standard deviation s in (5) is sometimes replaced by the sample range (see Definition 34.4.2).

35.13. Sequential Acceptance Sampling

Sequential analysis is exploited not only in acceptance sampling procedures but generally for statistical hypothesis testing (see e.g. [40],[481]). The advantage of the sequential analysis consists in the fact that the size of a random sample for the construction of a desired decision need not be unnecessarily large. Namely, this size is not prescribed in advance but it is determined during the procedure that is carried out successively in steps. The result of each step is exactly one of the three possible decisions: (a) accept the null hypothesis H_0; (b) accept the alternative H_1; (c) make an additional observation (or observations).

Sequential acceptance sampling procedures with the producer's risk α for the admissible fraction defective ϑ_0 and with the consumer's risk β for the inadmissible fraction defective ϑ_1 operate with the test statistic

$$v_k = \frac{p_{1k}}{p_{0k}} \tag{1}$$

for the k-th step, where p_{0k} and p_{1k} are the probability functions or probability densities that relate to the observations in the considered sample of size k under the

null hypothesis H_0: $\vartheta \leqq \vartheta_0$ and under the alternative H_1: $\vartheta \geqq \vartheta_1$, respectively. Then the decision rule in the k-th step is

(a) accept the whole lot if $v_k \leqq \beta/(1-\alpha)$;
(b) reject the whole lot if $v_k \geqq (1-\beta)/\alpha$;
(c) make an additional observation if $\beta/(1-\alpha) < v_k < (1-\beta)/\alpha$.

The procedure stops in the cases (a) and (b) while it has to be repeated for $k+1$ observations in the case (c).

Example 1. If one approximates the hypergeometric distribution with parameters N, M, k by the binomial distribution with parameters k, $\vartheta = M/N$ ($k/N < 0{\cdot}1$, see § 33.6) in a sequential acceptance sampling procedure by attributes, then

$$p_{0k} = \binom{k}{x} \vartheta_0^x (1-\vartheta_0)^{k-x}, \quad p_{1k} = \binom{k}{x} \vartheta_1^x (1-\vartheta_1)^{k-x}, \tag{2}$$

where x is the number of defective items in a sample of size k. According to (1) one has

$$v_k = \left(\frac{\vartheta_1}{\vartheta_0}\right)^x \left(\frac{1-\vartheta_1}{1-\vartheta_0}\right)^{k-x}. \tag{3}$$

Taking the logarithm of the right-hand side of (3) and putting

$$a = \ln \vartheta_1 - \ln \vartheta_0, \quad b = \ln(1-\vartheta_1) - \ln(1-\vartheta_0),$$

the decision rule for the k-th step of the procedure has the form:

(a) accept the whole lot if $ax + b(k-x) \leqq \ln\beta - \ln(1-\alpha)$;
(b) reject the whole lot if $ax + b(k-x) \geqq \ln(1-\beta) - \ln\alpha$;
(c) make an additional observation if
$\ln\beta - \ln(1-\alpha) < ax + b(k-x) < \ln(1-\beta) - \ln\alpha$.

In particular, for the situation in Fig. 35.3 ($\vartheta_0 = 0{\cdot}01$, $\vartheta_1 = 0{\cdot}05$, $\alpha = 0{\cdot}05$, $\beta = 0{\cdot}15$) one obtains:

(a) accept the whole lot if $1{\cdot}609x - 0{\cdot}041(k-x) \leqq -1{\cdot}846$;
(b) reject the whole lot if $1{\cdot}609x - 0{\cdot}041(k-x) \geqq 2{\cdot}833$;
(c) make an additional observation if $-1{\cdot}846 < 1{\cdot}609x - 0{\cdot}041(k-x) < 2{\cdot}833$.

36. STOCHASTIC PROCESSES

By TOMÁŠ CIPRA

References: [9], [11], [18], [48], [54], [59], [75], [80], [81], [96], [102], [122], [138], [141], [164], [183], [186], [192], [195], [206], [207], [237], [240], [256], [257], [265], [269], [300], [301], [335], [377], [399], [404], [406], [439], [440], [441], [499], [503], [506], [509].

36.1. Classification of Stochastic Processes

The concept of *stochastic process* (or *random process* or briefly *process*) is used if the random variable X depends on time. We write $X(t)$ for $t \in T$. Observing a stochastic process $X(t)$, one obtains the *realization* (or *trajectory, path, sample function*) of this stochastic process denoted by $x(t)$ that is already a real (deterministic) function of the argument t. Practical examples of realizations of stochastic processes are meteorological records, electroencephalograms, records of mechanical vibrations, seismograms, time registrations of failures in reliability tests, etc. The role of probability theory and mathematical statistics in the analysis of stochastic processes consists in the calculation of characteristics describing the behaviour of processes as a whole, in the construction of models that enable us to generate the corresponding processes, in filtering, in prediction, etc.

If T is an interval on the real axis, then one speaks of the *stochastic process in continuous time* (*random function*). If T is a set of discrete real values $t_1, t_2, \ldots$, then one speaks of the *stochastic process in discrete time* (*random sequence, time series*); the time moments $t_1, t_2, \ldots$ are often equidistant and then one usually writes X_n instead of $X(t_n)$. If the random variables $X(t)$ are continuous, then one speaks of the *stochastic process with continuous states*. If $X(t)$ are discrete, then one speaks of the *stochastic process with discrete states* (the process with discrete states 0, 1, 2, ... is often called the *counting process* since it usually registers the number of certain events in time). If random vectors $\boldsymbol{X}(t)$ stand instead of random variables $X(t)$, then one speaks of the *multivariate stochastic process* instead of the *univariate stochastic process*.

Example 1. We shall give some examples of stochastic processes in accord with the above given classification:

1. *Discrete time and discrete states.* An example is the *branching process* (*Galton-Watson process*) $\{X_n: n = 0, 1, 2, \dots\}$ that describes the numbers X_0, X_1, X_2, ... of items in particular generations. Each item of an arbitrary generation gives rise to j items (descendants) of the next generation with the probability p_j $(j = 0, 1, 2, \dots)$ independently of the behaviour of other items in its generation. If the 0-th generation contains k_0 items (i.e. $X_0 = k_0$), then the mean number of items in the n-th generation is

$$\mathrm{E}(X_n) = k_0 m^n \quad (n = 0, 1, 2, \dots), \tag{1}$$

where

$$m = \sum_{j=1}^{\infty} j p_j \tag{2}$$

is the mean number of the direct descendants of each item. The branching processes are employed, e.g., in modelling particle fission.

2. *Discrete time and continuous states.* An example is the sequence $\{X_t: t = \dots, -1, 0, 1, \dots\}$ of uncorrelated random variables with the distribution $N(0, \sigma^2)$. A sequence of arbitrary random variables, for which

$$\mathrm{E}(X_t) = 0, \quad \mathrm{var}(X_t) = \sigma^2 > 0, \quad \mathrm{cov}(X_s, X_t) = 0 \quad (s \neq t) \tag{3}$$

holds, is called the *white noise.* The white noise is important as a basic element for generating more complicated processes.

3. *Continous time and discrete states.* An example is the *Poisson process* $\{X(t): t \geqq 0\}$ (see also § 36.3), where $X(t)$ describes the number of occurrences of an observed event in the time interval $[0, t]$ (e.g., the number of phone calls coming to a telephone exchange during this interval). In addition, one assumes that

(a) $X(0) = 0$;

(b) the lengths of the intervals between occurrences of the observed event are independent random variables;

(c) the lengths in the assumption (b) have the exponential distribution with the probability density

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{for } x > 0, \\ 0 & \text{for } x \leqq 0, \end{cases} \tag{4}$$

where $\lambda > 0$ is a parameter (the so-called *intensity of Poisson process* or *intensity of flow*).

On these assumptions, the random variable $X(t)$ has the Poisson distribution with the parameter λt, i.e.

$$P(X(t) = i) = \exp(-\lambda t)(\lambda t)^i / i! \quad (i = 0, 1, \dots). \tag{5}$$

In particular, the mean number of occurrences of the event in the interval $[0, t]$ is $\mathrm{E}[X(t)] = \lambda t$. Thus the intensity gives an average number of occurrences of the event in the unit time interval. The Poisson process is a "stochastic process without memory" (see also Theorem 33.7.4), i.e. the fact that the observed event did not occur in a certain interval does not affect the probability of its occurrences in the next interval. The Poisson process is frequently used in queueing theory (see § 36.5 and § 36.6).

4. *Continuous time and continuous states.* An example is the *Wiener process* $\{W(t): t \geqq 0\}$ for which

(a) $W(0) = 0$ and $W(t)$ has continuous trajectories;

(b) the increments $W(t_2) - W(t_1)$, $W(t_3) - W(t_2)$, ..., $W(t_n) - W(t_{n-1})$ are independent random variables for arbitrary moments $0 \leqq t_1 < t_2 < \ldots < t_n$;

(c) an arbitrary increment $W(t+s) - W(t)$ $(0 \leqq t < t+s)$ has the distribution $N(0, s)$.

On these assumptions, the random variable $W(t)$ has the distribution $N(0, t)$ and

$$\operatorname{cov}[W(s), W(t)] = \min(s, t) \quad (s, t \geqq 0). \tag{6}$$

The Wiener process is employed, e.g., in modelling diffusion phenomena or Brownian motion and in asymptotic statistics.

A. MARKOV PROCESSES

36.2. Concept of Markov Processes

Let $\{X(t): t \geqq 0\}$ be a stochastic process in continuous time and with discrete states from the set $I = \{0, 1, 2, \ldots\}$ (the states are denoted by non-negative integers, for simplicity).

Definition 1. The stochastic process $\{X(t): t \geqq 0\}$ is called the *Markov process* if

$$\begin{aligned} P(X(\tau) = j \mid X(t) = i, X(t_n) = i_n, \ldots, X(t_1) = i_1) = \\ = P(X(\tau) = j \mid X(t) = i) \end{aligned} \tag{1}$$

for arbitrary $0 \leqq t_1 < t_2 < \cdots < t_n < t \leqq \tau$ and $i_1, \ldots, i_n, i, j \in I$ (the so-called *Markov property*). The probabilities of the type (1) are called *transition probabilities.* If these probabilities do not depend on particular values of t and τ

but only on their difference, then such Markov process is called *homogeneous* and the transition probabilities are denoted by

$$p_{ij}(\tau - t) = P(X(\tau) = j \mid X(t) = i). \tag{2}$$

REMARK 1. If t denotes the current time moment, then the Markov property (1) means that the probability behaviour of the Markov process in an arbitrary future time moment τ $(\tau \geqq t)$ depends only on the current state and not on past states. If the process is homogeneous, then the transition probabilities depend only on the distances of corresponding time moments and are invariant with respect to shifts in time. In particular,

$$p_{ij}(0) = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases} \tag{3}$$

Futher important characteristics of a Markov process are the probabilities

$$p_i(t) = P(X(t) = i) \tag{4}$$

that describe the probability distribution of the process at time t. If $t = 0$, then one speaks of the *initial distribution* of the Markov process. For increasing t, the behaviour of a Markov process sometimes becomes stabilized so that it can be described by the so-called *stationary distribution* with stationary probabilities of the form

$$\pi_i = \lim_{t \to \infty} p_i(t) \quad (i = 0, 1, \ldots). \tag{5}$$

Theorem 1. *In a homogeneous Markov process, the following equations hold*:

(i)

$$p_{ij}(t_1 + t_2) = \sum_{k=0}^{\infty} p_{ik}(t_1) p_{kj}(t_2) \quad (t_1, t_2 \geqq 0; \quad i, j = 0, 1, \ldots) \tag{6}$$

(*the so-called Chapman-Kolmogorov equations*);

(ii)

$$p_i(t) = \sum_{j=0}^{\infty} p_j(0) p_{ji}(t) \quad (t \geqq 0; \quad i = 0, 1, \ldots). \tag{7}$$

Theorem 2 (*Prospective Kolmogorov Differential Equations*). *Let a homogeneous Markov process fulfil the conditions*:

(a) *there exist limits*

$$q_{ii} = \lim_{h \to 0+} \frac{p_{ii}(h) - 1}{h} \quad (i = 0, 1, \ldots); \tag{8}$$

(b) *there exist limits*

$$q_{ij} = \lim_{h \to 0+} \frac{p_{ij}(h)}{h} \quad (i \neq j;\ i, j = 0, 1, \dots); \tag{9}$$

(c) *the convergence in* (9) *is uniform with respect to* i *for an arbitrary fixed* j.

Then

$$p'_{ij}(t) = \sum_{k=0}^{\infty} p_{ik}(t) q_{kj} \quad (t > 0;\ i, j = 0, 1, \dots) \tag{10}$$

holds for the transition probabilities and

$$p'_i(t) = \sum_{k=0}^{\infty} p_k(t) q_{ki} \quad (t > 0;\ i = 0, 1, \dots) \tag{11}$$

for the probability distribution of the process.

REMARK 2. The quantities q_{ij} $(i, j = 0, 1, \dots)$ are called the *transition intensities.* They fulfil the relation

$$p_{ii}(h) = 1 + q_{ii}h + o(h), \quad p_{ij}(h) = q_{ij}h + o(h) \quad (i \neq j), \tag{12}$$

where the symbol $o(h)$ represents a function f of argument h such that

$$\lim_{h \to 0} \frac{f(h)}{h} = 0 \tag{13}$$

(see § 11.4).

REMARK 3. The Kolmogorov differential equations are an important tool for the treatment of Markov processes. Besides the prospective (forward) equations one can also use the *retrospective* (*backward*) *Kolmogorov differential equations* of the form

$$p'_{ij}(t) = -\sum_{k=0}^{\infty} q_{ik} p_{kj}(t) \quad (t > 0;\ i, j = 0, 1, \dots) \tag{14}$$

(then one has to interchange i and j in the assumption (c) of Theorem 2).

REMARK 4. According to (11), the equations

$$0 = \sum_{k=0}^{\infty} \pi_k q_{ki} \quad (i = 0, 1, \dots) \tag{15}$$

hold for the stationary probabilities (5).

36.3. Examples of Markov Processes

Poisson Process. *Interpretation*: $X(t)$ describes the number of occurrences of an observed event in the time interval $[0, t]$. One assumes that the observed event occurs in the interval $(t, t+h)$ of a small length h exactly once with the probability $\lambda h + o(h)$ (i.e. with the probability proportional to the interval length) and more than once with the probability $o(h)$ independently of the number of occurrences till the time t. An equivalent interpretation of the Poisson process is given in the third point of Example 36.1.1.

Transition intensities:

$$q_{ii} = -\lambda, \quad q_{i,i+1} = \lambda, \quad q_{ij} = 0, \quad (j \neq i;\ j \neq i+1). \tag{1}$$

System of Kolmogorov differential equations (36.2.11):

$$\begin{aligned} p_0'(t) &= -\lambda p_0(t), \\ p_i'(t) &= \lambda p_{i-1}(t) - \lambda p_i(t) \quad (i = 1, 2, \dots). \end{aligned} \tag{2}$$

Initial conditions:

$$p_0(0) = 1, \quad p_i(0) = 0 \quad (i = 1, 2, \dots) \tag{3}$$

(they follow from the requirement $X(0) = 0$).

Solution (see also (36.1.5)):

$$p_i(t) = \exp(-\lambda t)(\lambda t)^i / i! \quad (i = 0, 1, \dots). \tag{4}$$

The efficient estimator (see § 34.6) of the intensity λ of the Poisson process is

$$\hat{\lambda} = \frac{n}{T}, \tag{5}$$

where n is the number of occurrences of the observed event during the time interval of length T. The corresponding confidence interval with confidence level $1 - \alpha$ is

$$\left(\chi^2_{\alpha/2}(2n+2)/2T,\ \chi^2_{1-\alpha/2}(2n+2)/2T\right). \tag{6}$$

Example 1. One has registered $n = 13$ failures of a device during $T = 1000$ working hours. Modelling the number of failures by the Poisson process, we estimate its intensity (the so-called *failure intensity*) by

$$\hat{\lambda} = \frac{n}{T} = \frac{13}{1000} = 0{\cdot}013.$$

The lower and upper bounds of the 95 % confidence interval are

$$\chi^2_{0\cdot025}(2n+2)/2T = \chi^2_{0\cdot025}(28)/2000 = 0\cdot0077\,,$$
$$\chi^2_{0\cdot975}(2n+2)/2T = \chi^2_{0\cdot975}(28)/2000 = 0\cdot0222$$

so that the intensity λ lies with the 95 % confidence in the interval $[0\cdot0077, 0\cdot0222]$.

Yule Process . *Interpretation*: $X(t)$ describes the population size at the time t. One assumes that, in the interval $(t, t+h)$ of a small length h, each item of the population gives rise to exactly one item with the probability $\lambda h + o(h)$ and to more than one item with the probability $o(h)$ independently of the behaviour of the other items. In contrast to the branching process discussed in the first point of Example 36.1.1, the Yule process is defined in continuous time.

Transition probabilities:

$$q_{ii} = 1 - i\lambda, \quad q_{i,i+1} = i\lambda, \quad q_{ij} = 0, \quad (j \neq i;\ j \neq i+1)\,. \tag{7}$$

System of Kolmogorov differential equations (36.2.11):

$$p_i'(t) = (i-1)\lambda p_{i-1}(t) - i\lambda p_i(t) \quad (i = 1, 2, \dots)\,. \tag{8}$$

Initial conditions:

$$p_{k_0}(0) = 1, \quad p_i(0) = 0 \quad (i \neq k_0) \tag{9}$$

(they follow from the requirement $X(0) = k_0$, i.e., the population starts with k_0 items at time $t = 0$).

Solution:

$$p_i(t) = \begin{cases} \binom{i-1}{i-k_0} \exp(-k_0\lambda t)\,[1 - \exp(-\lambda t)]^{i-k_0} & \text{for } i \geqq k_0\,, \\ 0 & \text{for } i < k_0\,. \end{cases} \tag{10}$$

Birth-and-Death Process . *Interpretation*: $X(t)$ describes the population size at the time t. One assumes that, in the interval $(t, t+h)$ of a small length h, the population containing i items at the time t increases by exactly one item with the probability $\lambda_i h + o(h)$ $(i = 0, 1, \dots)$, decreases by exactly one item with the probability $\mu_i h + o(h)$ $(i = 1, 2, \dots)$ and increases or decreases by more than one item with the probability $o(h)$ independently of the behaviour of the other items.

Transition probabilities:

$$\begin{aligned} &q_{ii} = 1 - \lambda_i - \mu_i \quad (\mu_0 = 0), \quad q_{i,i+1} = \lambda_i, \quad q_{i,i-1} = \mu_i, \\ &q_{ij} = 0 \quad (j \neq i-1;\ j \neq i;\ j \neq i+1)\,. \end{aligned} \tag{11}$$

System of equations (36.2.15):

$$\begin{aligned} 0 &= -\lambda_0\pi_0 + \mu_1\pi_1\,, \\ 0 &= \lambda_{i-1}\pi_{i-1} - (\lambda_i + \mu_i)\pi_i + \mu_{i+1}\pi_{i+1} \quad (i = 1, 2, \dots)\,. \end{aligned} \tag{12}$$

Solution:

$$\pi_0 = \left[1 + \sum_{n=1}^{\infty} \prod_{k=1}^{n} \frac{\lambda_{k-1}}{\mu_k}\right]^{-1}, \quad \pi_i = \pi_0 \sum_{k=1}^{n} \frac{\lambda_{k-1}}{\mu_k} \quad (i = 1, 2, \dots)\,. \tag{13}$$

REMARK 1. The solution (13) is a probability distribution if the series in the formula for π_0 converges (e.g., the convergence is guaranteed if $\lambda_i = 0$ for at least one index i). One can also write the system of Kolmogorov differential equations (36.2.11) that has a unique solution in the form of a probability distribution provided that the coefficients of the individual equations form bounded sequences. The Poisson and Yule processes are special cases of the birth-and-death process.

36.4. Markov Chains

Markov chains are an analogy of Markov processes in discrete time. Let again $I = \{0, 1, 2, \dots\}$ denote the set of discrete states.

Definition 1. The stochastic process (random sequence) $\{X_n\colon n = 0, 1, \dots\}$ is called the *Markov chain* if

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = j \mid X_n = i) \tag{1}$$

for arbitrary $i_0, i_1, \dots, i_{n-1}, i, j \in I$ (*Markov property*). If the transition probabilities (1) do not depend on n, then such Markov chain is called *homogeneous* and the *transition probabilities* (1) are denoted by

$$p_{ij} = P(X_{n+1} = j \mid X_n = i)\,. \tag{2}$$

Further important characteristics of a Markov chain are *k-step transition probabilities*

$$p_{ij}(k) = P(X_{n+k} = j \mid X_n = i) \quad (k = 0, 1, \dots) \tag{3}$$

and probabilities describing the probability distribution of the chain at time n,

$$p_i(n) = P(X_n = i) \quad (i = 0, 1, \dots)\,. \tag{4}$$

It is usual to arrange the transition probabilities in the form of the so-called *transition matrix* $\boldsymbol{P} = (p_{ij})$ and *k-step transition matrix* $\boldsymbol{P}(k) = (p_{ij}(k))$.

Theorem 1. *For a homogeneous Markov chain we have*

(i)

$$p_{ij}(k_1 + k_2) = \sum_{k=0}^{\infty} p_{ik}(k_1) p_{kj}(k_2) \tag{5}$$

(the so-called *Chapman-Kolmogorov equations*);

(ii)

$$\boldsymbol{P}(k) = \boldsymbol{P}^k; \tag{6}$$

(iii)

$$p_i(k) = \sum_{j=0}^{\infty} p_j(0) p_{ji}(k) \quad (k = 0, 1, \ldots). \tag{7}$$

Example 1. Let us consider Bernoulli trials with the probability of success p and the probability of failure $q = 1 - p$ (see § 33.6). Let $X_n = k$ provided that the last failure with respect to the n-th trial occurred in the $(n-k)$-th trial which was followed by a series of k successes $(k = 0, 1, \ldots, n)$. Then $\{X_n : \ n = 1, 2, \ldots\}$ is the Markov chain with the transition matrix

$$\boldsymbol{P} = \begin{bmatrix} q, & p, & 0, & 0, & 0, & \ldots \\ q, & 0, & p, & 0, & 0, & \ldots \\ q, & 0, & 0, & p, & 0, & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix},$$

i.e. $p_{00} = q$, $p_{01} = p$, $p_{02} = 0$, etc. Further

$$\boldsymbol{P}(k) = \boldsymbol{P}^k = \begin{bmatrix} q, & qp, & qp^2, & \ldots, & qp^{k-1}, & p^k, & 0, & 0, & \ldots \\ q, & qp, & qp^2, & \ldots, & qp^{k-1}, & 0, & p^k, & 0, & \ldots \\ q, & qp, & qp^2, & \ldots, & qp^{k-1}, & 0, & 0, & p^k, & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix}.$$

As $k \to \infty$, $\boldsymbol{P}(k)$ converges to the matrix in which all the elements of the j-th column are equal to qp^{j-1} $(j = 1, 2, \ldots)$.

If we observe a homogeneous Markov chain with the finite set of states $I = \{0, 1, \ldots, m\}$ and if n_{ij} denotes the number of observed transitions from the state i to the state j, then the maximum likelihood estimates (see § 34.7) of the transition probabilities are

$$\hat{p}_{ij} = n_{ij} \Big/ \sum_{k=0}^{m} n_{ik} \qquad (i, j = 0, 1, \ldots, m). \tag{8}$$

If the numerator in (8) is zero, then we put $\hat{p}_{ij} = 0$. Further aspects of the theory of Markov chains including the classification of states and problems connected with stationary distributions are given e.g. in [80], [81], [141].

B. QUEUEING THEORY

36.5. Service Systems

Queueing theory deals with the following situation: customers come to a *service system* where service channels (servers) provide service of a certain type, the customers are either served immediately or must wait for service, and then they leave the service system. Examples of service systems are telephone exchanges, petrol filling stations, computers shared by several users, and others. The objective of queueing theory is to obtain basic characteristics of a service system, e.g. the probability distribution of the number of the customers who are in the system at time t, distribution of the *waiting time* (i.e. the time that a customer spends in the system waiting for the service), distribution of the *service time*, distribution of the *waiting time in system* (i.e. the time that a customer spends in the system on the whole), the distribution of *busy periods* (the service channel does not work between these periods due to the lack of customers) and others. It is also important to investigate under which conditions the system achieves the so-called *stationary traffic* with stabilized behaviour.

TABLE 36.1

Symbol	X	Y
M	Poisson arrival process (i.e. exponential distribution of the times between arrivals)	exponential distribution of service time
E	Erlang distribution of the times between arrivals	Erlang distribution of service time
D	the constant times between arrivals	constant service time
G	general arrival process	general distribution of service time
GI	recurrent arrival process	

In order to distinguish among various service systems, one uses the so-called *Kendall's classification* $X/Y/c$ where X denotes the type of the stochastic process that describes the arrival of customers to the service system (the so-called *arrival process*), Y denotes the type of the probability distribution of the service time and

c is a positive integer (or ∞) that gives the number of service channels. Tab. 36.1 surveys the most frequent symbols for X and Y. In particular, the symbol GI is used if the times between the arrivals of customers are independent random variables with an identical (but otherwise unspecified) probability distribution.

Besides the basic Kendall's classification one distinguishes *loss systems*, in which the customer does not wait for service and leaves the system if the service channels are engaged. In case of waiting, there are various queue disciplines (first come-first served discipline, last come-first served discipline, discipline with priorities, discipline with balking, and others). The corresponding theory can be found e.g. in [96], [195], [406].

36.6. Examples of Service Systems

System $M/M/n$. *Interpretation*: The arrival of customers to the system is the Poisson process with the intensity λ (λ is the average number of customers who enter the system in a time unit). If some of the service channels is not engaged, then the customer is served immediately, the service time having the exponential distribution with the mean value $1/\mu$ (i.e. μ customers are served in a time unit on the average). In the opposite case the customers wait for service in a queue of unbounded length.

Condition for the existence of stationary traffic:

$$\varrho < 1\,, \tag{1}$$

where $\varrho = \beta/n$, $\beta = \lambda/\mu$ (ϱ is called the *traffic intensity*). For the stationary traffic, let π_i ($i = 0, 1, \dots$) denote the probability of the event that the system contains i customers (who wait in a queue or are served), W denote the waiting time and R denote the waiting time in system.

Probability distribution of the number of customers:

$$\pi_i = \begin{cases} \left[\sum_{j=0}^{n-1} \frac{\beta^j}{j!} + \frac{n^n}{n!} \sum_{j=n}^{\infty} \varrho^j\right]^{-1} & \text{for } i = 0, \\ \pi_0 \frac{\beta^i}{i!} & \text{for } i = 1, \dots, n, \\ \pi_0 \frac{n^n}{n!} \varrho^i & \text{for } i = n+1, n+2, \dots . \end{cases} \tag{2}$$

Probability of immediate service (without waiting):

$$\Pi_1 = \sum_{i=0}^{n-1} \pi_i = \pi_0 \sum_{i=0}^{n-1} \frac{\beta^i}{i!}\,. \tag{3}$$

Probability of waiting for service:

$$\Pi_2 = \sum_{i=n}^{\infty} \pi_i = \frac{\pi_n}{1-\varrho}. \tag{4}$$

Average number of customers waiting for service:

$$P_1 = \sum_{i=1}^{\infty} i\pi_{n+i} = \frac{\pi_n \varrho}{(1-\varrho)^2}. \tag{5}$$

Average number of engaged service channels:

$$P_2 = \sum_{i=1}^{n} i\pi_i + \sum_{i=1}^{\infty} n\pi_{n+i} = \beta. \tag{6}$$

Average number of customers in system:

$$P_3 = \sum_{i=1}^{\infty} i\pi_i = P_1 + P_2 = \frac{\pi_n \varrho}{(1-\varrho)^2} + \beta. \tag{7}$$

Average waiting time:

$$\mathrm{E}(W) = \frac{\pi_n}{(1-\varrho)(n\mu - \lambda)}. \tag{8}$$

Average waiting time in system:

$$\mathrm{E}(R) = \mathrm{E}(W) + 1/\mu. \tag{9}$$

REMARK 1. The system $M/M/n$ is a special case of the birth-and-death process (see § 36.3), e.g., one can put $\lambda_i = \lambda \quad (i = 0, 1, \dots)$ and $\mu_i = \mu \quad (i = 1, 2, \dots)$ for $n = 1$.

REMARK 2. The system $M/M/\infty$ with an unbounded number of service channels acheives the stationary traffic for arbitrary λ and μ, and

$$\pi_i = \mathrm{e}^{-\beta}\frac{\beta^i}{i!} \quad (i = 0, 1, \dots). \tag{10}$$

It means that the stationary distribution is the Poisson distribution with parameter β.

Example 1. Approximately $\lambda = 4$ customers enter a system $M/M/3$ in one hour. The service time for one customer is approximately half an hour so that $\mu = 2$ and

$$\beta = \frac{\lambda}{\mu} = \frac{4}{2} = 2, \quad \varrho = \frac{\beta}{n} = \frac{2}{3}.$$

Since the condition (1) is fulfilled we obtain the following characteristics of the stationary traffic by the successive application of the formulae (2)–(9):

$$\begin{aligned} \pi_0 &= 0{\cdot}111, \quad \pi_1 = 0{\cdot}222, \quad \pi_2 = 0{\cdot}222, \quad \pi_3 = 0{\cdot}148, \quad \text{etc.}, \\ \Pi_1 &= 0{\cdot}556, \quad \Pi_2 = 0{\cdot}444, \\ P_1 &= 0{\cdot}889, \quad P_2 = 2, \quad P_3 = 2{\cdot}889 \quad \text{(customers)}, \\ \mathrm{E}(W) &= 0{\cdot}222, \quad \mathrm{E}(R) = 0{\cdot}722 \quad \text{(hours)}. \end{aligned}$$

System $M/M/n$ with the Bounded Length r of Queue. *Interpretation*: The interpretation is similar to that of the system $M/M/n$ above. The only difference consists in the fact that the system is closed for further customers if the length of the queue achieves the value r.

Condition for the existence of stationary traffic:

The system achieves the stationary traffic for arbitrary λ and μ.

Probability distribution of the number of customers:

$$\pi_i = \begin{cases} \left[\sum_{j=0}^{n} \frac{\beta^j}{j!} + \frac{\beta^n}{n!} \sum_{j=1}^{r} \left(\frac{\beta}{n}\right)^j\right]^{-1} & \text{for } i = 0, \\ \pi_0 \frac{\beta^i}{i!} & \text{for } i = 1, \ldots, n, \\ \pi_0 \frac{\beta^i}{n! n^{i-n}} & \text{for } i = n+1, \ldots, n+r, \\ 0 & \text{for } i = n+r+1, n+r+2, \ldots. \end{cases} \tag{11}$$

Average waiting time (provided that a customer can enter the system):

$$\mathrm{E}(W) = \frac{(\beta^n/n!) \sum_{i=0}^{r-1} (i+1)\varrho^i}{n\mu\left[\sum_{i=0}^{n-1} (\beta^i/i!) + (1/n!)(1-\varrho^r)/(1-\varrho)\right]}. \tag{12}$$

REMARK 3. In particular, if $r = 0$, then no queues can arise. In this case

$$\pi_i = \begin{cases} (\beta^i/i!) \Big/ \sum_{j=0}^{n} (\beta^j/j!) & \text{for } i = 0, 1, \ldots, n, \\ 0 & \text{for } i = n+1, n+2, \ldots. \end{cases} \tag{13}$$

System $M/D/1$. *Interpretation*: The interpretation is similar to that of the system $M/M/1$ with the only exception that the service time of each customer is equal to a constant value $1/\mu$ (i.e., exactly μ customers are served in a time unit). The corresponding process $\{X(t)\colon t \geqq 0\}$ that describes the number of customers in the system is not the Markov process in this case.

Condition for the existence of stationary traffic:

$$\varrho = \lambda/\mu < 1\,. \tag{14}$$

Probability distribution of the number of customers:

$$\pi_i = \begin{cases} 1-\varrho & \text{for } i = 0, \\ (1-\varrho)(\mathrm{e}^{\varrho}-1) & \text{for } i = 1, \\ (1-\varrho)\sum\limits_{j=1}^{i-1}(-1)^{i-j}\mathrm{e}^{j\varrho}\left[\dfrac{(j\varrho)^{i-j}}{(i-j)!}+\dfrac{(j\varrho)^{i-j-1}}{(i-j-1)!}\right]+(1-\varrho)\mathrm{e}^{i\varrho} & \text{for } i = 2,3\ldots\,. \end{cases} \tag{15}$$

Average waiting time:

$$\mathrm{E}(W) = \frac{\varrho}{2\mu(1-\varrho)}\,. \tag{16}$$

C. STATIONARY PROCESSES

36.7. Correlation Properties of Stationary Processes

Stationary stochastic processes, whose characteristics are invariant with respect to a shift in time, play an important role among stochastic processes both from the theoretical and practical point of view (see, e.g., [9], [11], [54], [59], [75], [102], [164], [183], [186], [192], [206], [207], [257], [265], [377], [399], [404], [503], [506]). From now on, let $\{X(t)\colon t \in T\}$ be a stochastic process where T is a subset of the real axis and the random variables $X(t)$ can attain real values.

Definition 1. If $\mathrm{E}[|X(t)|] < \infty$ for all $t \in T$, then the function

$$\mu(t) = \mathrm{E}[X(t)] \quad (t \in T) \tag{1}$$

is called the *mean of stochastic process* (or *mean function*).

Definition 2. If $\mathrm{E}[X^2(t)] < \infty$ for all $t \in T$, then the function

$$R(s,t) = \operatorname{cov}[X(s), X(t)] = \mathrm{E}\{[X(s)-\mu(s)][X(t)-\mu(t)]\} \quad (s,t \in T) \tag{2}$$

is called the *autocovariance function* of stochastic process and the function

$$B(s, t) = \varrho[X(s), X(t)] = R(s, t)/[R(s, s)R(t, t)]^{1/2} \quad (s, t \in T) \tag{3}$$

is called the *autocorrelation function* of stochastic process.

REMARK 1. The autocovariance and autocorrelation functions generalize the concept of the covariance and correlation matrix, respectively (see Definition 33.5.7). Obviously $R(t, t) = \operatorname{var}[X(t)]$.

Theorem 1 (*Properties of Autocovariance Function*). *For the autocovariance function of a stochastic process we have*

(i)

$$|R(s, t)| \leqq [R(s, s)R(t, t)]^{1/2} \quad (s, t \in T) \quad (\textit{Schwarz Inequality}); \tag{4}$$

(ii)

$$R(s, t) = R(t, s) \quad (s, t \in T); \tag{5}$$

(iii) *the autocovariance function is non-negative definite, i.e.*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j R(t_i, t_j) \geqq 0 \tag{6}$$

holds for arbitrary $c_1, \ldots, c_n$ *real,* $t_1, \ldots, t_n \in T$, n *a positive integer.*

Properties of the autocorrelation function are analogous.

Definition 3. Let $\mathrm{E}[X^2(t)] < \infty$ for all $t \in T$. Then the stochastic process $\{X(t): t \in T\}$ is called *stationary* if its mean function $\mu(t)$ is constant and its autocovariance function $R(s, t)$ depends only on the difference $t-s$ of its arguments. In this case one writes

$$\mu(t) = \mu, \quad R(s, t) = R(t - s) \quad (s, t \in T). \tag{7}$$

REMARK 2. The autocovariance and autocorrelation functions of a stationary process depend only on the lag and not on the absolute time position (they are invariant with respect to shifts in time); therefore they can be written as functions $R(t)$ and $B(t)$ of one argument. In particular, $R(0) = \operatorname{var}[X(t)]$ and $B(t) = = R(t)/R(0)$. According to (5) one has $R(t) = R(-t)$. Sometimes in this case one speaks of a *weak stationarity* in contrast to a *strict stationarity* where not only the first and second moments but even the probability distribution of the process (i.e. the probability distribution of all random vectors of the type $(X(t_1), \ldots, X(t_n))'$)

is invariant with respect to shifts in time. If $\mathrm{E}\left[X^2(t)\right] < \infty$ for all $t \in T$, then the strict stationarity implies the weak stationarity. In case of a *normal stochastic process*, where all random vectors of the type $(X(t_1), \ldots, X(t_n))'$ have the multivariate normal distribution, both the types of stationarity are equivalent.

Example 1.

1. *White noise* (see point 2 of Example 36.1.1). It is a stationary random sequence with the zero mean function and the autocovariance function of the form

$$R(t) = \begin{cases} \sigma^2 & \text{for } t = 0, \\ 0 & \text{for } t \text{ an integer, } t \neq 0. \end{cases} \tag{8}$$

2. *Moving average process MA(q)*. It is a stationary random sequence $\{X_t\colon t \text{ an integer}\}$ that has the model

$$X_t = Y_t + \beta_1 Y_{t-1} + \cdots + \beta_q Y_{t-q} \quad (t \text{ an integer}), \tag{9}$$

where $\{Y_t\colon t \text{ an integer}\}$ is a white noise with the variance $\mathrm{var}(Y_t) = \sigma^2$ and $\beta_1, \ldots, \beta_q$, σ^2 $(\beta_q \neq 0,\ \sigma^2 > 0)$ are real parameters. The positive integer q is called the *order of moving average*. The mean function of the process $MA(q)$ is zero and its autocovariance function is

$$R(t) = \begin{cases} \sigma^2 \sum\limits_{i=0}^{q-|t|} \beta_{i+|t|}\beta_i & \text{for } |t| = 0, 1, \ldots, q, \\ 0 & \text{for } t \text{ an integer, } |t| > q. \end{cases} \tag{10}$$

3. *Autoregressive process AR(p)*. It is a stationary random sequence $\{X_t\colon t \text{ an integer}\}$ that has the model

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + Y_t \quad (t \text{ an integer}), \tag{11}$$

where $\{Y_t\colon t \text{ an integer}\}$ is a white noise with the variance $\mathrm{var}(Y_t) = \sigma^2$ and $\alpha_1, \ldots, \alpha_p$, σ^2 $(\alpha_p \neq 0,\ \sigma^2 > 0)$ are real parameters such that the roots $\lambda_1, \ldots$ $\ldots, \lambda_p$ of the polynomial

$$\lambda^p - \alpha_1 \lambda^{p-1} - \cdots - \alpha_{p-1}\lambda - \alpha_p \tag{12}$$

fulfil the condition

$$|\lambda_i| < 1 \quad (i = 1, \ldots, p). \tag{13}$$

The positive integer p is called the *order of autoregression*. The mean function of the process $AR(p)$ is zero and its autocovariance function fulfils the so-called

Yule-Walker equations of the form

$$R(t) = \begin{cases} \alpha_1 R(t-1) + \cdots + \alpha_p R(t-p) & \text{for } t = 1, 2, \ldots, \\ \alpha_1 R(1) + \cdots + \alpha_p R(p) + \sigma^2 & \text{for } t = 0. \end{cases} \tag{14}$$

In particular, the process $AR(1)$ has

$$R(t) = \frac{\sigma^2}{1-\alpha_1^2}\alpha_1^{|t|} \quad (t \text{ an integer}). \tag{15}$$

4. *Mixed process* $ARMA(p, q)$. It is a stationary random sequence $\{X_t\colon t \text{ an integer}\}$ that has the model

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + Y_t + \beta_1 Y_{t-1} + \cdots + \beta_q Y_{t-q} \quad (t \text{ an integer}), \tag{16}$$

where the notation and assumptions used for the processes $MA(q)$ and $AR(p)$ above remain valid.

5. *Harmonic process.* It is a stationary random function $\{X(t)\colon t \text{ real}\}$ that has the model

$$X(t) = \sum_{i=1}^{K} [A_i \cos(\lambda_i t) + B_i \sin(\lambda_i t)] \quad (t \text{ real}), \tag{17}$$

where A_i, B_i are random variables, for which

$$\begin{aligned} &\mathrm{E}(A_i) = \mathrm{E}(B_i) = 0, \quad \mathrm{var}(A_i) = \mathrm{var}(B_i) = \sigma_i^2, \\ &\mathrm{E}(A_i A_j) = \mathrm{E}(B_i B_j) = 0 \quad (i \neq j), \\ &\mathrm{E}(A_i B_j) = 0, \end{aligned} \tag{18}$$

and $\sigma_i^2 > 0$, λ_i $(i = 1, \ldots, K)$ are real parameters. The mean function of the harmonic process is zero and its autocovariance function is

$$R(t) = \sum_{i=1}^{K} \sigma_i^2 \cos(\lambda_i t) \quad (t \text{ real}). \tag{19}$$

REMARK 3. Let $\{\boldsymbol{X}(t)\colon t \in T\}$ be a p-variate stochastic process, i.e. $\boldsymbol{X}(t) = (X_1(t), \ldots, X_p(t))'$ are p-variate random vectors. Let $\mathrm{E}[X_k^2(t)] < \infty$ for all $t \in T$ and $k = 1, \ldots, p$. Then one defines the *mean* of this process (or *mean function*) as

$$\boldsymbol{\mu}(t) = (\mu_1(t), \ldots, \mu_p(t))' = \mathrm{E}[\boldsymbol{X}(t)] \quad (t \in T) \tag{20}$$

and the *matrix autocovariance function* as

$$\boldsymbol{R}(s, t) = (R_{ij}(s, t)) = \mathrm{E}\{[\boldsymbol{X}(s) - \boldsymbol{\mu}(s)][\boldsymbol{X}(t) - \boldsymbol{\mu}(t)]'\} \quad (s, t \in T), \tag{21}$$

where $\boldsymbol{R}(s, t)$ is a $p \times p$ matrix and

$$R_{ij}(s, t) = \mathrm{E}\{[X_i(s) - \mu_i(s)][X_j(t) - \mu_j(t)]\} \quad (s, t \in T; \quad i, j = 1, \ldots, p) \tag{22}$$

is the so-called *cross-covariance function* of the processes $\{X_i(t): t \in T\}$ and $\{X_j(t): t \in T\}$. In the stationary case one writes $\boldsymbol{\mu}$, $\boldsymbol{R}(t)$ and $R_{ij}(t)$.

Estimation of Correlation Characteristics . As the estimation of correlation characteristics of a stationary stochastic process is concerned, we mostly have at our disposal only a single realization $\{x_t: t = 1, \ldots, T\}$ in the case of a random sequence or $\{x(t): 0 \leqq t \leqq T\}$ in the case of a random function. Then the mean μ, autocovariance function $R(t)$ and autocorrelation function $B(t)$ can be estimated by

$$\hat{\mu} = \frac{1}{T} \sum_{t=1}^{T} x_t , \tag{23}$$

$$\hat{R}(t) = \frac{1}{T} \sum_{i=1}^{T-|t|} (x_i - \hat{\mu})(x_{i+|t|} - \hat{\mu}) \quad (t \text{ an integer}), \tag{24}$$

$$\hat{B}(t) = \hat{R}(t)/\hat{R}(0) \quad (t \text{ an integer}), \tag{25}$$

respectively, for the random sequence, or

$$\hat{\mu} = \frac{1}{T} \int_0^T x(t)\, \mathrm{d}t , \tag{26}$$

$$\hat{R}(t) = \frac{1}{T} \int_0^{T-|t|} [x(\tau) - \hat{\mu}]\, [x(\tau + |t|) - \hat{\mu}]\, \mathrm{d}\tau \quad (t \text{ real}), \tag{27}$$

$$\hat{B}(t) = \hat{R}(t)/\hat{R}(0) \quad (t \text{ real}) \tag{28}$$

respectively, for the random function.

REMARK 4. The estimates (23) and (26) are unbiased and the estimates (24) and (27) have only a small bias (see Definition 34.6.1). The coefficient $1/T$ in (24) and (27) is sometimes replaced by $1/(T - |t|)$ (this reduces the bias of these estimates but their mean square error increases). The estimates (23)–(28) have suitable properties for the so-called *ergodic processes* (see e.g. [9], [11], [122], [164], [377], [404]).

36.8. Spectral Properties of Stationary Processes

The *spectral* (or *Fourier*) *analysis* of stochastic processes considers a process as a mixture of periodic components (e.g. sine or cosine waves) with different

frequencies and, among others, it enables us to obtain a conception of intensity with which particular frequencies are contained in the investigated stochastic process (one speaks of the so-called *spectrum* of the stochastic process). From now on, let $\{X(t): t \in T\}$ be a stationary stochastic process with the zero mean function and autocovariance function $R(t)$.

Spectral Decomposition of Autocovariance Function.

Theorem 1 (*Spectral Decomposition of Autocovariance Function*).

(i) *Let T be the set of all integers. Then the autocovariance function $R(t)$ can be expressed in the form*

$$R(t) = \int_{-\pi}^{\pi} \cos(\lambda t)\, \mathrm{d}F(\lambda) \quad (t \text{ an integer}), \tag{1}$$

where $F(\lambda)$ is a non-decreasing function that is continuous from the right, and $F(-\pi) = 0$ and $F(\pi) = R(0)$. Moreover, the function $F(\lambda)$ with these properties is determined uniquely.

(ii) *Let T be the real axis. Let the autocovariance function $R(t)$ be continuous at the point $t = 0$. Then $R(t)$ can be expressed in the form*

$$R(t) = \int_{-\infty}^{\infty} \cos(\lambda t)\, \mathrm{d}F(\lambda) \quad (t \text{ real}), \tag{2}$$

where $F(\lambda)$ is a non-decreasing function that is continuous from the right, and $\lim_{\lambda \to -\infty} F(\lambda) = 0$ and $\lim_{\lambda \to \infty} F(\lambda) = R(0)$. Moreover, the function $F(\lambda)$ with these properties is determined uniquely.

Definition 1. The function $F(\lambda)$ from the decomposition (1) or (2) is called the *spectral distribution function.* If it can be written as

$$F(\lambda) = \int_{-\pi}^{\lambda} f(x)\, \mathrm{d}x \quad (-\pi \leqq \lambda \leqq \pi) \tag{3}$$

in the case (1) or as

$$F(\lambda) = \int_{-\infty}^{\lambda} f(x)\, \mathrm{d}x \quad (-\infty \leqq \lambda \leqq \infty) \tag{4}$$

in the case (2), then the function $f(\lambda)$ is called the *spectral density.*

REMARK 1. One has

$$\operatorname{var}(X_t) = R(0) = \int_{-\pi}^{\pi} \mathrm{d}F(\lambda) \tag{5}$$

and

$$\operatorname{var}[X(t)] = R(0) = \int_{-\infty}^{\infty} \mathrm{d}F(\lambda)\,. \tag{6}$$

If the spectral density $f(\lambda)$ exists, then one can write

$$R(t) = \int_{-\pi}^{\pi} \cos(\lambda t) f(\lambda)\,\mathrm{d}\lambda = 2\int_{0}^{\pi} \cos(\lambda t) f(\lambda)\,\mathrm{d}\lambda \quad (t \text{ an integer}) \tag{7}$$

and

$$R(t) = \int_{-\infty}^{\infty} \cos(\lambda t) f(\lambda)\,\mathrm{d}\lambda = 2\int_{0}^{\infty} \cos(\lambda t) f(\lambda)\,\mathrm{d}\lambda \quad (t \text{ real}). \tag{8}$$

The spectral density can be chosen so that it is an even function, i.e., $f(\lambda) = f(-\lambda)$.

Theorem 2.

(i) *Let T be the set of all integers. If*

$$\sum_{t=0}^{\infty} |R(t)| < \infty\,, \tag{9}$$

then the spectral density $f(\lambda)$ exists. If $f(\lambda)$ is continuous in $[-\pi, \pi]$, then

$$f(\lambda) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} R(t)\cos(\lambda t) = \frac{1}{2\pi}\left[R(0) + 2\sum_{t=1}^{\infty} R(t)\cos(\lambda t)\right] \quad (-\pi \leqq \lambda \leqq \pi). \tag{10}$$

(ii) *Let T be the real axis. Let $R(t)$ be continuous at the point $t = 0$. If*

$$\int_{0}^{\infty} |R(t)|\,\mathrm{d}t < \infty\,, \tag{11}$$

then the spectral density $f(\lambda)$ exists and it holds

$$f(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty} R(t)\cos(\lambda t)\,\mathrm{d}t = \frac{1}{\pi}\int_{0}^{\infty} R(t)\cos(\lambda t)\,\mathrm{d}t \quad (-\infty < \lambda < \infty). \tag{12}$$

REMARK 2. The formulae (10) and (12) are called the *inverse formulae.* They enable us to derive the form of the spectral density on the ground of the autocovariance function.

Spectral Decomposition of Stationary Process . The *spectral decomposition of a stationary process* is closely related to the spectral decomposition of an autocovariance function. For example the spectral decomposition of a random sequence $\{X_t\colon t \text{ an integer}\}$ has the form

$$X_t = \int_{-\pi}^{\pi} \cos(\lambda t)\,\mathrm{d}V(\lambda) + \int_{-\pi}^{\pi} \sin(\lambda t)\,\mathrm{d}W(\lambda) \quad (t \text{ an integer}), \tag{13}$$

where the symbols $\mathrm{d}V(\lambda)$ and $\mathrm{d}W(\lambda)$ can be considered as stochastic processes that are labelled with a continuous index λ. A simplified interpretation of the relation (13) is such that the random sequence can be expressed as a mixture of periodic components of the form

$$\mathrm{d}V(\lambda)\cos(\lambda t) + \mathrm{d}W(\lambda)\sin(\lambda t)\,. \tag{14}$$

The amplitudes of these periodic components are random and

$$\begin{gathered} \mathrm{E}\,[\mathrm{d}V(\lambda)] = \mathrm{E}\,[\mathrm{d}W(\lambda)] = 0, \\ \mathrm{E}\,[\mathrm{d}V(\lambda_1)\,\mathrm{d}V(\lambda_2)] = \mathrm{E}\,[\mathrm{d}W(\lambda_1)\,\mathrm{d}W(\lambda_2)] = 0 \quad (\lambda_1 \neq \lambda_2), \\ \mathrm{E}\,[\mathrm{d}V(\lambda_1)\,\mathrm{d}W(\lambda_2)] = 0 \end{gathered} \tag{15}$$

($V(\lambda)$ and $W(\lambda)$ are called the *stochastic processes with independent increments*). The variance of the periodic component (14) is

$$\operatorname{var}\,[\mathrm{d}V(\lambda)\cos(\lambda t) + \mathrm{d}W(\lambda)\sin(\lambda t)] = \mathrm{d}F(\lambda)\,. \tag{16}$$

If compared with (5), $\mathrm{d}F(\lambda)$ (or $f(\lambda)\,\mathrm{d}\lambda$) expresses the intensity (in terms of variances) with which the periodic component (14) corresponding to the frequency λ is contained in the decomposition (13).

Example 1. We shall give spectral characteristics of the stochastic processes from Example 36.7.1.

1. *White noise.* According to (36.7.8) and to the inverse formula (10) one obtains

$$f(\lambda) = \frac{1}{2\pi}\left[R(0) + 2\sum_{t=1}^{\infty} R(t)\cos(\lambda t)\right] = \frac{\sigma^2}{2\pi} \quad (-\pi \leqq \lambda \leqq \pi)\,. \tag{17}$$

The spectral density of the white noise is a constant function, i.e., particular frequencies are contained in the spectrum of this process with the same intensities.

2. *Moving average process* $MA(q)$:

$$f(\lambda) = \frac{\sigma^2}{2\pi}\,|\mathrm{e}^{\mathrm{i}q\lambda} + \beta_1 \mathrm{e}^{\mathrm{i}(q-1)\lambda} + \cdots + \beta_q|^2 \quad (-\pi \leqq \lambda \leqq \pi)\,. \tag{18}$$

3. *Autoregressive process AR(p)*:

$$f(\lambda) = \frac{\sigma^2}{2\pi} \frac{1}{\left|e^{ip\lambda} - \alpha_1 e^{i(p-1)\lambda} - \cdots - \alpha_p\right|^2} \quad (-\pi \leqq \lambda \leqq \pi). \tag{19}$$

In particular, the process $AR(1)$ has

$$f(\lambda) = \frac{\sigma^2}{2\pi} \frac{1}{1 + \alpha_1^2 - 2\alpha_1 \cos\lambda} \quad (-\pi \leqq \lambda \leqq \pi). \tag{20}$$

If $0 < \alpha_1 < 1$, then the function (20) is decreasing in $[0, \pi]$ (i.e., lower frequencies prevail in the spectrum, see Fig. 36.1a for $\alpha_1 = 0{\cdot}75$); if $-1 < \alpha_1 < 0$, then the function (20) is increasing in $[0, \pi]$ (i.e., higher frequencies prevail in the spectrum, see Fig. 36.1b for $\alpha_1 = -0{\cdot}75$).

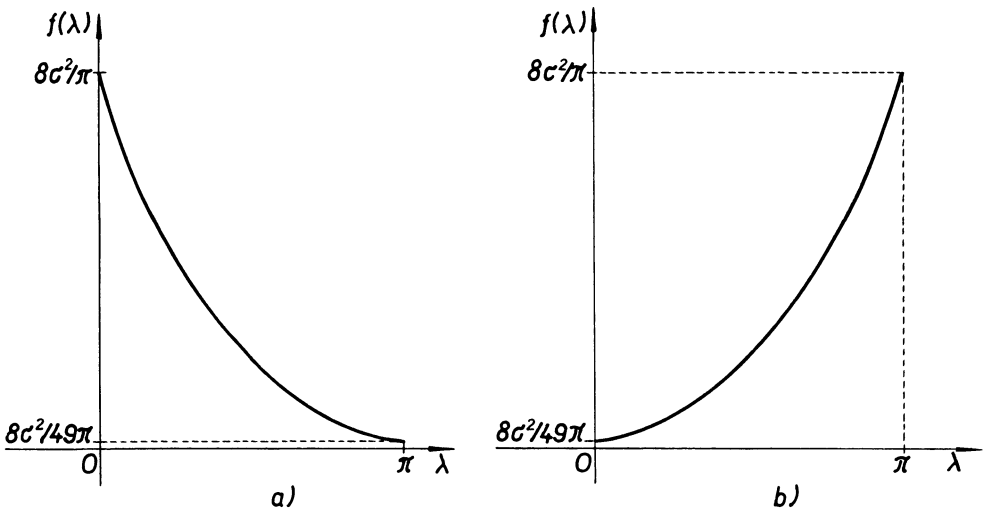


Fig. 36.1 a, b. Spectral density of the process $AR(1)$ for (a) $\alpha_1 = 0{\cdot}75$ and (b) $\alpha_1 = -0{\cdot}75$ (the function $f(\lambda)$ is even and therefore its graph for $-\pi \leqq \lambda < 0$ is omitted).

4. *Mixed process ARMA(p, q)*:

$$f(\lambda) = \frac{\sigma^2}{2\pi} \frac{\left|e^{iq\lambda} + \beta_1 e^{i(q-1)\lambda} + \cdots + \beta_q\right|^2}{\left|e^{ip\lambda} - \alpha_1 e^{i(p-1)\lambda} - \cdots - \alpha_p\right|^2} \quad (-\pi \leqq \lambda \leqq \pi). \tag{21}$$

5. *Harmonic process.* The spectral distribution function is a jump function with jumps of magnitudes $\sigma_i^2/2$ at the points $\pm\lambda_i$ $(i = 1, \ldots, K)$. The spectral density does not exist. The defining relation (36.7.17) expresses directly the spectral decomposition of this process.

REMARK 3. Spectral analysis is also used for multivariate stochastic processes. It enables us to obtain relations among spectra of the corresponding univariate processes (see e.g. [9], [207], [265], [377]).

Periodogram . Let $\{x_t\colon\ t = 1, \ldots, T\}$ be a realization of a stationary random sequence or let $\{x(t)\colon\ 0 \leqq t \leqq T\}$ be a continuous realization of a stationary random function.

Definition 2. The function

$$I(\lambda) = \frac{1}{2\pi T}\left\{\left[\sum_{t=1}^{T} x_t \cos(\lambda t)\right]^2 + \left[\sum_{t=1}^{T} x_t \sin(\lambda t)\right]^2\right\} \quad (-\pi \leqq \lambda \leqq \pi) \tag{22}$$

or

$$I(\lambda) = \frac{1}{2\pi T}\left\{\left[\int_0^T x(t)\cos(\lambda t)\,dt\right]^2 + \left[\int_0^T x(t)\sin(\lambda t)\,dt\right]^2\right\} \quad (-\infty < \lambda < \infty) \tag{23}$$

is called the *periodogram.*

REMARK 4. The periodogram can be expressed equivalently as

$$I(\lambda) = \frac{1}{2\pi}\left[\hat{R}(0) + 2\sum_{t=1}^{T-1} \hat{R}(t)\cos(\lambda t)\right] \quad (-\pi \leqq \lambda \leqq \pi) \tag{24}$$

or

$$I(\lambda) = \frac{1}{\pi}\int_0^T \hat{R}(t)\cos(\lambda t)\,dt \quad (-\infty < \lambda < \infty), \tag{25}$$

where $\hat{R}(t)$ is the estimate of the autocovariance function according to (36.7.24) or (36.7.27).

REMARK 5. The periodogram plays an important role in estimation of spectral densities. It is also used if we treat random sequences of the form

$$X_t = \sum_{i=1}^{r} [\alpha_i \cos(\lambda_i t) + \beta_i \sin(\lambda_i t)] + Y_t, \tag{26}$$

where α_i, β_i, λ_i $(0 < \lambda_i \leqq \pi)$ are unknown parameters and Y_t is a white noise with the variance $\sigma^2 > 0$ (or a more general stationary process). Since the periodogram constructed for a realization of (26) attains relatively large values at the points $\lambda_1, \ldots, \lambda_r$ it enables us to find "hidden frequencies" $\lambda_1, \ldots, \lambda_r$ as well as the number r. The recommended statistical procedure is called *Fisher's test of periodicity* (see e.g. [9], [377]). The remaining parameters α_i, β_i $(i = 1, \ldots, r)$ and σ^2 can be estimated as parameters of the linear regression model (see § 35.2).

Estimation of Spectral Density . Let $\{x_t: t = 1, \ldots, T\}$ be a realization of a stationary random sequence or let $\{x(t): 0 \leqq t \leqq T\}$ be a continuous realization of a stationary random function. Let the spectral density $f(\lambda)$ exist. Despite the fact that the periodogram $I(\lambda)$ is the asymptotically unbiased estimator (see Definition 34.6.3) of the spectral density $f(\lambda)$ it is not, in general, its consistent estimator (see Definition 34.6.4). Therefore the periodogram is usually modified with the help of a system of suitable weights w_t or $w(t)$ so that the spectral density can be estimated by

$$\hat{f}(\lambda) = w_0 \hat{R}(0) + 2 \sum_{t=1}^{T-1} w_t \hat{R}(t) \cos(\lambda t) \tag{27}$$

or

$$\hat{f}(\lambda) = \int_0^T w(t) \hat{R}(t) \cos(\lambda t)\, dt\,. \tag{28}$$

Example 2. The following estimates of the spectral density of a stationary random sequence are frequently used in practice:

Parzen's estimator: see (27) with weights

$$w_t = \begin{cases} \dfrac{1}{2\pi}\left[1 - \dfrac{6t^2}{m^2}\left(1 - \dfrac{t}{m}\right)\right] & \text{for } t = 0, 1, \ldots, \dfrac{m}{2}\,, \\ \dfrac{1}{\pi}\left(1 - \dfrac{t}{m}\right)^3 & \text{for } t = \dfrac{m}{2} + 1, \ldots, m\,, \\ 0 & \text{for } t = m + 1, \ldots, T - 1\,; \end{cases} \tag{29}$$

Tukey-Hanning estimator:

$$\hat{f}(\lambda) = \frac{1}{2}\hat{g}(\lambda) + \frac{1}{4}\hat{g}\left(\lambda - \frac{\pi}{m}\right) + \frac{1}{4}\hat{g}\left(\lambda + \frac{\pi}{m}\right)\,, \tag{30}$$

where

$$\hat{g}(\lambda) = \frac{1}{2m}\hat{R}(0) + \frac{1}{m}\sum_{t=1}^{m} \hat{R}(t) \cos(\lambda t)\,. \tag{31}$$

In both cases the estimates are constructed only at the points $\lambda = \pi j/m$ $(j = 0, 1, \ldots, m)$. It is recommended to choose the number m between $T/6$ and $T/5$. Other types of estimates of spectral density are given, e.g., in [9], [265], [377].

Filters . From the mathematical point of view, filtering is the construction of special transformations of an original process. The most frequent filters are the so-called *linear filters* that can be written as

$$z_t = \sum_{k=-\infty}^{\infty} \delta_k x_{t-k} \quad (t \text{ an integer}) \tag{32}$$

or

$$z(t) = \int_{-\infty}^{\infty} \delta(\tau)\, x(t-\tau)\, \mathrm{d}\tau \quad (t \text{ real}). \tag{33}$$

The sequence δ_k or the function $\delta(\tau)$ is chosen in such a way that the output process $Z(t)$ which results from the filtering of the input process $X(t)$ has some desired properties. Under general assumptions, the relation

$$f_Z(\lambda) = |\psi(\lambda)|^2 f_X(\lambda) \tag{34}$$

holds between spectral densities $f_X(\lambda)$ and $f_Z(\lambda)$ of the processes $X(t)$ and $Z(t)$, where

$$\psi(\lambda) = \sum_{k=-\infty}^{\infty} \delta_k \mathrm{e}^{-\mathrm{i}k\lambda} \quad \text{or} \quad \psi(\lambda) = \int_{-\infty}^{\infty} \delta(\tau)\mathrm{e}^{-\mathrm{i}\tau\lambda}\, \mathrm{d}\tau \tag{35}$$

is the so-called *transfer function* of the filter. General filter theory (and closely related prediction theory) is given, e.g., in [9], [48], [164], [269], [377].

Example 3. *Low-pass filter*, through which only low frequencies pass, has the transfer function $\psi(\lambda)$ that has to fulfil, due to (34), the relation

$$|\psi(\lambda)| = 0 \quad \text{for } |\lambda| > \varepsilon, \tag{36}$$

where ε is a small positive number.

37. LINEAR PROGRAMMING

By FRANTIŠEK NOŽIČKA

References: [16], [35], [42], [72], [105], [108], [166], [169], [247], [268], [355], [356], [371], [372], [423], [428], [500].

INTRODUCTORY REMARK. Linear programming is a branch of mathematics belonging to mathematical optimization methods. It provides a very useful mathematical tool for solving a lot of problems in economy where a great number of variables as well as many conditions representing restrictions for these variables are involved. The methods of linear programming can be efficiently applied also to many problems in engineering.

In the introductory paragraphs of this chapter, we state the problem considered, present some examples of practical problems leading to linear programming, and survey the fundamental theoretical procedures and results. How to proceed practically when employing the simplex method is demonstrated in Example 37.9.1. In the concluding paragraphs, we then refer to some further linear programming methods.

Throughout this chapter, we use the usual set notation: For example, $\{1, 2, 3\}$ denotes the set of numbers 1, 2 and 3. Writing $I \subset \{1, 2, 3\}$ we mean that I is a subset of the set $\{1, 2, 3\}$, e.g. the set $\{1, 2\}$. The cases $I = \{1, 2, 3\}$ or $I = \emptyset$ (i.e., I is the empty set) are not excluded. If $I = \{1, 2\}$, then $\{1, 2, 3\} \setminus I$ is the complement of the set I, i.e. $\{1, 2, 3\} \setminus I = \{3\}$ (thus the set consisting of the number 3). The notation

$$M = \{x \in E_3 \mid a_1x_1 + a_2x_2 + a_3x_3 = b\}$$

means that M is the set of all points $x = (x_1, x_2, x_3)$ of the Euclidean space E_3 for which the relation $a_1x_1 + a_2x_2 + a_3x_3 = b$ holds. (In case that at least one of the numbers a_1, a_2 and a_3 is non-zero, this relation represents the set of points of E_3 that lie in the plane $a_1x_1 + a_2x_2 + a_3x_3 = b$.)

Speaking about numbers in this chapter, we always mean *real numbers.*

37.1. Formulation of the General Problem of Linear Programming

Let

$$b_1, \ldots, b_m$$

and

$$c_1, \ldots, c_n$$

be given numbers (m and n are positive integers, $n \geqq 2$) and let

$$\begin{bmatrix} a_{11}, & \ldots, & a_{1n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_{m1}, & \ldots, & a_{mn} \end{bmatrix}$$

be a given matrix whose entries are numbers. In E_n, let us consider the set M of all the points $x = (x_1, \ldots, x_n)$ (we will also call them n-tuples $(x_1, \ldots, x_n)$) that satisfy the system of equations and inequalities (the *linear constraints*)

$$\begin{aligned} &\sum_{i=1}^{n} a_{ji}x_i = b_j \quad (j \in I), \\ &\sum_{i=1}^{n} a_{ji}x_i \leqq b_j \quad (j \in \{1, \ldots, m\} \setminus I), \end{aligned} \tag{1}$$

where $I \subset \{1, \ldots, m\}$. (For the notation see the beginning of this chapter.) We thus have the set

$$M = \Big\{x \in E_n \mid \sum_{i=1}^{n} a_{ji}x_i = b_j \quad (j \in I), \quad \sum_{i=1}^{n} a_{ji}x_i \leqq b_j \quad (j \in \{1, \ldots, m\} \setminus I)\Big\}. \tag{2}$$

Our task is to find at least one point $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ *in the set* M *for which the* (*given*) *linear function*

$$f(x) = \sum_{i=1}^{n} c_i x_i \tag{3}$$

assumes its extremum (i.e. minimum or maximum) *on* M.

The optimization problem just presented is often written briefly in the form

$$\max_{x \in M}\Big\{\sum_{i=1}^{n} c_i x_i\Big\}! \quad \text{or} \quad \min_{x \in M}\Big\{\sum_{i=1}^{n} c_i x_i\Big\}!. \tag{4}$$

Definition 1. The linear function (3) is called the *objective function*, the set M is called the *set of feasible points* (or the *feasible set*) corresponding to the *linear*

optimization problem given. Every point $\tilde{x} \in M$ that yields the extremum of the objective function (3) on the feasible set M is called the *solution* or the *optimal point* of the linear optimization problem given.

REMARK 1. If there is no such point $\tilde{x}$ (this situation occurs, for example, if the linear constraints (1) are incompatible, i.e., if $M = \emptyset$, or if the objective function (3) is not bounded from below (in case of minimum) or from above (in case of maximum) on M), then we say that the given *optimization problem has no solution* (see also § 37.6).

REMARK 2. The above formulated general linear programming problem can be properly interpreted as a geometrical problem if we consider the fact that each equation in (1) with a non-zero vector $\boldsymbol{a}_j = (a_{j1}, \ldots, a_{jn})$ describes a certain hyperplane in E_n and that each inequality in (1) with a non-zero vector $\boldsymbol{a}_j$ describes a certain closed half-space in E_n. In accordance with (2), the feasible set M is then the intersection of a finite number of closed half-spaces and hyperplanes in E_n. Every such non-empty set is called a *convex polyhedron* in E_n. In case $M \neq \emptyset$, the feasible set M can thus be geometrically interpreted as a convex polyhedron in E_n. For any choice of a number k and a non-zero vector $\boldsymbol{c} = (c_1, \ldots, c_n)$, the set

$$R_k = \left\{ x \in E_n \mid \sum_{i=1}^{n} c_i x_i = k \right\}$$

is also a hyperplane in E_n so that considering all the choices of k, we obtain a system of parallel hyperplanes in E_n. The objective function (3) can thus be geometrically interpreted as a system of parallel hyperplanes in E_n (see Fig. 37.1 for the case $n = 2$).

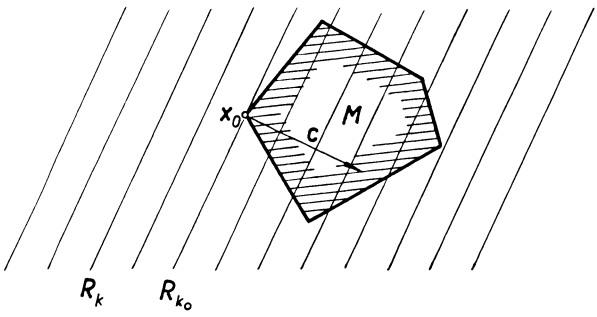


Fig. 37.1.

The optimization problem (4) can then be geometrically interpreted in the following way: Such a point $x_0 = (x_{01}, \ldots, x_{0n})$ of the set M is to be found that the

hyperplane with the description

$$R_{k_0} = \left\{ x \in E_n \mid \sum_{i=1}^{n} c_i x_i = k_0 \right\}, \quad \text{where } k_0 = \sum_{i=1}^{n} c_i x_{0i},$$

is "supporting" hyperplane of the polyhedron M (see the figure). Further, in case of minimization, the vector $\boldsymbol{c}$ with the initial point x_0 points to that of closed half-spaces with the boundary hyperplane R_{k_0}, in which the polyhedron M lies (see Fig. 37.1), while in case of maximization it points to the complementary half-space.

In this chapter, we will often speak (in case $M \neq \emptyset$) *about the corresponding polyhedron M instead of the set M of feasible points.*

37.2. Linear Optimization Problem in Normal Form

Let a (general) optimization problem (37.1.4) with the set (37.1.2) of its feasible points be given. Further let at least one equation occur in the description of M (i.e., let $I \neq \emptyset$). If now the system of equations

$$\sum_{i=1}^{n} a_{ji} x_i = b_j \quad (j \in I) \tag{1}$$

has no solution, then $M = \emptyset$ and the optimization problem considered has no solution, either. If the system (1) is solvable, then it has a certain rank s where $1 \leqq s \leqq n$. Assuming $s = n$, we have a trivial optimization problem where M is either a single-point set or the empty set. If $s < n$, then s variables can be expressed from the system (1) as linear function of the other $n - s$ variables. Without loss of generality we can suppose that these s variables are $x_1, \ldots, x_s$ and thus

$$x_i = \sum_{l=1}^{n-s} e_{il} x_{s+l} + d_i \quad (i = 1, \ldots, s). \tag{2}$$

Substituting this into all the inequalities in the description (37.1.2) of the set M, we obtain certain linear inequalities

$$\sum_{l=1}^{n-s} \tilde{a}_{jl} x_{s+l} \leqq \tilde{b}_j \quad (j \in \{1, \ldots, m\} \setminus I).$$

If some of these inequalities reduce to the form $0 \leqq \tilde{b}_j$ and if at least one of the numbers $\tilde{b}_j$ is negative, then $M = \emptyset$ and the optimization problem considered thus

has no solution. If all the numbers $\tilde{b}_j$ in the inequalities $0 \leqq \tilde{b}_j$ are non-negative, then these inequalities can be omitted. In this case, we put

$$I' = \left\{ j \in \{1, \ldots, m\} \setminus I \mid \sum_{l=1}^{n-s} |\tilde{a}_{jl}| > 0 \right\}$$

and arrive at the inequalities

$$\sum_{l=1}^{n-s} \tilde{a}_{jl} x_{s+l} \leqq \tilde{b}_j \quad (j \in I'). \tag{3}$$

Since each of the numbers x_{s+l} can be written as a difference of two non-negative numbers, i.e.

$$x_{s+l} = x_{s+l}^{+} - x_{s+l}^{-}, \quad \text{where } x_{s+l}^{+} \geqq 0 \text{ and } x_{s+l}^{-} \geqq 0 \quad (l = 1, \ldots, n-s), \tag{4}$$

the substitution into (3) yields the system of inequalities

$$\begin{aligned} &\sum_{l=1}^{n-s} \tilde{a}_{jl} x_{s+l}^{+} + \sum_{l=1}^{n-s} (-\tilde{a}_{jl}) x_{s+l}^{-} \leqq \tilde{b}_j \quad (j \in I'), \\ &x_{s+l}^{+} \geqq 0, \quad x_{s+l}^{-} \geqq 0 \quad (l = 1, \ldots, n-s). \end{aligned} \tag{5}$$

Substituting further (2) into the objective function f in (37.1.3), we obtain a certain linear function

$$\sum_{l=1}^{n-s} \tilde{c}_l x_{s+l}$$

and, with respect to (4), the linear function

$$\sum_{l=1}^{n-s} \tilde{c}_l x_{s+l}^{+} + \sum_{l=1}^{n-s} (-\tilde{c}_l) x_{s+l}^{-} \tag{5'}$$

of variables x_{s+l}^{+}, x_{s+l}^{-} $(l = 1, \ldots, n-s)$.

In case that no equation occurs in the description (37.1.2) of the set M (i.e. in case $I = \emptyset$), we will take into account only the substitution (4) (putting $s = 0$) and substitute both into all the inequalities in the description of the set M and into the objective function (37.1.3).

In this way, we can either transform the original (general) linear programming problem into an equivalent optimization problem with the objective function (5') and the feasible set described by (5) (which is a particular case of linear optimization problems of Definition 1) or we find in the course of this procedure that the original problem has no solution.

Definition 1. The linear optimization problem with the description

$$M = \left\{ x \in E_n \mid \sum_{i=1}^{n} a_{ji} x_i \leqq b_j \quad (j = 1, \ldots, m), \quad x_i \geqq 0 \quad (i = 1, \ldots, n) \right\} \tag{6}$$

of the set of feasible points is called the *linear optimization problem in normal form.*

REMARK 1. The procedure that preceded Definition 1 thus shows that either it is possible to transform a general linear programming problem into an equivalent linear optimization problem in normal form or the procedure reveals that the problem has no solution. Here, the *equivalence of two linear optimization problems* means such their property that existence of the solution of one of the problems implies existence of the solution of the other.

37.3. Linear Optimization Problem in Equality Form

Like in § 37.2, consider a (general) linear optimization problem (37.1.4) with the description (37.1.2) of the feasible set M. In the case that at least one inequality occurs in this description (i.e. $\{1, \ldots, m\} \setminus I \neq \emptyset$), we introduce further variables

$$\xi_j = b_j - \sum_{i=1}^{n} a_{ji} x_i \quad (j \in \{1, \ldots, m\} \setminus I) \tag{1}$$

that are subject to the conditions

$$\xi_j \geqq 0 \quad (j \in \{1, \ldots, m\} \setminus I)$$

in accordance with the inequalities from the description (37.1.2) of the set M. The optimization problem with the original objective function

$$f(x) = \sum_{i=1}^{n} c_i x_i ,$$

with the feasible set M described by

$$\begin{aligned} \sum_{i=1}^{n} a_{ji} x_i &= b_j \quad (j \in I) , \\ \sum_{i=1}^{n} a_{ji} x_i + \xi_j &= b_j \quad (j \in \{1, \ldots, m\} \setminus I) , \end{aligned} \tag{2}$$

and with the condition $\xi_j \geqq 0$ $(j \in \{1, \ldots, m\} \setminus I)$ is apparently equivalent (cf. Remark 37.2.1) to the original optimization problem. Moreover, putting $x_i = x_i^+ - x_i^-$ in (2), where $x_i^+ \geqq 0$ and $x_i^- \geqq 0$ $(i = 1, \ldots, n)$, we again arrive at an equivalent optimization problem with the variables x_i^+, x_i^- $(i = 1, \ldots, n)$ and ξ_j $(j \in \{1, \ldots, m\} \setminus I)$ occurring in the linear objective function (here, the variables ξ_j have zero coefficients) and in the feasible set whose description involves only linear equations together with the conditions of non-negativity of the variables.

Definition 1. The quantities ξ_j $(j \in \{1, \ldots, m\} \setminus I)$ defined by (1), which occur in the optimization problem assigned to the original linear optimization problem in the above presented way, are called *slack variables.*

Definition 2. The linear optimization problem with the description

$$M = \left\{ x \in E_n \mid \sum_{i=1}^{n} a_{ji} x_i = b_j \quad (j = 1, \ldots, m), \quad x_i \geqq 0 \quad (i = 1, \ldots, n) \right\}$$

of the set of feasible points is called the *linear optimization problem in equality form.*

REMARK 1. As we have just shown, a general linear optimization problem, whose feasible set has inequalities in its description, can be transformed into an equivalent linear optimization problem in equality form.

37.4. Examples of Linear Optimization Problems Solved in Practice

(a) Classical Transportation Problem

Let $P_1, \ldots, P_m$ be the sites of production (*production centres*) of some product p (e.g. coal mines) and let a_i $(a_i > 0)$ be the amount of the product p produced at the centre P_i $(i \in \{1, \ldots, m\})$ in some time period (the same for all the centres). The consumption of this product is planned at the sites of consumption $C_1, \ldots, C_n$ (e.g. in certain towns) in such a way that the amount b_j $(b_j > 0)$ corresponds to the site C_j $(j \in \{1, \ldots, n\})$. In the considered time period, the total production of the product is, moreover, assumed to be equal to its total consumption ("*economic balance*"), i.e.

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j .$$

Let x_{ij} $(i = 1, \ldots, m;\ j = 1, \ldots, n)$ be the unknown amounts of the product p in some appropriate units (e.g. weight units) to be transported from the production

centre P_i to the site of consumption C_j. The following conditions are thus to be fulfilled:

$$x_{ij} \geqq 0 \quad (i = 1, \ldots, m;\ j = 1, \ldots, n), \tag{1a}$$

$$\sum_{j=1}^{n} x_{ij} = a_i \quad (i = 1, \ldots, m), \tag{1b}$$

$$\sum_{i=1}^{m} i x_{ij} = b_j \quad (j = 1, \ldots, n). \tag{1c}$$

From the viewpoint of economy, the condition (1a) is evident. The condition (1c) represents the fact that the consumption b_j at the site C_j is to be met by the delivery from the production centres $P_1, \ldots, P_m$ considered. The condition (1b) expresses the economic requirement that the whole amount a_i of the product produced at the centre P_i be delivered to the sites $C_1, \ldots, C_n$ of its consumption. Further, positive numbers c_{ij} $(i = 1, \ldots, m;\ j = 1, \ldots, n)$ are given that correspond to the cost of transportation of the product unit from the centre P_i to the site C_j.

Our task is to determine the transported amounts x_{ij} $(i = 1, \ldots, m;\ j = 1, \ldots, n)$ in such a way that the total transportation cost, expressed by the linear function

$$f = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}, \tag{2}$$

be, under the conditions (1a), (1b), and (1c), minimal. The linear optimization problem thus consists in the minimization of the function (2) on the set of all points

$$x = (x_{11}, \ldots, x_{1n}, x_{21}, \ldots, x_{2n}, \ldots, x_{m1}, \ldots, x_{mn})$$

which is described by the linear constraints (1a), (1b), and (1c). In literature, this linear optimization problem, given apparently in equality form, is called the *classical transportation problem.*

(b) Blending Problem

Assume that a certain number of petrol sorts, that differ from each other both in quality and in price, are available. Denote these sorts by $B_1, \ldots, B_n$ and let p_i $(p_i > 0)$ be the price of a litre of petrol B_i. Suppose that $A_1, \ldots, A_m$ are the substantial components of all the petrol sorts considered whose amounts in a particular petrol sort (contained in a litre of it) determine its quality. Suppose further that a_{ij} $(a_{ij} \geqq 0, \sum_{j=1}^{n} a_{ij} > 0)$ are known numbers representing the amount of component A_j in a litre of petrol B_i. These data are well arranged in Tab. 37.1:

TABLE 37.1

	A_1	A_2	...	A_j	...	A_m
B_1	a_{11}	a_{12}	...	a_{1j}	...	a_{1m}
B_2	a_{21}	a_{22}	...	a_{2j}	...	a_{2m}
⋮	.....					
B_i	a_{i1}	a_{i2}	...	a_{ij}	...	a_{im}
⋮	.....					
B_n	a_{n1}	a_{n2}	...	a_{nj}	...	a_{nm}

Our task is to get, blending the petrols $B_1, \ldots, B_n$, a new petrol B_0, whose one litre contains at least the amount c_j $(c_j > 0)$ of the component A_j, in such a manner that the price of this new petrol be minimal.

Denoting by x_i $(i = 1, \ldots, m)$ the unknown amount of petrol B_i contained in a litre of the blend prepared, we have the following conditions for these quantities:

$$\begin{gathered} x_i \geqq 0 \quad (i = 1, \ldots, n), \\ \sum_{i=1}^{n} x_i = 1, \quad \sum_{i=1}^{n} a_{ij} x_i \geqq c_j \quad (j = 1, \ldots, m). \end{gathered} \tag{3}$$

The price of a litre of the petrol blend is

$$f(x) = \sum_{i=1}^{n} p_i x_i . \tag{4}$$

We can thus state the mathematical formulation (the mathematical model) of the problem considered:

On the set M of all the n-tuples $(x_1, \ldots, x_n)$ satisfying the constraints (3), we are to find such an n-tuple $x_0 = (x_{01}, \ldots, x_{0n})$ from M for which

$$f(x_0) = \min_{x \in M} f(x)$$

holds. It is thus a linear optimization problem with the objective function (4) and the feasible set M described by (3). In literature, this optimization problem is referred to as the *blending problem.*

(c) Production Planning

A major plant produces, in its sectors $S_1, \ldots, S_q$, certain products $P_1, \ldots, P_r$ that are assumed to be divisible into relatively small amounts (e.g. different sorts of beer in a brewery, different sorts of tinned meat in a canning factory). It is not supposed that each product must be necessarily manufactured in each sector of the plant. Each of the products passes through definite stages (production stages) of successive processing. For this purpose, there are certain departments $A_1, \ldots, A_p$ in the plant. For the sake of simplicity, suppose that these departments exist in each sector. Let p_i denote the minimal amount of the product P_i ($i \in \{1, \ldots, r\}$) planned to be manufactured in the plant during some time period (e.g. a year). Further let a_{ji}^k denote the maximal amount of the product P_i ($i \in \{1, \ldots, r\}$) that can be processed by the department A_j ($j \in \{1, \ldots, p\}$) of the sector S_k ($k \in \{1, \ldots, q\}$) during the time period given. If the product P_i does not pass through the department A_j of the sector S_k, we put $a_{ji}^k = 0$, otherwise $a_{ji}^k > 0$. It is assumed that no department can process two products at the same time and, moreover, that after passing through a certain department, the product does not re-enter it. Denote further by s_{ji}^k the cost of production of a unit amount of the product P_i in the department A_j of the sector S_k. The cost of production of a unit amount of the product P_i in the sector S_k is then the sum

$$\sum_{j=1}^{p} s_{ji}^k .$$

Let x_{ki} ($i = 1, \ldots, r$; $k = 1, \ldots, q$) be the unknown amount of the product P_i to be manufactured in the sector S_k. Since a_{ji}^k is the maximal amount of the product P_i that can be processed in the department A_j of the sector S_k during the given time period (this time period will be assigned the unit length), the number $1/a_{ji}^k$ (in case $a_{ji}^k > 0$) represents the fraction of this time period needed to process a unit amount of the product P_i in the corresponding department and sector. A fraction of the time period equal to x_{ki}/a_{ji}^k then corresponds to the amount x_{ki} of the product P_i in the department A_j of the sector S_k. This fraction is called the *capacity load* of the department A_j in sector S_k by the product P_i. Introducing the index sets

$$I(k, j) = \{i \in \{1, \ldots, r\} \mid a_{ji}^k > 0\} \quad (k \in \{1, \ldots, q\},\ j \in \{1, \ldots, p\}),$$

we see that the sum

$$\sum_{i \in I(k,j)} \frac{x_{ki}}{a_{ji}^k}$$

represents such a fraction of the time period that is needed to process all products in the department A_j of the sector S_k. This fraction is at most equal to the given

total time period (which has been assigned the unit length), so that

$$\sum_{i \in I(k,j)} \frac{x_{ki}}{a^k_{ji}} \leqq 1 \quad (k \in \{1, \ldots, q\},\ j \in \{1, \ldots, p\}). \tag{4a}$$

The conditions

$$x_{ki} \geqq 0 \quad (k \in \{1, \ldots, q\},\ i \in \{1, \ldots, r\}) \tag{4b}$$

are quite natural from the viewpoint of economy.

Putting

$$c_{ki} = \begin{cases} 1 & \text{if the product } P_i \text{ is manufactured in the sector } S_k, \\ 0 & \text{otherwise}, \end{cases}$$

we find that the sum

$$\sum_{k=1}^{q} c_{ki} x_{ki}$$

represents the total production of the product P_i in the plant and the conditions

$$\sum_{k=1}^{q} c_{ki} x_{ki} \geqq p_i \quad (i = 1, \ldots, p) \tag{4c}$$

are required to be fulfilled. The expression

$$f = \sum_{i=1}^{r} \sum_{k=1}^{q} \sum_{j=1}^{p} s^k_{ij} x_{ki} \tag{5}$$

represents the total cost of production of all the considered products during the given time period. The natural requirement is that this total cost is minimal under the conditions presented.

The linear optimization problem thus consists in the minimization of the objective function (5) of the variables x_{ki} $(k = 1, \ldots, q;\ i = 1, \ldots, r)$ satisfying the linear constraints (4a), (4b), and (4c).

37.5. Decomposition of a Convex Polyhedron into its Interior and Faces

An arbitrary convex polyhedron M is a non-empty intersection of a finite number of closed half-spaces and hyperplanes in E_n (see Remark 37.1.2) and it thus represents a closed convex set in E_n. That means: $\overline{M} = M$ and $\lambda_1 x + \lambda_2 y \in M$ holds for every pair of points $x \in M$, $y \in M$ and every pair of numbers $\lambda_1 \geqq 0$,

$\lambda_2 \geqq 0$ such that $\lambda_1 + \lambda_2 = 1$. In this sense, every linear manifold (straight line, plane, hyperplane) in E_n (as well as a point and the space E_n itself) is also a convex polyhedron in E_n.

We say that the *dimension of a convex polyhedron* M in E_n is d (and write $\dim M = d$) if the linear manifold of the smallest dimension, that contains the convex polyhedron M, has dimension d.

For example, a triangle in E_2 has dimension $d = 2$ since the linear manifold of smallest dimension, that contains it, is the plane E_2. The segment described by the relations $y = 3x + 1$, $-1 \leqq x \leqq 1$, has dimension $d = 1$ as the linear manifold of smallest dimension, that contains it, is the straight line in E_2 described by the equation $y = 3x + 1$.

Instead of the linear manifold of the smallest dimension containing the polyhedron M, we will often speak briefly about the *linear span of this polyhedron.*

Let M be a convex polyhedron of dimension d ($1 \leqq d \leqq n$) in E_n and L_d be its linear span. Every point $x \in M$, to which there exists $\varepsilon > 0$ so that its ε-neighbourhood in L_d (i.e. the set

$$U_d(x; \varepsilon) = \{y \in L_d \mid \varrho(y, x) < \varepsilon\},$$

where $\varrho(y, x)$ is the distance between the points y and x) belongs to M, is called the *interior point of the polyhedron* M. The set of all such points $x \in M$ is called the *interior of the polyhedron* M and is denoted by $\operatorname{rel.int} M$, in case $d = n$ also $\operatorname{int} M$. If the convex polyhedron consists of a single point, i.e. $M = \{x_0\}$, we put $\operatorname{rel.int} M = \{x_0\}$.

The set

$$\partial M = M \setminus \operatorname{rel.int} M$$

is called the *boundary of a convex polyhedron* M in E_n. If M is a linear manifold in E_n, then apparently $\operatorname{rel.int} M = M$ and thus $\partial M = \emptyset$.

Let M be a convex polyhedron in E_n, which is not a linear manifold. In this case, it can thus be described as an intersection of a certain linear manifold L and a finite number of closed half-spaces $\overline{H}_1, \ldots, \overline{H}_p$, i.e.

$$M = L \cap \overline{H}_1 \cap \cdots \cap \overline{H}_p, \tag{1}$$

where $L \cap H_1 \cap \cdots \cap H_p \neq \emptyset$ holds for the corresponding open half-spaces $H_1, \ldots, H_p$. We can easily verify that then

$$\begin{gathered} \operatorname{rel.int} M = L \cap H_1 \cap \cdots \cap H_p, \\ \dim M = \dim L. \end{gathered} \tag{1'}$$

Since $\overline{H}_i = H_i \cup R_i$ where R_i is the boundary hyperplane of the half-space H_i, we obtain from this and (1), and using the well-known operations with sets, that

$$\begin{aligned} M &= L \cap (H_1 \cup R_1) \cap \cdots \cap (H_p \cup R_p) = \\ &= (L \cap H_1 \cap \cdots \cap H_p) \cup \sum L \cap R_{i_1} \cap \cdots \cap R_{i_l} \cap H_{i_{l+1}} \cap \cdots \cap H_{i_p}, \end{aligned} \tag{2}$$

where the symbol $\sum$ represents the union of such non-empty sets

$$S^l_{i_1,\ldots,i_p} = L \cap R_{i_1} \cap \cdots \cap R_{i_l} \cap H_{i_{l+1}} \cap \cdots \cap H_{i_p}, \tag{3}$$

for which $1 \leqq l \leqq p$ and $\{i_1, \ldots, i_l, i_{l+1}, \ldots, i_p\}$ is a permutation of the p-tuple $\{1, \ldots, p\}$ with the property $i_1 \leqq \cdots \leqq i_l$, $i_{l+1} \leqq \cdots \leqq i_p$. The relations (1′), (2), and (3) imply that

$$M = \text{rel. int}\, M \cup \sum S^l_{i_1,\ldots,i_p}. \tag{4}$$

Each of the non-empty sets $S^l_{i_1,\ldots,i_p}$ in (3) is called the *face of the convex polyhedron* M and the expression on the right-hand side of (4) is called the *decomposition of the convex polyhedron* M *into its interior and faces.* The closure $\overline{S}^l_{i_1,\ldots,i_p}$ of a face $S^l_{i_1,\ldots,i_p}$ of a convex polyhedron M is called a *closed face of this polyhedron* and it follows from (3) that

$$\overline{S}^l_{i_1,\ldots,i_p} = L \cap R_{i_1} \cap \cdots \cap R_{i_l} \cap \overline{H}_{i_{l+1}} \cap \cdots \cap \overline{H}_{i_p}$$

(since $\overline{L} = L$ and $\overline{R}_i = R_i$). A closed face of a convex polyhedron is a convex polyhedron as well and, moreover,

$$\overline{S}^l_{i_1,\ldots,i_p} \subset \partial M, \quad \dim \overline{S}^l_{i_1,\ldots,i_p} < \dim M.$$

The face of dimension 0 is called a *vertex*, the face of dimension 1 is called an *edge of a convex polyhedron* M.

The reader can readily form a geometrical interpretation of the presented concepts in E_3 if he takes the closed half-spaces, whose intersection is the closed first octant in E_3, for $\overline{H}_1$, $\overline{H}_2$, and $\overline{H}_3$, and a plane, whose part lying in this octant is the polyhedron M, for L.

REMARK 1. Two linear manifolds L_{d_i} $(i = 1, 2)$ in E_n of dimensions $d_i = \dim L_{d_i}$, where $1 \leqq d_i \leqq n - 1$ $(i = 1, 2)$, are called *parallel* in E_n if $L_{d_1} \cap L_{d_2} = \emptyset$ and $(L_{d_1} \cup L_{d_2}) \subset L_d$ where $d = \max(d_1, d_2) + 1$. The following statement then holds for convex polyhedra in general:

Let M be a convex polyhedron in E_n containing a linear manifold L_d of dimension d, where $1 \leqq d \leqq n - 1$. Then every linear manifold of the same dimension,

that passes through an arbitrary point $x \in M$ and is parallel to the manifold L_d, belongs to the polyhedron M.

Since the first octant in E_n, i.e. the set $E_n^+ = \{x \in E_n \mid x_i \geqq 0 \ (i = 1, \ldots, n)\}$, contains no linear manifold of the space E_n whose dimension is greater than or equal to 1, the preceding statement implies this corollary:

If the feasible set M of a linear optimization problem in equality form (see Definition 37.3.2) *is non-empty, then M is a convex polyhedron in E_n containing vertices (i.e. at least one vertex).*

37.6. The Set of Optimal Points of a Linear Optimization Problem

Let us consider a (general) linear optimization problem (37.1.4) with the feasible set described in (37.1.2). Let us denote by M_{opt} the set of all optimal points of the considered problem, i.e.

$$M_{\text{opt}} = \{\tilde{x} \in M \mid f(\tilde{x}) = \min_{x \in M} f(x)\}$$

or

$$M_{\text{opt}} = \{\tilde{x} \in M \mid f(\tilde{x}) = \max_{x \in M} f(x)\}.$$

If $M_{\text{opt}} \neq \emptyset$, we say that the *considered optimization problem is solvable.* In the case $M_{\text{opt}} = \emptyset$ (i.e. in the case that either $M = \emptyset$, or $M \neq \emptyset$ but, at the same time, the function $f(x)$ does not assume its minimum or maximum on the set M) we say that the *given problem has no solution.*

The following theorems hold:

Theorem 1. *If the general linear optimization problem* (37.1.4), (37.1.2) *is solvable, then the set M_{opt} of all its optimal points is either a closed face of the corresponding convex polyhedron M, or the whole polyhedron. The case $M_{\text{opt}} = M$ occurs if $f(x) = 0$ for all $x \in M$ (i.e. $c_i = 0$ for $i = 1, \ldots, n$), or if for some point $x_0 \in M$, the hyperplane*

$$R_0 = \left\{x \in E_n \mid \sum_{i=1}^{n} c_i x_i = \sum_{i=1}^{n} c_i x_{0i}\right\}$$

contains the whole polyhedron M.

Theorem 2. *If the general optimization problem* (37.1.4), (37.1.2) *is solvable and if the polyhedron M contains at least one face of dimension d and no face of dimension less than d, then there exists such a face S_d of dimension d that $\overline{S}_d \subset M_{\text{opt}}$.*

Theorem 2 and the second statement of Remark 37.5.1 imply the next theorem:

Theorem 3. *If the linear optimization problem in equality form* (see Definition 37.3.2) *is solvable, then there exists a vertex of the corresponding convex polyhedron M that is the optimal point of the optimization problem considered.*

37.7. The Concept of Feasible Basic Point

Consider a linear optimization problem in equality form (see Definition 37.3.2), i.e. a linear optimization problem with the feasible set

$$M = \left\{ x \in E_n \mid \sum_{j=1}^{n} a_{ji} x_i = b_j \quad (j = 1, \ldots, m), \quad x_i \geqq 0 \quad (i = 1, \ldots, n) \right\}, \quad (1)$$

and suppose

a) $m < n$;

b) the matrix

$$\boldsymbol{A} = \begin{bmatrix} a_{11}, & \ldots, & a_{1n} \\ \ldots\ldots\ldots\ldots\ldots & & \\ a_{m1}, & \ldots, & a_{mn} \end{bmatrix} \quad (2)$$

has the maximal possible rank, i.e. m.

REMARK 1. If $M \neq \emptyset$, then the assumptions a) and b) imply that the set

$$L_{n-m} = \left\{ x \in E_n \mid \sum_{j=1}^{n} a_{ji} x_i = b_j \quad (j = 1, \ldots, m) \right\}$$

is non-empty as well and it represents a linear manifold of dimension $n - m$ in E_n. Thus in case $M \neq \emptyset$, M is a convex polyhedron in E_n, which is an intersection of the linear manifold L_{n-m} with the first octant

$$E_n^+ = \{ x \in E_n \mid x_i \geqq 0 \quad (i = 1, \ldots, n) \}$$

of the space E_n, and $0 \leqq \dim M \leqq n - m$.

REMARK 2. With respect to the fact that the rank of the matrix $\boldsymbol{A}$ is m, there exists at least one square submatrix of order m of the matrix $\boldsymbol{A}$ whose determinant is non-zero. Let

$$\begin{vmatrix} a_{1i_1}, & \ldots, & a_{1i_m} \\ \ldots\ldots\ldots\ldots\ldots & & \\ a_{mi_1}, & \ldots, & a_{mi_m} \end{vmatrix} \neq 0, \quad (3)$$

where $\{i_1, \ldots, i_m\}$ is a certain m-element subset of the set $\{1, \ldots, n\}$. Then we can express the variables $x_{i_1}, \ldots, x_{i_m}$ from the system of equations

$$\sum_{i=1}^{n} a_{ji} x_i = b_j \quad (j = 1, \ldots, m) \tag{4a}$$

as linear functions of the other $n - m$ variables $x_{i_{m+1}}, \ldots, x_{i_n}$, i.e.

$$x_{i_s} = d_{i_s} - \sum_{l=m+1}^{n} d_{i_s i_l} x_{i_l} \quad (s \in \{1, \ldots, m\}). \tag{4b}$$

Putting $x_{i_l} = 0$ for $l = m+1, \ldots, n$, we obtain $x_{i_s} = d_{i_s}$ $(s = 1, \ldots, m)$ from (4b) so that the point $\hat{x} \in E_n$ with coordinates

$$\hat{x}_{i_s} = d_{i_s}, \quad \hat{x}_{i_l} = 0 \quad (s = 1, \ldots, m;\ l = m+1, \ldots, n) \tag{5}$$

satisfies equations (4a) and it is thus a point of the linear manifold L_{n-m} from Remark 1. Conversely, assigning zero values to $n-m$ variables x_{i_l} $(l = m+1, \ldots, n)$ of the total number of n variables $x_1, \ldots, x_n$ in the system of equations (4a), we obtain the system of m linear equations

$$\sum_{s=1}^{m} a_{ji_s} x_{i_s} = b_j \quad (j = 1, \ldots, m). \tag{6}$$

If this system has a unique solution $x_{i_s} = d_{i_s}$ $(s = 1, \ldots, m)$ (which occurs if and only if (3) holds), then the point x with the coordinates given in (5) is a uniquely determined point of the variety L_{n-m}.

In the presented way we can thus find particular points of the linear variety L_{n-m}, which are characterized in the following definition:

Definition 1. Let $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_n)$ be a point satisfying (under the assumptions a) and b)) the system of equations (4a) from the description (1) of the set M and fulfilling the following conditions:

1. The set $I(\hat{x}) = \{i \in \{1, \ldots, n\} \mid \hat{x}_i = 0\}$ contains at least $n - m$ elements.
2. From the set $I(\hat{x})$, we can extract such its subset $\{i_{m+1}, \ldots, i_n\}$ of $n - m$ elements that the system of equations (6) has the unique solution $\hat{x}_{i_1}, \ldots, \hat{x}_{i_m}$, where we denoted by $\{i_1, \ldots, i_m\} = \{1, \ldots, n\} \setminus \{i_{m+1}, \ldots, i_n\}$ the complement of the subset $\{i_{m+1}, \ldots, i_n\}$.

The point $\hat{x}$ is then called the *basic solution of the equations* (4a), and the variables $x_{i_1}, \ldots, x_{i_m}$ and $x_{i_{m+1}}, \ldots, x_{i_n}$ are called *basic* and *non-basic variables*, respectively, corresponding to the basic solution $\hat{x}$.

Definition 2. If $\hat{x}$ is such a basic solution of the equations (4a) that the point $\hat{x}$ has more than $n - m$ zero coordinates, we say that $\hat{x}$ is a *degenerate basic solution of the equations* (4a). If the point $\hat{x}$ has just $n - m$ zero coordinates, we speak about a *non-degenerate* (or *regular*) *basic solution of the equations* (4a).

Definition 3. A basic solution $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_n)$ of the system of equations (4a) possessing the property $\hat{x}_i \geqq 0$ $(i = 1, \ldots, n)$ (i.e. the property $\hat{x} \in M$) is called a *feasible basic point of the optimization problem considered.* If a feasible basic point has more than $n - m$ zero coordinates, we speak about *degeneracy at this point.* If it has just $n - m$ zero coordinates, then it is called *regular.*

The fact that, in case $M \neq \emptyset$, there exists (under the assumptions a) and b)) at least one feasible basic point of the linear optimization problem, follows from Theorem 37.6.3 and the following Theorem 1, which, in turn, is a consequence of the definition of a vertex of the polyhedron M in § 37.5:

Theorem 1. *Each feasible basic point of the linear optimization problem considered is a vertex of the polyhedron M. Conversely, each vertex of the polyhedron M from* (1) *is a feasible basic point of the linear optimization problem considered, to which either a unique index basis (in case of non-degeneracy) or several index bases (in case of degeneracy) correspond.*

Example 1. Let the set

$$M = \{x \in E \mid x_1 - 2x_2 + x_3 = 1\,,\ 2x_1 + 3x_2 + x_3 = 2\,,\ x_i \geqq 0\ (i = 1, 2, 3)\}$$

be a feasible set of a certain linear optimization problem in three variables x_1, x_2 and x_3. We thus have $n = 3$, $m = 2$, and $n - m = 1$. Putting $x_1 = 0$ in the system of equations

$$x_1 - 2x_2 + x_3 = 1\,, \quad 2x_1 + 3x_2 + x_3 = 2\,, \tag{7}$$

we obtain the equations

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

with the unique solution $x_2 = \frac{1}{5}$ and $x_3 = \frac{7}{5}$. According to Definitions 1 and 2, the point $x_0 = (0, \frac{1}{5}, \frac{7}{5})$ is a non-degenerate basic solution of the system of equations (7) with index basis $\{2, 3\}$. Since all its coordinates are non-negative, it is a regular feasible basic point of the considered linear optimization problem (Definition 3).

Choosing $x_2 = 0$ in the equations (7), we obtain the equations

$$x_1 + x_3 = 1\,, \quad 2x_1 + x_3 = 2$$

with the unique solution $x_1 = 1$ and $x_3 = 0$. The point $y_0 = (1, 0, 0)$ is thus a degenerate basic solution of the system of equations (7) with index basis $\{1, 3\}$

and, due to the non-negativity of its coordinates, it is a feasible basic point of the considered linear optimization problem at which degeneracy occurs.

Finally, putting $x_3 = 0$ in the equations (7), we arrive at the equations

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

with the unique solution $x_1 = 1$ and $x_2 = 0$. Therefore, the previous point y_0 is the solution of the equations (7) in this case as well and it now represents their degenerate basis solution with index basis $\{1, 2\}$.

According to Theorem 1, this implies that the points x_0 and y_0 represent all the vertices of the polyhedron M, which is apparently a closed segment with end points x_0 and y_0.

37.8. Exchange of Basic Variables. Optimality Criterion. The Degenerate Case

Preparing an algorithm for the solution of the linear optimization problem in equality form considered in § 37.7 under the assumptions a) and b) presented there, we must first consistently describe the individual steps of the procedure. This is with which we will be concerned in the present paragraph.

Let $x_0 = (x_{01}, \ldots, x_{0n})$ be a known feasible basic point of the given linear optimization problem in equality form and with the feasible set from (37.7.1). Let this point corresponds to the basic solution of the system of equations (37.7.4a) with index basis $\{i_1, \ldots, i_m\}$. Without loss of generality we can assume $\{i_1, \ldots, i_m\} = \{1, \ldots, m\}$ (which can be always achieved by a suitable renumbering of variables $x_1, \ldots, x_n$). Then the equations (37.7.4b) have the form

$$x_i = d_{i0} - \sum_{j=m+1}^{n} d_{ij} x_j \quad (i = 1, \ldots, m), \tag{1}$$

where $d_{i0} = x_{i0} \geqq 0$ $(i = 1, \ldots, m)$. Substituting (1) into the objective function $f(x) = \sum_{i=1}^{n} i c_i x_i$, we obtain

$$f(x) = c_0 + \sum_{i=1}^{n} (c_i - z_i) x_i, \tag{2}$$

where

$$\begin{aligned} c_0 &= \sum_{i=1}^{m} c_i d_{i0}\,, \\ z_l &= \sum_{i=1}^{m} c_i d_{il} \quad (l = m+1, \ldots, n)\,, \\ z_l &= c_l \quad (l = 1, \ldots, m)\,. \end{aligned} \tag{2'}$$

The relevant data can be conveniently arranged in Tab. 37.2, where $x_1, \ldots, x_m$ are basic variables corresponding to index basis $\{1, \ldots, m\}$. The inner product of the row vector $(x_1, \ldots, x_n)$ and the row vector in the i-th row of the bold framed matrix in Tab. 37.2 (the so-called *basis matrix*) yields, by (1), the value of d_{i0}. The inner product of the column vector $(c_1, \ldots, c_m)$ on the left-hand side of Tab. 37.2 and the i-th column vector in the basis matrix gives the value of z_i $(i \in \{1, \ldots, n\})$. The number c_0 is the inner product of the column vector $(c_1, \ldots, c_m)$ and the column vector $(d_{10}, \ldots, d_{m0})$, and represents the value of the objective function at the point x_0.

Theorem 1, which follows, yields an optimality criterion for the point x_0 while Theorem 2 a criterion for the non-existence of solution of the given optimization problem

$$\min_{x \in M} \Big\{ \sum_{i=1}^{n} c_i x_i \Big\} ! \quad \text{or} \quad \max_{x \in M} \Big\{ \sum_{i=1}^{n} c_i x_i \Big\} ! \,. \tag{3}$$

Theorem 1. *If $c_i - z_i \geqq 0$ for $i = 1, \ldots, n$ or $c_i - z_i \leqq 0$ for $i = 1, \ldots, n$, then the feasible basic point x_0 is the optimal point of the minimization or maximization problem* (3), *respectively.*

Theorem 2. *If, in the minimization or maximization problem* (3), *there exists such an index $j_0 \in \{m+1, \ldots, n\}$ that*

$$c_{j_0} - z_{j_0} < 0\,, \quad d_{ij_0} \leqq 0 \quad \text{for } i = 1, \ldots, m$$

or

$$c_{j_0} - z_{j_0} > 0\,, \quad d_{ij_0} \leqq 0 \quad \text{for } i = 1, \ldots, m\,,$$

then the objective function $\sum_{i=1}^{n} i c_i x_i$ is not bounded from below or from above, respectively, on the set M.

REMARK 1. If none of the cases described in Theorems 1 and 2 takes place then, in the minimization or maximization problem (3), the set

$$I_1 = \{ j \in \{m+1, \ldots, n\} \mid c_j - z_j < 0 \}$$

TABLE 37.2

		x_1	$\cdots$	x_i	$\cdots$	x_m	x_{m+1}	$\cdots$	x_s	$\cdots$	x_n	
		c_1	$\cdots$	c_i	$\cdots$	c_m	c_{m+1}	$\cdots$	c_s	$\cdots$	c_n	
c_1	x_1	1	$\cdots$	0	$\cdots$	0	$d_{1,m+1}$	$\cdots$	d_{1s}	$\cdots$	d_{1n}	d_{10}
$\cdots$	$\cdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\cdots$
c_i	x_i	0	$\cdots$	1	$\cdots$	0	$d_{i,m+1}$	$\cdots$	d_{is}	$\cdots$	d_{in}	d_{i0}
$\cdots$	$\cdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\cdots$
c_m	x_m	0	$\cdots$	0	$\cdots$	1	$d_{m,m+1}$	$\cdots$	d_{ms}	$\cdots$	d_{mn}	d_{m0}
		z_1	$\cdots$	z_i	$\cdots$	z_m	z_{m+1}	$\cdots$	z_s	$\cdots$	z_n	c_0
		$c_1 - z_1$	$\cdots$	$c_i - z_i$	$\cdots$	$c_m - z_m$	$c_{m+1} - z_{m+1}$	$\cdots$	$c_s - z_s$	$\cdots$	$c_n - z_n$	

or

$$I_2 = \{j \in \{m+1, \ldots, n\} \mid c_j - z_j > 0\},$$

respectively, is non-empty and, to each index $j \in I_1$ or $j \in I_2$, respectively, there exists an index $i_j \in \{1, \ldots, m\}$ with the property $d_{i_j j} > 0$.It means that in the bottom row of Tab. 37.2 (in the so-called *characteristic row*), there is at least one negative or positive entry $c_j - z_j$, respectively, and in the column of the basis matrix above each such entry, there is at least one positive entry. The process, leading from the original feasible basic point x_0 with index basis $\{i_1, \ldots, i_m\} = \{1, \ldots, m\}$ to the feasible basic point whose index basis differs from the basis $\{i_1, \ldots, i_m\}$ by one index, is called the *index basis change* and represents a proper step of an algorithm we will describe later.

We begin with a feasible basic point that corresponds to the starting (above considered) vertex x_0 of a convex polyhedron M and with the relevant data (1), (2) and (2') arranged in Tab. 37.3 below. If the vertex x_0 is not an optimal feasible point of the considered linear optimization problem, there exist indices k and s, $1 \leqq k \leqq m$ and $m+1 \leqq s \leqq n$, such that

$$d_{ks} \neq 0\,.$$

Then we can calculate the variable x_s from the k-th equation in (1), according to the formula

$$x_s = \frac{1}{d_{ks}}\left(d_{k0} - \sum_{\substack{j=m+1 \\ j \neq s}}^{n} d_{kj} x_j - x_k\right).$$

Substituting this into the remaining equations (1), we obtain

$$x_i = d_{i0} - \frac{d_{is} d_{k0}}{d_{ks}} - \sum_{\substack{j=m+1 \\ j \neq s}}^{n} \left(d_{ij} - \frac{d_{is} d_{kj}}{d_{ks}}\right) x_j + \frac{d_{ik}}{d_{ks}} x_k \quad (i = 1, \ldots, m;\ i \neq k)$$

and, substituting into the objective function (2) with its coefficients given in (2'), we further have

$$f = c_0 + (c_s - z_s)\frac{d_{k0}}{d_{ks}} + \sum_{\substack{j=m+1 \\ j \neq s}}^{n} \left[(c_j - z_j) - (c_s - z_s)\frac{d_{kj}}{d_{ks}}\right] x_j + \frac{c_s - z_s}{d_{ks}} x_k\,.$$

Writing the obtained equations briefly as

$$x_s + \sum_{\substack{j=m+1 \\ j \neq s}}^{n} d'_{sj} x_j + d'_{sk} x_k = d'_{s0}\,,$$

$$x_i + \sum_{\substack{j=m+1 \\ j\neq s}}^{n} d'_{ij}x_j + d'_{ik}x_k = d'_{i0} \quad (i = 1, \ldots, m;\ i \neq k),$$

$$f = c'_0 + \sum_{j=1}^{n}(c_j - z'_j)x_j$$

and comparing the coefficients, we find that

$$\begin{aligned} &d'_{s0} = \frac{d_{k0}}{d_{ks}}, \quad d'_{sj} = \frac{d_{kj}}{d_{ks}} \quad (j = m+1, \ldots, n;\ j \neq s), \quad d'_{sk} = \frac{1}{d_{ks}}, \\ &d'_{i0} = d_{i0} - \frac{d_{is}d_{k0}}{d_{ks}}, \quad d'_{ij} = d_{ij} - \frac{d_{is}d_{kj}}{d_{ks}}, \quad d'_{ik} = -\frac{d_{is}}{d_{ks}} \\ &\qquad (i = 1, \ldots, m;\ i \neq k \text{ and } j = m+1, \ldots, n;\ j \neq s), \end{aligned} \tag{4}$$

$$\begin{aligned} &z'_j = z_j + (c_s - z_s)d'_{sj} \quad (j = m+1, \ldots, n;\ j \neq s), \\ &z'_k = c_k + (c_s - z_s)d'_{sk}, \quad c'_0 = c_0 + (c_s - z_s)d'_{s0}, \\ &z'_i = c_i \quad (i = 1, \ldots, m;\ i \neq k), \quad z'_s = c_s. \end{aligned} \tag{5}$$

From this we further have

$$\begin{aligned} &c_j - z'_j = (c_j - z_j) - (c_s - z_s)d'_{sj} \quad (j = m+1, \ldots, n;\ j \neq s), \\ &c_k - z'_k = -(c_s - z_s)d'_{sk}, \\ &c_i - z'_i = 0 \quad (i = 1, \ldots, m;\ i \neq k), \quad c_s - z'_s = 0, \end{aligned} \tag{6}$$

$$\begin{aligned} &\sum_{\substack{i=1 \\ i\neq k}}^{m} c_i d'_{ij} + c_s d'_{sj} = z'_j \quad (j = m+1, \ldots, n;\ j \neq s), \\ &\sum_{\substack{i=1 \\ i\neq k}}^{m} c_i d'_{ik} + c_s d'_{sk} = z'_k, \quad \sum_{\substack{i=1 \\ i\neq k}}^{m} c_i d'_{i0} + c_s d'_{s0} = c'_0. \end{aligned} \tag{7}$$

The elimination just described together with the corresponding transformation relations (4), (5), (6), and (7) represent the transition from a feasible basic point to a new basic solution of the equations (37.7.4a) which, however, need not be a feasible basic point of the given optimization problem, in general (see Definitions 37.7.1 and 37.7.2). The systems of basic variables of the original and the new basic solution differ just by a single variable. The transformation formulae (4), (5), (6), and (7) describe the calculations necessary in the successive steps of the desired algorithm. These transformations can be carried out in a routine way with the help of Tab. 37.3. This "tableau transformation" is performed as follows: We enclose the above considered non-zero entry d_{ks} (it is called a *pivot* or *central element*) in a box. The pivot lies in the k-th row (called a *pivot row*) and in the s-th column

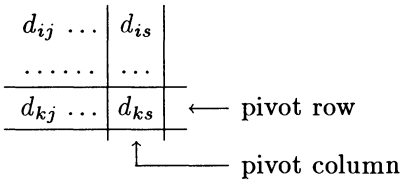
(called a *pivot column*) of the basis matrix in Tab. 37.3. The last row of Tab. 37.3, which contains the values of $c_j - z_j$ $(j = 1, \ldots, n)$, is called the *characteristic row* and it is the key for the decision whether the corresponding feasible basic point (the vertex x_0 of the polyhedron M in the case considered) is an optimal point of the given linear optimization problem of § 37.7. From Tab. 37.3, we come to the next tableau (see Tab. 37.4), proceeding in accord with the following rules describing the transformation:

1° We exchange the variable x_k with the variable x_s in the second column and the number c_k with the number c_s in the first row of Tab. 37.3.

2° We divide all the entries of the pivot row of the basis matrix as well as the last entry d_{k0} of this row by the pivot d_{ks}.

3° We put all the entries of the pivot column except for the pivot element d_{ks} equal to zero.

4° We calculate all the entries of the basis matrix in Tab. 37.4 that lie neither in the k-th row nor in the s-th column, using the so-called "*cross rule*". Let us calculate e.g. the entry d'_{ij} in Tab. 37.4 that corresponds to the entry d_{ij} (where $i \neq k$ and $j \neq s$) at the same position in Tab. 37.3. In Tab. 37.3, consider the four entries



and subtract the product of the entries d_{kj} and d_{is} of the pivot row and column, divided by the pivot element d_{ks}, from the original entry d_{ij} at the considered position in Tab. 37.3. Similarly we calculate the entries of the last column in Tab. 37.4 except for the entry d'_{s0}.

5° We obtain the entry z'_j as an inner product of the first column of Tab. 37.4 and the j-th column of the basis matrix in this tableau. We calculate c'_0 analogously.

6° The quantity $c_j - z'_j$ $(j = 1, \ldots, n)$ can be calculated in two different ways, either as the difference of the second row and the last but one row of Tab. 37.4, or by the cross rule. It is advantageous to use both the methods and compare the results.

When the procedure described is implemented in a computer, the quantities computed are stored at the same locations but the above rules are applied in the sequence 1°, 4°, 2°, 3°, 5° and 6°.

The presented calculation rules completely describe the transition from a certain basic solution to another one. These two basic solutions may represent either two

TABLE 37.3

		x_1	$\dots$	x_k	$\dots$	x_m	x_{m+1}	$\dots$	x_s	$\dots$	x_n	
		c_1	$\dots$	c_k	$\dots$	c_m	c_{m+1}	$\dots$	c_s	$\dots$	c_n	
c_1	x_1	1	$\dots$	0	$\dots$	0	$d_{1,m+1}$	$\dots$	d_{1s}	$\dots$	d_{1n}	d_{10}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_k	x_k	0	$\dots$	1	$\dots$	0	$d_{k,m+1}$	$\dots$	$\boxed{d_{ks}}$	$\dots$	d_{kn}	d_{k0}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_m	x_m	0	$\dots$	0	$\dots$	1	$d_{m,m+1}$	$\dots$	d_{ms}	$\dots$	d_{mn}	d_{m0}
		z_1	$\dots$	z_k	$\dots$	z_m	z_{m+1}	$\dots$	z_s	$\dots$	z_n	c_0
		$c_1 - z_1$	$\dots$	$c_k - z_k$	$\dots$	$c_m - z_m$	$c_{m+1} - z_{m+1}$	$\dots$	$c_s - z_s$	$\dots$	$c_n - z_n$	

TABLE 37.4

	x_1	$\cdots$	x_k	$\cdots$	x_m	x_{m+1}	$\cdots$	x_s	$\cdots$	x_n	
	c_1	$\cdots$	c_k	$\cdots$	c_m	c_{m+1}	$\cdots$	c_s	$\cdots$	c_n	
c_1	x_1	d'_{11}	$\cdots$	d'_{1k}	$\cdots$	d'_{1m}	$d'_{1,m+1}$	$\cdots$	0	$\cdots$	d'_{1n}	d'_{10}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_{k-1}	x_{k-1}	$d'_{k-1,1}$	$\cdots$	$d'_{k-1,k}$	$\cdots$	$d'_{k-1,m}$	$d'_{k-1,m+1}$	$\cdots$	0	$\cdots$	$d'_{k-1,n}$	$d'_{k-1,0}$
c_s	x_s	d'_{s1}	$\cdots$	d'_{sk}	$\cdots$	d'_{sm}	$d'_{s,m+1}$	$\cdots$	1	$\cdots$	d'_{sn}	d'_{s0}
c_{k+1}	x_{k+1}	$d'_{k+1,1}$	$\cdots$	$d'_{k+1,k}$	$\cdots$	$d'_{k+1,m}$	$d'_{k+1,m+1}$	$\cdots$	0	$\cdots$	$d'_{k+1,n}$	$d'_{k+1,0}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_m	x_m	d'_{m1}	$\cdots$	d'_{mk}	$\cdots$	d'_{mm}	$d'_{m,m+1}$	$\cdots$	0	$\cdots$	d'_{mn}	d'_{m0}
		z'_1	$\cdots$	z'_k	$\cdots$	z'_m	z'_{m+1}	$\cdots$	z'_s	$\cdots$	z'_n	c'_0
		$c_1 - z'_1$	$\cdots$	$c_k - z'_k$	$\cdots$	$c_m - z'_m$	$c_{m+1} - z'_{m+1}$	$\cdots$	$c_s - z'_s$	$\cdots$	$c_n - z'_n$	

different points, or, in case of degeneracy, sometimes also a single point. In order that the above described procedure may lead to an algorithm for the computation of the optimal basic point of the considered linear optimization problem from § 37.7, we have to choose the pivot element from the tableau form of the feasible basic solution in each step in such a way that the transition to a new basic solution yields a feasible basic point with a better (or at least with the same) value of the objective function. This choice is considered in the following two theorems that refer to the data in Tab. 37.3 and correspond to a starting feasible basic point x_0.

Theorem 3. *If there exists an index pair* (k, s), $1 \leqq k \leqq m$, $m+1 \leqq s \leqq n$, *possessing the properties*

(i) $c_s - z_s < 0$ *(or* $c_s - z_s > 0$*)*,

(ii) $d_{k0} > 0$, $d_{ks} > 0$, $\dfrac{d_{k0}}{d_{ks}} = \min\limits_{i \in I'} \dfrac{d_{i0}}{d_{is}}$,

where

$$I' = \{i \in \{1, \ldots, m\} \mid d_{is} > 0\},$$

then the variables $x_1, \ldots, x_{k-1}, x_s, x_{k+1}, \ldots, x_m$ *are basic variables of the feasible basic point* x' *which is represented by the tableau form* 37.4 *and for which*

$$f(x') < f(x_0) \text{ or } f(x') > f(x_0),$$

respectively, holds.

Theorem 4. *If there exists an index pair* (k, s), $1 \leqq k \leqq n$, $m+1 \leqq s \leqq n$, *possessing the properties*

(i) $c_s - z_s < 0$ *(or* $c_s - z_s > 0$*)*,

(ii) $\dfrac{d_{k0}}{d_{ks}} = \min\limits_{i \in I'} \dfrac{d_{i0}}{d_{is}} = 0$,

where

$$I' = \{i \in \{1, \ldots, m\} \mid d_{is} > 0\} \neq \emptyset,$$

then (with the notation introduced in Theorem 3) $x' = x_0$ *and* $f(x') = f(x_0)$.

REMARK 2. Theorems 3 and 4 are useful if none of the cases described in Theorems 1 and 2 occurs, i.e., if we have to consider the case from Remark 1. Theorems 3 and 4 give a procedure for the transition to a feasible basic point with a different index basis by the exchange of a single basic variable, with the help of the appropriate choice of a pivot element and with the aim to reach a better value of the objective function. If the hypothesis of Theorem 3 is fulfilled, this end is gained. If, however, the case of Theorem 4 takes place, then the value of the objective function does not change and there are certain troubles to locate a next feasible basic point, that are connected — from the geometrical viewpoint — with the question how

to come from the starting vertex of the polyhedron M to its neighbouring vertex which yields a better value of the objective function. After a next change of basic variables, the case of Theorem 4 can, however, occur for the corresponding basic point if, under the hypothesis of Theorem 3, the choice of pivot element from the formula

$$\frac{d_{k0}}{d_{ks}} = \min_{i \in I'} \frac{d_{i0}}{d_{is}}$$

is ambiguous.

REMARK 3. The case of Theorem 4 occurs if the feasible basic solution (see Definition 37.7.1) is degenerate, what is briefly called a *degeneracy* in the theory of linear programming. The occurrence of degeneracy is connected with the structure of the feasible set M in (37.7.1) and there are methods to avoid this case. One of the approaches consists in assigning, to the given linear optimization problem, another one (called the *ε-perturbed problem*) with the same objective function $\sum_{i=1}^{n} c_i x_i$ but with the feasible set

$$M(\varepsilon) = \left\{ x \in E_n \mid \sum_{i=1}^{n} a_{ji} x_i = b_j + \sum_{i=1}^{n} a_{ji} \varepsilon^i \ (j = 1, \ldots, m), \ x_i \geqq 0 \ (i = 1, \ldots, n) \right\}$$

where ε is a very small positive number from the next theorem:

Theorem 5. *To every linear optimization problem in equality form for which the assumptions* a) *and* b) *from* § 37.7 *are satisfied, there exists such a number* ε_0, *with* $0 < \varepsilon_0 < 1$, *that every feasible basic solution of the assigned* ε*-perturbed optimization problem is non-degenerate for any* $\varepsilon \in (0, \varepsilon_0)$. *Moreover, the original and the assigned* ε*-perturbed optimization problems are either both solvable, or both have no solution. If* $x_0(\varepsilon_0)$ *is the optimal point of the assigned* ε*-perturbed problem, then*

$$x_0 = \lim_{\varepsilon \to 0+} x_0(\varepsilon)$$

is the optimal point of the original optimization problem.

37.9. Simplex Method. An Example

The algorithm we are going to present (currently called the *simplex method in the theory of linear programming*) corresponds to the following geometrical idea: We start with a known vertex x_0 of the polyhedron M and — if x_0 is not the optimal point of the given linear optimization problem from § 37.7 — we proceed along the edges of the polyhedron M from a vertex to a neighbouring vertex until we reach a vertex that is the optimal point of the optimization problem considered.

Since the polyhedron M has a finite number of vertices, the number of steps of the algorithm exploiting the geometrical idea presented is finite as well.

Consider the linear optimization problem in equality form from § 37.7, i.e. the problem to find the minimum (or maximum) of the given linear function

$$f = \sum_{i=1}^{n} c_i x_i \quad (\boldsymbol{c} = (c_1, \ldots, c_n) \text{ is a non-zero vector})$$

on the set

$$M = \left\{ x \in E_n \mid \sum_{i=1}^{n} a_{ji} x_i = b_j \quad (j = 1, \ldots, m),\ x_i \geqq 0\ (i = 1, \ldots, n) \right\}, \qquad (1)$$

or, more exactly, the problem of finding at least one point $\tilde{x} \in M$ at which the function f assumes its minimum (or maximum) on M.

The assumptions:

1° $1 \leqq m < n$.

2° The rank of the matrix $\boldsymbol{A}$ of the coefficients of the system of equations in the description (1) of the set M is m.

3° The point $x_0 = (x_{01}, \ldots, x_{0n})$ is a known feasible basic point with the corresponding system of basic variables $x_1, \ldots, x_m$ (this can always be achieved by renumbering the variables).

Initialization Step of the Algorithm: We solve the system of equations from the description (1) of the set M for the basic variables $x_1, \ldots, x_m$,

$$x_i = d_{i0} - \sum_{j=m+1}^{n} d_{ij} x_j \quad (i = 1, \ldots, m).$$

In the way presented in § 37.8, we now form Tab. 37.5 ("simplex tableau") that corresponds to the initial feasible basic point x_0.

If the hypothesis of Theorem 37.8.1 is satisfied for the data on the bottom (characteristic) row of Tab. 37.5, the point x_0 is the optimal point and the algorithm ends.

If the assumptions of Theorem 37.8.2 are satisfied, the given problem has no solution and the algorithm ends as well.

If none of the two cases occurs, the first step of the algorithm follows.

1*st Step*: In the characteristic (last) row of Tab. 37.5 (where we have $c_i - z_i = 0$ for $i = 1, \ldots, m$ in accordance with (37.8.2')), we choose an index j_0 with the

TABLE 37.5

	x_1	$\cdots$	x_i	$\cdots$	x_m	x_{m+1}	$\cdots$	x_j	$\cdots$	x_n	
	c_1	$\cdots$	c_i	$\cdots$	c_m	c_{m+1}	$\cdots$	c_j	$\cdots$	c_n	
c_1	x_1	1	$\cdots$	0	$\cdots$	0	$d_{1,m+1}$	$\cdots$	d_{1j}	$\cdots$	d_{1n}	d_{10}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_i	x_i	0	$\cdots$	1	$\cdots$	0	$d_{i,m+1}$	$\cdots$	d_{ij}	$\cdots$	d_{in}	d_{i0}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_m	x_m	0	$\cdots$	0	$\cdots$	1	$d_{m,m+1}$	$\cdots$	d_{mj}	$\cdots$	d_{mn}	d_{m0}
		z_1	$\cdots$	z_i	$\cdots$	z_m	z_{m+1}	$\cdots$	z_j	$\cdots$	z_n	c_0
		$c_1 - z_1$	$\cdots$	$c_i - z_i$	$\cdots$	$c_m - z_m$	$c_{m+1} - z_{m+1}$	$\cdots$	$c_j - z_j$	$\cdots$	$c_n - z_n$	

property

$$c_{j_0} - z_{j_0} = \min_{l \in \{1,\dots,n\}} (c_l - z_l) \quad \text{or} \quad c_{j_0} - z_{j_0} = \max_{l \in \{1,\dots,n\}} (c_l - z_l). \tag{2}$$

If there are more indices with the property (2), we choose e.g. the smallest of them. This choice is based on no theoretical foundation; it represents only a unique choice from the set of indices j_0 with the property $c_{j_0} - z_{j_0} < 0$ (or $c_{j_0} - z_{j_0} > 0$). The uniqueness of the choice can be also achieved e.g. in such a way that we choose the first negative (or positive) entry from the left in the characteristic row of Tab. 37.5. The particular decision guaranteeing uniqueness of the choice is completely immaterial for the algorithm itself. To the uniquely chosen index j_0 with the property $c_{j_0} - z_{j_0} < 0$ (or $c_{j_0} - z_{j_0} > 0$), we introduce the index set

$$I_{j_0} = \{i \in \{1, \dots, m\} \mid d_{ij_0} > 0\}$$

that is non-empty in accordance with Remark 37.8.1. Further we introduce the index set

$$I^0_{j_0} = \left\{ l \in I_{j_0} \mid \frac{d_{l0}}{d_{lj_0}} = \min_{i \in I_{j_0}} \frac{d_{i0}}{d_{ij_0}} \right\}. \tag{3}$$

If the set (3) contains exactly one element i_0, we choose the entry $d_{i_0 j_0}$ in Tab.37.5, which is apparently positive, for the pivot element, the i_0-th row for the pivot row, and the j_0-th column for the pivot column (see Tab. 37.6).

TABLE 37.6

		x_1	$\dots$	x_α	$\dots$	x_{j_0}	$\dots$	
		c_1	$\dots$	c_α	$\dots$	c_{j_0}	$\dots$	
c_1	x_1	$\dots\dots\dots$		$d_{1\alpha}$	$\dots$	d_{1j_0}	$\dots$	d_{10}
$\vdots$	$\vdots$			$\vdots$		$\vdots$		$\vdots$
c_i	x_i	$\dots\dots\dots$		$d_{i\alpha}$	$\dots$	d_{ij_0}	$\dots$	d_{i0}
$\vdots$	$\vdots$			$\vdots$		$\vdots$		$\vdots$
c_{i_0}	x_{i_0}	$\dots\dots\dots$		$d_{i_0\alpha}$	$\dots$	$\boxed{d_{i_0 j_0}}$	$\dots$	$d_{i_0 0}$
$\vdots$	$\vdots$			$\vdots$		$\vdots$		$\vdots$
c_m	x_m	$\dots\dots\dots$		$d_{m\alpha}$	$\dots$	d_{mj_0}	$\dots$	d_{m0}
		z_1	$\dots$	z_α	$\dots$	z_{j_0}	$\dots$	c_0
		$c_1 - z_1$	$\dots$	$c_\alpha - z_\alpha$	$\dots$	$c_{j_0} - z_{j_0}$	$\dots$	

If the index set $I_{j_0}^0$ in (3) contains more indices, we introduce a further index set

$$I_{j_0}^1 = \left\{ l \in I_{j_0}^0 \mid \frac{d_{l1}}{d_{lj_0}} = \min_{i \in I_{j_0}^0} \frac{d_{i1}}{d_{ij_0}} \right\}.$$

If the index i_0 is its single element, we choose the entry $d_{i_0 j_0}$ for the pivot element, the i_0-th row for the pivot row, and the j_0-th column for the pivot column. In case that even the set $I_{j_0}^1$ contains more than one index, we introduce a further index set

$$I_{j_0}^2 = \left\{ l \in I_{j_0}^1 \mid \frac{d_{l2}}{d_{lj_0}} = \min_{i \in I_{j_0}^1} \frac{d_{i2}}{d_{ij_0}} \right\}.$$

If the set $I_{j_0}^2$ contains a single index i_0, then we choose the entry $d_{i_0 j_0}$ in Tab. 37.5 for the pivot element, end the i_0-th row and the j_0-th column for the pivot ones. Otherwise we again introduce a further index set by a general formula

$$I_{j_0}^k = \left\{ l \in I_{j_0}^{k-1} \mid \frac{d_{lk}}{d_{lj_0}} = \min_{i \in I_{j_0}^{k-1}} \frac{d_{ik}}{d_{ij_0}} \right\} \quad (k \geqq 1).$$

A theoretical analysis of degeneracy shows (see e.g. [355], p. 104–119) that the presented process ends with a certain set $I_{j_0}^s$ containing one and only one element i_0 that determines the pivot element $d_{i_0 j_0}$ and the corresponding pivot row and column in Tab. 37.5 uniquely. Having uniquely chosen the pivot element $d_{i_0 j_0}$ (marked in Tab. 37.6), we change the original system $x_1, \ldots, x_m$ of basic variables and arrive at a new system $x_1, \ldots, x_{i_0-1}, x_{j_0}, x_{i_0+1}, \ldots, x_m$ of basic variables using the procedure described by 1° to 6° of § 37.8. In the way completely analogous to the transition from Tab. 37.3 to Tab. 37.4, we thus obtain Tab. 37.7 corresponding either to a new feasible basic point $x^{(1)} \neq x_0$ or — in case of degeneracy — again to the starting feasible basic point x_0 with a different index basis. The above described procedure for a unique choice of pivot element prevents the algorithm from *cycling*, i.e. the case, when, after some finite number of steps, we come back to a point obtained already in some previous step. According to Theorems 37.8.1 and 37.8.2, we find out from the characteristic row of Tab. 37.7 whether the algorithm ends (i.e. whether the point $x^{(1)}$ is the optimal point of the problem or whether the problem has no solution) or whether the next, *second step* of the algorithm is to be considered. In this step, we proceed completely analogously to the first step, the only difference consisting in the fact that the point $x^{(1)}$ is now taken for the starting feasible basic point. Proceeding in this way, we obtain, step by step, a finite sequence $x_0, x^{(1)}, \ldots, x^{(s)}$ of vertices of the polyhedron M. Moreover, if we consider the data in the relevant characteristic row of the tableau corresponding to the feasible basic solution $x^{(s)}$, we can decide in the last, s-th step of the algorithm, whether the point $x^{(s)}$ is optimal or whether the given optimization problem has no solution.

TABLE 37.7

		x_1	$\dots$	x_α	$\dots$	x_{j_0}	$\dots$	x_β	$\dots$	x_n	
		c_1	$\dots$	c_α	$\dots$	c_{j_0}	$\dots$	c_β	$\dots$	c_n	
c_1	x_1	${}^{(1)}d_{11}$	$\dots$	${}^{(1)}d_{1\alpha}$	$\dots$	$0 = {}^{(1)}d_{1j_0}$	$\dots$	${}^{(1)}d_{1\beta}$	$\dots$	${}^{(1)}d_{1n}$	${}^{(1)}d_{10}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_{i_0-1}	x_{i_0-1}	${}^{(1)}d_{i_0-1,1}$	$\dots$	${}^{(1)}d_{i_0-1,\alpha}$	$\dots$	$0 = {}^{(1)}d_{i_0-1,j_0}$	$\dots$	${}^{(1)}d_{i_0-1,\beta}$	$\dots$	${}^{(1)}d_{i_0-1,n}$	${}^{(1)}d_{i_0-1,0}$
c_{j_0}	x_{j_0}	${}^{(1)}d_{j_01}$	$\dots$	${}^{(1)}d_{j_0\alpha}$	$\dots$	$1 = {}^{(1)}d_{j_0j_0}$	$\dots$	${}^{(1)}d_{j_0\beta}$	$\dots$	${}^{(1)}d_{j_0n}$	${}^{(1)}d_{j_00}$
c_{i_0+1}	x_{i_0+1}	${}^{(1)}d_{i_0+1,1}$	$\dots$	${}^{(1)}d_{i_0+1,\alpha}$	$\dots$	$0 = {}^{(1)}d_{i_0+1,j_0}$	$\dots$	${}^{(1)}d_{i_0+1,\beta}$	$\dots$	${}^{(1)}d_{i_0+1,n}$	${}^{(1)}d_{i_0+1,0}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_m	x_m	${}^{(1)}d_{m1}$	$\dots$	${}^{(1)}d_{m\alpha}$	$\dots$	$0 = {}^{(1)}d_{mj_0}$	$\dots$	${}^{(1)}d_{m\beta}$	$\dots$	${}^{(1)}d_{mn}$	${}^{(1)}d_{m0}$
		${}^{(1)}z_1$	$\dots$	${}^{(1)}z_\alpha$	$\dots$	${}^{(1)}z_{j_0} = c_{j_0}$	$\dots$	${}^{(1)}z_\beta$	$\dots$	${}^{(1)}z_n$	${}^{(1)}c_0$
		$c_1 - {}^{(1)}z_1$	$\dots$	$c_\alpha - {}^{(1)}z_\alpha$	$\dots$	$0 = c_{j_0} - {}^{(1)}z_{j_0}$	$\dots$	$c_\beta - {}^{(1)}z_\beta$	$\dots$	$c_n - {}^{(1)}z_n$	

Example 1. Find the maximum of the objective function

$$f = 6x + 6y + z$$

on the feasible set

$$\begin{gathered} 6x - 6y + z \leqq 6, \\ 6x + 6y + z \leqq 12, \\ 6x - 6y - z \geqq -6, \\ 6x + 6y + z \geqq 0, \\ 0 \leqq x \leqq 1, \quad 0 \leqq y \leqq 1, \quad 0 \leqq z \leqq 3. \end{gathered}$$

Introducing slack variables $\xi_1, \ldots, \xi_7$ (see Definition 37.3.1), we obtain a linear optimization problem in equality form with the same characteristic function

$$f = 6x + 6y + z + 0 \cdot \xi_1 + 0 \cdot \xi_2 + \cdots + 0 \cdot \xi_7 \tag{4}$$

and the feasible set M described by the relations

$$\begin{array}{rrrrrrrrrrrr} 6x & -6y & +z & +\xi_1 & & & & & & & = & 6, \\ 6x & +6y & +z & & +\xi_2 & & & & & & = & 12, \\ -6x & +6y & +z & & & +\xi_3 & & & & & = & 6, \\ 6x & +6y & +z & & & & -\xi_4 & & & & = & 0, \\ x & & & & & & & +\xi_5 & & & = & 1, \\ & y & & & & & & & +\xi_6 & & = & 1, \\ & & z & & & & & & & +\xi_7 & = & 3, \end{array} \tag{4'}$$

$$x \geqq 0, \quad y \geqq 0, \quad z \geqq 0, \quad \xi_r \geqq 0 \quad (r = 1, \ldots, 7).$$

The assumptions 1° and 2° from the beginning of this paragraph are obviously satisfied for the set M described by (4'). The equations in (4') can be easily solved for the variables z, ξ_1, ξ_2, ξ_3, ξ_4, ξ_5 and ξ_6:

$$\begin{aligned} z &= 3 - \xi_7 \\ \xi_1 &= 6 - 6x + 6y - z = 6 - 6x + 6y - (3 - \xi_7) = 3 - 6x + 6y + \xi_7, \\ \xi_2 &= 12 - 6x - 6y - z = 12 - 6x - 6y - (3 - \xi_7) = 9 - 6x - 6y + \xi_7, \\ \xi_3 &= 6 + 6x - 6y - z = 6 + 6x - 6y - (3 - \xi_7) = 3 + 6x - 6y + \xi_7, \\ \xi_4 &= 6x + 6y + z = 6x + 6y + (3 - \xi_7) = 3 + 6x + 6y - \xi_7, \\ \xi_5 &= 1 - x, \\ \xi_6 &= 1 - y, \end{aligned} \tag{5}$$

Putting $x = x_0 = 0$, $y = y_0 = 0$ and $\xi_7 = \xi_{07} = 0$ in the equations (5), we obtain the unique solution $z = z_0 = 3$, $\xi_1 = \xi_{01} = 3$, $\xi_2 = \xi_{02} = 9$, $\xi_3 = \xi_{03} = 3$, $\xi_4 = \xi_{04} = 3$,

$\xi_5 = \xi_{05} = 1$ and $\xi_6 = \xi_{06} = 1$. The point

$$x^{(0)} = (x_0, y_0, z_0, \xi_{01}, \xi_{02}, \xi_{03}, \xi_{04}, \xi_{05}, \xi_{06}, \xi_{07}) = (0, 0, 3, 3, 9, 3, 3, 1, 1, 0)$$

thus represents a feasible basic point of the linear optimization problem with the objective function f from (4) and with the feasible set M described by (4'). The system $\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6, z$ of basic variables corresponds to this point; the variables ξ_7, x and y are non-basic. The point $x^{(0)}$ is thus a known feasible basic point. The assumption 3° from the beginning of this paragraph is thus satisfied, too, and we can apply the simplex method. Note that the choice of the starting feasible basic point was easy here. For a general case see § 37.10.

According to our discussion of the subject, we now construct a tableau (see Tab. 37.8) from (4) and (4'). Do not pay attention to the frame, in which the entry in the first row and the eight column of the basis matrix is enclosed, for the moment. As we are interested in the maximum of the objective function we consider only positive entries in the characteristic row of Tab. 37.8. They are only in the x- and y-columns and their values are 6 and 6. Since not all of the entries of the characteristic row are less than or equal to zero we cannot use Theorem 37.8.1 to conclude whether the point $x^{(0)}$ is optimal for the considered problem. The characteristic row of Tab. 37.8 does not imply that the studied problem has no solution, either, since there are positive numbers in the columns above all the positive entries of the characteristic row (see Theorem 37.8.2). The algorithm thus continues.

TABLE 37.8

		ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6	z	x	y	ξ_7	
		0	0	0	0	0	0	1	6	6	0	
0	ξ_1	1	0	0	0	0	0	0	**6**	−6	−1	3
0	ξ_2	0	1	0	0	0	0	0	6	6	−1	9
0	ξ_3	0	0	1	0	0	0	0	−6	6	−1	3
0	ξ_4	0	0	0	1	0	0	0	−6	−6	1	3
0	ξ_5	0	0	0	0	1	0	0	1	0	0	1
0	ξ_6	0	0	0	0	0	1	0	0	1	0	1
1	z	0	0	0	0	0	0	1	0	0	1	3
		0	0	0	0	0	0	1	0	0	1	3
		0	0	0	0	0	0	0	6	6	−1	

1st Step. We find the largest entry of the objective row in Tab. 37.8. It is the number 6 which lies both in the x- and the y-column. We choose the x-column for

the pivot column since in our ordering of all the variables

$$x_1 = \xi_1,\ x_2 = \xi_2,\ x_3 = \xi_3,\ x_4 = \xi_4,\ \ x_5 = \xi_5,$$
$$x_6 = \xi_6,\ x_7 = z,\ \ x_8 = x,\ \ x_9 = y,\ \ x_{10} = \xi_7,$$

the index of the variable x is the least of all the assigned indices for which the corresponding column of Tab. 37.8 contains positive entries in the characteristic row. The entries of the last column of Tab. 37.8, which lie in the rows containing positive entries in the x-column of the basis matrix, are divided by the corresponding positive numbers. Thus we obtain the ratios

$$\frac{3}{6},\quad \frac{9}{6},\quad \frac{1}{1}$$

and determine uniquely

$$\frac{3}{6} = \min\left(\frac{3}{6}, \frac{9}{6}, \frac{1}{1}\right)$$

where the ratio $\frac{3}{6}$ comes from the first row of the basis matrix. We choose this row for the pivot one. Then the pivot element is determined uniquely (it is enclosed in box in Tab. 37.8). Exchanging the basic variables in the manner described by the operations 1° to 6° in § 37.8, we arrive at the following simplex tableau 37.9 which corresponds to the feasible basic point $x^{(1)}$. The values of the basic variables are $x = \frac{1}{2}$, $\xi_2 = 6$, $\xi_3 = 6$, $\xi_4 = 6$, $\xi_5 = \frac{1}{2}$, $\xi_6 = 1$ and $z = 3$ (they can be found in the last column of Tab. 37.9), the values of the non-basic variables ξ_1, y and ξ_7 are zeroes. The feasible basic point is non-degenerate and has exactly $n - m = 3$ zero coordinates. Since the objective row in Tab. 37.9 contains a single positive entry,

TABLE 37.9

		ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6	z	x	y	ξ_7	
		0	0	0	0	0	0	1	6	6	0	
6	x	$\frac{1}{6}$	0	0	0	0	0	0	1	-1	$-\frac{1}{6}$	$\frac{1}{2}$
0	ξ_2	-1	1	0	0	0	0	0	0	12	0	6
0	ξ_3	1	0	1	0	0	0	0	0	0	-2	6
0	ξ_4	1	0	0	1	0	0	0	0	-12	0	6
0	ξ_5	$-\frac{1}{6}$	0	0	0	1	0	0	0	**1**	$\frac{1}{6}$	$\frac{1}{2}$
0	ξ_6	0	0	0	0	0	1	0	0	1	0	1
1	z	0	0	0	0	0	0	1	0	0	1	3
		1	0	0	0	0	0	1	6	-6	0	6
		-1	0	0	0	0	0	0	0	12	0	

number 12, and since not all of the entries of the corresponding y-column are less than or equal to zero, the algorithm continues.

2nd Step. The column of Tab. 37.9 containing entry 12 in the characteristic row is uniquely the pivot column. Positive entries in this column of the basis matrix are (from top to bottom) numbers 12, 1 and 1 that lie in the second, fifth, and sixth row of the matrix, respectively. The entries of the last column of Tab. 37.9, which lie in these rows, are divided (in this order) by these numbers. We thus obtain the ratios $\frac{6}{12}$, $\frac{\frac{1}{2}}{1}$ and $\frac{1}{1}$, i.e. $\frac{1}{2}$, $\frac{1}{2}$ and 1, from which the smallest value is $\frac{1}{2}$ and the corresponding row index is determined ambiguously (the ratio $\frac{1}{2}$ corresponds to both the ξ_2-row and the ξ_5-row). In the next course of the above presented algorithm, we calculate further ratios as follows: The entry in the ξ_1-column and the ξ_2-row of the basis matrix is divided by the entry in the pivot y-column and the ξ_2-row. Similarly the entry in the ξ_1-column and the ξ_5-row is divided by the entry in the pivot y-column and the ξ_5-row. We obtain the ratios $-\frac{1}{12}$ and $\frac{-\frac{1}{6}}{1}$ (in this order), i.e. $-\frac{1}{12}$ and $-\frac{1}{6}$, and the smallest of them, namely $-\frac{1}{6}$, corresponds uniquely to the ξ_5-row which is chosen for the pivot one. This determines the pivot element (enclosed in box in Tab. 37.9) uniquely. Changing the basis (basic variables), we arrive at a new tableau (see Tab. 37.10) that implies neither the optimality of the corresponding feasible basic point $x^{(2)}$ (the values of the basic variables are $x = 1$, $\xi_2 = 0$, $\xi_3 = 6$, $\xi_4 = 12$, $y = \frac{1}{2}$, $\xi_6 = \frac{1}{2}$ and $z = 3$; the values of the non-basic variables ξ_1, ξ_5 and ξ_7 are zero; the point $x^{(2)}$ is degenerate feasible basic point as it has more than $n - m$, i.e. more than 3, zero coordinates), nor the conclusion that the given optimization problem has no solution. A next step of the algorithm thus follows.

TABLE 37.10

		ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6	z	x	y	ξ_7	
		0	0	0	0	0	0	1	6	6	0	
6	x	0	0	0	0	1	0	0	1	0	0	1
0	ξ_2	**1**	1	0	0	−12	0	0	0	0	−2	0
0	ξ_3	1	0	1	0	0	0	0	0	0	−2	6
0	ξ_4	−1	0	0	1	12	0	0	0	0	2	12
6	y	$-\frac{1}{6}$	0	0	0	1	0	0	0	1	$\frac{1}{6}$	$\frac{1}{2}$
0	ξ_6	$\frac{1}{6}$	0	0	0	−1	1	0	0	0	$-\frac{1}{6}$	$\frac{1}{2}$
1	z	0	0	0	0	0	0	1	0	0	1	3
		−1	0	0	0	12	0	1	6	6	2	12
		1	0	0	0	−12	0	0	0	0	−2	

3rd Step. Since the characteristic row of Tab. 37.10 contains a single positive entry (number 1 in the ξ_1-column of the tableau) the pivot columns is determined by it uniquely. The pivot ξ_1-column contains positive entries in the ξ_2-, ξ_3- and ξ_6-row, namely numbers 1, 1 and $\frac{1}{6}$. By them, the entries of the last column of Tab. 37.10, which lie in these rows, are divided (in this order). We thus obtain the ratios $\frac{0}{1}$, $\frac{6}{1}$ and $\frac{1/2}{1/6}$, the smallest of them being equal to zero and corresponding uniquely to the ξ_2-row. This uniquely determines the pivot element (enclosed in box in Tab. 37.10). Changing the basis, we arrive at next tableau (see Tab. 37.11) whose characteristic row contains only non-positive entries. According to Theorem 37.8.1, the corresponding feasible basic point $x^{(3)}$ with the value $x = 1$, $\xi_1 = 0$, $\xi_3 = 6$, $\xi_4 = 12$, $y = \frac{1}{2}$, $\xi_6 = \frac{1}{2}$ and $z = 3$ of basic variables and the values $\xi_2 = \xi_5 = \xi_7 = 0$ of non-basic variables is the optimal point of the linear maximization problem with the objective function (4) and the feasible set (4'). This implies that the point $(\tilde{x}, \tilde{y}, \tilde{z}) = (1, \frac{1}{2}, 3)$ is the optimal point of the original optimization problem considered. Further, the value of the objective function at this optimal point is equal to 12 (the last entry in the last column of Tab. 37.11).

TABLE 37.11

		ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6	z	x	y	ξ_7	
		0	0	0	0	0	0	1	6	6	0	
6	x	0	0	0	0	1	0	0	1	0	0	1
0	ξ_1	1	1	0	0	−12	0	0	0	0	−2	0
0	ξ_3	0	−1	1	0	12	0	0	0	0	0	6
0	ξ_4	0	1	0	1	0	0	0	0	0	0	12
6	y	0	$\frac{1}{6}$	0	0	−1	0	0	0	1	$-\frac{1}{6}$	$\frac{1}{2}$
0	ξ_6	0	$-\frac{1}{6}$	0	0	1	1	0	0	0	$\frac{1}{6}$	$\frac{1}{2}$
1	z	0	0	0	0	0	0	1	0	0	1	3
		0	1	0	0	0	0	1	6	6	0	12
		0	−1	0	0	0	0	0	0	0	0	

37.10. Finding a Feasible Basic Point

The simplex method described in § 37.9 for linear optimization problems in equality form has been derived under the assumption that the number of equations in the feasible set M from (37.9.1) is less than the number of variables and that a starting feasible basic point is known. In practice (economic, primarily), we treat mostly large optimization problems of the type considered, where the number of equality

constraints may be larger than the number of variables, where (in general) the starting feasible basic point (i.e. the initial vertex of the polyhedron M from (37.9.1)) is not known, and where it is also very hard to determine the rank of the matrix $\boldsymbol{A}$ of the coefficients in the equality constraints from the description (37.9.1) of the feasible set. We can, however, use a method based on the simplex method, whose application either provides the information that the feasible set M is empty, or leads to finding a starting feasible basic point. To this end, consider the given linear optimization problem with the feasible set M from (37.9.1) in whose description we can assume that

$$b_j \geqq 0 \quad (j = 1, \ldots, m) \tag{1}$$

(this can be achieved if we multiply the equations with negative right-hand sides by -1). To the given problem, we assign an auxiliary optimization problem

$$\min\{\varphi(u)\}\,! \qquad (u = (u_1, \ldots, u_m)) \tag{2}$$

with the objective function

$$\varphi(u) = \sum_{j=1}^{m} u_j$$

and the feasible set

$$\tilde{M} = \Big\{(x, u) \in E_{n+m} \mid \sum_{i=1}^{n} a_{ji}x_i + u_j = b_j \quad (j = 1, \ldots, m)\,, \quad x_i \geqq 0 \quad (i = 1, \ldots, n)\,, \quad u_j \geqq 0 \quad (j = 1, \ldots, m)\Big\} \tag{2'}$$

The set $\tilde{M}$ is obtained from the set M described by (37.9.1) if we introduce a new variable u_j $(j = 1, \ldots, m)$ in the j-th equation in the way apparent from the description (2') of $\tilde{M}$. The variables u_j $(j = 1, \ldots, m)$ are called *artificial variables.* For the auxiliary linear optimization problem (2) we see, on the one hand, that (1) holds (since this has been assumed) and, on the other hand, that the number of equality constraints in the description (2') of the set $\tilde{M}$ is less than the number $n+m$ of variables $x_1, \ldots, x_n, u_1, \ldots, u_m$. Moreover, these equations can be solved for the variables $u_1, \ldots, u_m$,

$$u_j = b_j - \sum_{i=1}^{n} a_{ji}x_i \quad (j = 1, \ldots, m)\,.$$

Their basic solution (x_0, u_0), where $x_{0i} = 0$ for $i = 1, \ldots, n$ and $u_j = b_j \geqq 0$ for $j = 1, \ldots, m$, represents a feasible basic point for the optimization problem (2) in equality form and the assumptions from § 37.9 for the application of the simplex method are thus satisfied. Since the objective function $\varphi(u)$ is bounded from below

and continuous on the closed set $\tilde{M}$, there always exists the optimal feasible point $(\tilde{x}, \tilde{u})$ of the problem (2). The following theorem holds:

Theorem 1. *If* $(\tilde{x}, \tilde{u}) = (\tilde{x}_1, \ldots, \tilde{x}_n, \tilde{u}_1, \ldots, \tilde{u}_m)$ *is the optimal point of the auxiliary optimization problem* (2), *then* $M = \emptyset$ *in the case* $\sum_{i=1}^{m} j\tilde{u}_j > 0$ (*and thus the original optimization problem with the feasible set* M *from* (39.9.1) *has no solution*), *or the point* $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ *is a feasible basic point of the original optimization problem with the feasible set* M *from* (37.9.1) *in the case* $\tilde{u}_j = 0$ $(j = 1, \ldots, m)$.

REMARK 1. The next Theorem 2, which can be deduced from the so-called theory of *parametric linear programming* (see, e.g., [356]), enables us to omit the auxiliary problem (2) for finding a starting feasible basic point when we solve the given problem

$$\min_{x \in M} \Big\{ \sum_{i=1}^{n} c_i x_i \Big\} ! \quad (\text{or } \max_{x \in M} \Big\{ \sum_{i=1}^{n} c_i x_i \Big\} ! \,) \tag{3}$$

with the feasible set M from (37.9.1). Instead of the originally considered problem (3) we solve another one by the simplex method. It is the assigned problem

$$\min_{(x,u) \in \tilde{M}} \Big\{ \sum_{i=1}^{n} c_i x_i + \mu \sum_{j=1}^{m} u_j \Big\} !$$

or (4)

$$\max_{(x,u) \in \tilde{M}} \Big\{ \sum_{i=1}^{n} c_i x_i - \mu \sum_{j=1}^{m} u_j \Big\} ! \,)$$

with the feasible set $\tilde{M}$ from (2') where μ is a parameter whose value is taken larger than any other number encountered in the computation. This method is called the *parametric μ-method* and the following theorem holds:

Theorem 2. *If there exists a number* μ_0 *such that for no* $\mu > \mu_0$ *there is an optimal feasible basic point* $(\tilde{x}, \tilde{u})$ *of the problem* (4) *with the property* $\tilde{u}_j = 0$ *for* $j = 1, \ldots, m$, *then the original problem* (3) *has no solution. Otherwise there exists a sufficiently large positive number* μ *such that the point* $\tilde{x}$ *is the optimal point of the original problem* (3) (*as long as the point* $(\tilde{x}, \tilde{u})$ *with the property* $\tilde{u}_j = 0$ $(j = 1, \ldots, m)$ *is the optimal basic point of the problem* (4)).

37.11. Duality Principle

Consider a linear maximization problem in normal form (see Definition 37.2.1), i.e. the problem

$$\max_{x \in M} \Big\{ \sum_{i=1}^{n} c_i x_i \Big\} ! \tag{1}$$

with the feasible set

$$M = \Big\{ x \in E_n \mid \sum_{i=1}^{n} a_{ji} x_i \leqq b_j \quad (j = 1, \ldots, m), \quad x_i \geqq 0 \quad (i = 1, \ldots, n) \Big\}. \tag{1'}$$

The integers m and n are chosen arbitrarily and there are no requirements on the rank of the matrix

$$\boldsymbol{A} = \begin{bmatrix} a_{11}, & \ldots, & a_{1n} \\ \ldots\ldots\ldots\ldots\ldots & & \\ a_{m1}, & \ldots, & a_{mn} \end{bmatrix}.$$

To the maximization problem (1), we assign the linear minimization problem

$$\min_{u \in N} \Big\{ \sum_{j=1}^{m} b_j u_j \Big\} ! \tag{2}$$

with the feasible set

$$N = \Big\{ u \in E_m \mid \sum_{j=1}^{m} a_{ji} u_j \geqq c_i \quad (i = 1, \ldots, n), \quad u_j \geqq 0 \quad (j = 1, \ldots, m) \Big\}. \tag{2'}$$

In the literature on linear programming, the pair of the optimization problems (1) and (2) is often written symbolically in the concise matrix form

$$\begin{aligned} &\textit{Problem} \text{ I} : && \max\{\boldsymbol{c}^{\mathrm{T}} \boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} \leqq \boldsymbol{b},\ \boldsymbol{x} \geqq \boldsymbol{o}\}, \\ &\textit{Problem} \text{ II} : && \min\{\boldsymbol{b}^{\mathrm{T}} \boldsymbol{u} \mid \boldsymbol{A}^{\mathrm{T}} \boldsymbol{u} \geqq \boldsymbol{c},\ \boldsymbol{u} \geqq \boldsymbol{o}\}. \end{aligned} \tag{3}$$

From the following Tab. 37.12, we can readily deduce the structure of the pair of linear optimization problems (3), from which the first one is always a maximization problem and the second one a minimization problem.

TABLE 37.12

	x_1	...	x_i	...	x_n	
u_1	a_{11}	...	a_{1i}	...	a_{1n}	$\leqq b_1$
⋮	⋮		⋮		⋮	⋮
u_j	a_{j1}	...	a_{ji}	...	a_{jn}	$\leqq b_j$
⋮	⋮		⋮		⋮	⋮
u_m	a_{m1}	...	a_{mi}	...	a_{mn}	$\leqq b_m$
	$\geqq c_1$	...	$\geqq c_i$	...	$\geqq c_n$	

We obtain the system of constraints $\boldsymbol{A}\boldsymbol{x} \leqq \boldsymbol{b}$ in Problem I formally from Tab. 37.12 if we multiply the row $(x_1, \ldots, x_n)$ by each of the rows $(a_{j1}, \ldots, a_{jn})$ and put the inner product obtained less than or equal to b_j $(j = 1, \ldots, m)$ as is marked in Tab. 37.12. In a similar way, we get the system of constraints $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{u} \geqq \boldsymbol{c}$ in Problem II if we multiply the column $(u_1, \ldots, u_m)$ by each of the columns $(a_{1i}, \ldots, a_{mi})$ and put the result greater than or equal to c_i $(i = 1, \ldots, n)$. The pair of the linear optimization problems I and II (Problem I is a maximization problem, Problem II is a minimization one) is called a *pair of dual linear optimization problems in normal form* and Problem II is called *dual* to the *original* (*primal*) *Problem* I. The following theorems hold:

Theorem 1. *If the feasible sets M and N of the pair of dual Problems* I *and* II *in* (3) *are non-empty, then we have, for an arbitrary point $x \in M$ and an arbitrary point $u \in N$* (where $x = (x_1, \ldots, x_n)$ and $u = (u_1, \ldots, u_m)$),

$$\sum_{i=1}^{n} c_i x_i \leqq \sum_{j=1}^{m} b_j u_j$$

(*i.e.* $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \leqq \boldsymbol{b}^{\mathrm{T}}\boldsymbol{u}$ *in matrix notation*).

Theorem 2 (*Duality Principle*). *If the feasible sets M and N of the dual Problems* I *and* II *in* (3) *are non-empty, then there exist an optimal point x_0 of Problem* I *and an optimal point u_0 of Problem* II, *and the values of the objective functions of both these optimization problems are equal to each other at these optimal points, i.e.* $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}_0 = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{u}_0$.

Theorem 3. *Problem* I *is solvable if and only if Problem* II *is solvable.*

REMARK 1. If x is an arbitrary feasible point of Problem I and u is an arbitrary feasible point of Problem II, then there exist optimal points x_0 and u_0 of Problems I and II according to Theorem 2, and Theorem 1 implies that

$$\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \leqq \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}_0 = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{u}_0 \leqq \boldsymbol{b}^{\mathrm{T}}\boldsymbol{u}\,.$$

We obtain from this that the feasible points found in the individual steps of the algorithm for solving Problems I and II determine an interval containing the optimal value of the objective function. Moreover, the length of this interval decreases as the number of steps grows. Another advantage of the dual problem consists in the fact that, for a great number of practical linear programming problems, solving of the dual problem is simpler and less time-consuming than solving the primal problem.

REMARK 2. A general problem of linear programming stated in § 37.1 can also be assigned a certain dual problem. Moreover, the pair of these problems is such that the statements of Theorems 1, 2, and 3 hold (see, e.g., [355], Chap. 3).

CONCLUDING REMARK. In addition to the just discussed "classical" simplex method, which represents a *finite* algorithm for solving linear programming problems, there exist several its modifications like e.g. the *revised simplex method*, the *dual simplex method*, and the *primal-dual algorithm*. The class of linear programming algorithms, which represent limit (and thus *infinite*) processes, includes some *gradient methods* (see, e.g., [514]) and the so-called *centroid method*. There are two methods important, in the first place, from the theoretical point of view: The *Khachiyan ellipsoid method* and, in particular, the *Karmarkar method of successive projections* (see Karmarkar, N., Combinatorica 1984, pp. 373–395) that converge "in polynomial time". It means that if the problem involves n variables, the time estimate for its solution is $O(n^k)$ where k is an integer. An algorithm possessing this property is said to belong to the class of *polynomial time algorithms*. The simplex method does not have the property mentioned. The estimate for this method is $O(\mathrm{e}^{kn})$. Also the last two of the mentioned algorithms have been modified in different manners and their efficiency has been compared with the simplex method.

A special class of optimization problems belonging to linear programming is the class of the so-called *integer* (*discrete*, in general) *problems of linear programming*. In their solution, the algorithms called *cutting plane methods* prevail. Among them, the so-called *Gomory algorithm* is well-known (Gomory, R.E., Bull. Amer. Math. Society 64(1958)).

REFERENCES

[1] Adams, R.A.: Single Variable Calculus. Cambridge (Mass.), Addison-Wesley 1993.

[2] Andree, R.V.: Selections from Modern Abstract Algebra. Holt, Rinehart and Winston 1962.

[3] Angot, A.: Compléments de Mathématiques à l'usage des ingénieurs de l'électronique et des télécomunications, 3iéme ed. revue et augmentée. Edit. de la Revue d'optique 1957.

[4] Apostol, T.M.: Calculus, Vols 1, 2. Waltham (Mass.), Blaisdell 1991.

[5] Arscott, F.M.: Periodic Differential Equations. Oxford, Pergamon Press 1964.

[6] Artobolevskii, I.I.: Mechanisms for the Generation of Plane Curves. Oxford, Pergamon Press 1964.

[7] Askey, R.A.: Special Functions: Group Theoretical Aspects and Applicatins. Dordrecht, Reidel (Kluwer) 1984.

[8] Askey, R.A.: Ortogonal Polynomials and Special Functions. London, Hayden Books 1975

[9] Askwith, E.H.: Analytical Geometry of the Conic sections. London, Adam and Charles Black 1950.

[10] Auvil, D.L.: Poluga, C.: Elementary Algebra. Cambridge (Mass.), Addison-Wesley 1978.

[11] Ball, R.W.: Principles of Abstract Algebra. Holt, Rinehart and Winston 1963.

[12] Banchoff, T., Wermer, J.: Linear Algebra through Geometry. Berlin, Springer 1963.

[13] Beaumont, R.A., Ball, A.W.: Introduction to Modern Algebra and Matrix Theory. Holt, Rinehart and Winston 1954.

[14] Bell, R.J.T.: Coordinate Solid Geometry. London, Macmillan 1954.

[15] Bermant, A.F.: Course of Mathematical Analysis. Oxford, Pergamon Press 1963.

[16] Bickley, W.G.: Via Vector to Tensor. London, English University Press 1962.

[17] Binmore, K.G.: Mathematical Analysis. 2nd ed. Cambridge, Cambridge Univ. Press 1982.

[18] Birkhoff, G., Bartee, T.G.: Modern Applied Algebra. New York, McGraw-Hill 1968.

[19] Blaschke, W.: Differentialgeometrie I. Berlin, Springer 1930.

[20] Bloom, D.H.: Linear Alebra and Geometry. Cambridge, Cambridge Univ. Press 1979.

[21] Borůvka, O.: Grundlagen der Gruppoid- und Gruppentheorie. Berlin, D. Verlag der Wiss. 1960.

[22] Bosch, S., Guentzer, U.: Remmert, R;: Non-Archimedes Analysis. A Systematic Approch to Rigid Geometry. Berlin, Springer 1984.

[23] Bowen, R.M., Wang, C.C.: Introduction to Vectors and Tensors, I, II. New York, Plenum 1976.

[24] Bowman, F.: Introduction to Determinats and Matrices. London, English University Press 1962.

[25] Brand, L.: Advanced Calculus. New York, Wiley and Sons 1955.

[26] Bronstein, I.N.: Semediaev, K.A.: Taschenbuch der Mathematik für Ingenieure und Studenten der Technischen Hochschulen, 6. Aufl. Leipzig, Teubner 1963.

[27] Burdette, A.C.: An Introduction to Analytic Geometry and Calculus, revised edition. New York, Academic Press 1973.

[28] Burington, R.S.: Handbook of Mathematical Tables and Formulae. New York, McGraw-Hill 1965.

[29] Burkill, J.C.: The Lebesgue Integral. Cambridge, Cambridge Univ. Press 1951.

[30] Burkill, J.C.: A First Course in Mathematical Analysis. Cambridge, Cambridge Univ. Press 1962.

[31] Burkill, J.C.: Burkill, H.: A Second Course in Mathematical Analysis. Cambridge, Cambridge Univ. Press 1970.

[32] Carlson, B.C.: Special Functions of Applied Mathematics. New York, Academic Press 1977.

[33] Chevalley, C.: Fundamental Concepts of Algebra. New York, Academic Press 1956.

[34] Churchill, R.V.: Fourier Series and Boundary Value Problems. New York, McGraw-Hill 1963.

[35] Coburn, N.: Vector and Tensor Analysis. London, Macmillan 1955.

[36] Cohn, P.M.: Universal Algebra. Dordrecht, Reidel (Kluwer) 1980.

[37] Copson, E.T.: Asymptotic Expansions. Cambridge, Cambridge Univ. Press 1964.

[38] Courant, R.: Hilbert, D.: Methods of Mathematical Physics, Vols 1, ,2. New York, Interscience 1962.

[39] Craven, B.D.: Lebesgue Measure and Integral. London, Pitman 1982.

[40] Dallmann, H., Elster, K.H.: Einführung in die höhere Mathematik für Naturwissenschaftler und Ingenieure. Leipzig, Fisher 1987.

[41] Davis, P.J., Rabinowitz, P.: Methods of Numerial Integration. New York, Academic Press 1974.

[42] Dodson, C.T.J.: Tensor Goemetry. London, Pitman 1978.

[43] Durell, C.: A New Trigonometry for Schools. London, Bell and Sons 1963.

[44] Edwards, R.E.: Fourier Series. 2nd ed. Berlin, Springer, Part I 1979, Part II 1982.

[45] Eisenhart, L.P.: An Introduction to Differential Geometry. Princeton Univ. Press 1940.

[46] Eisenreich, G.: Lineare Algebra und analytische Geometrie. Berlin, Akademieverlag 1980.

[47] Engels, H.: Numerical Quadrature and Cubature. Aachen, Inst. für Geometrie und praktische Mathematik 1980.

[48] Erdélyi, A: Asymptotic Expansions. New York, Dover 1956.

[49] Erdélyi, A., Magnus, W., Oberhettinger, F., Tricomi, F.G.: Higher Trancedental Functions, Vols 1, 2, 3. New York, McGraw-Hill 1953-5

[50] Fadeeva, V.N.: Computational Methods of Linear Algebra. London, Constable 1959.

[51] Fadeev, D.K.: Fadeeva, V.N.: Computational Methods of Linear algebra. San Francisco, Freemans 1963.

[52] Ferrar, M.: Special Matrices and Their Applications. London, Oxford Univ. Press 1960

[53] Fiedler, M.: Special Matrices and Their Applications in Numerical Analysis. Dordrecht, Reidel (Kluwer) 1986.

[54] Fikhtengol'ts, G.M.: Fundamentals of Mathematical Analysis, Parts 1, 2. Oxford, Pergamon Press 1965.

[55] Fisher, R.C.: Calculus and Analytic Geometry. Englewood Cliffs (N.J.), Prentice-Hall 1961.

[56] Flanders, H., Price, J.J.: Trigonometry. New York, Academic Press 1975.

[57] Fleming, W.M.: Functions of Several Variables. Cambridge (Mass.), Addison-Wesley 1965.

[58] Fraissé, R.: Course of Mathematical Logic, Vol. I. Dordrecht, Reidel (Kluwer) 1973.

[59] Fraleigh, J.B.: Calculus with Analytic Geometry. Cambridge (Mass.), Addison-Wesley 1980.

[60] Franklin, F.: A Treatise on Advanced Calculus. New York, Wiley and Sons 1940.

[61] Frazer, R.A., Duncan, W.I., Collar, A.R.: Elementary Matrices. Cambridge, Cambridge Univ. Press 1946.

[62] Fuller, G.: Analytic Geometry. 5th ed. Cambridge (Mass.), Addison-Wesley 1979.

[63] Gantmacher, F.R.: Applications of the Theory of Matrices. New York, Interscience 1959.

[64] Goodbody, A.M.: Cartesian Tensors. Chicherster, Horwood 1982.

[65] Graustein, W.C.: Differential Geometry. New York, Dover 1947.

[66] Graves, L.M.: Theory of Functions of Real Variables. New York, McGraw-Hill 1956.

[67] Green, S.L.: Algebraic Solid Geometry. Cambridge, Cambridge Univ. Press 1961.

[68] Grossman, S.I.: Calculus. 3rd ed. New York, Academic Press 1984.

[69] Grove, L.C.: Algebra. New York, Academic Press 1984.

[70] Grzegorczyk, A.: An Outline of Mathematical Logic. Dordrecht, Reidel (Kluwer) 1974.

[71] Hall, F.M.: Introduction to Abstract Algebra. Cambridge, Cambridge Univ. Press 19, Vol. I 1969, Vol. II 1972.

[72] Halmos, P.R.: Measure Theory. London, Van Nostrand 1950.

[73] Halmos, P.R.: Naive Set Theory. London, Van Nostrand 1960.

[74] Hardy, G.H.: A Course of Pure Mathematics. Cambridge, Cambridge Univ. Press 1952.

[75] Hardy, G.H.: Divergent Series. London, Oxford Univ. Press 1949.

[76] Hartman, S., Mikusiński, J.: Theory of Lebesgue Measure and Integration. Oxford, Pergamon Press 1961.

[77] Head, J.W.: Mathematical Techniques in Electronics and Engineering Analysis. London, Van Nostrand 1964.

[78] Henstock, R.: Theory of Integration. London, Butterworths 1963.

[79] Hermes. H.: Introduction to Mathematical Logic. Berlin, Springer 1972.

[80] Hewitt, E.: Stromberg, K.: Real and Abstract Analysis. Berlin, Springer 1975.

[81] Hildebrandt, T.H.: Introduction to the Theory of Integration. New York, Academic Press 1963.

[82] Hlawiczka, P.: Matrix Algebra for Electronic Engineering. London, Hayden Books 1965.

[83] Hungerford, T.W.: Algebra. Berlin, Springer 1980.

[84] Hyslop, J.M.: Infinite series. Edinburgh, Oliver and Boyd 1954.

[85] Ibragimov, N.H.: Trasformation Group Applied to Mathematical Physics. Dordrecht, Reidel (Kluwer) 1985.

[86] Jackson, D.: Fourier Series and Orthogonal Polynomials. 6th ed. 1963.

[87] Jacobs, K.: Measure and Integral. New York, Academic Press 1978.

[88] Jacobson, N.: Basic Algebra. San Francisco, Freemans, Part I. 1974, Part II. 1980.

[89] Jeffreys, H.: Assymptotic Approximations. London, Oxford Univ. Press 1962.

[90] Joos, G.: Theoretical Physics. 3rd Ed. London, Blackie and Sons 1958.

[91] Kaplan, W.: Advanced Calculus. 2nd ed. Cambridge (Mass.), Addison-Wesley 1973.

[92] Kaplan, W.: Advanced Mathematics for Engineers. Cambridge (Mass.), Addison-Wesley 1981.

[93] Keedy, M.L., Bittinger, M.L.: Arithmetic. 4th ed. Cambridge (Mass.), Addison-Wesley 1983.

[94] Kelley, J.L., Srinivasan, T.P.: Measure and Integral. Berlin, Springer 1988.

[95] Kestelman, H.: Modern Theories of Integration. New York, Dover 1960.

[96] Khinchin, A.Y.: A Course of Mathematical Analysis. New York, Gordon and Breach 1961.

[97] Khinchin, A.Y.: Continued Fractions. Chicago, Chicago Univ. Press 1964.

[98] Klingenberg, W.: Lineare Algebra und Geometrie. Berlin, Springer 1984.

[99] Knopp, K.: Theory and Applications of Infinite Series. London, Blackie and Sons 1951.

[100] Kochendörfer, R.: Inroduction to Algebra. Dordrecht, Reidel (Kluwer) 1972.

[101] Koecher, M.: Lineare Algebra und Analytiche Geometrie, 2. Aufl., Berlin, Springer 1985.

[102] Kolman, B.: Calculus for the Management, Life and Social Sciences. New York, Academic Press 1981.

[103] Kolman, B., Shapiro A.: Precalculus: Functions and Graphs. New York, Academic Press 1984.

[104] Kreyszig, E.: Differential Geometry. London, Oxford Univ. Press 1959.

[105] Kuhnert, F.: Vorlesungen über lineare Algebra. Berlin, D. Verlag der Wiss. 1976.

[106] Kurzweil, J.: Nichtabsolut konvergente Integrale. Leipzig, Teubner 1980.

[107] Lancaster, P., Tismenetsky, M.: The Theory of Matrices. New York, Academic Press 1985.

[108] Lanczos, C.: Applied Analysis. London, Pitman 1957.

[109] Landau, E.: Differential and Integral Calculus. London, Chelsea Publ. Co. 1951.

[110] Lane, E.P.: Metric Differential Geometry of Curves and Surfaces. Chicago, Chicago Univ. Press 1940.

[111] Lang, S.: A First Course in Calculus. 4th ed. Cambridge (Mass.), Addison-Wesley 1978.

[112] Lang, S.: Calculus of Several Variables. 2nd ed. Cambridge (Mass.), Addison-Wesley 1979.

[113] Lang, S.: Linear Algebra. 3rd ed. Berlin, Springer 1987.

[114] Lawden, D.F.: An Introduction to Tensor Calculus, Relativity and Cosmology. New York, Wiley and Sons 1982.

[115] Levin, M., Girshowich, J.: Optimal Quadrature Formulas. Leipzig, Teubner 1979.

[116] Lichnerowicz, A.: Elements of Tensor Calculus. London, Methuen 1962.

[117] Lichnerowicz, A.: Geometry of Groups of Transformations. Dordrecht, Noordhoff 1977.

[118] Lockwood, E.H.: A Book of Curves. Cambridge, Cambridge Univ. Press 1961.

[119] Loomis, L.H.: Calculus. 3rd ed. Cambridge (Mass.), Addison-Wesley 1982.

[120] Marek, I., Žitný, K.: Matrix Analysis for Applied Sciences. Leipzig, Teubner 1986.

[121] Mainardi, P.: Barkan, H.: Calculus and its Applications. Oxford, Pergamon Press 1963.

[122] Marsden, J., Weistein, A.: Calculus. 2nd Ed. Berlin, Springer 1985.

[123] Maurin, K.: Analysis. Dordrecht, Reidel (Kluwer), Part I. 1976, Part II. 1980, Part III. 1982.

[124] McCrea, W.H.: Analytical Geometry of Three Dimensions. Edinburgh, Oliver and Boyd 1953

[125] McHale, T.A., Witzke, P.T.: Applied Trigonometry. Cambridge (Mass.), Addison-Wesley 1983.

[126] McKeague, C.P.: Trigonometry. New York, Academic Press 1984.

[127] McShane, E.J. (ed.): Unified Integration. New York, Academic Press 1983.

[128] Michal, A.D.: Matrix and Tesor Calculus. New York, Wiley and Sons 1947.

[129] Miller, K.S.: Elements of Modern Abstract Algebra. London, Harper and Row 1984.

[130] Mostow, G.D., Sampson, J.H., Meyer, J.P.: Fundamental Structures of Algebra. New York, McGraw-Hill 1963.

[131] Munroe, M.E.: Introduction to Measure and Integration. Cambridge (Mass.), Addison-Wesley 1953.

[132] Nash, C., Sen, S.: Topology and Geometry for Physicists. New York, Academic Press 1983.

[133] Natanson. I.P.: Theory of Functions of a Real Variable. London, Constable 1961.

[134] Nicholson, M.M.: Fundamentals and Techniques of mathematics for scientists. Harlow, Longmans 1961.

[135] Nobbs, C.: Trigonometry. London, Oxford Univ. Press 1962.

[136] Oberhettinger, F.: Fourier Expansions: A Collection of Formulas. New York, Academic Press 1973.

[137] Olmsted, J.M.H.: Real Variables. New York, Aplleton-Century-Crofts 1959.

[138] O'Neil, B.: Elementary Differential Geometry. New York, Academic Press 1966.

[139] Osgood, W.F.: Functions of Real and Complex Variables. London, Chelsea Publ. Co. 1935.

[140] Parker, W.V., Eaves, J.C.: Matrices. New York, Ronald Press 1960.

[141] Pease, M.C.: Methods of Matrix Algebra. New York, Academic Press 1960.

[142] Plumpton, C., Chirgwin, B.H.: Course of Mathematics for Engineers and Scientists. Oxford, Pergamon Press 1963.

[143] Primrose, E.J.F.: Plane Algebraic Curves. London, Macmillan 1955.

[144] Ralston, A.: A First Course in Numerical Analysis. New York, McGraw-Hill 1965.

[145] Robetrs, A.W.: Introductory Calculus. 2nd ed. New York, Academic Press 1972.

[146] Rogosinski, W.W.: Volume and Integral. Edinburgh, Oliver and Boyd 1962.

[147] Romanovski, P.I.: Mathematical Methods for Engineers and Technogists. Oxford, Pergamon Press 1961.

[148] Rothe, R.: Höhere Mathematik für Mathematiker, Physiker und Ingenieure, I-V, 13. Aufl. Leipzig, Teubner 1954.

[149] Rutherford, D.E.: Vector Methods. Edinburgh, Oliver and Boyd 1954.

[150] Salmon, G.: A Treatise on Conic Sections. London, Chelsea Publ. Co. 1954.

[151] Sawyer, W.W.: An Engineering Approach to Linear Algebra. Cambridge, Cambridge Univ. Press 1972.

[152] Schmeisser, H.J., Triebel, H.: Topics in Fourier Analysis and Function Spaces. New York, Wiley and Sons 1986.

[153] Schöne, W.: Differentialgeometrie. Leipzig, Teubner 1987.

[154] Schouten, J.A.: Tensor Analysis for Physicists. London, Oxford Univ. Press 1951.

[155] Sample, J.G., Kneebone, G.T.: Algebraic Curves. London, Oxford Univ. Press 1959.

[156] Silverman, R.A.: Modern Calculus and Analytic Geometry. London, Macmillan 1969.

[157] Simmonds, J.G.: A Brief on Tensor Analysis. Berlin, Springer 1982.

[158] Smirnov, V.I.: Course of Higher Mathematics, 5 Vols. Oxford, Pergamon Press 1964.

[159] Smith, L.: Linear Algebra. 2nd ed. Berlin, Springer 1984.

[160] Sokolnikoff, I.S.: Advanced Calculus. New York, McGraw-Hill 1939.

[161] Sokolnikoff, I.S.: Tensor Analysis. New York, Wiley and Sons, London, Chapman and Hall 1951.

[162] Sommerville, D.M.Y.: Analytical Conics. London, Bell and Sons 1951.

[163] Sommerville, D.M.Y.: Analytical Geometry of Three Dimensions. Cambridge, Cambridge Univ. Press 1947.

[164] Spiegel, M.R.: Theory and Problems of Advanced Calculus. Berlin, Schaum Publ. Co. 1963.

[165] Spain, B.: Analytical Conics. Oxford, Pergamon Press 1957.

[166] Spain, B.: Tensor Calculus. Edinburgh, Oliver and Boyd 1953.

[167] Spain, B.: Vector Analysis. London, Van Nostrand 1957.

[168] Spanier, J., Oldham, K.B.: An Atlas of Functions. Berlin, Springer 1987.

[169] Steward, C.A.: Advanced Calculus. London, Methuen 1940.

[170] Steward, G.W.: Introduction. New York, Academic Press 1973.

[171] Stigant, S.A.: Elements of Determinats, Matrices and Tensors for Engineers. London, Macmillan 1959.

[172] Strang, S.A.: Linear Algebra and its Applications. New York, Academic Press 1980.

[173] Švec, A.: Global Differential Geometry of Surfaces. Dordrecht, Reidel (Kluwer) 1981.

[174] Synge, J.A., Schild, A.: Tensor Calculus. Univ. Toronto, Toronto Press 1949.

[175] Tarasov. N.P.: A Course of Advanced Mathematics for Technical Schools. Oxford, Pergamon Press 1961.

[176] Taylor, S.J.: Introduction to Measure and Integration. Cambridge, Cambridge Univ. Press 1966.

[177] Turnbull, H.W.: Theory of Determinants, Matrices and Invariants. New York, Dover 1970.

[178] Vance, E.P.: Modern Algebra and Trigonometry. 3rd ed. Cambridge (Mass.), Addison-Wesley 1973.

[179] Waerden, B.L. van der: Algebra I, II. Berlin, Springer 1960.

[180] Wawrzyńczyk, A.: Group Representation and Special Functions. Dordrecht, Reidel (Kluwer) 1986.

[181] Weatherburn, C.E.: Avanced Vector Analysis. London, Bell and Sons 1949.

[182] Weir, A.J.: Lebesgue Integration and Measure. Cambridge, Cambridge Univ. Press 1973.

[183] Whittaker, E.T., Watson, G.N.: A Course of Modern Analysis. 4th ed. Cambridge, Cambridge Univ. Press 1958.

[184] Wolf, J.A., Cahen, M., De Wilde, M.: Harmonic Analysis and Representations of Semi-Simple Lie Groups. Dordrecht, Reidel (Kluwer) 1981.

[185] Wylie, C.R.: Plane Trigonometry. New York, McGraw-Hill 1955.

[186] Yates, R.C.: Analytic Geometry with Calculus. Englewood Cliffs (N.J.), Prentice-Hall 1961.

[187] Yates, R.C.: Curves and their Properties. London, J.B. Edwards 1959.

[188] Zaanen, A.C.: An introduction to the Theory of Integration. Amsterdam, North Holland 1958.

[189] Zygmund, A.: Trigometric Series, Vols 1, 2. Cambridge, Cambridge Univ. Press 1959.

INDEX

I, or II means Volume I, or Volume II, respectively. The symbol ff means "and following pages". For example, "Abstract functions, II 364 ff" means that abstract functions are treated in Volume II, starting with page 364. Articles (definite, indefinite) have been omitted whenever it was possible.

REFERENCES

[1] Afifi, A.A., Azen, S.P.: Statistical Analysis. A Computer Oriented Approach. New York, Academic Press 1979.

[2] Agnew, R.P.: Differential Equations. New York, McGraw-Hill 1960.

[3] Ahlfors, L.V.: Complex Analysis. New York, McGraw-Hill 1953.

[4] Aizenberg, L.A., Yuzhakov, A.P.: Integral Representations and Residues in Multidimensional Complex Analysis. Providence, Amer. Math. Soc. 1983.

[5] Akhiezer, N.I.: Theory of Approximations. New York, Ungar Publ. Comp. 1956.

[6] Akhiezer, N.I.: The Calculus of Variations. New York, Blaisdell 1962.

[7] Akhiezer, N.I.: Lectures on Integral Transforms. Providence, Amer. Math. Soc. 1988.

[8] Akhiezer, N.I., Glazman, I.M.: Theory of Linear Operators in Hilbert Space, I, II. London, Pitman 1981.

[9] Anděl, J.: Statistische Analyse von Zeitreihen. Berlin, Akademie-Verlag 1984.

[10] Anderson, T.W.: An Introduction to Multivariate Statistical Analysis. New York, Wiley 1958.

[11] Anderson, T.W.: The Statistical Analysis of Time Series. New York, Wiley 1971.

[12] Angot, A.: Compléments de Mathématiques à l'usage des ingénieurs de l'électrotechnique et des télecommunications. 3ième ed. revue et augmentée. Paris, Édition de la Revue d'optique 1957.

[13] Ansorge, R.: Differenzenapproximationen partieller Anfangswertaufgaben. Stuttgart, Teubner 1976.

[14] Antontsev, A.N., Kazhiktov, A.V., Monakhov, V.N.: Boundary Value Problems in Mechanics of Non-homogeneous Fluids. Amsterdam, Elsevier 1990.

[15] Arnold, S.F.: The Theory of Linear Models and Multivariate Analysis. New York, Wiley 1981.

[16] Arrow, K.J., Hurwitz, L., Uzawa, H.: Studies in Linear and Nonlinear Programming. Standford (Cal.), Standford Univ. Press 1958.

[17] Arscott, F.M.: Periodic Differential Equations. Oxford, Pergamon Press 1964.

[18] Ash, R.B., Gardner, M.F.: Topics in Stochastic Processes. New York, Academic Press 1975.

[19] Axelsson, O., Barker, V.A.: Finite Element Solution of Boundary Value Problems. Theory and Computation. New York, Academic Press 1984.

[20] Aziz, A.K., Babuška, I.: Mathematical Foundations in the Finite Element Method. New York, Academic Press 1972.

[21] Babuška, I., Práger, M., Vitásek, E.: Numerical Processes in Differential Equations. London, Wiley 1966.

[22] Babuška, I., Rektorys, K., Vyčichlo, F.: Mathematische Elastizitätstheorie der ebenen Probleme. Berlin, Akademie-Verlag 1960.

[23] Babuška, I., Szabó, B.: Finite Element Analysis. New York, Wiley 1991.

[24] Bailey, P.B., Shampine, L.F., Waltman, P.E.: Nonlinear Two-Point Boundary Value Problems. New York, Academic Press 1968.

[25] Banerjee, P.K., Butterfield, R.: Boundary Elements Methods in Egineering Science. London, McGraw-Hill 1981.

[26] Barbu, V.: Nonlinear Semigroups and Differential Equations in Banach Spaces. Dordrecht, Sijthoff and Noordhoff 1976.

[27] Barbu, V., Precupanu, Th.: Convexity and Optimization in Banach Spaces. Dordrecht, Reidel 1985.

[28] Bateman, H. et al.: Tables of Integral Transforms. New York, McGraw-Hill 1954.

[29] Bateman, H.: Partial Differential Equations of Mathematical Physics. London, Cambridge Univ. Press 1959.

[30] Bathe, K.J., Wilson, E.L.: Numerical Methods in Finite Element Analysis. Englewood Cliffs (N.J.), Prentice-Hall 1976.

[31] Bauer, H.: Wahrscheinlichkeitstheorie und Grundzüge der Massentheorie. Berlin, W. de Gruyter 1974.

[32] Beauzamy, B.: Introduction to Banach Spaces and Their Geometry. Amsterdam, Elsevier 1985.

[33] Beckenbach, E.F.: Modern Mathematics for the Engineer. New York, McGraw-Hill, 1st series 1956, 2st series 1961.

[34] Becker, E.B., Carey, G.F., Oden, J.T.: Finite Elements. An Introduction. Englewood Cliffs (N.J.), Prentice-Hall 1981.

[35] Beer, K.: Lösung grosser linearer Optimierungsaufgaben. Berlin, D. Verlag der Wiss. 1977.

[36] Bellman, R.: Stability Theory of Differential Equations. New York, McGraw-Hill 1953.

[37] Bellman, R., Adomian, G.: Partial Differential Equations. Dordrecht, Reidel 1984.

[38] Bellman, R.E., Roth, R.S.: Methods in Approximation. Dordrecht, Reidel 1986.

[39] Beresin, I.S., Shidkow, N.P.: Numerische Methoden. Berlin, D. Verlag der Wiss., Bd. I 1970, Bd. II 1971.

[40] Berger, J.O: Statistical Decision Theory. New York, Springer 1980.

[41] Bers, L., John, G., Schechter, M.: Partial Differential Equations. Providence, Amer. Math. Soc. 1981.

[42] Best, M.J.: Linear Programming: Active Set Analysis and Computer Programms. Englewood Cliffs (N.J.), Prentice-Hall 1985.

[43] Bhat, B.R.: Modern Probality. An Introductory Textbook. New York, Wiley 1985.

[44] Bieberbach, L.: Theorie der gewöhnlichen Differentialgleichungen. Berlin, Springer 1953.

[45] Bifurcation Analysis. Contributions from purely mathematical application to applications in many other fileds (ed. by M. Hazewinkel). Dordrecht, Reidel 1985.

[46] Birman, M.S., Solomjak, M.Z.: Spectral Theory of Self-Adjoint Operators in Hilbert Space. Dordrecht, Kluwer 1987.

[47] Bliss, G.A.: Calculus of Variations. Mathematical Association of America 1944.

[48] Bloomfield, P.: Fourier Analysis of Time Series. An Introduction. New York, Wiley 1976.

[49] Bohl, E.: Finite Modelle gewöhnlicher Randwertaufgaben. Stuttgart, Teubner 1981.

[50] Böhmer, K.: Spline-Funktionen. Stuttgart, Teubner 1974.

[51] Bolza, O.: Lectures on the Calculus of Variations. New York, Dover 1961.

[52] Boor, C. de: Practical Guide to Splines. New York, Springer 1978.

[53] Borůvka, O.: Linear Differential Transformations of the Second Order. London, English. Univ. Press 1971.

[54] Box, G.E.P., Jenkins, G.M.: Time Series Analysis, Forecasting and Control. San Francisco, Holden Day 1976.

[55] Braess, D.: Nonlinear Approximation Theory. Berlin, Springer 1986.

[56] Brandt, A.: Multi-Level Adaptive Solutions to Boundary-Value Problems. Mathematics of Computation 31 (1977), 333–390.

[57] Brandt, S.: Statistical and Computational Methods in Data Analysis. Amsterdam, North Holland 1970.

[58] Brebbia, C.A., Telles, J., Wrobel, L.C.: Boundary Element Techniques. Theory and Applications in Engineering. Berlin, Springer 1984.

[59] Brockwell, P.J., Davis, R.A.: Time Series: Theory and Methods. New York, Springer 1987.

[60] Burgers, J.M.: The Nonlinear Diffusion Equation. Dordrecht, Reidel 1982.

[61] Burkill, J.C., Burkill, H.: A Second Course in Mathematical Analysis. Cambridge, Cambridge Univ. Press 1970.

[62] Burnside, W.S., Panton, A.W.: Theory of Equations. Dublin, Dublin Univ. Press 1935.

[63] Burr, I.W.: Statistical Quality Control Methods. New York, Dekker 1976.

[64] Bury, K.V.: Statistical Models in Applied Science. New York, Wiley 1975.

[65] Carathéodory, C.: Variationsrechnung und partielle Differentialgleichungen erster Ordnung. 2. Aufl. Leipzig, Teubner 1956.

[66] Carathéodory, C.: Theory of Functions, Vols I, II. 2nd ed. New York, Chelsea 1958.

[67] Carroll, R.: Mathematical Physics. Amsterdam, Elsevier 1988.

[68] Carslaw, H.S., Jaeger, J.C.: Operational Methods in Applied Mathematics. London, Oxford Univ. Press 1948.

[69] Cesari, L.: Asymptotic Behaviour and Stability Problems in Differential Equations. Berlin, Springer 1963.

[70] Černý, I.: Foundations of Analysis in the Complex Domain. Praha, Academia – Chichester, Horwood 1991.

[71] Chambers, J.M.: Computational Methods for Data Analysis. New York, Wiley 1977.

[72] Charnes, A.: Optimality and Degeneracy in Linear Programming. Econometrica 20 (1952).

[73] Chase, C.I.: Elementary Statistical Procedures. New York, McGraw-Hill 1984.

[74] Chase, W., Brown, F.: General Statistics. New York, Wiley 1983.

[75] Chatfield, C.: The Analysis of Time Series: Theory and Practice. London, Chapman and Hall 1975.

[76] Chazarain, J., Piriou, A.: Introduction to the Theory of Linear Partial Differential Equations. Amsterdam, North-Holland 1982.

[77] Churchill, R.V.: Operational Mathematics in Engineering, 2nd ed. New York, McGraw-Hill 1958.

[78] Churchill, R.V.: Complex Variable and Applications. New York, McGraw-Hill 1960.

[79] Churchill, R.V.: Fourier Series and Boundary Value Problems. New York, McGraw--Hill 1963.

[80] Chung, K.L.: Markov Chains with Stationary Transition Probabilities. New York, Springer 1967.

[81] Chung, K.L.: Elementary Probability Theory with Stochastic Processes. New York, Wiley 1974.

[82] Chung, K.L.: A Course in Probability Theory. New York, Wiley 1974.

[83] Ciarlet, P.G.: Finite Element Method for Elliptic Problems. Amsterdam, Elsevier 1978.

[84] Ciarlet, P.G.: Mathematical Elasticity, Vol I. Amsterdam, Elsevier 1988.

[85] Clements, D.L.: Boundary Value Problems Governed by Second Order Elliptic Systems. London, Pitman 1981.

[86] Cochran, W.G.: Sampling Techniques. New York, Wiley 1963.

[87] Coddington, E.A., Levinson, N.: Theory of Ordinary Differential Equations. New York, McGraw-Hill 1955.

[88] Collatz, L.: Eigenwertaufgaben mit technischen Anwendungen. Leipzig, Akademische Verlagsgesellschaft 1949.

[89] Collatz, L.: The Numerical Treatment of Differential Equations. Berlin, Springer 1960.

[90] Collatz, L.: Functional Analysis and Numerical Mathematics. New York, Academic Press 1966.

[91] Computational Aspects of Complex Analysis (Ed. by H. Werner et al.). Dordrecht, Reidel 1982.

[92] Conover, W.J.: Practical Nonparametric Statistics. New York, Wiley 1971.

[93] Conway, J.B.: A Course in Functional Analysis. 2nd ed. Berlin, Springer 1990.

[94] Cooke, R.G.: Linear Operators. London, Macmillan 1953.

[95] Cooley, W.W., Lohnes, P.R.: Multivariate Data Analysis. New York, Wiley 1971.

[96] Cooper, R.B.: Introduction to Queueing Theory. New York, North Holland 1981.

[97] Copson, E.T.: Metric Spaces. Cambridge, Cambridge Univ. Press 1968.

[98] Copson, E.T.: Partial Differential Equations. Cambridge, Cambridge Univ. Press 1974.

[99] Courant, R., Hilbert, D.: Methods of Mathematical Physics, Vols I, II. New York, Interscience 1962.

[100] Cox, C.P.: A Handbook of Introductory Statistical Methods. New York, Wiley 1986.

[101] Cramér, H.: Mathematical Methods of Statistics. Princeton, Princeton Univ. Press 1946.

[102] Cramér, H., Leadbetter, M.R.: Stationary and Related Stochastic Processes. New York, Wiley 1967.

[103] Dacorogna, B.: Direct Methods in the Calculus of Variations. Berlin, Springer 1989.

[104] Dahlquist, G., Björck, A.: Numerical Methods. Englewood Cliffs (N.J.), Prentice-Hall 1974.

[105] Dan, S: Linear Programming in Industry. Theory and Applications. Berlin, Springer 1974.

[106] Daniel, C.: Applications of Statistics to Industrial Experimentation. New York, Wiley 1976.

[107] Daniel, J.W., Moore, R.E.: Computation and Theory in Ordinary Differential Equations. San Francisco, Freeman 1970.

[108] Dantzig, G.B.: Lineare Programmierung und Erweiterungen. Berlin, Springer 1966.

[109] Dautray, L., Lions, J.L.: Mathematical Analysis and Numerical Methods for Science and Technology, Vols 1–6. Berlin, Springer 1988–1990.

[110] Day, W.D.: Tables of Laplace Transforms. London, Iliffe 1965.

[111] Deimling, K.: Nonlinear Functional Analysis. Berlin, Springer 1985.

[112] Dettman, J.W.: Mathematical Methods in Physics and Engineering. New York, McGraw-Hill 1962.

[113] Dezin, A.A.: Partial Differential Equations. Berlin, Springer 1987.

[114] Dillon, W.R., Goldstein, M.: Multivariate Analysis. Methods and Applications. New York, Wiley 1984.

[115] Discretization in Differential Equations and Enclosures (ed. by E. Adams et al.). Berlin, Akademieverlag 1987.

[116] Ditkin, V.A., Prudnikov, A.P.: Operational Calculus in Two Variables and its Applications. Oxford, Pergamon Press 1963.

[117] Dixon, W.J., Massey, F.J.: Introduction to Statistical Analysis. New York, McGraw-Hill 1983.

[118] Doetsch, G.: Handbuch der Laplace-Transformation. Basel, Birkhäuser 1950.

[119] Doetsch, G.: Guide to the Applications of Laplace Transforms. London, van Nostrand 1961.

[120] Dodge, H.F., Roming, H.G.: Sampling Inspection Tables. New York, Wiley 1959.

[121] Dongarra, J.J. et al.: LINPACK Users' Guide. Philadelphia, SIAM 1979.

[122] Doob, J.L.: Stochastic Processes. New York, Wiley 1953.

[123] Dowdy, S., Wearden, S.: Statistics for Research. New York, Wiley 1983.

[124] Draper, N.R., Smith, H.: Applied Regression Analysis. New York, Wiley 1966.

[125] Duff, G.F.D.: Partial Differential Equations. Toronto, Univ. of Toronto Press – London, Oxford Univ. Press 1956.

[126] Dunn, J., Clark, V.A.: Applied Statistics: Analysis of Variance and Regression. New York, Wiley 1974.

[127] Du Toit, S.H.C., Steyn, A.G.W., Stumpf, R.H.: Graphical Exploratory Data Analysis. New York, Springer 1986.

[128] Eisenreich, G.: Vorlesungen über Funktionentheorie mehrerer Variabler. Leipzig, Teubner 1980.

[129] Elsgol'ts, L.E.: Differential Equations. New York, Gordon and Breach 1961.

[130] Elsgol'ts, L.E.: Calculus of Variations. Oxford, Pergamon Press 1963.

[131] Encyclopaedia of Mathematics, 10 Vols. Dordrecht, Reidel, 1987–1990.

[132] Epstein, B.: Partial Differential Equations. New York, McGraw-Hill 1962.

[133] Erdélyi, A., Magnus, W., Oberhettinger, F., Tricomi, F.G.: Higher Transcendental Functions, Vols 1, 2, 3. New York, McGraw-Hill 1953.

[134] Erdélyi, A., Magnus, W., Oberhettinger, F., Tricomi, F.G.: Tables of Integral Transforms, 2 Vols. New York, McGraw-Hill 1954.

[135] Erdélyi, L.: Operator Theory and Functional Analysis. London, Pitman 1979.

[136] Evans, L.C.: Weak Convergence Methods for Nonlinear Partial Differential Equations. Providence, Amer. Math. Soc. 1990.

[137] Fabian, V., Hannan, E.J.: Introduction to Probability and Mathematical Statistics. New York, Wiley 1985.

[138] Fahrmeir, L.: Rekursive Algorithmen für Zeitreihenmodelle. Göttingen, Vanderhoeck und Ruprecht 1981.

[139] Feistauer, M., Ženíšek, A.: Finite Element Solution of Nonlinear Elliptic Problems. Numer. Math. 50 (1987), 451–475.

[140] Feistauer, M., Ženíšek, A.: Compactness Method in the Finite Element Theory of Nonlinear Problems. Numer. Math. 52 (1988), 147–163.

[141] Feller, W.: An Introduction to Probability Theory and its Applications. New York, Wiley 1966.

[142] Fenner, R.T.: Finite Element Methods for Engineers. New York, Macmillan 1975.

[143] Fenyö, S., Stolle, H.W.: Theorie und Praxis der linearen Integralgleichungen. Basel, Birkhäuser, I 1982, II 1983, III 1983, IV 1984.

[144] Fichera, G.: Numerical and Quantitative Analysis, Vols 1, 2. London, Pitman 1978.

[145] Fiedler, M.: Special Matrices and their Applications in Numerical Mathematics. Dordrecht, Nijhoff 1986.

[146] Field, M.: Several Complex Variables and Complex Manifolds, Vols 1, 2. Cambridge, Cambridge Univ. Press 1982.

[147] Fletcher, A., Miller, J.C.P., Rosenhead, L.: Index of Mathematical Tables, 2 Vols. 2nd ed. Oxford, Blackwell's 1962.

[148] Fodor, G.: Laplace Transform in Engineering. Budapest, Hungarian Acad. of Sciences 1965.

[149] Forsyth, A.R.: Calculus of Variations. New York, Dover 1960.

[150] Forsythe, G.E., Malcolm, M.A., Moler, C.B.: Computer Methods for Mathematical Computations. Englewood Cliffs (N.J.), Prentice-Hall 1977.

[151] Forsythe, G.E., Rosenbloom, P.C.: Numerical Analysis and Partial Differential Equations. New York, Wiley 1958.

[152] Forsythe, G.E., Wasow, W.R.: Finite Difference Methods for Partial Differential Equations. London, Wiley 1960.

[153] Fortin, M., Glowinski, R.: Augmented Lagrangian Methods: Applications to Numerical Solution of Boundary Value Problems. Amsterdam, Elsevier 1983.

[154] Fox, C.: An Introduction to the Calculus of Variations. London, Oxford Univ. Press 1954.

[155] Fox, L.: Numerical Solution of Ordinary and Partial Differential Equations. Oxford, Pergamon Press 1962.

[156] Friedman, A.: Partial Differential Equations of Parabolic Type. Englewood Cliffs (N.J.), Prentice-Hall 1964.

[157] Friedman, A.: Mathematics in Industrial Problems. Berlin, Springer, I 1988, II 1989, III 1990.

[158] Fuchs, V.A., Shabat, B.V.: Functions of a Complex Variable. Oxford, Pergamon Press 1964.

[159] Fučík, S.: Solvability of Nonlinear Equations and Boundary Value Problems. Dordrecht, Reidel 1981.

[160] Fučík, S., Kufner, A.: Nonlinear Differential Equations. Amsterdam, Elsevier 1980.

[161] Fučík, S., Nečas, J., Souček, V.: Einführung in die Variationsrechnung. Lepzig, Teubner 1977.

[162] Fučík, S., Nečas, J., Souček, J., Souček, V.: Spectral Analysis of Nonlinear Operators. Berlin, Springer 1973.

[163] Fuks, B.A.: Theory of Analytic Functions of Several Complex Variables. Providence, Amer. Math. Soc. 1975.

[164] Fuller, W.A.: Introduction to Statistical Time Series. New York, Wiley 1976.

[165] Gajewski, H., Gröger, K., Zacharias, K.: Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen. Berlin, Akademie-Verlag 1974.

[166] Gál, T.: Postoptimal Analysis. Parametric Programming and Related Topics. New York, McGraw-Hill 1979.

[167] Gallagher, R.H.: Finite Element Analysis. Englewood Cliffs (N.J.), Prentice-Hall 1975.

[168] Gantmacher, F.R.: Applications of the Theory of Matrices. New York, Interscience 1959.

[169] Gass, S.: Linear Programming: Methods and Applications, 2nd ed. New York, McGraw-Hill 1964.

[170] Gear, C.W.: Numerical Initial Value Problems in Ordinary Differential Equations. Englewood Cliffs (N.J.), Prentice-Hall 1971.

[171] Gelfand, I.M., Fomin, S.V.: Calculus of Variations. Englewood Cliffs (N.J.), Prentice-Hall 1963.

[172] Gelfand, I.M., Schilow, G.E., Wilenkin, N.J.: Verallgemeinerte Funktionen (Distributionen). Berlin, D. Verlag der Wiss., I 1967, II 1969, III 1974, IV 1974.

[173] George, A., Liu, J.W.H.: Computer Solution of Large Sparse Positive Definite Systems. Englewood Cliffs (N.J.), Prentice-Hall 1981.

[174] Gilbert, R.P., Weinacht, R.J.: Function Theoretic Methods in Differential Equations. London, Pitman 1976.

[175] Gill, P.E., Murray, W., Wright, M.H.: Practical Optimization. London, Academic Press 1981.

[176] Girault, V., Raviart, P.A.: Finite Element Methods for Navier-Stokes Equations. Berlin, Springer 1986.

[177] Glowinski, R., Lions, J.L., Tremolières, R.: Numerical Analysis of Variational Inequalities. Amsterdam, North-Holland 1981.

[178] Gnadesikan, R.: Methods for Statistical Data Analysis of Multivariate Observations. New York, Wiley 1977.

[179] Gnedenko, B.V., Beljajev, Ju. K., Solovjev, A.D.: Mathematische Methoden der Zuverlässigkeitstheorie. Berlin, Akademie-Verlag 1968.

[180] Godunov, S.K., Ryabenkij, V.S.: Difference Schemes. Amsterdam, North-Holland 1987.

[181] Goldenveizer, A.L.: Theory of Elastic Thin Shells. Oxford, Pergamon Press 1961.

[182] Golub, G.H., Loan, C.F. van: Matrix Computations. Baltimore, Hopkins 1989.

[183] Gottman, J.M.: Time Series Analysis: A Comprehensive Introduction for Social Scientists. Cambridge, Cambridge Univ. Press 1981.

[184] Gould, S.H.: Variational Methods for Eigenvalue Problems, 2nd ed. London, Oxford Univ. Press 1966.

[185] Graffi, D.: Nonlinear Partial Differential Equations in Physical Problems. London, Pitman 1980.

[186] Granger, C.W.J, Newbold, P.: Forecasting Economic Time Series. New York, Academic Press 1977.

[187] Grauert, A., Fritsche, K.: Several Complex Variables. New York, Springer 1976.

[188] Graybill, F.A.: Theory and Application of the Linear Models. North Scituate, Duxbury Press 1976.

[189] Greenspan, D.: Theory and Solutions of Ordinary Differential Equations. New York, Macmilan 1960.

[190] Greenspan, D.: Introduction to Partial Differential Equations. New York, McGraw-Hill 1961.

[191] Greguš, M.: Third-Order Linear Differential Equations. Dordrecht, Reidel 1987.

[192] Grenander, U., Rosenblatt, M.: Statistical Analysis of Stationary Time Series. New York, Wiley 1957.

[193] Grigorieff, R.D.: Numerik gewöhnlicher Differentialgleichungen, 1, 2. Stuttgart, Teubner 1977.

[194] Groenevald, R.A.: An Introduction to Probability and Statistics Using BASIC. New York, Dekker 1976.

[195] Gross, D., Harris, C.M.: Fundamentals of Queueing Theory. New York, Wiley 1985.

[196] Guest, P.G.: Numerical Methods of Curve Fitting. New York, Cambridge Univ. Press 1961.

[197] Gunning, R.C., Rossi, H.: Analytic Functions of Several Complex Variables. Englewood Cliffs (N.J.), Prentice-Hall 1965.

[198] Guttman, I., Wilks, S.S., Hunter, J.S: Introductory Engineering Statistics. New York, Wiley 1982.

[199] Hackbusch, W.: Multigrid Methods and Applications. Berlin, Springer 1985.

[200] Hahn, G.J., Shapiro, S.S.: Statistical Models in Engineering. New York, Wiley 1967.

[201] Hájek, J., Šidák, Z.: Theory of Rank Tests. Praha, Academia 1967.

[202] Hald, A.: Statistical Tables and Formulas. New York, Wiley 1952.

[203] Halmos, P.R.: Finite-Dimensional Vector Spaces. New York, van Nostrand 1958.

[204] Handbook of Numerical Analysis. Vol 1: Finite Difference Methods, Vol. 2: Finite Element Methods (ed. by P.G. Ciarlet and J.L. Lions). Amsterdam, Elsevier 1990.

[205] Handbook of Tables for Probability and Statistics. Cleveland (Ohio), Chemical Rubber Company 1968.

[206] Hannan, E.J.: Time Series Analysis. London, Chapman and Hall 1967.

[207] Hannan, E.J.: Multiple Time Series. New York, Wiley 1970.

[208] Hartley, H.: The Modified Gauss-Newton Method for the Fitting of Nonlinear Regression Functions by Least Squares. Technometrics 3 (1961), 269–280.

[209] Hartree, D.R.: Numerical Analysis. London, Oxford Univ. Press 1955.

[210] Haslinger, J., Neittaanmäki, P.: Finite Element Approximation for Optimal Shape Design. Theory and Applications. Chichester, Wiley 1988.

[211] Heading, J.: Mathematical Methods in Science and Engineering, 2nd ed. London, Arnold 1970.

[212] Henrici, P.: Discrete Variable Methods in Ordinary Differential Equations. New York, Wiley 1962.

[213] Henrici, P.: Applied and Computational Complex Analysis. New York, Wiley: I 1974, II 1977, III 1986.

[214] Hestenes, M.R.: Calculus of Variations and Optimal Control Theory. New York, Wiley 1956.

[215] Hildebrand, F.B.: Introduction to Numerical Analysis. New York, McGraw-Hill 1956.

[216] Hlaváček, I., Haslinger, J., Nečas, J., Lovíšek, J.: Solution of Variational Inequalities in Mechanics. Berlin, Springer 1988.

[217] Hoel, P.G.: Elementary Statistics. New York, Wiley 1976.

[218] Hoel, P.G.: Introduction to Mathematical Statistics. New York, Wiley 1984.

[219] Hollander, M., Wolfe, D.W.: Nonparametric Statistical Methods. New York, Wiley 1973.

[220] Hörmander, L.: Linear Partial Differential Equations. Berlin, Springer 1964.

[221] Hörmander, L.: Introduction to Complex Analysis in Several Complex Variables. 3rd revised ed. Amsterdam, North-Holland 1991.

[222] Hörmander, L.: Analysis of Linear Partial Differential Operators. Berlin, Springer, I 1983, II 1983, III 1985, IV 1985.

[223] Hort, W., Thoma, A.: Die Differentialgleichungen der Technik und Physik. 5. Aufl. Leipzig, Barth 1954.

[224] Householder, A.S: Principles of Numerical Analysis. New York, McGraw-Hill 1953.

[225] Huber, P.J.: Robust Statistics. New York, Wiley 1981.

[226] Huebner, K.H.: The Finite Element Method for Engineers. New York, Wiley 1975.

[227] Hurwitz, A., Courant, R.: Funktionentheorie. Berlin, Springer 1964.

[228] Ibragimov, I.A., Hasminskii, R.Z.: Statistical Estimation. Asymptotic Theory. New York, Springer 1981.

[229] Ikeda, T.: Maximum Principles in Finite Element Models for Convection-Diffusion Phenomena. Amsterdam, Elsevier 1983.

[230] Iman, R.L., Conover, W.J.: A Modern Approach to Statistics. New York, Wiley 1983.

[231] IMSL Inc., Houston, TX: IMSL Library Reference Manual ed. 8. 1980.

[232] Integral Equations – a Reference Text (ed. by P.P. Zabreyko et al.). Dordrecht, Sijthoff and Noordhoff 1975.

[233] Ioffe, A.D., Tichomirov, V.H.: Theorie der Extremalaufgaben. Berlin, D. Verlag der Wiss. 1979.

[234] Isaacson, E., Keller, H.B.: Analysis of Numerical Methods. New York, Wiley 1966.

[235] Isaru, L.: Numerical Methods for Differential Equations and Applications. Dordrecht, Reidel 1984.

[236] Istratescu, V.: Fixed Point Theory. Dordrecht, Reidel 1981.

[237] Jazwinski, A.H.: Stochastic Processes and Filtering Theory. New York, Academic Press 1970.

[238] Jeffreys, H.: Operational Methods in Mathematical Physics. London, Cambridge Univ. Press 1927.

[239] Jeffreys, H., Jeffreys, B.S.: Mathematical Physics. London, Cambridge Univ. Press 1956.

[240] Jenkins, G.M., Watts, D.G.: Spectral Analysis and its Applications. San Francisco, Holden Day 1968.

[241] Jerri, A.J.: Introduction to Integral Equations with Applications. New York, Dekker 1987.

[242] Johnson, C.: Numerical Solution of Partial Differential Equations by the Finite Element Method. Cambridge, Cambridge Univ. Press 1988.

[243] Johnson, N.I., Kotz, S.: Distributions in Statistics. Discrete Distributions. New York, Wiley 1970.

[244] Johnson, N.I., Kotz, S.: Distributions in Statistics. Continuous Univariate Distributions. New York, Wiley 1970.

[245] Johnson, N.I., Kotz, S.: Distributions in Statistics. Continuous Multivariate Distributions. New York, Wiley 1972.

[246] Joos, G.: Theoretical Physics. 3rd ed. London, Blackie 1958.

[247] Judin, D.B., Goldstein, E.G.: Lineare Optimierung I. Berlin, Akademieverlag 1968.

[248] Kačur, J.: Method of Rothe in Evolution Equations. Leipzig, Teubner 1985.

[249] Kalbfleisch, J.G.: Probability and Statistical Inference. New York, Springer 1979.

[250] Kamke, E.: Differentialgleichungen, Lösungsmethoden und Lösungen. 3. Aufl. Leipzig, Akademische Verlagsgesellschaft 1956.

[251] Kamke, E.: Differentialgleichungen II, Partielle Differentialgleichungen. 4. Aufl. Leipzig, Goest und Portig 1964.

[252] Kantorovich, L.V., Krylov, V.I.: Approximate Methods of Higher Analysis. Groningen, Noordhoff 1958.

[253] Kaplan, W.: Lectures on Functions of a Complex Variable. Ann Arbor, Univ. of Michigan Press 1955.

[254] Kaplan, W.: Ordinary Differential Equations. Cambridge (Mass.), Addison-Wesley 1958.

[255] Keller, B.H.: Numerical Methods for Two-Point Boundary Value Problems. Waltham (Mass.), Blaisdell 1968.

[256] Kemeny, J.G., Snell, J.L.: Finite Markov Chains. Princeton, Van Nostrand 1960.

[257] Kendall, M.: Time Series. London, Griffin 1973.

[258] Kendall, M.G., Stuart, A.: The Advanced Theory of Statistics. London, Griffin 1966.

[259] Kitahara, M.: Boundary Integral Equation Methods in Eigenvalue Problems of Elastodynamics and Thin Plates. Amsterdam, Elsevier 1985.

[260] Klötzler, R.: Mehrdimensionale Variationsrechnung. Berlin, D. Verlag der Wiss. 1971.

[261] Knops, R.J.: Nonlinear Analysis and Mechanics. London, Pitman 1977.

[262] Kober, H.: Dictionary of Conformal Representations. New York, Dover 1957.

[263] Kolmogorov, A.N., Fomin, S.V.: Elements of the Theory of Functional Analysis 1, 2. New York, Graylock 1957.

[264] Komkov, V.: Variational Principles of Continuum Mechanics with Engineering Applications. Dordrecht, Reidel, Vol. 1 1986, Vol. 2 1988.

[265] Koopmans, L.H.: The Spectral Analysis of Time Series. New York, Academic Press 1974.

[266] Koosis, D.J.: Statistics. New York, Wiley 1985.

[267] Koppenfels, W., Stallmann, F.: Praxis der konformen Abbildung. Berlin, Springer 1959.

[268] Korbut, A.A., Finkelstein, J.J.: Diskrete Optimierung. Berlin, Akademieverlag 1971.

[269] Koroljuk, V.S., Portenko, N.I., Skorochod, A.V., Turbin, A.F.: Handbook of Probability Theory and Mathematical Statistics. Moscow, Nauka 1985 (in Russian).

[270] Krall, A.M.: Applied Analysis. Dordrecht, Reidel 1986.

[271] Krantz, S.G.: Function Theory of Several Complex Variables. New York, Wiley 1982.

[272] Krasnoselskij, M.A., Vainikko, G.M., Zabreyko, R.P.: Approximate Solution of Operator Equations. Dordrecht, Sijthoff and Nordhoff 1972.

[273] Krasnoselskij, M.A. et al.: Integral Operators in Spaces of Summable Functions. Dordrecht, Sijthoff and Noordhoff 1976.

[274] Krein, M.G.: Topics in Differential and Integral Equations and Operator Theory. Dordrecht, Reidel 1983.

[275] Kres, H.: Statistical Tables for Multivariate Analysis. A Handbook with References to Applications. Berlin, Springer 1983.

[276] Kress, R.: Linear Integral Equations. Berlin, Springer 1989.

[277] Kreyszig, E.: Advanced Engineering Mathematics. 2nd ed. New York, Wiley 1967.

[278] Kreyszig, E.: Introductory Mathematical Statistics. New York, Wiley 1970.

[279] Křížek, M.: An Equilibrium Finite Element Method in Three Dimensional Elasticity. Apl. Mat. 27 (1982), 46–75.

[280] Křížek, M., Neittaanmäki, P.: Finite Element Approximation of Variational Problems and Applications. Harlow, Longman 1990.

[281] Krylov, N.V.: Nonlinear Elliptic and Parabolic Equations of the Second Order. Dordrecht, Reidel 1986.

[282] Kubíček, M., Marek, M.: Computational Methods in Bifurcation Theory and Dissipative Structures. Berlin, Springer 1983.

[283] Kufner, A.: Weighted Sobolev Spaces. Leipzig, Teubner 1980.

[284] Kufner, A., John, O., Fučík, S.: Function Spaces. Praha, Academia 1977.

[285] Kufner, A., Sändig, A.M.: Some Applications of Weighted Sobolev Spaces. Leipzig, Teubner 1987.

[286] Kurzweil, J.: Ordinary Differential Equations. Amsterdam, Elsevier 1986.

[287] Ladyzhenskaya, O.A.: The Boundary Value Problems of Mathematical Physics. Berlin, Springer 1984.

[288] Ladyzhenskaya, O.A., Solonnikov, V.A., Uralceva, N.N.: Linear and Quasilinear Equations of Parabolic Type. Providence, Amer. Math. Soc. 1988.

[289] Lakshmikantham, V., Leela, S., Martinyuk, A.A.: Stability Analysis of Nonlinear Systems. New York, Dekker 1988.

[290] Lambe, C.G., Tranter, C.J.: Differential Equations for Engineers and Scientists. London, English Univ. Press 1961.

[291] Lambert, J.D.: Computational Methods in Ordinary Differential Equations. London, Wiley 1973.

[292] Lanczos, C.: Applied Analysis. Englewood Cliffs (N.J.), Prentice-Hall 1956.

[293] Lapidus, L., Seinfeld, J.H.: Numerical Solution of Ordinary Differential Equations. New York, Academic Press 1971.

[294] Larson, H.J.: Introduction to Probability and Statistical Inference. New York, Wiley 1982.

[295] Laurent, P.J.: Approximation et Optimisation. Paris, Hermann 1972.

[296] Lavrentyev, M.A., Shabat, B.V.: Methods of the Theory of Functions of a Complex Variable. Moscow, Gostekhizdat 1958 (in Russian).

[297] Lehmann, E.L.: Testing Statistical Hypotheses. New York, Wiley 1959.

[298] Lehmann, E.L.: Theory of Point Estimation. New York, Wiley 1983.

[299] Lions, J.L., Magenes, E.: Non-homogeneous Boundary Value Problems and Applications I, II. Berlin, Springer 1972.

[300] Liptser, R.S., Shiryayev, A.N.: Statistics of Random Processes I. General Theory. New York, Springer 1984.

[301] Liptser, R.S., Shiryayev, A.N.: Statistics of Random Processes II. Applications. New York, Springer 1978.

[302] Loève, M.: Probability Theory. Princeton, Van Nostrand 1955.

[303] Lukeš, J., Malý, J., Zajíček, L.: Fine Topology Methods in Real Analysis and Potential Theory. Berlin, Springer 1986.

[304] Lusternik, L.A.: Shortest Way; Variational Problems. London, Macmillan 1964.

[305] Lusternik, L.A., Sobolev, V.I.: Elements of Functional Analysis. New York, Halstadt 1974.

[306] Madansky, A.: Prescriptions for Working Statisticians. New York, Springer 1988.

[307] Maddox, I.J.: Elements of Functional Analysis. Cambridge, Cambridge Univ. Press 1970.

[308] Maindonald, J.H.: Statistical Computation. New York, Wiley 1984.

[309] Malgrange, B.: Theory of Functions of Several Complex Variables. Berlin, Springer 1984.

[310] Mann, N.R., Schafer, R.E., Singpurwalla, N.D.: Methods for Statistical Analysis of Reliability and Life Data. New York, Wiley 1974.

[311] Marden, M.: The Geometry of Zeros of a Polynomial. Providence, Amer. Math. Soc. 1969.

[312] Marek, I., Žitný, K.: Matrix Analysis for Applied Sciences. Leipzig, Teubner 1989.

[313] Markushevich, A.I.: Theory of Analytic Functions. Moscow, Gostekhizdat 1950 (in Russian).

[314] Marsal, D.: Finite Differenzen und Elemente. Berlin, Springer 1988.

[315] Martin, H.C., Carey, G.F.: Introduction to Finite Element Analysis. New York, McGraw-Hill 1973.

[316] Maurin, K.: Methods of Hilbert Spaces. Warszawa, PWN 1967.

[317] Mazja, W.: Einbettungssätze für Sobolewsche Räume. Leipzig, Teubner, I 1979, II 1980.

[318] McLachtan, N.W.: Ordinary Nonlinear Differential Equations. London, Oxford Univ. Press 1950.

[319] McLachtan, N.W.: Modern Operational Calculus. New York, Dover 1962.

[320] Meinhold, P., Wagner, E.: Partielle Differentialgleichungen. 5. Aufl. Leipzig, Teubner 1987.

[321] Meis, T., Morowitz, U.: Numerical Solution of Partial Differential Equations. New York, Springer 1981.

[322] Meyer, G.H.: Initial Value Methods for Boundary Value Problems; Theory and Application of Invariant Imbedding. New York, Academic Press 1973.

[323] Mikhlin, S.G.: Integral Equations and Applications. Oxford, Pergamon Press 1957.

[324] Mikhlin, S.G.: Variational Methods in Mathematical Physics. New York, Pergamon Press 1963.

[325] Mikusiński, J.: Operational Calculus. Oxford, Pergamon Press 1959.

[326] Miles, J.: Integral Transforms in Applied Mathematics. Cambridge, Cambridge Univ. Press 1971.

[327] Miller, K.S.: Engineering Mathematics. London, Constable 1956.

[328] Milne, R.D.: Applied Functional Analysis. London, Pitman 1980.

[329] Milojević, P.S.: Nonlinear Functional Analysis. New York, Dekker 1988.

[330] Milton, J.S., Arnold, J.C.: Probability and Statistics in the Engineering and Computing Sciences. New York, McGraw-Hill 1986.

[331] Milton, J.S., Colbert, J.J.: Applied Statistics with Probability. New York, Van Nostrand 1979.

[332] Miranker, W.L.: Numerical Methods for Stiff Equations and Singular Perturbation Problems. Dordrecht, Reidel 1980.

[333] Mitchell, A.R.: Computational Methods in Partial Differential Equations. London, Wiley 1969.

[334] Mitchell, A.R., Wait, R.: The Finite Element Method in Partial Differential Equations. Chichester, Wiley 1977.

[335] Montgomery, D.C., Johnson, L.A.: Forecasting and Time Series Analysis. New York, McGraw-Hill 1976.

[336] Montgomery, D.C., Peck, E.A.: Introduction to Linear Regression Analysis. New York, Wiley 1982.

[337] Morosanu, G.H.: Nonlinear Evolution Equations and Applications. Dordrecht, Kluwer 1988.

[338] Morrison, D.F.: Multivariate Statistical Methods. New York, McGraw-Hill 1976.

[339] Morse, P.M., Feshbach, H.: Methods of Mathematical Physics I, II. New York, McGraw-Hill 1953.

[340] Mukherjea, A., Pothoven, K.: Real and Functional Analysis. 2nd ed. New York, Plenum: I 1984, II 1985.

[341] Murray, D.A.: Introductory Course in Differential Equations for Students in Classical and Engineering Colleges. New York, Longmans 1954.

[342] Muskhelishvili, N.I.: Singular Integral Equations. Dordrecht, Sijthoff and Noordhoff 1977.

[343] Myint, U.T., Debnath, L.: Partial Differential Equations for Scientists and Engineers. 3rd ed. Amsterdam, North-Holland 1987.

[344] NAG, Oxford, U.K.: NAG Fortran Library Manual Mark 8. 1980.

[345] Naiman, A., Rosenfeld, R., Zirkel, G.: Understanding Statistics. New York, McGraw-Hill 1983.

[346] Najmark, M.A.: Lineare Differentialoperatoren. Berlin, Springer 1960.

[347] Narasimhan, R.: Complex Analysis in One Variable. Boston, Birkhäuser 1985.

[348] Nečas, J.: Les méthodes directes en théorie des équations elliptiques. Praha, NČSAV 1967.

[349] Nečas, J.: Introduction to the Theory of Nonlinear Elliptic Equations. Leipzig, Teubner 1982.

[350] Nečas, J., Hlaváček, I.: Mathematical Theory of Elastic and Elasto-Plastic Bodies: An Introduction. Amsterdam, Elsevier 1981.

[351] Nedelec, J.C.: Approximation des équations intégrales en mécanique et en physique. Cours de l'école d'été C.E.A. (IRIA) E.D.F., 1977.

[352] Nehari, Z.: Conformal Mapping. New York, McGraw-Hill 1952.

[353] Nonlinear Partial Differential Equations in Applied Science (ed. by H. Fujita, P.D. Lax and G. Strang). Amsterdam, Elsevier 1982.

[354] Norrie, D.H., Vries, G.H.: An Introduction to Finite Element Analysis. New York, Academic Press 1978.

[355] Nožička, F., Guddat, J., Hollatz, H.: Theorie der linearen Optimierung. Berlin, Akademieverlag 1972.

[356] Nožička, F., Guddat, J., Hollatz, H.: Theorie der linearen parametrischen Optimierung. Berlin, Akademieverlag 1974.

[357] Numerical Conformal Mapping (ed. by L.N. Trefethen). Amsterdam, Elsevier 1986.

[358] Numerical Solution of Boundary Value Problems for Ordinary Differential Equations (ed. by A.K. Aziz). New York, Academic Press 1975.

[359] Nürnberger, G.: Approximation by Spline Functions. Berlin, Springer 1989.

[360] Olver, F.W.J.: The Evaluation of Zeros of High-Degree Polynomials. Phil. Trans. Roy. Soc. London, A 244 (1952), 385–415.

[361] Ostrowski, A.M.: Solution of Equations and Systems of Equations. New York, Academic Press 1960.

[362] Owen, D.B.: Handbook of Statistical Tables. Reading (Mass.), Addison-Wesley 1962.

[363] Pars, L.A.: Calculus of Variations. London, Heinemann 1962.

[364] Parzen, E.: Modern Probability Theory and its Applications. New York, Wiley 1960.

[365] Pascali, D., Sburlan, S.: Nonlinear Mapping of Monotone Type. Dordrecht, Sijthoff and Noordhoff 1980.

[366] Patel, J.K., Kapadla, C.H., Owen, D.B.: Handbook of Statistical Distributions. New York, Dekker 1976.

[367] Patil, G.P., Yoshi, S.W., Rao, C.R.: A Dictionary and Bibliography of Discrete Distributions. Edinburg, Oliver and Boyd 1968.

[368] Pearson, E.S., Hartley, H.O.: Biometrica Tables for Statisticians. Volume I. Cambridge, Cambridge Univ. Press 1956.

[369] Petrovskij, I.G.: Lectures on Partial Differential Equations. London, Interscience 1954.

[370] Petrovskij, I.G.: Ordinary Differential Equations. Englewood Cliffs (N.J.), Prentice-Hall 1986.

[371] Piehler, J.: Ganzzahlige Optimierung. Leipzig, Teubner 1970.

[372] Piehler, J.: Algebraische Methoden in der ganzzahligen Optimierung. Leipzig, Teubner 1983.

[373] Pissanetzky, S.: Sparse Matrix Technology. London, Academic Press 1984.

[374] Pol, B. van der, Bremmer, H.: Operational Calculus Based on the Two-Sided Laplace Integral. Cambridge, Cambridge Univ. Press 1950.

[375] Pontryagin, L.S.: Ordinary Differential Equations. Cambridge (Mass.), Addison-Wesley 1962.

[376] Press, W.H. et al.: Numerical Recipes. The Art of Scientific Computing. Cambridge, Cambridge University Press 1988.

[377] Priestley, M.B.: Spectral Analysis and Time Series. London, Academic Press 1981.

[378] Ralston, A., Rabinowitz, P.: A First Course in Numerical Analysis. 2nd ed. New York, McGraw-Hill 1978.

[379] Ramm, A.G.: Scattering by Obstacles. Dordrecht, Reidel 1986.

[380] Randles, R.H., Wolfe, D.A.: Introduction to the Theory of Nonparametric Statistics. New York, Wiley 1979.

[381] Range, R.M.: Holomorphic Functions and Integral Representations in Several Complex Variables. Berlin, Springer 1986.

[382] Rao, C.R.: Linear Statistical Inference and its Applications. New York, Wiley 1965.

[383] Recent Developments in Several Complex Variables (ed. by J.E. Fornaes). Princeton, Princeton Univ. Press 1980.

[384] Recent Topics in Nonlinear Partial Differential Equations (ed. by M. Mimura, T. Nishida, K. Masuda, T. Suzuki). Amsterdam, North-Holland: I 1984, II 1985, III 1988.

[385] Reddy, B.D.: Fourier Analysis and Boundary Value Problems: An Introductory Treatment. London, Pitman 1986.

[386] Redish, K.A.: An Introduction to Computational Methods. London, English Univ. Press 1961.

[387] Reinhardt, H.J.: Analysis of Approximate Methods for Differential and Integral Equations. Berlin, Springer 1986.

[388] Reissig, R.G., Sansone, G., Conti, R.: Nonlinear Differential Equations of Higher Order. Dordrecht, Sijthoff and Noordhoff 1974.

[389] Rektorys, K.: Variational Methods in Mathematics, Science and Engineering. 2nd ed. Dordrecht, Reidel 1980.

[390] Rektorys, K.: The Method of Discretization in Time and Partial Differential Equations. Dordrecht, Reidel 1982.

[391] Rényi, A.: Probability Theory. Budapest, Akadémia Kiadó 1970.

[392] Richtmyer, R.D., Morton, K.N.: Difference Methods for Initial Value Problems. New York, Wiley 1967.

[393] Riedrich, T.: Vorlesungen über nichtlineare Operatorengleichungen. Leipzig, Teubner 1976.

[394] Riesz, F., Nagy, Sz.: Vorlesungen über Funktionalanalysis. Berlin, D. Verlag der Wiss. 1982.

[395] Roberts, S.M., Shipman, J.S.: Two-Point Boundary Value Problems: Shooting Methods. New York, Elsevier 1972.

[396] Rohatgi, V.K.: Statistical Inference. New York, Wiley 1984.

[397] Rolewicz, S.: Metric Linear Spaces. Dordrecht, Reidel 1985.

[398] Romanovskij, P.I.: Mathematical Methods for Engineers and Technologists. Oxford, Pergamon Press 1961.

[399] Rosenblatt, M.: Stationary Sequences and Random Fields. Boston, Birkhäuser 1985.

[400] Rosinger, E.E.: Generalized Solutions of Nonlinear Partial Differential Equations. Amsterdam, Elsevier 1987.

[401] Rothe, R.: Höhere Mathematik für Mathematiker, Physiker und Ingenieure 1–5. 13. Aufl. Leipzig, Teubner 1954.

[402] Rotschild, V., Lagothetis, N.: Probability Distributions. New York, Wiley 1985.

[403] Rouche, N., Mawhin, J.: Ordinary Differential Equations: Stability and Periodic Solutions. London, Pitman 1980.

[404] Rozanov, Ju.A.: Stationary Random Processes. San Francisco, Holden Day 1967.

[405] Ruge, J., Stüben, K.: Efficient Solution of Finite Difference and Finite Element Equations by Algebraic Multigrid (AMG). GMD-Studien Nr. 89, St. Augustin, GMD 1984.

[406] Saaty, T.L.: Elements of Queueing Theory with Applications. New York, McGraw-Hill 1961.

[407] Sachs, L.: Applied Statistics. A Handbook of Techniques. Berlin, Springer 1984.

[408] Saks, S., Zygmund, A.: Analytic Functions. Warszawa, Polskie Towarzystwo Matematyczne 1952.

[409] Salvadori, M.G., Schwarz, R.J.: Differential Equations in Engineering Problems. Englewood Cliffs (N.J.), Prentice-Hall 1954.

[410] Samarskij, A.A., Nikolajev, E.S.: Methods of Solution of Finite Difference Equations. Moskva, Nauka 1978 (in Russian).

[411] Sanders, D.H., Eng, R.J., Murph, A.F.: Statistics: A Fresh Approach. New York, McGraw-Hill 1985.

[412] Sard, A.: Linear Approximations. Providence, Amer. Math. Soc. 1982.

[413] Schechter, M.: Spectra of Partial Differential Operators. Amsterdam, North-Holland 1986.

[414] Scheffé, H.: The Analysis of Variance. New York, Wiley 1963.

[415] Schmeidler, W.: Integralgleichungen mit Anwendungen in Physik und Technik. Leipzig, Akademieverlag 1950.

[416] Schmid, C.F.: Statistical Graphics: Design Principles and Practices. New York, Wiley 1983.

[417] Schneeweiss, W.: Zuverlässigkeitstheorie. Berlin, Springer 1977.

[418] Schwabik, Š., Tvrdý, M., Vejvoda, O.: Differential and Integral Equations: Boundary Value Problems and Adjoints. Dordrecht, Reidel 1978.

[419] Schwarz, H.R.: Methode der finiten Elemente. Stuttgart, Teubner 1984.

[420] Scott, M.R.: Invariant Imbedding and Its Applications for Ordinary Differential Equations: An Introduction. Reading (Mass.), Addison-Wesley 1973.

[421] Searle, S.R.: Linear Models. New York, Wiley 1971.

[422] Seber, G.A.F.: Linear Regression Analysis. New York, Wiley 1977.

[423] Seifart, E., Manteufel, K.: Lineare Optimierung. Leipzig, Teubner 1985.

[424] Serfling, R.J.: Approximation Theorems of Mathematical Statistics. New York, Wiley 1980.

[425] Shampine, L.F., Gordon, M.K.: Computer Solution of Ordinary Differential Equations. San Francisco, Freeman 1975.

[426] Showalter, R.E.: Hilbert Space Methods for Partial Differential Equations. London, Pitman 1977.

[427] Siegel, S., Castellan, N.J.: Nonparametric Statistics. New York, McGraw-Hill 1988.

[428] Simonnard, M.: Programation linéaire. Paris, Dunod 1962.

[429] Singh, S.P.: Nonlinear Functional Analysis and Its Applications. Dordrecht, Reidel 1986.

[430] Sirovich, L.: Introduction to Applied Mathematics. Berlin, Springer 1988.

[431] Skrypnik, I.V.: Nonlinear Elliptic Boundary Value Problems. Leipzig, Teubner 1986.

[432] Smirnov, V.I.: Course of Higher Mathematics, 5 Vols. Oxford, Pergamon Press 1964.

[433] Smith, B.T. et al.: Matrix Eigensystem Routines – EISPACK Guide. Lecture Notes in Computer Science Vol. 6. 2nd ed. Berlin, Springer 1976.

[434] Smithies, F.: Integral Equations. Cambridge, Cambridge Univ. Press 1962.

[435] Sneddon, I.N.: Fourier Transforms. New York, McGraw-Hill 1951.

[436] Sneddon, I.N.: Elements of Partial Differential Equations. New York, McGraw-Hill 1957.

[437] Sneddon, I.N.: Introduction to Partial Differential Equations. New York, McGraw--Hill 1957.

[438] Sobolev, S.L.: Partial Differential Equations of Mathematical Physics. Oxford, Pergamon Press 1964.

[439] Späth, H.: Algorithmen für elementare Ausgleichsmodelle. München, Oldenburg 1973.

[440] Srinivasan, S.K., Mehata, K.M.: Probability and Random Processes. New York, McGraw-Hill 1981.

[441] Srinivasan, S.K., Mehata, K.M.: Stochastic Processes. New York, McGraw-Hill 1988.

[442] Statistics Technical Committee of Amer. Soc. for Quality Control. Glossary and Tables for Statistical Quality Control. Milwaukee, Wisconsin 1973.

[443] Steffensen, J.F.: Interpolation. New York, Chelsea 1950.

[444] Stepanov, W.W.: Lehrbuch der Differentialgleichungen. 6. Aufl. Berlin, D. Verlag der Wiss. 1989.

[445] Stetter, H.J.: Analysis of Discretization Methods for Ordinary Differential Equations. Berlin, Springer 1973.

[446] Stoer, J., Bulirsch, R.: Introduction to Numerical Analysis. New York, Springer 1980.

[447] Stoll, W.: Holomorphic Functions of Finite Order in Several Variables. Providence, Amer. Math. Soc 1974.

[448] Stoodley, K.D.C.: A First Course in Applied and Computational Statistics. Chichester, Horwood 1984.

[449] Strang, G., Fix, G.: An Analysis of the Finite Element Method. Englewood Cliffs (N.J.), Prentice-Hall 1973.

[450] Struble, R.A.: Nonlinear Differential Equations. New York, McGraw-Hill 1962.

[451] Swarztrauber, P.N.: The Methods of Cyclic Reduction, Fourier Analysis and the FACR Algorithm for the Discrete Solution of Poisson's Equation on a Rectangle. SIAM Review 19 (1977), 490–501.

[452] Swarztrauber, P.N., Sweet, R.A.: Algorithm 541. Efficient Fortran subprograms for the solution of separable elliptic partial differential equations. ACM Transactions on Mathematical Software 5 (1979), 352–371.

[453] Swoboda, H.: Knaurs Buch der modernen Statistik. München, Knaur 1971.

[454] Systems of Nonlinear Partial Differential Equations (ed. by J.M. Ball). Dordrecht, Reidel 1983.

[455] Szmydt, Z.: Fourier Transformation and Linear Differential Equations. Dordrecht, Reidel 1977.

[456] Taufer, J.: Lösung der Randwertprobleme von linearen Differentialgleichungen. Rozpravy ČSAV, řada mat. a přírod. věd 83, Praha, Academia 1973.

[457] Taylor, A.E.: Introduction to Functional Analysis. New York, Wiley 1958.

[458] Teman, R.: Navier Stokes Equations. 2nd ed. Amsterdam, Elsevier 1985.

[459] Tewarson, R.P.: Sparse Matrices. New York, Academic Press 1973.

[460] The State of the Art in Numerical Analysis (ed. by D.A.H. Jacobs). London, Academic Press 1977.

[461] The State of the Art in Numerical Analysis (ed. by A. Iserles and M.J.D. Powell). Oxford, Clarendon Press 1987.

[462] Thirzing, W.: Lehrbuch der mathematischen Physik. Berlin, Springer: I 1988, II 1989.

[463] Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems. Berlin, Springer 1984.

[464] Tichomirov, V.H.: Grundprinzipien der Theorie der Extremalaufgaben. Leipzig, Teubner 1982.

[465] Titchmarsh, E.C.: Eigenfunction Expansions, 2 Vols. London, Oxford Univ. Press 1958.

[466] Toutenburg, H.: Prior Information in Linear Models. New York, Wiley 1982.

[467] Tranter, C.J.: Integral Transforms in Mathematical Physics. London, Wiley 1954.

[468] Traub, L.F.: Iterative Methods for Solution of Equations. Englewood Cliffs (N.J.), Prentice-Hall 1964.

[469] Tricomi, F.G.: Integral Equations. New York, Interscience 1957.

[470] Tricomi, F.G.: Differential Equations. London, Blackie 1961.

[471] Triebel, H.: Analysis and Mathematical Physics. Dordrecht, Reidel 1986.

[472] Tykhonov, A.N., Samarskij, A.A.: Partial Differential Equations of Mathematical Physics. San Francisco, Holden Day 1964.

[473] Ullman, N.R.: Elementary Statistics: An Applied Approach. New York, Wiley 1978.

[474] Vainikko, G.: Funktionalanalysis der Diskretisierungsmethoden. Leipzig, Teubner 1976.

[475] Varga, R.S.: Matrix Iterative Analysis. Englewood Cliffs (N.J.), Prentice-Hall 1962.

[476] Verhulst, F.: Nonlinear Equations and Dynamic Systems. Berlin, Springer 1990.

[477] Vladimirov, V.S.: Methods of the Theory of Functions of Several Complex Variables. Cambridge (Mass.), Massachusetts Institute of Technology Press 1966.

[478] Voelker, D., Doetsch, G.: Die zweidimensionale Laplace-Transformation. Basel, Birkhäuser 1950.

[479] Vulich, B.Z.: Introduction to Functional Analysis for Scientists and Technologists. Oxford, Pergamon Press 1963.

[480] Wachspress, E.L.: Iterative Solution of Elliptic Systems. Englewood Cliffs (N.J.), Prentice-Hall 1966.

[481] Wald, A.: Sequential Analysis. New York, Wiley 1947.

[482] Wall, H.S.: Continued Fractions. London, van Nostrand 1948.

[483] Walter, W.: Gewöhnliche Differentialgleichungen. Berlin, Springer 1990.

[484] Wasan, M.T.: Parametric Estimation. New York, McGraw-Hill 1976.

[485] Watson, G.N.: Bessel Functions. London, Cambridge Univ. Press 1966.

[486] Webster, A.G.: Partial Differential Equations of Mechanics and Physics. New York, Stechert 1933.

[487] Weisberg, S.: Applied Linear Regression. New York, Wiley 1985.

[488] Wendland, W., Stephan, E.: BIEM for Mixed Boundary Value Problems. In: Innovative Numerical Analysis in the Engineering Sciences (ed. by R. Shaw). Univ. Press of Virginia 1979, 543–554.

[489] Wenzel, H.: Gewöhnliche Differentialgleichungen I (5. Aufl.), II (4. Aufl.). Leipzig, Teubner 1987.

[490] Werner, H., Arndt, H.: Gewöhnliche Differentialgleichungen. Berlin, Springer 1987.

[491] Wetzel, W., Jöhnk, M.D., Naeve, P.: Statistische Tabellen. Berlin, W. de Gruyter 1967.

[492] White, R.E.: An Introduction to the Finite Element Method with Applications to Nonlinear Problems. Chichester, Wiley 1985.

[493] Whiteman, J.R.: A Bibliography for Finite Elements. London, Academic Press 1975.

[494] Whittaker, E.T., Watson, G.N.: A Course of Modern Analysis. 4th ed. London, Cambridge Univ. Press 1958.

[495] Wilcox, C.H.: Asymptotic Solution of Differential Equations. London, Wiley 1964.

[496] Wilkinson, J.H.: Rounding Errors in Algebraic Processes. London, HMSO 1963.

[497] Wilkinson, J.H.: The Algebraic Eigenvalue Problem. New York, Oxford Univ. Press 1965.

[498] Wilkinson, J.H., Reinsch, C.: Linear Algebra. Handbook for Automatic Computation, Vol. 2. New York, Springer 1971.

[499] Wilks, S.S.: Mathematical Statistics. New York, Wiley 1962.

[500] Williams, H.P.: Model Building in Mathematical Programming. Chichester, Wiley 1978.

[501] Wladimirov, W.S.: Gleichungen der mathematischen Physik. Berlin, D. Verlag der Wiss. 1972.

[502] Wlasov, W.S.: Allgemeine Schalentheorie und ihre Anwendung in der Technik. Berlin, Akademieverlag 1958.

[503] Wold, H.: A Study in the Analysis of Stationary Time Series. Uppsala, Almquist and Wiksell 1954.

[504] Wonnacott, R.J., Wonnacott, T.H.: Introductory Statistics. New York, Wiley 1985.

[505] Wonnacott, R.J., Wonnacott, T.H.: Statistics: Discovering its Power. New York, Wiley 1982.

[506] Yaglom, A.M.: An Introduction to the Theory of Stationary Random Functions. Englewood Cliffs (N.J.), Prentice-Hall 1962.

[507] Yosida, K.: Functional Analysis. Berlin, Springer 1965.

[508] Young, N.: An Introduction to Hilbert Space. Cambridge, Cambridge Univ. Press 1988.

[509] Young, P.: Recursive Estimation and Time-Series Analysis. Berlin, Springer 1984.

[510] Zeidler, E.: Nonlinear Functional Analysis and Its Applications, I–V. Berlin, Springer 1984–1988.

[511] Ženíšek, A.: Nonlinear Elliptic and Evolution Problems and Their Finite Element Approximations. London, Academic Press 1990.

[512] Zienkiewicz, O.C.: The Finite Element Method in Engineering Science. 3rd ed. London, McGraw-Hill 1977.

[513] Zlámal, M.: On the Finite Element Method. Numer. Math. 12 (1968), 394–409.

[514] Zoutendijk, G.: Methods of Feasible Directions. A Study in Linear and Nonlinear Programming. Amsterdam, Elsevier 1960.

[515] Zuily, C.: Problems in Distributions and Partial Differential Equations. Amsterdam, Elsevier 1988.

Some other recent publications of interest:

Arnold, V.I.: Singularities of Caustics and Wave Fronts. Dordrecht, Kluwer 1990.

Benilan, P., Chipot, M., Evans, L.C., Pierre, M.: Recent Advances in Nonlinear Elliptic and Parabolic Problems. London, Pitman 1989.

Brudnyi, Yu.A., Krugljak, N.Ya.: Interpolation Functors and Interpolation Spaces. Amsterdam, Elsevier 1991.

Chirka, E.M.: Complex Analytic Sets. Dordrecht, Kluwer 1989.

Gamkrelidzde, R.V.: Analysis. Berlin, Springer, I (Integral Represention and Asymptotic Methods) 1989, II (Convex Analysis and Approximation Theory) 1990.

Gindikin, S.G., Khenkin, G.M, Vitushkin, A.G.: Several Complex Analysis. Berlin, Springer, I (1990), II (to appear), III (1989), IV (1990).

Goldstein, V., Reshetnyak, Yu.: Quasiconformal Mappings and Sobolev Spaces. Dordrecht, Kluwer 1990.

Griffiths, D.F., Watson, G.A.: Numerical Analysis. Pitman 1990.

Hämmerlin, G., Hoffmann, K.H.: Numerical Mathematics. Springer, Berlin 1991.

Lakshmikantham, V., Matrosov, V.M., Sivasundaram, S.: Vector Lyapunov Functions and Stability Analysis of Nonlinear Systems. Dordrecht, Kluwer 1991.

Leung, A.W.: Systems of Nonlinear Partial Differential Equations. Dordrecht, Kluwer 1989.

Levitan, B.M., Sargsjan, I.S.: Sturm-Liouville and Dirac Operators. Dordrecht, Kluwer 1991.

Marino, A., Murthy, M.K.V.: Nonlinear Variational Problems. London, Pitman 1989.

Maz'ya, V.G., Nikol'skij, S.M.: Analysis IV (Linear and Boundary Integral Equations). Berlin, Springer 1991.

Mennicken, M., Moeller, M.: Spectral Theory and Boundary Value Problems. Amsterdam, North-Holland 1992.

Mlak, W.: Hilbert Space and Operator Theory. Dordrecht, Kluwer 1991.

Nikol'skij, S.M.: Analysis III (Spaces of Differentiable Functions). Berlin, Springer 1990.

Paivarinta, P.: Function Spaces. London, Pitman 1989.

Perko, L.: Differential Equations and Dynamical Systems. Berlin, Springer 1991.

Petkov, V.: Scattering Theory for Hyperbolic Operators. Amsterdam, North-Holland 1989.

Rabinovich, M.I., Trubetskov, D.I.: Oscillations and Waves. Dordrecht, Kluwer 1989.

Rauch, J.: Partial Differential Equations. Berlin, Springer 1991.

Rosinger, E.E.: Nonlinear Partial Differential Equations. Amsterdam, North-Holland 1990.

Sachdev, P.L.: Nonlinear Ordinary Differential Equations and their Applications. New York, Dekker 1990.

Samoilenko, Y.S.: Spectral Theory of Families of Self-Adjoint Operators. Dordrecht, Kluwer 1991.

Schaaf, R.: Global Solution Branches of Two-Point Boundary Value Problems. Berlin, Springer 1990.

Skorohod, A.V.: Random Processes with Independent Increments. Dordrecht, Kluwer 1991.

Sleeman, B.D., Jarvis, R.J.: Ordinary and Partial Differential Equations. London, Pitman 1989.

Sobczyk, K.: Stochastic Differential Equations. Dordrecht, Kluwer 1991.

Wen, G.C., Begehr, H.G.W.: Boundary Value Problems for Elliptic Equations and Systems. London, Pitman 1990.

Wentzel, A.D.: Limit Theorems on Large Deviations for Markov Stochastic Processes. Dordrecht, Kluwer 1990.

Ziemer, W.P.: Weakly Differentiable Functions. Berlin, Springer 1989.

INDEX

I, or II means Volume I, or Volume II, respectively. The symbol ff means "and following pages". For example, "Abstract functions, II 364 ff" means that abstract functions are treated in Volume II, starting with page 364. Articles (definite, indefinite) have been omitted whenever it was possible.