Taylor
Mann

ADVANCED CALCULUS

Third Edition

Wiley

Angus E. Taylor W. Robert Mann

ADVANCED CALCULUS

Third Edition

ADVANCED CALCULUS

Third Edition

ANGUS E. TAYLOR
University of California

W. ROBERT MANN
University of North Carolina

JOHN WILEY & SONS, INC.
New York / Chichester / Brisbane / Toronto / Singapore

Copyright © 1955, 1972, 1983, by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of
this work beyond that permitted by Sections
107 and 108 of the 1976 United States Copyright
Act without the permission of the copyright
owner is unlawful. Requests for permission
or further information should be addressed to
the Permissions Department, John Wiley & Sons.

Library of Congress Cataloging in Publication Data:

Taylor, Angus Ellis, 1911–
Advanced calculus.

Includes index.
1. Calculus. I. Mann, W. Robert (William Robert), 1920– II. Title.
QA303.T2 1982 515 81-16141
ISBN: 0-471-02566-6 AACR2

Printed in the United States of America

10

Scanned on Nov/25th/2004

PREFACE

The concept of a course in advanced calculus has a long history in American colleges and universities. Quite naturally, the concept has not remained static. Nor does it have exactly the same connotation to all teachers, students, and users of mathematics at any given time. There are a variety of needs to be served by a textbook in advanced calculus. In planning this edition of our book, we have retained the basic qualities that contributed to user satisfaction with the first two editions while changing and introducing some new features that we believe to be important as well as appropriate in the light of current conditions.

We believe that the scope of a book on advanced calculus should be broad enough to:

1. Build a bridge from elementary calculus to higher mathematical analysis suitable for use by students of two sorts, both those intending to specialize in mathematics and those who need more advanced mathematics as a tool in other studies or in their employment.
2. Provide a thorough treatment of the calculus of functions of several variables and of vector functions of vector variables.
3. Provide a firm grounding in the fundamentals of analysis, embracing at least the following topics: point set theory on the line and in Euclidean space, continuous functions and mappings, uniform convergence, the Riemann integral, and infinite series.
4. Pay due attention to some topics important for applied mathematics, including some theory of curves and surfaces; vector fields; such notions as gradient, divergence, and curl and their occurrence in integral theorems; some notions about numerical methods and the potential for use of simple computers in problems calling for approximation or minimization; and improper integrals.

Our mode of treatment of the material is based on our belief that the book should be more than a skeletal framework of tersely stated assumptions, definitions, theorems, and proofs. The exposition is designed to make the book readable and understandable by a student through his/her own efforts if he/she will read carefully and learn to verify or carry through for himself/herself the steps of reasoning that are either given or indicated. An important part of an advanced calculus course is the training that it entails in deductive reasoning from explicit assumptions and definitions. The best way of assuring that a student will

benefit from this training is to make the process interesting by providing suitable motivation and enough guidance to assure to the diligent student the pleasure of success. With confidence thus gained the student may proceed to become an ever more independent learner.

Our book is written on the assumption that students using it have normal skill in the formal aspects of elementary calculus and that they can draw freely on the standard subject matter of algebra, trigonometry, analytic geometry, and elementary calculus. In Chapter 1, we present a systematic overview of the more theoretical side of elementary calculus of functions of one real variable, but without a full treatment of those topics that depend in a crucial way on a rigorous exposition of the properties of the real number system. Such an exposition is given in Chapter 2, followed in Chapter 3 by a rigorous presentation of those properties of continuous real functions of a real variable that are essential to the theoretical structure of differential calculus of such functions. These chapters form a part of the bridge to higher mathematical analysis. The extent to which Chapter 1 will need to be formally included in a course of advanced calculus will depend on the level of preparation of the students and judgment of the teacher. Even if not extensively used for regular assignments, Chapter 1 is useful for reference and supplementary reading and study and can provide motivation for the work in Chapters 2 and 3.

Chapter 4 is on several special topics somewhat apart from the mainstream of a course on advanced calculus. The results are occasionally used in important ways here and there in the book; they are available for reference and can be taken up by the individual student or by the teacher if the need arises.

Parts of Chapter 6 and 13 will be to some extent familiar to students from work in more elementary courses. It is feasible to omit Chapter 13 almost entirely from a course based on the book except insofar as some parts early in Chapter 13 may be needed for reference in connection with §14.6 and §§18.6 and 18.61.

A thorough treatment of the differential calculus of real functions of several variables and vector functions of vector variables is provided in Chapters 6, 7, 8, 9, and 12, with important supporting material on vectors, matrices, and linear transformations in Chapters 10 and 11. Two different approaches to implicit function theorems and inversions of mappings are provided: *without* vectors in Chapters 8 and 9, *with* vectors in Chapter 12. The vector approach is the modern way, and it has many advantages. But the older, classical, way of dealing with implicitly defined real functions of several variables (Theorem I, §8,1) is simple, elegant, and very instructuve; it deserves to be remembered. We think the local inversion of mappings should be understood and appreciated in the context of mappings from $\mathbb{R}^2$ to $\mathbb{R}^2$, as in Chapter 9, before dealing with the general inversion theorem in §12.6.

The amount of linear algebra needed to support the vector differential calculus of Chapter 12 is presented in Chapter 11 (and, to some extent, in parts of Chapter 10). Chapter 11 also contains the minimum amount of material on norms and metrics that is needed to discuss continuity and differentiability of vector functions of vectors. Our discussion is, for the most part, limited to finite

dimensional vector spaces with the standard orthonormal system of basis vectors associated with a Cartesian system of coordinates. Thus we regularly use the Euclidean norm and metric in $\mathbb{R}^n$. We also use the matrix representation of linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$ that goes along with the use of standard orthonormal bases. Students using the book are expected to have some familiarity with matrices and elementary linear algebra. Some glimpses using more abstract points of view are offered.

In Chapter 12 we present with care the definitions of the differential and the derivative (taken in that order) of a vector function from $\mathbb{R}^n$ to $\mathbb{R}^m$. This is a subject that has needed some clarification in the textbook literature. If $\mathbf{f}(\mathbf{x})$ is a function from a part of $\mathbb{R}^n$ to $\mathbb{R}^m$, differentiable for certain values of $\mathbf{x}$, the differential $d\mathbf{f}$ of $\mathbf{f}$ is a function of two variables $\mathbf{x}$ and $d\mathbf{x}$ that is linear in $d\mathbf{x}$ (where $d\mathbf{x}$ varies over all of $\mathbb{R}^n$) and has values in $\mathbb{R}^m$. For a given $\mathbf{x}$ at which $\mathbf{f}$ is differentiable the linear transformation that takes $d\mathbf{x}$ into the value of $d\mathbf{f}$ at $(\mathbf{x}, d\mathbf{x})$ is called the derivative of $\mathbf{f}$ at $\mathbf{x}$ and denoted by $\mathbf{f}'(\mathbf{x})$. Thus the value of $d\mathbf{f}$ is $\mathbf{f}'(\mathbf{x})\, d\mathbf{x}$, where the juxtaposition means that $\mathbf{f}'(\mathbf{x})$ acts on $d\mathbf{x}$. Observe that the value of $\mathbf{f}'$ is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, whereas the value of $d\mathbf{f}$ is a vector in $\mathbb{R}^m$. Thus the derivative and the differential are quite different functions.

A third important block of material for courses in advanced calculus is provided by Chapters 16, 17, 20, and at least the first part of Chapter 18. These provide essential material for the study of limits, convergence, and continuity for functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ (including the special cases when $n = 1$ or $m = 1$, or both), together with a completion of the theoretical structure of elementary integral calculus. To this block may be added, if time permits, selections from Chapters 19, 21 and 22, depending on the interests of the class and the teacher.

Most of the main topics mentioned thus far in this survey of the book are important for applied as well as for pure mathematics. The same is true for other topics yet to be mentioned, although the treatment of these topics is aimed rather more at those interested in applications than at those interested in pure theory. We are referring to the later part of Chapter 10 (on scalar and vector fields), most of Chapters 14 and 15 and parts of Chapter 22. Some selection from this body of material will very likely be appropriate and desirable in a course in advanced calculus with a broad clientele.

There is another way of looking at mathematical analysis, in which one classifies the content of a book, not by the various topics and different subjects, but according to the following categories: ideas and concepts, theorems and proofs, and specific problems, together with methods for dealing with them and the information provided by the solution. These categories are not completely separate from each other, of course. Concepts enter into the statements and proofs of theorems. Some theorems provide answers to interesting problems, and techniques of proof may (but do not in all cases) furnish explicit solutions to problems. We believe that the vitality of mathematics derives in a highly significant way from interest in the solution of problems that can be formulated mathematically. The generation of powerful methods for solving problems

necessitates the building of theories (which may be somewhat abstract or elaborate) to justify the methods and to amass a body of knowledge useful to those who apply the theories and the methods. We have endeavored to deal with our subject with a judicious (and, we hope, instructive) mixture of attention to concepts, theories, problems, and techniques of proofs and solutions.

Among the specific ways in which this edition differs from the second edition are these: substantial changes in the discussion of quadratic forms in §6.9, a number of changes in Chapter 7, especially in the discussion of critical points in §7.6, a considerable revision of Chapter 9, addition to Chapter 10 of material on vectors in space of n-dimensions, a number of revisions in Chapters 11 and 12 (including major revisions of the treatment of the differential and the proof of the inversion theorem) and a change in the discussion of Stirling's formula (§22.8) with a sharper result.

An important innovation is the use made of programmable pocket calculators. This is in Chapter 12, which some students regard as forbiddingly theoretical. We think this may be partly due to the difficulty in developing a feeling for the derivative of a function from $\mathbb{R}^n$ to $\mathbb{R}^m$. To alleviate this difficulty we have introduced Newton's method at the level of problems involving the derivative of functions from $\mathbb{R}^n$ to $\mathbb{R}^n$. We have also tried to present more clearly the gradient as a derivative by including some numerical applications of the method of steepest descent. Along similar lines we give a generalization of the elementary product formula for differentiation, which gives the derivative of the scalar product of two vector-valued functions. This formula is then used to arrive at the method of least squares and the idea of a generalized matrix inverse. These practical applications cannot, of course, make the basic theory easier, but according to our experience, they are helpful to the students' understanding. It is still too early to discern exactly what the optimal role of abudant, cheap, and easy computational capability in teaching advanced calculus is going to be. We believe however that the present state of calculator technology offers opportunities to combine theory and practice in a way that illuminates both.

In conclusion we want to express, as we did in the second edition, our debt of gratitude to the students we have taught, from the teaching of whom we have learned much. We have enjoyed our teaching, partly because it has deepened our own learning, but especially when we have seen our students growing in understanding and appreciation of mathematical analysis. The questions and comments of students have often led us to new insights both in the subject and in our ways of teaching. We hope that other students and other teachers will find that this book opens the doors to understanding and enjoyment.

Angus E. Taylor
W. Robert Mann

CONTENTS

1 / FUNDAMENTALS OF ELEMENTARY CALCULUS

1. Introduction 1
1.1 Functions 2
1.11 Derivatives 12
1.12 Maxima and Minima 20
1.2 The Law of the Mean (The Mean-Value Theorem for Derivatives) 26
1.3 Differentials 32
1.4 The Inverse of Differentiation 35
1.5 Definite Integrals 38
1.51 The Mean-Value Theorem for Integrals 45
1.52 Variable Limits of Integration 46
1.53 The Integral of a Derivative 49
1.6 Limits 53
1.61 Limits of Functions of a Continuous Variable 54
1.62 Limits of Sequences 58
1.63 The Limit Defining a Definite Integral 67
1.64 The Theorem on Limits of Sums, Products, and Quotients 67

2 / THE REAL NUMBER SYSTEM

2. Numbers 72
2.1 The Field of Real Numbers 72
2.2 Inequalities. Absolute Value 74
2.3 The Principle of Mathematical Induction 75
2.4 The Axiom of Continuity 77
2.5 Rational and Irrational Numbers 78
2.6 The Axis of Reals 79
2.7 Least Upper Bounds 80
2.8 Nested Intervals 82

3 / CONTINUOUS FUNCTIONS

3. Continuity 85
3.1 Bounded Functions 86
3.2 The Attainment of Extreme Values 88
3.3 The Intermediate-Value Theorem 90

4 / EXTENSIONS OF THE LAW OF THE MEAN

4. Introduction 95
4.1 Cauchy's Generalized Law of the Mean 95
4.2 Taylor's Formula with Integral Remainder 97
4.3 Other Forms of the Remainder 99
4.4 An Extension of the Mean-Value Theorem for Integrals 105
4.5 L'Hospital's Rule 106

5 / FUNCTIONS OF SEVERAL VARIABLES

5. Functions and Their Regions of Definition 116
5.1 Point Sets 117
5.2 Limits 122
5.3 Continuity 125
5.4 Modes of Representing a Function 127

6 / THE ELEMENTS OF PARTIAL DIFFERENTIATION

6. Partial Derivatives 130
6.1 Implicit Functions 132
6.2 Geometrical Significance of Partial Derivatives 135
6.3 Maxima and Minima 138
6.4 Differentials 144
6.5 Composite Functions and the Chain Rule 154
6.51 An Application in Fluid Mechanics 162
6.52 Second Derivatives by the Chain Rule 164
6.53 Homogeneous Functions. Euler's Theorem 168
6.6 Derivatives of Implicit Functions 172
6.7 Extremal Problems with Constraints 177
6.8 Lagrange's Method 182
6.9 Quadratic Forms 189

7 / GENERAL THEOREMS OF PARTIAL DIFFERENTIATION

7. Preliminary Remarks 196
7.1 Sufficient Conditions for Differentiability 197

7.2 Changing the Order of Differentiation 199
7.3 Differentials of Composite Functions 202
7.4 The Law of the Mean 204
7.5 Taylor's Formula and Series 207
7.6 Sufficient Conditions for a Relative Extreme 211

8 / IMPLICIT-FUNCTION THEOREMS

8. The Nature of the Problem of Implicit Functions 222
8.1 The Fundamental Theorem 224
8.2 Generalization of the Fundamental Theorem 227
8.3 Simultaneous Equations 230

9 / THE INVERSE FUNCTION THEOREM WITH APPLICATIONS

9. Introduction 237
9.1 The Inverse Function Theorem in Two Dimensions 241
9.2 Mappings 247
9.3 Successive Mappings 252
9.4 Transformations of Co-ordinates 255
9.5 Curvilinear Co-ordinates 258
9.6 Identical Vanishing of the Jacobian. Functional Dependence 263

10 / VECTORS AND VECTOR FIELDS

10. Purpose of the Chapter 268
10.1 Vectors in Euclidean Space 268
10.11 Orthogonal Unit Vectors in $\mathbb{R}^3$ 273
10.12 The Vector Space $\mathbb{R}^n$ 274
10.2 Cross Products in $\mathbb{R}^3$ 280
10.3 Rigid Motions of the Axes 283
10.4 Invariants 286
10.5 Scalar Point Functions 291
10.51 Vector Point Functions 293
10.6 The Gradient of a Scalar Field 295
10.7 The Divergence of a Vector Field 300
10.8 The Curl of a Vector Field 305

11 / LINEAR TRANSFORMATIONS

11. Introduction 309
11.1 Linear Transformations 312

11.2 The Vector Space $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ 313
11.3 Matrices and Linear Transformations 313
11.4 Some Special Cases 316
11.5 Norms 318
11.6 Metrics 319
11.7 Open Sets and Continuity 320
11.8 A Norm on $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ 324
11.9 $\mathscr{L}(\mathbb{R}^n)$ 327
11.10 The Set of Invertible Operators 330

12 / DIFFERENTIAL CALCULUS OF FUNCTIONS FROM $\mathbb{R}^n$ TO $\mathbb{R}^m$

12. Introduction 335
12.1 The Differential and the Derivative 336
12.2 The Component Functions and Differentiability 340
12.21 Directional Derivatives and the Method of Steepest Descent 343
12.3 Newton's Method 347
12.4 A Form of the Law of the Mean for Vector Functions 350
12.41 The Hessian and Extreme Values 352
12.5 Continuously Differentiable Functions 354
12.6 The Fundamental Inversion Theorem 355
12.7 The Implicit Function Theorem 361
12.8 Differentiation of Scalar Products of Vector Valued Functions of a Vector Variable 366

13 / DOUBLE AND TRIPLE INTEGRALS

13. Preliminary Remarks 376
13.1 Motivations 376
13.2 Definition of a Double Integral 379
13.21 Some Properties of the Double Integral 381
13.22 Inequalities. The Mean-Value Theorem 382
13.23 A Fundamental Theorem 383
13.3 Iterated Integrals. Centroids 384
13.4 Use of Polar Co-ordinates 390
13.5 Applications of Double Integrals 395
13.51 Potentials and Force Fields 401
13.6 Triple Integrals 404
13.7 Applications of Triple Integrals 409
13.8 Cylindrical Co-ordinates 412
13.9 Spherical Co-ordinates 413

14 / CURVES AND SURFACES

14. Introduction 417
14.1 Representations of Curves 417
14.2 Arc Length 418
14.3 The Tangent Vector 421
14.31 Principal normal. Curvature 423
14.32 Binormal. Torsion 425
14.4 Surfaces 428
14.5 Curves on a Surface 433
14.6 Surface Area 437

15 / LINE AND SURFACE INTEGRALS

15. Introduction 445
15.1 Point Functions on Curves and Surfaces 445
15.12 Line Integrals 446
15.13 Vector Functions and Line Integrals. Work 451
15.2 Partial Derivatives at the Boundary of a Region 455
15.3 Green's Theorem in the Plane 457
15.31 Comments on the Proof of Green's Theorem 463
15.32 Transformations of Double Integrals 465
15.4 Exact Differentials 469
15.41 Line Integrals Independent of the Path 474
15.5 Further Discussion of Surface Area 478
15.51 Surface Integrals 480
15.6 The Divergence Theorem 484
15.61 Green's Identities 492
15.62 Transformation of Triple Integrals 494
15.7 Stokes's Theorem 499
15.8 Exact Differentials in Three Variables 505

16 / POINT-SET THEORY

16. Preliminary Remarks 512
16.1 Finite and Infinite Sets 512
16.2 Point Sets on a Line 514
16.3 The Bolzano–Weierstrass Theorem 517
16.31 Convergent Sequences on a Line 518
16.4 Point Sets in Higher Dimensions 520
16.41 Convergent Sequences in Higher Dimensions 521
16.5 Cauchy's Convergence Condition 522
16.6 The Heine–Borel Theorem 523

17 / FUNDAMENTAL THEOREMS ON CONTINUOUS FUNCTIONS

17. Purpose of the Chapter 527
17.1 Continuity and Sequential Limits 527
17.2 The Boundedness Theorem 529
17.3 The Extreme-Value Theorem 529
17.4 Uniform Continuity 529
17.5 Continuity of Sums, Products, and Quotients 532
17.6 Persistence of Sign 532
17.7 The Intermediate-Value Theorem 533

18 / THE THEORY OF INTEGRATION

18. The Nature of the Chapter 535
18.1 The Definition of Integrability 535
18.11 The Integrability of Continuous Functions 539
18.12 Integrable Functions with Discontinuities 540
18.2 The Integral as a Limit of Sums 542
18.21 Duhamel's Principle 545
18.3 Further Discussion of Integrals 548
18.4 The Integral as a Function of the Upper Limit 548
18.41 The Integral of a Derivative 550
18.5 Integrals Depending on a Parameter 551
18.6 Riemann Double Integrals 554
18.61 Double Integrals and Iterated Integrals 557
18.7 Triple Integrals 559
18.8 Improper Integrals 559
18.9 Stieltjes Integrals 560

19 / INFINITE SERIES

19. Definitions and Notation 566
19.1 Taylor's Series 569
19.11 A Series for the Inverse Tangent 572
19.2 Series of Nonnegative Terms 573
19.21 The Integral Test 577
19.22 Ratio Tests 579
19.3 Absolute and Conditional Convergence 581
19.31 Rearrangement of Terms 585
19.32 Alternating Series 587
19.4 Tests for Absolute Convergence 590
19.5 The Binomial Series 597
19.6 Multiplication of Series 600
19.7 Dirichlet's Test 604

20 / UNIFORM CONVERGENCE

20. Functions Defined by Convergent Sequences 610
20.1 The Concept of Uniform Convergence 613
20.2 A Comparison Test for Uniform Convergence 618
20.3 Continuity of the Limit Function 620
20.4 Integration of Sequences and Series 621
20.5 Differentiation of Sequences and Series 624

21 / POWER SERIES

21. General Remarks 627
21.1 The Interval of Convergence 627
21.2 Differentiation of Power Series 632
21.3 Division of Power Series 639
21.4 Abel's Theorem 643
21.5 Inferior and Superior Limits 647
21.6 Real Analytic Functions 650

22 / IMPROPER INTEGRALS

22. Preliminary Remarks 654
22.1 Positive Integrands. Integrals of the First Kind 656
22.11 Integrals of the Second Kind 661
22.12 Integrals of Mixed Type 664
22.2 The Gamma Function 666
22.3 Absolute Convergence 670
22.4 Improper Multiple Integrals. Finite Regions 673
22.41 Improper Multiple Integrals. Infinite Regions 678
22.5 Functions Defined by Improper Integrals 682
22.51 Laplace Transforms 690
22.6 Repeated Improper Integrals 693
22.7 The Beta Function 695
22.8 Stirling's Formula 699

ANSWERS TO SELECTED EXERCISES 709

INDEX 727

ADVANCED CALCULUS

1 / *FUNDAMENTALS OF ELEMENTARY CALCULUS*

1 / INTRODUCTION

A course in advanced calculus must build upon the presumption that students studying the subject have already gained some knowledge of elementary calculus. We shall therefore begin by taking a backward look over those parts of calculus with which the reader of this book should have facility and a measure of understanding. Our object in such a retrospect is not to conduct a systematic review. The purpose is, rather, to establish a common point of view for students whose training in calculus, up to this point, must inevitably reflect a wide variety of practices in teaching, choice of subject matter, and distribution of emphasis between the acquisitions of problem-solving skills and mastery of fundamental theory. As we survey the field of elementary calculus we shall stress the conceptual aspect of the subject: fundamental definitions and processes which underlie all the applications. In a first course in calculus it is often the case that the fundamental notions are introduced through the medium of particular geometrical or physical applications. Thus, to the beginner, the derivative may be typified by, or even identified with, the speed of a moving object, while the integral is thought of as the area under a curve. We now seek to take a more general, or abstract, view. Differentiation and integration are processes which are carried out upon functions. We need to have a clear understanding of the definitions of these processes, quite apart form their applications.

Another aspect of our survey will be our concern with the logical unfolding of the fundamental principles of calculus. Here again we strive to take a more mature point of view. We wish to indicate in what respects it is desirable and necessary to look more deeply into the derivations of rules and proofs of theorems. There are places in elementary calculus, as usually taught to beginners, where the development is necessarily inadequate from the standpoint of logic. In many places the reasoning leans heavily on intuition or on one sort or another of plausibility argument. That this state of affairs persists is partly due to a deliberate placing of emphasis: we make our primary goal the attainment of skill in the manipulative techniques of calculus which lend themselves readily to applications at an elementary level in physics, engineering, and the like. This kind of skill (up to a certain point) can be imparted without paying much attention to questions of logical rigor. But it is also true that there are logical inadequacies in a first course in calculus which cannot be made good entirely within the customary time limits of such a course (two or three semesters), even

where a reasonably heavy emphasis is laid upon "theory." At bottom the subject of calculus rests upon the real number system and the theory of limits. A full appreciation and understanding of this foundation material must come slowly, but the need for such understanding becomes more acute as we progress in learning. In advanced calculus we must make a deeper study of the real number system, of the theory of limits, and of the properties of continuous functions. In this way only can we proceed easily and with confidence to a mastery of many new concepts and processes of higher mathematics.

1.1 / FUNCTIONS

At the very outset we must discuss the mathematical concept of a *function*, for we shall constantly be talking about properties of functions and about processes which are applied to functions. The function concept has been very much generalized since the early development of calculus by Leibniz and Newton. At the present time the word "function" is used broadly to mean any determinate correspondence between two classes of objects.

Example 1. Consider the class of all plane polygons. If to each polygon we make correspond the number which is the perimeter of the polygon (in terms of some fixed unit of length), this correspondence is a function. Here the first class of objects is composed of certain figures, while the members of the second class are positive numbers.

To begin with, let us consider functions which are correspondences between sets of real numbers. Such functions are called *real functions of a real variable.* The first set of numbers is the *domain of definition* or simply the *domain* of the function. The second set, consisting of the values taken on by the function, is called the *range.* Once the domain, which we may call D, has been specified, the function is defined as soon as a definite rule of correspondence has been given, assigning to each number of D some corresponding number in the range. If x is a symbol which may be used to denote any member of D, we call x the *independent variable* of the function. In some situations it is very natural to have more than one number associated with a given value of x and to call such a correspondence a *multiple-valued function.* If each value of x corresponds to just one number in the range we have a *single-valued function*, which is what is properly meant by the term *function.* We usually find it possible to deal with multiple-valued functions by separating them into several (possibly infinitely many) single-valued functions. Hereafter we shall always assume that all functions referred to are single-valued, unless the situation explicitly indicates the contrary.

A function may be defined by an algebraic or trigonometric formula, but it need not be so.

Example 2. If x denotes any real number, let $[x]$ denote the algebraically

largest integer which does not exceed x; e.g.,

$$[-2.3] = -3, [-1] = -1, [0] = 0, [3.5] = 3, [7] = 7, [7.2] = 7.$$

The correspondence between x and $[x]$ defines a function. If we use f to denote this function, then we would say that f is defined by $f(x) = [x]$.

Example 3. Another simple function is defined by associating with x its *absolute value* $|x|$. The definition of $|x|$ is:

$$|x| = x \text{ if } x \geq 0, \qquad |x| = -x \text{ if } x < 0.$$

Thus $|7| = 7$, $|0| = 0$, $|-5| = 5$, $|3 - 10| = 7$. If we think of x as a point on a number scale (the x-axis), then $|x|$ is the numerical distance (always nonnegative) between x and the origin.

The concept and the symbolism of absolute value are quite important. The student will need to get accustomed to reading sentences that contain inequalities and absolute values. Thus, for instance, $|7 - 5| = 2$, $|-16 - (-10)| = 6$, and, in general, $|x_1 - x_2|$ is the distance between points x_1 and x_2 on the x-axis. As another example, $|x - 5| < 2$ means that the distance between x and 5 is less than 2; this is equivalent to saying that x lies between 3 and 7. We can write this in the form $3 < x < 7$. A general statement of the same sort is that $|x - a| < b$ (where $b > 0$) is equivalent to $a - b < x < a + b$.

We regard functions as mathematical entities, and represent them by symbols. The commonly used symbols are the Latin letters f, g, h, F, G, H, and the Greek letters ϕ, ψ, Φ, Ψ, but in principle any symbol may be used. If f is the symbol for a particular function, we use $f(x)$ to represent the number which the function makes correspond to any particular value of the independent variable x; this is called the *value* of the function at x.

Example 4. Let f be the symbol for the function which makes correspond to a positive number the natural logarithm of that number. Then $f(x) = \log_e x$. (We shall normally drop the subscript e and write $\log x$ in place of $\log_e x$.)

There is some ambiguity in the use of functional notation, for $f(x)$ is frequently used as a symbol for the function itself, as well as for the value of the function. Thus, for example, we speak of the "function $\sin x$," "the function $x^2 - 3x + 5$," or "the function $\phi(x)$." There is of course a difference between the function and the value of the function. If the symbol $f(x)$ appears, the context will usually make clear whether reference is being made to the function or to the value of the function. To avoid possible ambiguity we shall cultivate the practice of writing "the function f" instead of "the function $f(x)$." This usage is in accord with prevalent practice in current literature, and the student will do well to become familiar with it.

If y is a symbol for the value of the function f at x, we can write $y = f(x)$. Here y is called the *dependent variable*; we say that y is a function of x. In elementary calculus most of the stress is upon functions which are defined by means of fairly simple formulas connecting the independent variable x and the

dependent variable y. Here, however, we look toward understanding the principles of calculus as they apply to functions which are arbitrary except insofar as they are restricted by specified hypotheses.

We shall in due course have to deal with functions of more than one variable. The general notion of a function is still that of a correspondence. A real function F of two real variables x, y is a correspondence which assigns a number $F(x, y)$ as the value of the function corresponding to the pair of values x, y of the two independent variables. The use of functional notation and the designation of the function by the single letter F require no detailed comment, since the basic ideas are no different from those already explained.

The characteristic feature of calculus is its use of *limiting processes.* Differentiation and integration involve certain notions of passage to a limit. A fuller discussion of ideas about limits is presented later on in this chapter (§§1.6–1.64). Here we wish to touch on only one limit notion, that of the limit of a real function of one real variable. This notion is fundamental in the definition of a derivative.

Suppose f is a function which is defined for all values of x near the fixed value x_0, and possibly, though not necessarily, at x_0 as well. We wish to attach a clear meaning to the statement: $f(x)$ approaches A (or tends to the limit A) as x approaches x_0. The symbolic form of the statement is

$$\lim_{x \to x_0} f(x) = A. \tag{1.1–1}$$

The symbol A is understood to stand for some particular real number. The arrow is used as a symbol for the word "approaches." Sometimes (1.1–1) is expressed in the form $f(x) \to A$ as $x \to x_0$. Here are three typical examples of statements of this kind: (*a*) $x^3 \to 8$ as $x \to 2$, (*b*) $(x-1)^{1/2} \to 3$ as $x \to 10$, (*c*) $\log_{10} x \to 2$ as $x \to 100$.

Definition. *The assertion* (1.1–1) *means that we can insure that the absolute value* $|f(x) - A|$ *is as small as we please merely by requiring that the absolute value* $|x - x_0|$ *be sufficiently small, and different from zero. This verbal statement is expressible in terms of inequalities as follows: Suppose* ϵ *is any positive number. Then there is some positive number* δ *such that*

$$|f(x) - A| < \epsilon \quad \text{if} \quad 0 < |x - x_0| < \delta. \tag{1.1–2}$$

Note that $0 < |x - x_0|$ is the same as $x \neq x_0$. Note also that $|f(x) - A| < \epsilon$ is the same as $A - \epsilon < f(x) < A + \epsilon$, and $|x - x_0| < \delta$ the same as $x_0 - \delta < x < x_0 + \delta$.

We can give a geometrical portrayal of the inequalities (1.1–2). Let the points (x, y) with $y = f(x)$ be located on a rectangular co-ordinate system; also locate the point (x_0, A). For any $\epsilon > 0$ draw the two horizontal lines $y = A \pm \epsilon$. Now (1.1–1) means that, by choosing δ small enough, those points of the graph of $y = f(x)$ which lie between the two vertical lines $x = x_0 \pm \delta$ and not on the line $x = x_0$ will also lie between the horizontal lines $y = A \pm \epsilon$. Fig. 1 shows a

specimen of this situation. The diagram also shows how δ may have to be made smaller as ϵ becomes smaller.

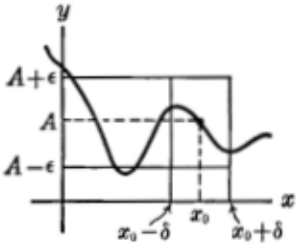


Fig. 1.

It is to be emphasized that (1.1–1) places no restrictions whatever on the value of f at x_0, in case it is defined at that point.

Appreciation of the formal definition of the meaning of (1.1–1) takes time and experience. The formal definition is the basis for exact reasoning on matters involving the limit concept. But it is also quite important to develop an intuitive understanding of the notion of a limit. This may be done by considering a large number of illustrative examples and by observing the way in which the limit concept is used in the development of calculus. One needs to learn by example how a function $f(x)$ may fail to approach a limit as x approaches x_0.

The variable x may approach x_0 from either of two sides. Let us use $x \to x_0+$ to indicate that x approaches x_0 from the right, and $x \to x_0-$ to indicate approach from the left. The conditions for $\lim_{x \to x_0} f(x) = A$ are then that $f(x) \to A$ as $x \to x_0+$ and also $f(x) \to A$ as $x \to x_0-$. *In terms of inequalities the meaning of* $f(x) \to A$ *as* $x \to x_0+$ *is this: to any* $\epsilon > 0$ *corresponds some* $\delta > 0$ *such that* $|f(x) - A| < \epsilon$ *if* $x_0 < x < x_0 + \delta_0$. The meaning of $f(x) \to A$ *as* $x \to x_0-$ may be expressed in a similar way.

Example 5. The limit of $f(x)$ as $x \to x_0$ may fail to exist because:

(a) The limits from right and left exist but are not equal. This is the case with

$$f(x) = 1 + \frac{|x|}{x},$$

where $f(x) \to 2$ as $x \to 0+$ and $f(x) \to 0$ as $x \to 0-$.

(b) The values of $f(x)$ may get larger and larger (tend to infinity) as $x \to x_0$ from one side or the other, or from both sides. This is the case with $f(x) = 1/x$ as $x \to 0$.

(c) The values of $f(x)$ may oscillate infinitely often, approaching no limit. This is the case with $f(x) = \sin(1/x)$ which oscillates infinitely often between -1 and $+1$ as $x \to 0$ from either side.

The graphs of the three foregoing functions are shown in Figs. 2a, 2b, and 2c, respectively.

Example 6. If $f(x) = e^{-1/x^2}$, then $\lim_{x \to 0} f(x) = 0$. To "see" the correctness of this result, one must have clearly in mind the nature of the exponential function. When x is near zero, $-1/x^2$ is large and negative; now e raised to a large negative power is a small positive number. Hence e^{-1/x^2} is nearly 0 when x is nearly 0, and $f(x) \to 0$ as $x \to 0$. This is an example of a rough intuitive argument leading to a conclusion about a certain limit.

It is instructive to see how the intuitive argument is made precise by

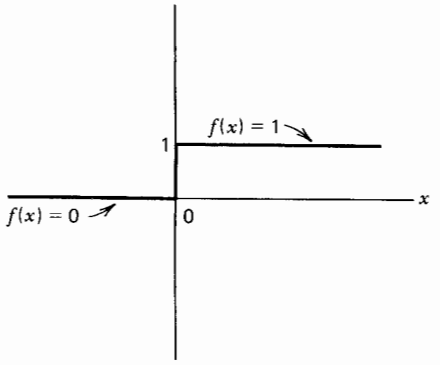


Fig. 2a.

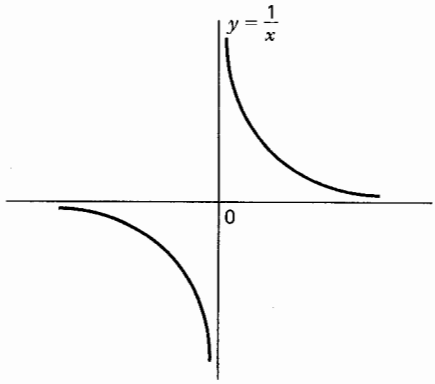


Fig. 2b.

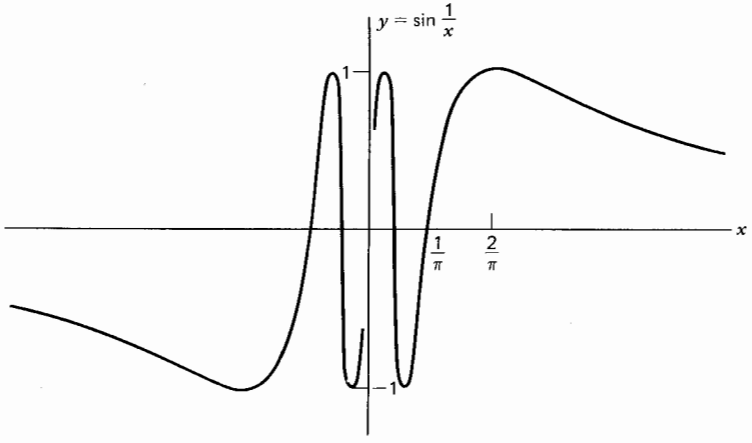


Fig. 2c.

reference to the definition of a limit. The statement $\lim_{x\to 0} e^{-1/x^2} = 0$ means that to each $\epsilon > 0$ corresponds some δ such that

$$|e^{-1/x^2} - 0| < \epsilon \text{ if } 0 < |x - 0| < \delta. \tag{1.1–3}$$

Let us see how we may find a suitable δ when ϵ is given. In doing this we take for granted the properties of the exponential and logarithmic functions.

Since e to any power is positive, $|e^{-1/x^2} - 0| < \epsilon$ is equivalent to $e^{-1/x^2} < \epsilon$. We rewrite this inequality in several successive equivalent forms:

$$\frac{1}{e^{1/x^2}} < \epsilon, \quad \frac{1}{\epsilon} < e^{1/x^2}, \quad \log\left(\frac{1}{\epsilon}\right) < \frac{1}{x^2}.$$

Let us suppose that $\epsilon < 1$, so that $\log(1/\epsilon) > 0$. Then further equivalent forms are

$$0 < x^2 < \frac{1}{\log(1/\epsilon)}, \quad 0 < |x| < \left[\frac{1}{\log(1/\epsilon)}\right]^{1/2}.$$

It now appears that, if $0 < \epsilon < 1$, we can choose

$$\delta = \left[\frac{1}{\log(1/\epsilon)}\right]^{1/2},$$

and then (1.1–3) will hold, as required.

Even in very obvious situations it is worth while to practice finding a δ corresponding to a given ϵ, just to drive home an appreciation of the meaning of the definition of a limit.

Example 7. Given $\epsilon > 0$, find δ so that $|f(x) - 4| < \epsilon$ if $0 < |x - 2| < \delta$, where $f(x) = x^3 - x^2$. This will show that $\lim_{x \to 2}(x^3 - x^2) = 4$. We have

$$x^3 - x^2 - 4 = (x - 2)(x^2 + x + 2).$$

To begin with, let us consider only values of x such that $|x - 2| < 1$, or $1 < x < 3$. For such x we certainly have $4 < x^2 + x + 2 < 14$, and hence

$$|x^3 - x^2 - 4| \leqq 14\,|x - 2|.$$

Now we see that $|x^3 - x^2 - 4| < \epsilon$ provided $14\,|x - 2| < \epsilon$, or $|x - 2| < \epsilon/14$. Hence we choose for δ any positive number such that both $\delta \leqq 1$ and $\delta \leqq \epsilon/14$. This choice meets the requirements.

Reasoning with limits is facilitated by various simple theorems. Among the most important such theorems are the following rules, which we state here informally:

Suppose that

$$\lim_{x \to x_0} f(x) = A \text{ and } \lim_{x \to x_0} g(x) = B; \text{ then}$$

$$\lim_{x \to x_0} [f(x) + g(x)] = A + B, \tag{1.1–4}$$

$$\lim_{x \to x_0} [f(x)g(x)] = AB, \tag{1.1–5}$$

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = \frac{A}{B}, \text{ provided } B \neq 0. \tag{1.1–6}$$

Formal proofs of the validity of these three rules are made in §1.64. Meanwhile we accept them and use them.

Closely related to the limit concept is the concept of *continuity.*

Definition. *Suppose the function f is defined at x_0 and for all values of x near x_0. Then the function is said to be continuous at x_0 provided that*

$$\lim_{x \to x_0} f(x) = f(x_0). \tag{1.1–7}$$

Most of the functions which we deal with in calculus are continuous; points of discontinuity are exceptional, but may occur. A function may fail to be continuous at x_0 either because $f(x)$ does not approach any limit at all as $x \to x_0$, or because it approaches a limit which is different from $f(x_0)$.

Example 8. The function $f(x) = [x]$ (see Example 2) is discontinuous at x_0 if x_0 is an integer, but is continuous at x_0 if x_0 is not an integer.

We observe in this case that $\lim_{x \to 2} f(x)$ does not exist, for when x is near 2, $f(x) = 1$ if $x < 2$ and $f(x) = 2$ if $x > 2$. The situation is similar at other integers.

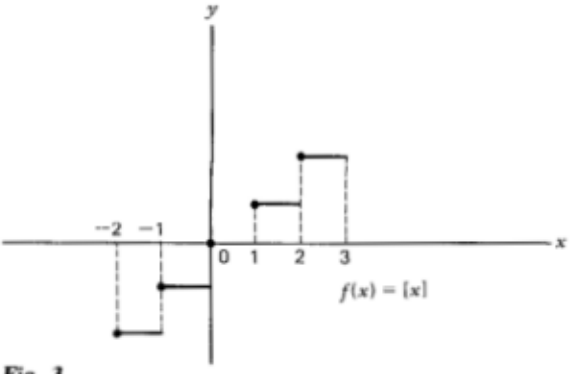


Fig. 3.

The graph of $y = f(x)$ is shown in Fig. 3. At the breaks in the graph when x is an integer n, the value of $f(n)$ is indicated by a heavy dot.

Example 9. Suppose we define a function by

$$f(x) = [x] + [2x - x] - 1.$$

Direct inspection shows the following:

$$f(0) = 1,$$
$$f(x) = 0 \quad \text{if} \quad 0 < x < 1,$$
$$f(1) = 1,$$
$$f(x) = 0 \quad \text{if} \quad 1 < x < 2,$$
$$f(2) = 1.$$

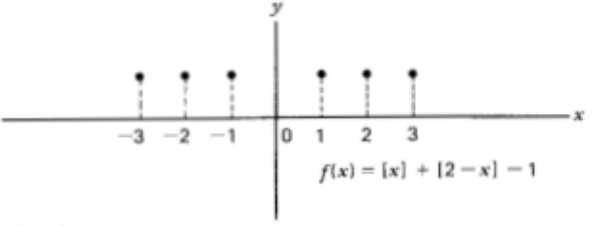


Fig. 4.

Consequently, $\lim_{x\to 1} f(x) = 0$; but $f(1) = 1$, and so f is not continuous at $x = 1$. The graph of $y = f(x)$ is indicated in Fig. 4. From the definition it may be seen that $f(n) = 1$ if n is an integer and $f(x) = 0$ if $n < x < n + 1$.

Example 10. Let us define $f(x) = (\sin x)/x$ if $x \neq 0$. This definition of $f(x)$ has no meaning if $x = 0$, since division by 0 is undefined. However, let us make the additional definition $f(0) = 1$. With this definition, f is continuous at $x = 0$. For, as we learn in elementary calculus,

$$\lim_{x\to 0} \frac{\sin x}{x} = 1. \tag{1.1–8}$$

Since we have defined $f(0) = 1$, (1.1–8) shows that $\lim_{x\to 0} f(x) = f(0)$; therefore f is continuous at $x = 0$, by the definition.

We have based the concept of continuity directly upon the concept of a limit. A condition for continuity of a function may be given directly in terms of inequalities, just as we defined a limit in terms of inequalities. Thus, if f is defined throughout some interval containing x_0 and all points near x_0, f is continuous at x_0 if to each positive ϵ corresponds some positive δ such that

$$|f(x) - f(x_0)| < \epsilon \quad \text{whenever} \quad |x - x_0| < \delta. \tag{1.1–9}$$

This form of the condition for continuity is equivalent to the original definition.

Many common words are used in mathematics in a specialized way. Usually the mathematical meaning of a word has some relation to the common meaning of the word; but mathematical meanings are precise, whereas common meanings are broad or variable. The adjective "continuous" is a word of this kind, with a restrictive and precise mathematical meaning. Experience shows that students tend to read more, in the way of preconceived notions about the meaning of the term, into the word "continuous" than is implied by the definition. In analytic geometry and calculus we become familiar with the graphs of many functions, and there is a tendency to associate the term "continuous function" with the picture of a smooth, unbroken curve. Now it is true that if f is continuous at each point of an interval, the corresponding part of the graph of $y = f(x)$ will be an unbroken curve. But it need not be smooth. Smoothness is related to differentiability; the more derivatives f has, the smoother is its graph. A function may be continuous without having a derivative. In that case the graph of $y = f(x)$ might be so crinkly, so devoid of smoothness, as to make correct visualization of it quite impossible.

EXERCISES

Where the square-bracket notation occurs in these exercises, $[f(x)]$ denotes the algebraically largest integer which is $\leq f(x)$ (see Example 2).

1. Find each of the limits indicated, using algebraic simplification and the rules (1.1–4)–(1.1–6).

(a) $\lim_{x\to 4} \dfrac{x^2-16}{x-4}$;

(b) $\lim_{x\to 1} \dfrac{x^{10}-1}{x-1}$;

(c) $\lim_{x\to 1} \dfrac{x^n-1}{x-1}$ (n a positive integer);

(d) $\lim_{x\to 0} \dfrac{(x+2)^2-4}{x}$;

(e) $\lim_{x\to 0} \dfrac{1}{x}\left(\dfrac{1}{2+x}-\dfrac{1}{2}\right)$;

(f) $\lim_{x\to 0} \dfrac{1}{x}\left(\dfrac{1}{(4+x)^2}-\dfrac{1}{16}\right)$;

(g) $\lim_{x\to 2} \dfrac{x^3+x^2-5x-2}{x^2-4}$;

(h) $\lim_{x\to -1} \dfrac{108(x^2+2x)(x+1)^3}{(x^3+1)^3(x-1)}$.

2. Find each of the following limits, using roughly quantitative arguments based on your knowledge of the various functions involved (somewhat as in the first paragraph of the discussion of Example 6).

(a) $\lim_{x\to 0} 2^{\log x^2}$;

(b) $\lim_{x\to 0} \cos(e^{-1/x^2})$;

(c) $\lim_{x\to 0} \dfrac{1+\cos x}{1+(\log x^2)^2}$;

(d) $\lim_{x\to \pi} \tan^{-1}\left(\tan^2 \dfrac{x}{2}\right)$, where the inverse tangent has its principal value, i.e., $v = \tan^{-1} u$ means $u = \tan v$ and $-\dfrac{\pi}{2} < v < \dfrac{\pi}{2}$;

(e) $\lim_{x\to 0} \log\left(\dfrac{\sin x}{x}\right)$.

3. Draw the graphs for each of the following functions and then answer the questions:

(a) $f(x) = 2x - 1$ if $x \leqq 1$, $f(x) = 6 - 5x$ if $x > 1$. Is f continuous at $x = 1$?

(b) $f(x) = (x^2/2) - 2$ if $0 < x < 2$, $f(x) = 2 - (8/x^2)$ if $2 < x$. Does $\lim_{x\to 2} f(x)$ exist? How should $f(2)$ be defined to make f continuous at $x = 2$?

(c) $f(x) = [1 - x^2]$. Consider only $-1 \leqq x \leqq 1$. Does $\lim_{x\to 0} f(x)$ exist? Is f continuous at $x = 0$?

(d) $f(x) = (x - 1)[x]$. Consider only $0 \leqq x \leqq 2$. Is f continuous at $x = 1$?

(e) $f(x) = x/|x|$ (undefined if $x = 0$). Does $\lim_{x\to 0} f(x)$ exist?

4. In each of the following cases f is defined by the given formula only if $x \neq 0$. How should $f(0)$ be defined to make f continuous at $x = 0$?

(a) $f(x) = \dfrac{\sin 2x}{x}$; (b) $f(x) = \dfrac{\tan x}{x}$; (c) $f(x) = \dfrac{(x+2)^3 - 8}{x}$; (d) $f(x) = 10^{-1/x^2}$.

Exercises 5–7 form a natural unit.

5. If $f(x) = cx$, where c is a constant, show that $\lim_{x\to x_0} f(x) = f(x_0)$ by applying the definition of a limit as expressed by (1.1–2). If $c \neq 0$ what can you take δ to be in terms of ϵ and c?

6. If c is constant and n is a positive integer, show that $\lim_{x\to x_0} cx^n = cx_0^n$. Use Exercise 5, mathematical induction, and (1.1–5).

7. By a *polynomial* in x we mean a function defined by an expression

$$P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_n$$

where the coefficients $a_0, a_1, \ldots, a_n$ are constants, and n is an integer $\geqq 0$. If $n = 0$, $P(x)$ is constant in value, and the degree of $P(x)$ is said to be zero. If $n \geqq 1$ and $a_0 \neq 0$, we say that the degree of polynomial is n. Prove that $P(x)$ is continuous at every point x_0. Use the result of Exercise 6. What other result about limits do you use?

8. By a rational function of x we mean a function defined by an expression

$$R(x) = \frac{p(x)}{P(x)},$$

where $p(x)$ and $P(x)$ are polynomials (see Exercise 7). The function is defined except when $P(x) = 0$. Show that it is continuous at x_0 if it is defined there. Use definition (1.1–7) and state exactly what appeal you make to facts about limits stated in the text or established in previous exercises.

9. If $f(x) = \sin\frac{1}{x}$, **(a)** find $f\left(\frac{1}{n\pi}\right)$, $n = 1, 2, \ldots$; **(b)** find $f\left(\frac{2}{n\pi}\right)$, $n = 1, 5, 9, \ldots$; **(c)** find $f\left(\frac{2}{n\pi}\right)$, $n = 3, 7, 11, \ldots$. **(d)** How does the derivative $f'(x)$ behave as $x \to 0$?

10. **(a)** How does $2^{-1/x}$ behave as $x \to 0+$? **(b)** as $x \to 0-$? **(c)** What can you say about $\lim_{x\to 0} \frac{1}{1+2^{-1/x}}$?

11. Graph each of the following functions: **(a)** $|x|$, **(b)** $|x-1|$, **(c)** $|x+2|$, **(d)** $|x^3|$, **(e)** $|1-x^2|$. Do any of these functions have any points of discontinuity?

12. Graph the function $x - [x]$ and discuss its discontinuities.

13. Which of the following functions is continuous at $x = 0$? **(a)** $[x^2+2]$, **(b)** $[4-x^2]$, **(c)** $[x^2-1]$. Graph each function when $-1 \leqq x \leqq 1$.

14. If $f(x) = \frac{[\frac{3}{2}+x]-[\frac{3}{2}]}{x}$, **(a)** find $f(1), f(-1), f(\frac{1}{2}), f(-\frac{1}{2}), f(\frac{1}{4}), f(-\frac{1}{4})$. **(b)** Without using the square brackets, write expressions for $f(x)$ if $0 < x < \frac{1}{2}$ and if $-\frac{1}{2} < x < 0$. **(c)** What is $\lim_{x\to 0} f(x)$?

15. If $f(x) = \frac{|x^2-1|}{x-1}$, **(a)** find $f(\frac{1}{2}), f(\frac{3}{4}), f(\frac{7}{8}), f(\frac{3}{2}), f(\frac{5}{4}), f(\frac{9}{8})$. **(b)** Express $f(x)$ without absolute values if $x > 1$; if $0 < x < 1$. **(c)** What can you say about $\lim_{x\to 1} f(x)$?

16. If $f(x) = \frac{|2+x|-|x|-2}{x}$, **(a)** find $f(1), f(-1), f(-2), f(2)$. **(b)** Write an expression for $f(x)$ without absolute values if $0 < x$; if $-2 < x < 0$. **(c)** What can you say about $\lim_{x\to 0} f(x)$?

17. If $f(x) = [7x^2 - 14]$, **(a)** find $f(0), f(1), f(2), f(\frac{1}{2}), f(\frac{3}{2}), f(\frac{5}{3})$. **(b)** Is f continuous at $x = 0$? **(c)** Is it continuous at $x = 1$? **(d)** at $x = \sqrt{2}$?

18. If $f(x) = [\sin x]$, **(a)** find $f(0), f(\pi/2), f(-\pi/2), f(\pi/4), f(-\pi/4)$. **(b)** Does $\lim_{x\to 0} f(x)$ exist? **(c)** What does $f(x)$ approach as $x \to (\pi/2)-$? **(d)** as $x \to 0-$?

19. Prove that $\lim_{x\to 1}(x^2+2x) = 3$ by finding δ in terms of a given positive ϵ so that $|x^2+2x-3| < \epsilon$ if $|x-1| < \delta$.

20. Show that $\left|\frac{1}{x^2+16} - \frac{1}{25}\right| \leqq \frac{7}{500}|x-3|$ if $2 < x < 4$. Hence, for any $\epsilon > 0$, find δ so that $\left|\frac{1}{x^2+16} - \frac{1}{25}\right| < \epsilon$ if $|x-3| < \delta$, thus proving directly that $\lim_{x\to 3} \frac{1}{x^2+16} = \frac{1}{25}$.

21. Show that $|(1+x^3)-1| \leqq 7|x|$ if $-1 < x < 1$. Then prove by the definition (1.1–2) that $\lim_{x\to 0}(1+x)^3 = 1$ (i.e., find a suitable δ for any given positive ϵ).

22. Show that $\left|\frac{1}{x}\left(\frac{1}{2+x}-\frac{1}{2}\right)+\frac{1}{4}\right| < \epsilon$ if $0<|x|<\delta$, where δ is the smaller of the numbers $1, 4\epsilon$ (ϵ being positive). Translate this situation into a statement of the form $\lim_{x\to x_0} f(x) = A$, specifying what you take for f, x_0, and A.

23. Show that, if $0<\epsilon<1$, $10^{-1/x} < \epsilon$ when $0 < x < \left(\log_{10}\frac{1}{\epsilon}\right)^{-1}$. What is the corresponding statement about a limit?

24. Does $\lim_{x\to 0}\frac{2^{1/x}+3}{2^{1/x}+1}$ exist?

25. Suppose that a function f is defined by setting $f(x) = 1/n$ if $x = 1/2^n$, where $n = 1, 2, 3, \ldots$, and $f(x) = 0$ for all other values of x. Is f continuous at $x = 0$?

1.11 / DERIVATIVES

Elementary calculus deals with the processes of differentiation and integration, the techniques of these processes as they pertain to various common functions, and the applications of the processes to problems of geometry, physics, and other sciences. Let us examine the concept of the derivative.

Consider a function f, defined for values of the variable x in the interval $a < x < b$. Let x_0 be any fixed point of the interval, and consider the ratio

$$\frac{f(x)-f(x_0)}{x-x_0}, \tag{1.11–1}$$

where $x \neq x_0$ and x is a variable point of the interval. The ratio (1.11–1) is called a difference quotient.

Definition. *If the difference quotient* (1.11–1) *approaches a limit as x approaches* x_0, *the limit is called the derivative of f at* $x = x_0$, *and is denoted by* $f'(x_0)$. *Thus, by definition,*

$$f'(x_0) = \lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0}, \tag{1.11–2}$$

provided the limit exists.

Quite likely the student is familiar with another notation in connection with this definition. Sometimes we write $x = x_0 + \Delta x$, and then (1.11–2) takes the form

$$f'(x_0) = \lim_{\Delta x\to 0}\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}. \tag{1.11–3}$$

Here the symbol Δx denotes an independent variable. For many algebraic calculations it is convenient to use h in place of Δx. Also, we may drop the zero subscript; we then have the definition

$$f'(x) = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}, \tag{1.11–4}$$

provided the limit exists.

In addition to the notation $f'(x)$ for the derivative we frequently use the notation $\frac{d}{dx} f(x)$.

Example 1. Using form (1.11–2), calculate $f'(x_0)$ if $f(x) = x^2$. Here

$$f(x) - f(x_0) = x^2 - x_0^2 = (x - x_0)(x + x_0);$$

$$f'(x_0) = \lim_{x \to x_0} (x + x_0) = 2x_0.$$

Example 2. Using form (1.11–4), calculate $f'(x)$ if $f(x) = 1/x$. Here

$$f(x + h) - f(x) = \frac{1}{x + h} - \frac{1}{x} = \frac{-h}{(x + h)x};$$

$$f'(x) = \lim_{h \to 0} \frac{-1}{(x + h)x} = -\frac{1}{x^2}.$$

Definition. *A function which has a derivative at a certain point is said to be differentiable at that point.*

In the definiton (1.11–2) we were assuming that f was defined in an interval extending some distance on each side of the point x_0. It is understood that x may approach x_0 from either side, and that the limit of the difference quotient is the same when x approaches x_0 from the left as when the approach is from the right.

It is useful to define *one-sided* derivatives. Using the notation for limits from the right and left, respectively, as explained just prior to Example 5 in §1.1, we define the right-hand derivative $f'_+(x_0)$ and the left-hand derivative $f'_-(x_0)$ as follows:

$$f'_+(x_0) = \lim_{x \to x_0+} \frac{f(x) - f(x_0)}{x - x_0}, \tag{1.11–5}$$

$$f'_-(x_0) = \lim_{x \to x_0-} \frac{f(x) - f(x_0)}{x - x_0}, \tag{1.11–6}$$

provided the limits exist.

If in discussing a function we confine our attention wholly to an interval $a \leqq x \leqq b$, then we shall understand that $f'(a)$ means $f'_+(a)$, and that $f'(b)$ means $f'_-(b)$. If $a < x_0 < b$, however, and if the function is differentiable at x_0, then we must have $f'_+(x_0) = f'_-(x_0)$. The derivative $f'(x_0)$ is then the common value of the two one-sided derivatives. For an example of a case in which the two one-sided derivatives exist but are unequal, see Exercise 12.

Example 3. Let $f(x) = \sqrt{2(1 - \cos 2x)}$. Show that this function is not differentiable at the points $x = 0, \pm\pi, \pm 2\pi, \ldots$ and find the one-sided derivatives at these points.

We recall the trigonometric identity

$$1 - \cos 2x = 2 \sin^2 x.$$

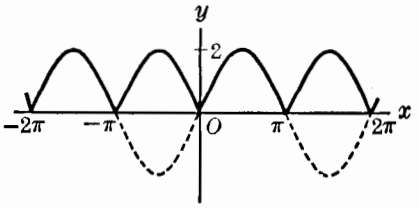


Fig. 5.

Thus*

$$f(x) = \sqrt{4 \sin^2 x} = 2|\sin x|.$$

This means that $f(x) = 2 \sin x$ when $\sin x \geqq 0$; but $f(x) = -2 \sin x$ when $\sin x < 0$. The graph of $f(x)$ is shown in Fig. 5. The dotted portions represent the function $2 \sin x$ when $\sin x < 0$. From the symmetry of the figure we see that the situation at all the points $x = n\pi$ ($n = 0, \pm 1, \ldots$) is the same. The right-hand derivative at each of these points is 2, and the left-hand derivative is -2:

$$f'_+(0) = \lim_{x \to 0+} \frac{2 \sin x - 0}{x - 0} = 2,$$

$$f'_-(0) = \lim_{x \to 0-} \frac{-2 \sin x - 0}{x - 0} = -2.$$

We take it for granted that readers of this book are acquainted with the interpretation of the derivative $f'(x)$ as the slope of the curve $y = f(x)$ when we employ a graphical representation in rectangular co-ordinates with equal scales on the two axes (see Fig. 6). To say that f is differentiable at a point means geometrically that the curve $y = f(x)$ has at that point a unique tangent line which is not parallel to the y-axis.

We also take it for granted that the student is familiar with the interpretation of the derivative as an instantaneous rate of change. Without going into detail we emphasize the fact that the concepts of velocity, acceleration, and all kinds of instantaneous rates of change find their precise mathematical formulation in terms of the notion of the derivative of a function.

Students beginning this book are expected to know the general rules of differentiation, including the rules for dealing with sums, products, and quotients.

*In this book we adhere to the standard convention that if $A \geqq 0$, $\sqrt{A}$ *means the nonnegative square root of* A. According to this convention $\sqrt{a^2} = a$ if $a \geqq 0$, but $\sqrt{a^2} = -a$ if $a < 0$. Both cases are covered by the formula $\sqrt{a^2} = |a|$. Finally, $A^{1/2}$ and $\sqrt{A}$ are merely different notations for the same thing.

It will be convenient to draw up a list of differentiation formulas

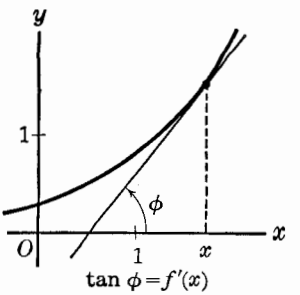


Fig. 6.

GENERAL RULES

$$\frac{d}{dx}[f(x)+g(x)] = f'(x)+g'(x). \tag{1.11–7}$$

$$\frac{d}{dx}[cf(x)] = cf'(x) \ (c \text{ constant}). \tag{1.11–8}$$

$$\frac{d}{dx}[f(x)g(x)] = f(x)g'(x)+f'(x)g(x). \tag{1.11–9}$$

$$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)f'(x)-f(x)g'(x)}{[g(x)]^2}. \tag{1.11–10}$$

Each of these rules is in fact a theorem. We understand that the functions $f(x)$, $g(x)$ are defined on some interval $a < x < b$. Rule (1.11–7), when stated more fully as a theorem, may be expressed as follows: *If $f(x)$ and $g(x)$ are differentiable for a particular value of x, then their sum is also differentiable for this value of x, and* (1.11–7) *holds.* The student should amplify each of the rules (1.11–8)–(1.11–10) into a formally stated theorem in the same manner. What special provision must be made in connection with (1.11–10)?

We mentioned at the end of § 1.1 that a function may be continuous and yet not differentiable. For instance, the function of Example 3 is continuous for all values of x, but it is not differentiable at the points $n\pi$, $n = 0, \pm 1, \pm 2, \ldots$. However, differentiability *does* imply continuity as the following theorem shows.

THEOREM 1. *If f is differentiable at x_0, it is continuous there.*

Proof. When $x \neq x_0$ we can write

$$f(x) = \frac{f(x)-f(x_0)}{x-x_0}(x-x_0)+f(x_0).$$

Then, by (1.1–4) and (1.1–5),

$$\lim_{x\to x_0} f(x) = f'(x_0)\cdot 0 + f(x_0) = f(x_0).$$

This completes the proof.

The student must already be accustomed to using the rule for differentiating a composite function (sometimes called the chain rule). By a *composite* function we mean a function formed by substitution of one function in place of the independent variable in another function:

$$F(t) = f[g(t)]. \tag{1.11–11}$$

To fix the ideas precisely, suppose that g is defined when $\alpha < t < \beta$, and that the values of the function satisfy the inequality $a < g(t) < b$. Suppose that f is defined when $a < x < b$. Then, replacing x by $g(t)$, we obtain the composite function F defined by (1.11–11).

THEOREM II. *Suppose g is differentiable at a point t_0 of the interval $\alpha < t < \beta$. Let $x_0 = g(t_0)$, and suppose that f is differentiable at x_0. Then the composite function F is differentiable at t_0, and*

$$F'(t_0) = f'(x_0)g'(t_0). \tag{1.11–12}$$

In elementary calculus this theorem is often expressed symbolically in a different way, by writing

$$y = f(x), x = g(t).$$

Then

$$\frac{dy}{dt} = \frac{dy}{dx} \cdot \frac{dx}{dt}.$$

Example 4. Suppose $f(x) = x^{17}$, $g(t) = t - t^2$. Then

$$F(t) = (t - t^2)^{17} \quad \text{and} \quad F'(t) = 17(t - t^2)^{16}(1 - 2t).$$

We accept Theorem II as known from elementary calculus. The proof is a somewhat delicate matter, however, and the student who wishes to study the proof will find a discussion of it in Exercise 26 at the end of this section.

Throughout calculus there are two aspects of the development of the subject. On the one hand we formulate concepts and rules applicable to *arbitrary* functions having certain properties. Theorems I and II are of this type. On the other hand there are the *particular* functions which we deal with as illustrations and in all practical applications, e.g.,

$$(1 - x^2)^{1/2}, \sin 2x, \tan^{-1} x, \log x, e^{-x^2/2},$$

and many others. We assume that the student knows the formulas for differentiation of the standard elementary functions, and in general we shall regard all such functions as being available for illustrative purposes.

In order to illustrate the possibility of various kinds of situations which do not ordinarily arise with the standard elementary functions, we sometimes resort to the contrivance of functions specifically defined so as to exhibit some peculiarity. Such specially contrived functions serve to help the student appreciate the generality of the concept of a function. They also teach him to be wary of tacitly assuming more than is implied in a given definition or hypothesis.

Example 5. Let a function be defined as follows:

$$f(x) = x^2 \sin \frac{1}{x} \text{ if } x \neq 0,$$

$$f(0) = 0.$$

We shall see that this function is differentiable for all values of x, but that the calculation of its derivative at $x = 0$ requires special attention.

We note first of all that the formula by which $f(x)$ is defined when $x \neq 0$ cannot be used when $x = 0$, since $1/x$ is then undefined. Hence the assignment of the value of $f(0)$ may be made as we choose. The value $f(0) = 0$ is chosen because this makes the function continuous at $x = 0$; that is,

$$\lim_{x \to 0} x^2 \sin \frac{1}{x} = 0. \tag{1.11–13}$$

The correctness of this result is seen from the inequality

$$\left| x^2 \sin \frac{1}{x} \right| \leqq x^2, \tag{1.11–14}$$

which holds since the value of the sine function never exceeds unity.

To find the derivative, we follow standard procedures in writing

$$f'(x) = x^2 \cos \frac{1}{x} \cdot \left(\frac{-1}{x^2} \right) + 2x \sin \frac{1}{x}$$

$$f'(x) = -\cos \frac{1}{x} + 2x \sin \frac{1}{x}. \tag{1.11–15}$$

This result is correct when $x \neq 0$. If $x = 0$, however, the foregoing procedure for finding $f'(x)$ is not valid, for it is based on the rule for the derivative of the product of two functions, namely x^2 and $\sin(1/x)$; the second of these functions is not defined at $x = 0$, and cannot be defined there so as to be differentiable.

As yet, then, we do not know whether $f(x)$ is differentiable at $x = 0$. Now, by definition,

$$f'(0) = \lim_{x \to 0} \frac{f(x) - f(0)}{x - 0},$$

provided the limit exists. Since $f(0) = 0$,

$$\frac{f(x) - f(0)}{x - 0} = \frac{f(x)}{x} = x \sin \frac{1}{x},$$

and we see that

$$f'(0) = \lim_{x \to 0} x \sin \frac{1}{x} = 0. \tag{1.11–16}$$

It is worth pointing out that $f'(x)$ is not continuous at $x = 0$; for from (1.11–15) we see that as $x \to 0$, $f'(x)$ approaches no limit but oscillates infinitely often from -1 to $+1$.

A graph is helpful in visualizing the nature of the function f. The student should construct such a graph, using the method of multiplication of ordinates. The curve $v = f(x)$ oscillates between the curves $y = x^2$, $y = -x^2$, crossing the axes at the points $\pm \frac{1}{\pi}, \pm \frac{1}{2\pi}, \pm \frac{1}{3\pi}, \ldots$.

In concluding this section we point out a certain principle of reasoning about limits which we used in arriving at (1.11–13) and (1.11–16). It is the following:

If two functions, F, G satisfy an inequality of the form

$$A \leqq F(x) \leqq G(x), \tag{1.11–17}$$

where A is some fixed number, and if $\lim_{x\to x_0} G(x) = A$, *then* $\lim_{x\to x_0} f(x) = A$ *also.*

For example, in applying this principle to arrive at (1.11–16) we put $F(x) = |x \sin(1/x)|$, $G(x) = |x|$, $A = 0$, $x_0 = 0$. The principle just stated is a special case of Theorem XII, §1.61. Its truth is an immediate consequence of the definition of a limit.

EXERCISES

1. (a) If $f(x) = x^n$, where n is a positive integer, compute $f'(x_0)$, using (1.11–2) and a factorization theorem. (b) Compute $f'(x)$, using (1.11–4) and the binomial theorem.

2. Suppose p and q are positive integers without a common factor. Let $f(x) = x^{p/q}$. Suppose x, x_0 are positive, and write $u = x^{1/q}$, $u_0 = x_0^{1/q}$. Verify that

$$\frac{f(x) - f(x_0)}{x - x_0} = \frac{u^p - u_0^p}{u^q - u_0^q}.$$

Proceed from here to show directly that $f'(x_0) = nx_0^{n-1}$, where $n = p/q$.

3. Let $f(x) = x^c$, where $x > 0$ and c is irrational. Since c cannot be expressed as a ratio of integers, the method of Exercise 2 is not available for calculating $f'(x)$. However, assuming as known the differentiability properties of the exponential and logarithmic functions, show that the formula $f'(x) = cx^{c-1}$ may be deduced from the fact that $f(x) = e^{c \log x}$.

4. Let $f(x) = \sin x$, $g(x) = \cos x$. Show that finding $f'(0)$ is the same as finding $\lim_{x\to 0} \frac{\sin x}{x}$, and that finding $g'(0)$ is the same as finding $\lim_{x\to 0} \frac{\cos x - 1}{x}$. Taking for granted that these limits have values 1, 0 respectively, deduce the formula $f'(x) = \cos x$, using (1.11–4) and the expansion formula for $\sin(x + h)$.

5. Show that the formula $g'(x) = -\sin x$ may be derived from the relations $f'(x) = \cos x$, $\cos x = \sin\left(\frac{\pi}{2} - x\right)$. ($f$ and g are defined in Exercise 4). What theorem do you use?

6. Let $f(x) = \log x$. Using only the definition of the derivative and standard properties of the logarithm function, explain why $f'(1) = \lim_{h\to 0} \log(1 + h)^{1/h}$.

7. If $f(x) = \log x$, show that

$$\frac{f(x + h) - f(x)}{h} = \frac{1}{x} \frac{f(1 + t) - f(1)}{t},$$

where $t = h/x$. Hence explain why f is differentiable at x if it is differentiable at 1, and show that $f'(x) = \frac{f'(1)}{x}$.

8. Let $f(x) = e^x$. Explain why $f'(0) = \lim_{x\to 0} \frac{e^x - 1}{x}$. Show that, if $f'(0) = 1$ is known, one can deduce $f'(x) = e^x$ with the aid of the laws of exponents. Start from (1.11–4).

9. The radius of a sphere is being increased at a variable rate. This rate is 2

centimeters per second when the radius is 5 centimeters. Find the rate of change of the volume of the sphere, in the cubic centimeters per second, at this particular moment.

10. A rocket is being launched straight upward from the earth. It burns liquid fuel at a variable rate, the rate being N gallons per mile when the rocket is 10 miles high. If the speed of the rocket at this time is 1000 miles per hour, what is the instantaneous rate of fuel consumption in gallons per hour? Let x be the altitude of the rocket t hours after launching. Suppose the rate of fuel consumption is $kx^{-1/2}$ gallons per mile and $3k(ct)^{1/2}$ gallons per hour where c and k are positive constants. Find a formula for the rate of rise of the rocket, and deduce the formula connecting x and t.

11. If f is the function of Example 5 and $F(t) = f(t^2 - 1)$, find $F'(1)$.

12. Show that $f(x) = |x|$ is not differentiable at $x = 0$. Is it continuous? Find $f'_+(0)$ and $f'_-(0)$.

13. Let $f(x) = e^{-|x|}$. **(a)** Graph this function. **(b)** Is it continuous at $x = 0$? **(c)** Is it differentiable there?

14. Let $f(x) = x|x|$. **(a)** Graph the function. **(b)** Find $f'(x)$ if $x > 0$; if $x = 0$; if $x < 0$. **(c)** Is the derivative f' differentiable at $x = 0$?

15. **(a)** If $f(x) = [x]$, compute $f'(\frac{3}{2})$ by (1.11–4). **(b)** How does Theorem I show that f is not differentiable at $x = 2$? **(c)** What is the value of $f'_+(2)$? **(d)** How does $\dfrac{f(2+h) - f(2)}{h}$ behave as $h \to 0-$?

16. Discuss the continuity and differentiability at $x = 0$ of f, where $f(x) = x \sin(1/x)$ when $x \neq 0$, $f(0) = 0$.

17. Show that the function defined as $f(x) = x^3 \sin(1/x)$ if $x \neq 0$, $f(0) = 0$, has a derivative for all values of x, and that f' is continuous at $x = 0$ but not differentiable there.

18. **(a)** For what values of the exponent n (an integer) will $f'(x)$ exist at $x = 0$ if $f(x) = x^n \sin(1/x^2)$ when $x \neq 0$, and $f(0) = 0$? **(b)** For what values of n will f' be continuous at $x = 0$? **(c)** For what values of n will f' be differentiable at $x = 0$?

19. Let $f(x) = \dfrac{x}{1 + e^{1/x}}$ if $x \neq 0$, $f(0) = 0$. Does $f'(0)$ exist? Sketch the graph near $x = 0$, showing the directions from which a point approaches the origin along the curve.

20. Discuss the differentiability at $x = 1$ of $f(x) = (x - 1)[x]$. Draw the graph when $0 \leqq x \leqq 2$.

21. If $f(x) = [x] + (x - [x])^{1/2}$, sketch the graph when $0 \leqq x \leqq 3$. What can you say about continuity and differentiability of f at $x = 1$ and $x = 2$? Write a formula for $f(x)$ without square brackets when $0 < x < 1$, and use it to find $f'(\frac{1}{2})$.

22. Let f be a function which is defined for all x, with the properties (i) $f(a + b) = f(a)f(b)$, (ii) $f(0) = 1$, (iii) f is differentiable at $x = 0$. Show that f is differentiable for all values of x, and that $f'(x) = f'(0)f(x)$.

23. Let functions f and g be defined for all x and possess the following properties: (i) $f(x + y) = f(x)g(y) + f(y)g(x)$, (ii) f and g are differentiable at $x = 0$, with $f(0) = 0$, $f'(0) = 1$, $g(0) = 1$, $g'(0) = 0$. Show that f is differentiable for all values of x, with $f'(x) = g(x)$. If it is also known that $g(x + y) = g(x)g(y) - f(x)f(y)$, show that g is differentiable for all values of x, with $g'(x) = -f(x)$.

24. Suppose $f\left(\dfrac{1}{2^n}\right) = \dfrac{1}{2^{2n}}$, $n = 1, 2, \ldots$, and $f(x) = 0$ for all other values of x. Is f differentiable at $x = 0$? What is the situation if instead we define $f\left(\dfrac{1}{2^n}\right) = \dfrac{1}{2^{n+1}}$?

25. Construct proofs for rules (1.11–7)–(1.11–10), using (1.11–4).

26. Given f, g, and F as in (1.11–11) and Theorem II, consider any nonzero value of Δt so small that $\alpha < t_0 + \Delta t < \beta$, and define

$$\Delta x = g(t_0 + \Delta t) - g(t_0), \quad \Delta y = f(x_0 + \Delta x) - f(x_0).$$

Then define

$$\epsilon = \frac{\Delta y}{\Delta x} - f'(x_0) \quad \text{if} \quad \Delta x \neq 0 \quad \text{and} \quad \epsilon = 0 \quad \text{if} \quad \Delta x = 0.$$

Show that

$$\frac{F(t_0 + \Delta t) - F(t_0)}{\Delta t} = \frac{\Delta y}{\Delta t} = [f'(x_0) + \epsilon]\frac{\Delta x}{\Delta t}.$$

Explain carefully why Δx and ϵ approach 0 as $\Delta t \to 0$, and then explain how you can carry through the proof of Theorem II.

1.12 / MAXIMA AND MINIMA

One of the important things about the derivative is that it helps us to locate the relative maxima and minima of a function. Let us formulate exactly what can be said about such things.

It is necessary to say what we mean by an *open interval* of the x-axis. If a and b are numbers such that $a < b$, all numbers x such that $a < x < b$ form what is called the open interval from a to b, or more briefly, the open interval (a, b). The end points a, b do not belong to the open interval. By contrast, the set of all numbers x such that $a \leqq x \leqq b$ is called a *closed interval*; here the end points belong to the interval. We shall denote closed intervals by the use of square brackets: $[a, b]$; for open intervals we shall use ordinary parentheses: (a, b). By a *neighborhood* of x_0 we mean an open interval containing x_0.

Definition. *Let f be a function which is defined on an open interval (a, b), and let x_0 be a point of the interval. We say that f has a relative maximum at x_0 if there is some neighborhood of x_0 (say (a_1, b_1), where $a_1 < x_0 < b_1$) contained in (a, b) and containing x_0 such that $f(x) \leqq f(x_0)$ if $a_1 < x < b_1$. This means that $f(x_0)$ is at least as large (algebraically) as $f(x)$ at all points x for some distance on either side of x_0 (see Fig. 7).*

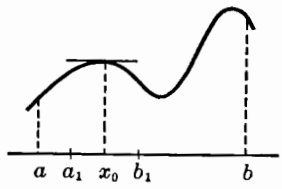


Fig. 7. *A relative maximum.*

A similar definition is made for a relative minimum, the inequality being reversed: $f(x) \geqq f(x_0)$ when x is near x_0.

Both a relative maximum and a relative minimum are covered by the term "a relative *extremum*."

THEOREM III. *Suppose that f has a relative extremum at the point x_0 of the open interval (a, b) and suppose that f is differentiable at x_0. Then $f'(x_0) = 0$.*

Proof. For definiteness assume that f has a relative maximum at x_0. Consider the one-sided derivatives at x_0, and bear in mind that $f(x) \leqq f(x_0)$ when x is sufficiently near x_0. Then $f(x) - f(x_0) \leqq 0$: accordingly

$$\frac{f(x) - f(x_0)}{x - x_0} \leqq 0 \text{ when } x > x_0$$

and so (see (1.11–5))

$$f'_+(x_0) \leqq 0.$$

Likewise

$$\frac{f(x) - f(x_0)}{x - x_0} \geqq 0 \text{ when } x < x_0,$$

so that $f'_-(x_0) \geqq 0$ (see (1.11–6)). But, since f is differentiable at x_0, we have $f'(x_0) = f'_+(x_0) = f'_-(x_0)$. Therefore $f'(x_0) = 0$, for it is neither positive nor negative. The proof for the case of a relative minimum is entirely similar.

It may happen, of course, that a function has a relative extremum at a point x_0, but is not differentiable there. This happens with $f(x) = 1 - x^{2/3}$ at $x = 0$, and with $f(x) = |x - 1|$ at $x = 1$. See Fig. 8, also.

The proof of Theorem III rests on the following reasoning about limits: *If a variable quantity is* $\leqq 0$ *and approaches the limit* A, *then* $A \leqq 0$; *likewise, if the variable quantity is* $\geqq 0$ *and approaches the limit* B, *then* $B \geqq 0$. This principle is considered further in §1.61 (Theorem X).

In defining a relative extremum we compared the value $f(x_0)$ with values $f(x)$ at points x *on both sides of* x_0. Theorem III applies only when x_0 is an *interior* point of the interval on which we are examining the values of f. By contrast, let us consider a function which is defined only when $a \leqq x \leqq b$, and suppose that $f(a)$ is greater than $f(x)$ when x is near a on the right. Then, if the right-hand derivative $f'_+(a)$ exists, we can infer that $f'_+(a) \leqq 0$, but not that $f'_+(a) = 0$ (see Fig. 9). We leave it for the student to draw the appropriate conclusion about $f'_-(b)$ if $f(x) \leqq f(b)$ when x is near b on the left.

Of course, the mere fact that $f'(x_0) = 0$ does not guarantee that f has a relative extremum at x_0. This is illustrated by $f(x) = x^3$ at $x = 0$, where the graph has a horizontal tangent but the function has neither a relative maximum nor a

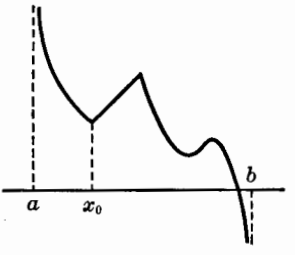


Fig. 8. A relative minimum.



Fig. 9.

relative minimum. If f is differentiable at x_0, $f'(x_0) = 0$ is a *necessary*, but not a *sufficient*, condition for f to have a relative extremum at x_0.

There are tests which are sufficient, but not necessary, for a relative extremum at x_0, and which at the same time provide a means of distinguishing a relative maximum from a relative minimum. We postpone consideration of such tests to later sections. See, for instance, Example 8, §1.2.

In many problems we are interested in a function which is defined over some given interval, and we wish to find the largest (or smallest) value which the function assumes on the given interval. The interval may be closed, e.g., $0 \leqq x \leqq 4$, or open, e.g., $1 < x < 3$, or neither. Examples of intervals which are neither open nor closed are furnished by inequalities such as $0 < x \leqq 10$ and $2 \leqq x < 8$ (open at one end and closed at the other). Also, we may be interested in finding the greatest or least value of $f(x)$ for an infinite range of values of x. For example, let x denote the altitude of a right circular cone circumscribed about a sphere of radius b, and let $f(x)$ denote the volume of the cone. One finds without much trouble that

$$f(x) = \frac{\pi}{3}\left(\frac{b^2x^2}{x - 2b}\right).$$

The significant values of x are those for which $x > 2b$, since if $x \leqq 2b$ there can be no cone of altitude x circumscribed about the sphere. For further consideration of this problem see Exercise 10, page 25.

There may or may not be an absolute maximum or an absolute minimum in a given situation. This will depend on the nature of the function and the interval which are involved.

Example 1. $f(x) = x$; interval $0 < x \leqq 10$.

Here there is no absolute minimum, since $f(x)$ can be as near 0 as we please, but never attains that value on the specified interval. There is an absolute maximum, occurring at $x = 10$.

Example 2. $f(x) = x^2$; interval $-1 \leqq x < 2$.

Here there is an absolute minimum at $x = 0$; there is no absolute maximum on the interval, for $f(x)$ can approach but never reach the value 4 when x is restricted to the specified interval.

Example 3. $f(x) = \tan x$; interval $-\pi/2 < x < \pi/2$.

Here there is neither absolute maximum nor absolute minimum.

There is a very important theorem to the following effect:

If f is defined and continuous at each point of the finite closed interval $a \leqq x \leqq b$, then at some point of the interval $f(x)$ attains an absolute maximum value. Likewise, at some point of the interval $f(x)$ attains an absolute minimum value.

This theorem is taken up carefully and proved in §3.2. For the present we shall accept the theorem and strive to appreciate its usefulness. The requirement that the interval be finite and closed is quite essential, for in the absence of these

limitations $f(x)$ might not attain any absolute extreme values, as we see by Examples 1–3.

In practice the functions we are interested in are usually differentiable at all points of the interval (there may sometimes be isolated exceptional points). If an *absolute* extreme value occurs at an interior point of the interval, it is also a *relative* extreme value in the sense of Theorem III, and therefore we must have $f'(x) = 0$ at the point, provided f is differentiable. We therefore have the following guiding principle in searching for points at which $f(x)$ can attain an absolute maximum or minimum value: *Suppose f is differentiable on the given interval, except perhaps at a finite number of points, and suppose it is known that an absolute maximum (or minimum) value is actually attained. Then the point of attainment is either*

(*a*) *a point where* $f'(x) = 0$,
(*b*) *a point at one end of the interval,*
or (*c*) *a point where f is not differentiable.*

In the common type of problem studied in elementary calculus, the solution is usually found under (*a*). In fact, it usually happens with physical or geometrical problems that there is only one interior point of the given interval where $f'(x) = 0$. Solutions under (*b*) do occur sometimes, even in physical problems, and a carefully reasoned solution should always take account of the situation at the ends of the interval, perhaps even before computing $f'(x)$. The situation (*c*) may occur also, but this will be more rare in common practice.

***Example* 4.** Find a number x between 0 and 1 such that $f(x) = \frac{2}{x} + \frac{8}{1-x}$ is as small as possible.

We observe that $f(x) > 0$ when $0 < x < 1$; f is continuous in the specified open interval. Also, $f(x)$ becomes very large (in fact $f(x) \to +\infty$) as x approaches either end of the interval. We conclude that the graph of $y = f(x)$ near the ends of the interval has an appearance somewhat as shown in Fig. 10. It follows from this reasoning that if we choose a closed interval $a \leqq x \leqq b$, with $a > 0$ and very near 0, and $b < 1$ and very near 1, the function f will have smaller values in the interior of the interval $[a, b]$ than it has in the rest of the interval $(0, 1)$. Since f is continuous on the finite closed interval $[a, b]$, it must attain a value at some point of $[a, b]$ which is an absolute minimum among all the values occurring on the interval. This absolute minimum will also be an absolute minimum among all the values of f occurring on the open interval $(0, 1)$. Now f is differentiable in $(0, 1)$; hence the required point of absolute minimum must be a point at which $f'(x) = 0$. We therefore proceed to compute the derivative and solve the equation $f'(x) = 0$:

$$f'(x) = \frac{-2}{x^2} + \frac{8}{(1-x)^2} = \frac{6x^2 + 4x - 2}{x^2(1-x)^2},$$

$$3x^2 + 2x - 1 = 0, x = \tfrac{1}{3} \text{ or } x = -1.$$

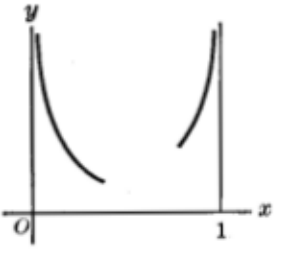


Fig. 10.

The solution in the interval $(0, 1)$ is $x = \frac{1}{3}$. We conclude that $f(x)$ attains its minimum value at $x = \frac{1}{3}$, the minimum value is 18. Observe that no test, by second derivatives or otherwise, is necessary to distinguish between a maximum and a minimum in this case, since we know that a minimum value exists, and there is only one point in the interval at which $f'(x) = 0$.

We emphasize that the assurance of the existence of an absolute minimum in Example 4 is based on use of the theorem cited following Example 3 on page 22. Likewise, in Example 5 (following), the existence of an absolute maximum is assured by the same theorem.

Example 5. If $f(x) = x(4-x)\{(x-2)^2+2\}$, find the absolute maximum value of $f(x)$ when $0 \leqq x \leqq 4$.

Here we see that f is differentiable (and continuous) for all values of x. Moreover, $f(0) = f(4) = 0$, and $f(x) > 0$ if $0 < x < 4$. There must be an absolute maximum for f somewhere on the closed interval $[0, 4]$, and it clearly does not occur at either end of the interval. Hence it must occur at some interior point where $f'(x) = 0$. A simple calculation shows that

$$f(x) = -x^4 + 8x^3 - 22x^2 + 24x,$$
$$f'(x) = -4x^3 + 24x^2 - 44x + 24,$$
$$f'(x) = -4(x-1)(x-2)(x-3).$$

There are three points where $f'(x) = 0$: $x = 1$, 2, and 3. Calculation shows that

$$f(1) = 9, f(2) = 8, f(3) = 9.$$

Hence the absolute maximum value of f on $[0, 4]$ is 9, occurring at $x = 1$ and $x = 3$. All of this shows up clearly on a graph, which the student should construct for himself. It is noteworthy, however, that the reasoning is conclusive without the graph.

EXERCISES

1. (**a**) Find all the points of relative maxima and minima and sketch the graph of $f(x) = (x+5)^2(x^3-10)$. (**b**) Find the absolute maximum and minimum values of $f(x)$ on the interval $-6 \leqq x \leqq -2$.

2. Find the algebraically largest and smallest values of $f(x) = 20x^3 - 2700x + 7000$ when $0 \leqq x \leqq 10$.

3. Find the absolute maximum of $f(x) = (x^2-75)/(x-10)$ for $0 \leqq x < 10$. Begin by sketching the graph enough to show why a maximum must be attained in this interval.

4. Consider $f(x) = x/(x^2+a^2)^{3/2}$ for $x \geqq 0$. Explain why $f(x)$ must attain an absolute maximum value at some point. Find the point

5. Find the absolute maximum of $\sin^2 2\theta(1+\cos 2\theta)$ for $0 \leqq \theta \leqq \pi/2$.

6. Discuss the possible absolute extrema of $f(x) = x^4 + (256/x^2)$ for $0 < x \leqq 4$, and find any that exist.

7. Consider the function $(27/\sin x) + (64/\cos x)$, $0 < x < \pi/2$. Why must there be an

absolute minimum in this open interval? Find where it occurs, and the corresponding value of the function.

8. Without drawing the graph, find the absolute maximum and minimum values of the function (**a**) $5\cos^3 x - 3\cos x$; (**b**) $2\sin x - 1 + 2\cos^2 x$.

9. A right circular cone of altitude x is inscribed in a sphere of radius c. Express the volume of the cone as a function of x. What open interval of values of x is of significance in considering nondegenerate cones? Explain why there must be an absolute maximum volume for some x in this interval, and find the x for which the maximum is attained.

10. Let V be the volume of a right circular cone of altitude x circumscribed about a sphere of radius b. Show that V attains its absolute minimum when $x = 4b$. Write the argument out fully and carefully after the manner of the discussion of Example 4.

11. A right circular cylinder with radius of base x is inscribed in a right circular cone with radius of base r and altitude h. (**a**) Express the total surface area of the cylinder (including ends) as a function of x. (**b**) What interval of x-values are of significance if a nondegenerate cylinder is wanted? (**c**) What inequality must be satisfied by r and h if there is to be a nondegenerate cylinder of maximum area? (**d**) What is the situation if $h = 6$, $r = 2$? if $h = 4$, $r = 2$? if $h = 4$, $r = 3$?

12. Consider $f(x) = \dfrac{\sin \pi x}{x(1-x)}$ if $0 < x < 1$. (**a**) Complete the definition of f at $x = 0$, 1 so as to make f continuous on $[0, 1]$. (**b**) Find the absolute maximum and minimum of $f(x)$ on $[0, 1]$ after completing (**a**). (**c**) Sketch the graph of $y = f(x)$.

13. Let $f(x) = 5\sqrt{16 + x^2} + 4\sqrt{(3-x)^2}$ (taking the positive square root in both cases, so that $\sqrt{(3-x)^2} = |3-x|$). Note that f is differentiable except when $x = 3$ and that $f(x) \to +\infty$ as $x \to +\infty$ or $x \to -\infty$. (**a**) How do you infer from this that f must attain an absolute minimum value? (**b**) Find formulas for $f'(x)$ according as $x < 3$ or $x > 3$, and show that $f'(x) < 0$ if $x < 3$, while $f'(x) > 0$ if $x > 3$. (**c**) What do you infer about the point of attainment of the minimum? (**d**) What is the minimum value of $f(x)$?

14. A spring is located at $(0, a)$, and a man's house is located at $(b, 0)$, where $a > 0, b > 0$. A pipeline is to be laid in two straight parts, the first part from the spring to $(x, 0)$, and the second part from $(x, 0)$ to the house. The two parts will cost c_1 and c_2 dollars per unit length, respectively.

(**a**) Show that the total cost of the pipeline, for *any* value of x, is

$$f(x) = c_1\sqrt{a^2 + x^2} + c_2|b - x|.$$

(**b**) Show that f has its absolute minimum value for some x such that $0 < x \leqq b$.
HINT: Consider the sign of $f'(x)$ when $x \leqq 0$ and $x > b$ respectively.

(**c**) Find the inequalities which must be satisfied by c_1, c_2, a, and b if f is to attain its minimum for an x such that $0 < x < b$.

Observe that Exercise 13 is a special case of this exercise. A contrasting special case is afforded by taking $c_1 = 5$, $c_2 = 4$, $a = 3$, $b = 5$. If one does these two special cases first, the general problem will be more interesting.

15. Discuss the intervals of definition of the function

$$f(x) = \{(16 - x^2)(x^2 - 9)\}^{1/2},$$

and find the absolute maximum of the function.

16. A man wishes to get from point A to point B, these points being diametrically opposite each other on the shores of a circular pond. The man can row $1\frac{1}{2}$ miles per hour and walk 5 miles per hour.

(a) If there is a boat available at A, what combination of rowing and walking will take him to B in the least possible time?

(b) Discuss the problem in case the rowing and walking speeds are, respectively, u and v miles per hour. Check your results carefully in the special case $u = 2$, $v = 4$.

17. Consider a and b as fixed, with $b < a$. Let c be a variable such that $b < c < a$. Let ϕ be the acute angle between the tangents to the circle $x^2 + y^2 = c^2$ and the ellipse $b^2x^2 + a^2y^2 = a^2b^2$ at a point of intersection. Find $\tan \phi$ when c is chosen so that ϕ is greatest.

18. Write out the proof of Theorem III for the case of a relative minimum.

Show that, if f has a relative minimum at x_0, and $g(x) = -f(x)$, then g has a relative maximum at x_0. Hence deduce the proof for the case of a minimum from the facts already established for a maximum.

1.2 / THE LAW OF THE MEAN (THE MEAN-VALUE THEOREM FOR DERIVATIVES)

The theorem which goes by the name of *the law of the mean* is one of the most important theoretical results in the subject of differential calculus. It is used as a tool in many places in the later developments of calculus, both differential and integral, particularly in connection with proofs. We wish to emphasize very strongly that the student of advanced calculus needs to gain an appreciation of the power of the law of the mean as an instrument of systematic reasoning. The first step should be to become thoroughly familiar with the content of the law itself.

THEOREM IV. *(The law of the mean.) Let f be a function which is continuous at each point of the closed interval $a \leqq x \leqq b$, and let it have a derivative at each point of the open interval $a < x < b$. Then there is a point $x = X$ in the open interval $(a < X < b)$ such that*

$$f(b) - f(a) = (b - a)f'(X). \tag{1.2–1}$$

The theorem has a geometrical interpretation. Represent the function graphically by the curve $y = f(x)$, and let A, B be the points on the curve corresponding to $x = a$, $x = b$, respectively. The formula (1.2–1) states that there is some point on the curve, with abscissa $x = X$, at which the tangent is parallel to the line AB. There may be more than one suitable value of X; the essential thing is that there is always at *least* one (see Fig. 11).

It is worth noting that (1.2–1) remains true if we exchange a and b, for both sides merely change sign when this is done. Thus, suppose x_1, x_2 are the end-points of an interval on which the conditions of the law of the mean are satisfied. Then we can write

$$f(x_2) - f(x_1) = (x_2 - x_1)f'(\xi), \tag{1.2–2}$$

where $x = \xi$ is some point between x_1 and x_2. In



Fig. 11.

writing this formula we do not need to know which of the numbers x_1, x_2 is the larger.

The geometrical interpretation of the law of the mean makes its truth plausible. A proof must be based on analytical reasoning, however. We follow the usual procedure of basing the proof on an auxiliary theorem named after the seventeenth-century mathematician Michel Rolle.

ROLLE'S THEOREM. *Let g be a function which satisfies the conditions of Theorem IV, and suppose further that $g(a) = g(b) = 0$. Then for some X such that $a < X < b$ it is true that $g'(X) = 0$.*

Proof. We distinguish two cases: (1) the case in which $g(x)$ is zero on the whole interval, and (2) the case in which $g(x)$ assumes values other than 0 at some points of the interval. In case (1) the derivative $g'(x)$ is identically zero and the existence of X is assured. In case (2) let M and m be the maximum and minimum values, respectively, of $g(x)$ on the closed interval $[a, b]$. At least one of the values M, m must be different from 0, and must therefore occur at point x of the *open* interval (a, b). By Theorem III (§1.12) we conclude that $g'(X) = 0$.

We have glossed over the main difficulty in this proof, namely the matter of the existence of the extreme values M and m. Here we appeal to the theorem which asserts that if a function is continuous on a finite closed interval, it actually attains its absolute extremal values at certain points of the interval. This theorem has been referred to before (see §1.12, following Example 3); we treat it formally as Theorem III in §3.2.

The law of the mean is deduced from Rolle's theorem by an artifice. The function f in Theorem IV need not vanish at $x = a$ and $x = b$. Suppose, however, that $y = F(x)$ is the equation of the straight line AB in Fig. 11, so that $F(a) = f(a)$, $F(b) = f(b)$. Let $g(x) = F(x) - f(x)$. Then g will be a function meeting the conditions of Rolle's theorem. The equation of the line in question is

$$y = \frac{f(b) - f(a)}{b - a}(x - a) + f(a).$$

Hence we set

$$g(x) = \frac{f(b) - f(a)}{b - a}(x - a) + f(a) - f(x).$$

Note that $g(a) = g(b) = 0$. The derivative is

$$g'(x) = \frac{f(b) - f(a)}{b - a} - f'(x).$$

The conclusion $g'(X) = 0$ of Rolle's theorem is now seen to be equivalent to the law of the mean in the form (1.2–1).

We now give some simple examples illustrating the law of the mean in particular instances.

Example 1. Suppose $f(x) = x^3$. Find a suitable value for X in (1.2–1) when $a = -1$ and $b = 2$.

Since $f'(x) = 3x^2$, the law of the mean takes the form $b^3 - a^3 = (b - a)3X^2$ in the present case. With $a = -1$, $b = 2$ we have $8 - (-1)^3 = [2 - (-1)]3X^2$, or $X^2 = 1$. Solving, we find $X = \pm 1$. We want a value of X such that $a < X < b$, i.e., $-1 < X < 2$. Hence $X = 1$ is the suitable value in question.

Example 2. If $f(x) = x^2$, show that the suitable value of X in the law of the mean is $X = (a + b)/2$.

We have $f'(x) = 2x$. Hence the law of the mean becomes $b^2 - a^2 = (b - a)2X$, or $X = (a + b)/2$. Where is this point located in relation to a and b?

Example 3. If $f(x) = \sin x$ and $x_1 = 0$, $x_2 = 5\pi/6$, find ξ such that $x_1 < \xi < x_2$ and (1.2–2) holds.

We have $f'(x) = \cos x$ and $\sin(5\pi/6) = \frac{1}{2}$. Hence (1.2–2) takes the form

$$\tfrac{1}{2} - 0 = \sin\frac{5\pi}{6} - \sin 0 = \frac{5\pi}{6}\cos\xi$$

or

$$\cos\xi = \frac{3}{5\pi} = 0.19099.$$

Since $0 < \xi < 5\pi/6$, we find $\xi = \cos^{-1}(0.19099) = 1.37863$. This is somewhat less than $\pi/2$, which is to be expected, as may be seen from a carefully drawn graph.

In actual practice we are seldom interested in the exact value of the X occurring in (1.2–1); the important thing is that X lies between $x = a$ and $x = b$. This enables us to obtain inequalities for estimating the value of $f(x)$.

Example 4. Show that $\frac{1}{3} < \log 1.5 < \frac{1}{2}$.

This may be done as follows: We take $f(x) = \log x$, $a = 1$, $b = 1.5$; then $f'(x) = 1/x$, and by the law of the mean

$$\log 1.5 - \log 1 = \log 1.5 = (1.5 - 1)\frac{1}{X} = \frac{0.5}{X},$$

where $1 < X < 1.5$. It follows that

$$\frac{0.5}{1.5} < \frac{0.5}{X} < \frac{0.5}{1}, \text{ or } \frac{1}{3} < \frac{0.5}{X} < \frac{1}{2}.$$

This gives the required result.

Example 5. Use the law of the mean to show that

$$\frac{\log x}{x} < \frac{\log a}{x} + \frac{1}{a} \tag{1.2–3}$$

if $0 < a < x$.

We use (1.2–1) with $f(x) = \log x$ and b replaced by x. Then

$$\log x - \log a = \frac{x - a}{X}, \quad a < X < x.$$

Hence

$$\frac{\log x}{x} = \frac{\log a}{x} + \frac{x-a}{x} \cdot \frac{1}{X}.$$

Now

$$\frac{x-a}{x} \cdot \frac{1}{X} < \frac{1}{X} < \frac{1}{a},$$

and so (1.2–3) is seen to be correct.

A much used variant form of the law of the mean is obtained from (1.2–1) in the following way: Let $h = b - a$, so that $b = a + h$. Then X may be written $X = a + \theta h$, where $0 < \theta < 1$, because any number between a and $a + h$ can be expressed in this latter form. Thus we have

$$f(a+h) = f(a) + hf'(a+\theta h),\ 0 < \theta < 1. \tag{1.2–4}$$

Example 6. Use (1.2–4) to show that

$$(1+h)^\alpha > 1 + \alpha h \tag{1.2–5}$$

if $h > 0$ and $\alpha > 1$.

We take $f(x) = x^\alpha$, $a = 1$. Then by (1.2–4),

$$(1+h)^\alpha = 1 + h\alpha(1+\theta h)^{\alpha-1}.$$

If $0 < h$ we have $(1+\theta h)^{\alpha-1} > 1$, since $1 + \theta h > 1$ and $\alpha > 1$. The inequality (1.2–5) follows at once.

One very important consequence of the law of the mean will be formally stated here as a theorem.

THEOREM V. *Let f be differentiable at each point of the open interval $a < x < b$, and suppose that $f'(x) = 0$ at each such point. Then the value of the function is constant on the interval.*

Proof. It follows by Theorem I, §1.11, that f is continuous at each point of the open interval (a, b). Consider any two distinct points of the interval, say x_1, x_2, where $a < x_1 < x_2 < b$. We may apply the law of the mean, obtaining formula (1.2–2). But $f'(\xi) = 0$, by hypothesis, and so $f(x_1) = f(x_2)$. We have now proved the theorem, for we have shown that f has the same value at any two points of the interval.

Theorem V plays an essential role in the explanation of the relation between differentiation and integration, as we shall see when we come to the proof of Theorem VIII, §1.53. Theorem V is also useful in dealing with the concept of the "general solution" of a certain elementary type of differential equation, as we shall see in §1.4.

The following example also affords an important application of the law of the mean.

Example 7. Suppose that f satisfies the conditions of the law of the mean on the interval $a \leqq x \leqq b$, and that $f'(x) > 0$ when $a < x < b$. Show that $f(x)$ increases as x increases.

We are to show that $x_1 < x_2$ implies $f(x_1) < f(x_2)$ whenever $a \leqq x_1 < x_2 \leqq b$. The law of the mean tells us that there is some ξ such that $x_1 < \xi < x_2$ and $f(x_2) - f(x_1) = (x_2 - x_1)f'(\xi)$. Since $(x_2 - x_1) > 0$ and $f'(\xi) > 0$, we infer that $f(x_2) - f(x_1) > 0$; this is equivalent to $f(x_1) < f(x_2)$.

Example 8. Suppose that f is defined and differentiable on an open interval containing the point x_0. Suppose that $f'(x_0) = 0$ and that for all x sufficiently near x_0, $f'(x) > 0$ when $x < x_0$ and $f'(x) < 0$ when $x > x_0$. Show that these conditions are sufficient to guarantee that f has a relative maximum at x_0.

The argument is based on Example 7. As x increases, $f(x)$ increases when $f'(x) > 0$. By similar reasoning $f(x)$ decreases when x increases if $f'(x) < 0$. In the present case we see that the given conditions imply that for some small number h, $f(x)$ is increasing as x goes from $x_0 - h$ to x_0, and decreasing as x goes from x_0 to $x_0 + h$. Hence $f(x)$ must attain a relative maximum at x_0.

From this argument it will be apparent to the student how one may formulate sufficient conditions for a relative minimum at x_0.

EXERCISES

1. Use the law of the mean to show that $\frac{1}{9} < \sqrt{66} - 8 < \frac{1}{8}$.

2. Prove that there is no value of m such that $x^3 - 3x + m = 0$ has two distinct roots in the interval $0 \leqq x \leqq 1$. Use Rolle's theorem.

3. If $f(x) = x^3 - 3x^2 + 2x$, $a = 0$, $h = \frac{1}{2}$, find a suitable value of θ in the formula (1.2–4).

4. For what values of C is $Cx - \sin x$ an increasing function of x (for *all* x)?

5. Show that $2/\pi < (\sin\theta)/\theta < 1$ if $0 < \theta < \pi/2$. HINT: Examine the sign of the derivative of $(\sin\theta)/\theta$.

6. Prove the following inequalities, using the law of the mean.

(a) $\sqrt{1+x} < 4 + (x - 15)/8$ if $x > 15$.

(b) $\tan^{-1} x < (\pi/4) + (x - 1)/2$ if $1 < x$.

(c) $\frac{\pi}{4} - \frac{1-x}{1+x^2} < \tan^{-1} x < \frac{\pi}{4} - \frac{1-x}{2}$ if $0 < x < 1$.

(d) $h/(1+h^2) < \tan^{-1} h < h$ if $0 < h$.

7. Prove that the inequality (1.2–5) also holds if $-1 < h < 0$. Explain the reasoning about inequalities with care, noting that if $0 < A < 1$ and $B < 0$, then $AB > B$.

8. Prove the following inequalities:

(a) $\frac{x}{1+x} < \log(1+x) < x$, if $-1 < x < 0$ or $0 < x$.

(b) $1 + \frac{x}{2\sqrt{1+x}} < \sqrt{1+x} < 1 + \frac{1}{2}x$, if $-1 < x < 0$ or $0 < x$.

(c) $1 - \frac{x}{2} < \frac{1}{(1+x)^{1/2}} < 1 - \frac{x}{2(1+x)^{3/2}}$, if $-1 < x < 0$ or $0 < x$.

(d) $p(x-1) < x^p - 1 < px^{p-1}(x-1)$ if $1 < x$, $1 < p$.

(e) $\dfrac{m(x-1)}{x^{1-m}} < x^m - 1 < m(x-1)$ if $0 < m < 1$, $1 < x$.

9. If $a > b > 0$ and m and n are positive numbers such that $m + n = 1$, show that $a^m b^n < ma + nb$. Use Exercise 8(*e*).

10. (a) Prove that $x^n + ax + b = 0$ (a, b real) has at most two distinct real roots if n is even, and at most three if n is odd.

(b) Prove that $x^n + ax^2 + b = 0$ (a, b real) has at most three distinct real roots if n is odd, and at most four if n is even.

11. Explain why $a_0x^4 + a_1x^3 + a_2x^2 + a_3x + a_4 = 0$ must have a root between 0 and 1 if $(a_0/5) + (a_1/4) + (a_2/3) + (a_3/2) + a_4 = 0$.

12. Let $F(x) = (f(x) - f(a))(g(b) - g(x))$, where f and g are continuous when $a \leqq x \leqq b$ and differentiable when $a < x < b$. Suppose further that $g'(x)$ is never zero. Show that there is a ξ between a and b such that

$$\frac{f'(\xi)}{g'(\xi)} = \frac{f(\xi) - f(a)}{g(b) - g(\xi)}.$$

13. Suppose $f''(x) > 0$ when $a \leqq x \leqq b$. Explain by an analytical argument why there can be at most one point of the interval at which $f'(x) = 0$. What is the geometrical interpretation as regards the curve $y = f(x)$?

14. Suppose f is continuous on the interval $a < x < b$, and that f is known to be differentiable on this interval except possibly at one point x_0. Suppose further that $\lim_{x \to x_0} f'(x)$ exists. Use the law of the mean to prove that f is differentiable at x_0 and that the derivative f' is continuous at that point.

15. Generalize the result of Illustrative Example 7 as follows: Suppose f is continuous on $a \leqq x \leqq b$ and differentiable on $a < x < b$. Suppose further that $f'(x) \geqq 0$ when $a < x < b$ and that $f'(x) > 0$ for at least one value of x. Prove that $f(a) < f(b)$. [It is easy to see that $f(a) \leqq f(b)$; what requires more care is to show that $f(a) \neq f(b)$.]

16. Suppose $f(x) = x^2 \sin(1/x) + (x/2)$ if $x \neq 0$, and define $f(0) = 0$. (a) Show that $f'(0) > 0$. (b) Show that, no matter how small the positive number h may be there are infinitely many points on both sides of $x = 0$ and within distance h of $x = 0$ at which $f'(x) = \frac{3}{2}$ and also infinitely many at which $f'(x) = -\frac{1}{2}$. This shows that there is no interval containing $x = 0$ in which $f(x)$ is always increasing as x increases, in spite of the fact that the slope of the curve $y = f(x)$ is positive at $x = 0$.

17. Suppose the following things: that f is continuous when $0 \leqq x \leqq a$, differentiable when $0 < x < a$; that $f(0) = 0$; and that $f'(x)$ increases as x increases. Show that $f(x)/x$ increases as x increases. (A suitable use of the law of the mean is indicated.) Examples are furnished by $f(x) = e^x - 1$ and by $f(x) = \tan x$, $0 \leqq x \leqq \pi/2$.

18. Let f be differentiable when $a < x < b$. Suppose x_1 and x_2 are distinct points of the interval, and let P_1 and P_2 be the corresponding points of the curve.

(a) Show that the condition for P_1 to lie above the line tangent to the curve at P_2 is

$$f(x_1) - f(x_2) - (x_1 - x_2)f'(x_2) > 0.$$

(b) If the condition in (a) holds for every pair of points on the curve, show that $f'(x)$ increases as x increases. In fact, if $x_1 < x_2$, show that

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

(c) Show, conversely, that if $f'(x)$ increases as x increases, the condition in (a) is satisfied whenever $x_1 \neq x_2$. (Use the law of the mean.)

(d) Under the conditions in (c) show that the curve $y = f(x)$ between any two of its points lies entirely below the chord joining those points. Begin by showing that an analytic expression of this state of affairs is

$$\frac{f(x)-f(x_1)}{x-x_1} < \frac{f(x_2)-f(x)}{x_2-x}$$

whenever $x_1 < x < x_2$.

1.3 / DIFFERENTIALS

The notion of a differential is closely related to that of a derivative. For functions of a single independent variable this relationship is very close indeed, and very simple. For functions of several independent variables the relationship is less simple. At this point we are concerned only with functions of a single variable. We presume that the student is acquainted with differentials and their uses in the formal procedures of elementary calculus. Our purpose here is mainly to define differentials carefully and demonstrate the fundamental property upon which much of the usefulness of differentials depends.

Suppose that f is a function of the independent variable x, and let us assume that f is differentiable for certain values of x (i.e., that $f'(x)$ exists for these values).

Definition. *Let dx denote an independent variable which may take on any value whatsoever. Then the function of x and dx whose value is $f'(x)dx$ is called the differential of f. Observe that the differential is a homogeneous linear function of dx; that is, for a fixed value of x, the differential has as its value a fixed multiple of dx.*

If we write $y = f(x)$, and if f is differentiable for a particular value of x, it is customary to write

$$dy = f'(x)\,dx, \tag{1.3–1}$$

so that dy is the value of the differential of f for assigned values of x and dx.

If we regard x as fixed, dy is a dependent variable whose value depends on the independent variable dx. The variables dx and dy are often referred to as the differentials of x and y, respectively. The adjacent Fig. 12 illustrates geometrically the functional dependence of dy on dx, as well as the relation to the function f itself. The xy-co-ordinate axes and the graph of the function $y = f(x)$ are shown in unbroken lines. A second co-ordinate system is showing with its origin at a

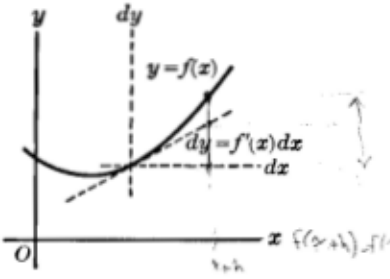


Fig. 12.

typical point (x, y) of the curve $y = f(x)$. The axes in this system are scales for the measurement of the variables dx, dy. The equation (1.3–1) has as its graph a straight line of slope $f'(x)$. This line is, of course, the tangent to the curve $y = f(x)$ at the origin of the dx-dy co-ordinate system.

From (1.3–1) we have the quotient relation

$$\frac{dy}{dx} = f'(x) \tag{1.3–2}$$

whenever $dx \neq 0$. The d-notations dx and dy go back to Leibniz's work in the seventeenth century, but Leibniz did not define the derivative by the limit of a quotient as we did in (1.11–4). It is to be emphasized that there is no need for dx and dy to be small in (1.3–2).

Example 1. If $y = f(x) = \sin x$, calculate the value of dy for $x = \pi/3$, $dx = \pi/6$. Here $f'(x) = \cos x$, so $dy = \cos x\, dx$. Evaluating, we obtain

$$dy = \left(\cos \frac{\pi}{3}\right)\frac{\pi}{6} = \frac{\pi}{12}.$$

Probably the most important feature of the formula (1.3–1) is that its truth is unaffected by the introduction of a new independent variable.

Example 2. Suppose $y = x^2$ and $x = t^3 + t$, so that $y = (t^3 + t)^2 = t^6 + 2t^4 + t^2$. If we regard x as an independent variable, then $dy = 2x\, dx$, by (1.3–1). Here dx is an independent variable. But if we regard t as an independent variable, then both x and y are dependent on t, and the notations dy, dx acquire new meanings:

$$x = t^3 + t,\ dx = (3t^2 + 1)\, dt,$$
$$y = t^6 + 2t^4 + t^2,\ dy = (6t^5 + 8t^3 + 2t)\, dt.$$

But even with these new meanings, it is still true that $dy = 2x\, dx$. We verify this by writing

$$2x\, dx = 2(t^3 + t)(3t^2 + 1)\, dt = (6t^5 + 8t^3 + 2t)\, dt = dy.$$

What we have verified here in a particular case may be demonstrated in general by appealing to the rule for differentiating a composite function (Theorem II, § 1.11). Suppose $y = f(x)$ and $x = g(t)$, so that $y = F(t)$, where $F(t) = f(g(t))$. Then, with t as independent variable,

$$dy = F'(t)\, dt,\ dx = g'(t)\, dt. \tag{1.3–3}$$

By Theorem II we have

$$F'(t) = f'(x)g'(t). \tag{1.3–4}$$

Hence combining (1.3–3) and (1.3–4),

$$dy = f'(x)g'(t)\, dt = f'(x)\, dx,$$

so that (1.3–1) holds, even though x and dx are no longer independent variables.

The use of differentials is a great convenience in algebraic manipulations

which are incidental to much work in calculus. Using differentials rather than derivatives, one is often enabled to retain a desirable symmetry by not forcing a decision as to which variable is independent. The differential formula for arc length of a plane curve illustrates this point. The formula is

$$ds^2 = dx^2 + dy^2; \qquad (1.3\text{–}5)$$

the co-ordinates (x, y) of a point on the curve are functions of some parameter, and the arc length s, measured from some chosen initial point on the curve, likewise depends on the parameter. But the formula (1.3–5) holds (granted suitable conditions on the curve) no matter what the parameter may be.

The fact that $f'(x)$ is the ratio of dy to dx no matter what variable is independent is of great usefulness when we wish to compute the slope of a curve defined parametrically.

EXERCISES

1. **(a)** If $y = (1 - x^2)/(1 + x^2)$, compute dy when $x = 1$, $dx = 2$.

(b) If $x = \tan(t/2)$, compute dx when $t = \pi/2$, $dt = 2$.

(c) If x in **(a)** is replaced by its value in terms of t from **(b)**, show that, on simplification, $y = \cos t$. From this formula compute dy when $t = \pi/2$, $dt = 2$. Compare the answers to **(b)** and **(c)** with your result in part **(a)**.

2. From $x = r\cos\theta$, $y = r\sin\theta$, and $ds^2 = dx^2 + dy^2$ derive the formula $ds^2 = dr^2 + r^2\,d\theta^2$.

3. **(a)** If $y = f(x)$, $y' = f'(x)$, and so on, why is $dy' = y''\,dx$?

(b) What are dy'' and $d(y')^2$ when expressed with dx as a factor?

(c) Suppose that $x = f(t)$, $y = g(t)$, and write $x' = f'(t)$, $y' = g'(t)$, and so on. Show that

$$\frac{d}{dt}\left(\frac{dy}{dx}\right) = \frac{x'y'' - y'x''}{(x')^2}.$$

4. For a plane curve C, construct the tangent at a typical point $P(x, y)$, and let angles ϕ, ψ, θ be as indicated on Fig. 13, so that for the general case $\phi = \theta + \psi + n\pi$ where n is an integer ($n = 0$ in Fig. 13). From this equation and the relations

$$\frac{dy}{dx} = \tan\phi, \frac{y}{x} = \tan\theta,$$

show that

$$\tan\psi = \frac{x\,dy - y\,dx}{x\,dx + y\,dy} = \frac{r\,d\theta}{dr}.$$

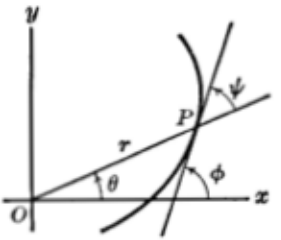


Fig. 13.

5. From $\operatorname{ctn}\psi = \dfrac{dr}{r\,d\theta}$ (see Exercise 4) find $d\psi$ in terms of r, r', r'', and $d\theta$, where $r' = \dfrac{dr}{d\theta}$, and $r'' = \dfrac{dr'}{d\theta}$.

6. The curvature of a plane curve $y = f(x)$ is defined as $K = d\phi/ds$, where $dy/dx = \tan\phi$ and $ds^2 = dx^2 + dy^2$. Derive the formula

$$K = \pm\frac{y''}{(1 + y'^2)^{3/2}}.$$

7. Using the definition of K in Exercise 6, and the relation between ϕ, ψ, and θ in Exercise 4, derive the formula

$$K = \pm \frac{r^2 + 2r'^2 - rr''}{(r^2 + r'^2)^{3/2}}.$$

(Exercise 5 should be worked first.)

8. Let $y = f(x)$ be the equation of a plane curve C. The center of curvature of the curve corresponding to a particular point (x, y) on C is a point (X, Y) a distance R from (x, y) along the normal to C at (x, y), in the direction toward the concave side of the curve; here R is the radius of curvature. It may be shown that

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

$$Y = y + \frac{1 + y'^2}{y''}.$$

It is assumed, of course, that $y'' \neq 0$.

The locus of (X, Y) as (x, y) moves along C, is a curve called the *evolute* of C. We may regard the above equations as parametric equations of the evolute, with x as parameter. Show that $dY/dX = -1/y'$. This proves that the normal to C at (x, y) is tangent to the evolute at (X, Y).

1.4 / THE INVERSE OF DIFFERENTIATION

Many applications of calculus, some of them quite elementary, require the determination of a function from the knowledge of its first or second derivative, together with supplementary data about the function for particular values of the independent variable. The first and simplest general problem of this kind may be put as follows:

Problem. *Given a continuous function $f(x)$, defined on a certain interval, find all functions defined on this interval and having $f(x)$ as derivative. In symbols, find y as a function of x such that*

$$\frac{dy}{dx} = f(x). \tag{1.4–1}$$

Directly out of his or her experience with differentiation, the student is able to solve many problems of this type. The process is one of using standard differentiation formulas in reverse. Let us examine the reasoning carefully in a typical case.

Example. Find y as a function of x such that $y = 1$ when $x = 0$ and

$$\frac{dy}{dx} = \frac{x}{1 + x^2}. \tag{1.4–2}$$

Here $f(x)$ is defined and continuous for all values of x, so we want the solution to be defined for all values of x.

The normal procedure is to write

$$dy = \frac{x\,dx}{1+x^2} = \frac{1}{2}\frac{d(1+x^2)}{1+x^2}; \tag{1.4–3}$$

$$y = \tfrac{1}{2}\log(1+x^2) + C; \tag{1.4–4}$$

$$1 = \tfrac{1}{2}\log(1+0) + C, \quad C = 1;$$

$$y = \tfrac{1}{2}\log(1+x^2) + 1. \tag{1.4–5}$$

The supporting argument (usually not made explicit in practice) runs as follows: *If* there is a function y satisfying (1.4–2), then it also satisfies (1.4–3). In this latter form, as finally written, we recognize that a y satisfying (1.4–2) is furnished by (1.4–4), where C is an arbitrary constant. The proper determination of C is made by substituting the given matched values for x and y. Thus we obtain (1.4–5), which is actually a solution of the problem, as may be checked.

One question remains: Is (1.4–5) the *only* solution to the problem? Once we obtain (1.4–4) it is clear that C is uniquely determined by the condition that $y = 1$ when $x = 0$. The question is then: Does (1.4–4) give *all* the functions y satisfying (1.4–2)? The answer is affirmative, and is supplied by Theorem V, § 1.2. For let y be *any* function which is differentiable for all values of x and satisfies (1.4–2). Then the derivative of the function

$$y - \tfrac{1}{2}\log(1+x^2)$$

is zero for all values of x; hence, by Theorem V, this function is constant, so that y is given by (1.4–4) for *some* value of C.

The equation (1.4–1) is the very simplest type of first-order differential equation. The problem which we have posed in connection with this equation may be called the problem of finding the "general solution" of (1.4–1). The general solution is the family of all functions $y = y(x)$ satisfying the equation. Any member of this family may be called "a particular solution." By appeal to Theorem V, §1.2, we obtain the following conclusion:

If $y_1(x)$ and $y_2(x)$ are any two particular solutions of (1.4–1), *they differ by a constant. Thus, if $y_1(x)$ is any particular solution, the general solution is given by*

$$y = y_1(x) + C, \tag{1.4–6}$$

where C is an arbitrary constant. In all this we assume, of course, that all the solutions considered are differentiable on the interval where $f(x)$ is defined.

The main problem is thus reduced to the finding of any one particular solution of (1.4–1). In many important simple problems such a particular solution may be found either by direct inspection or by various ingenious devices, all of which depend upon extensive familiarity with formulas of differentiation and manipulation of differentials. But it is not difficult to give examples in which no solution is forthcoming from the class of functions which a student meets and

learns to differentiate in elementary calculus. Thus, for example, the equations

$$\frac{dy}{dx} = e^{-x^2}, \frac{dy}{dx} = \sqrt{1+x^3}$$

have no solution within this class.

It is well to pause at this point and reflect upon the meaning of the word "function." In elementary differential calculus practically all our experience is with functions of a few basic types: algebraic, trigonometric and inverse trigonometric, exponential and logarithmic, and rather simple compounding of these types. It turns out that within this class of "elementary" functions, differentiation always leads to functions which are again in the class. Such is not the case with the inverse of differentiation, however; there are elementary functions which are not derivatives of elementary functions, e.g., e^{-x^2} and $\sqrt{1+x^3}$. Now the general theorems of calculus deal with functions which are arbitrary except for requirements of differentiability or continuity, and which certainly need not be elementary in the sense of the first part of this paragraph. Once we have rejected the limitation of our considerations to "elementary" functions, we may well ask: What nonelementary functions do we know? If e^{-x^2} is not the derivative of any elementary function, how are we to find solutions of the equation $dy/dx = e^{-x^2}$? Evidently it is necessary in some fashion to acquire a supply of nonelementary functions of which we know the derivatives.

There are several very important methods for building such a supply. One method is that of integration of known functions. Starting with a given continuous function $f(x)$ defined on some interval, we form

$$F(x) = \int_a^x f(t)\,dt, \tag{1.4–7}$$

where a and x belong to the interval, and a is kept fixed. Another method is that of forming infinite series whose terms are given functions of x:

$$F(x) = u_1(x) + u_2(x) + u_3(x) + \cdots.$$

We shall later be able to show that

$$\int_0^x \sqrt{1+t^3}\,dt$$

is a function $F(x)$ such that $F'(x) = \sqrt{1+x^3}$, and that

$$x - \frac{x^3}{3 \cdot 1!} + \frac{x^5}{5 \cdot 2!} - \frac{x^7}{7 \cdot 3!} + \cdots$$

is a function $F(x)$ such that $F'(x) = e^{-x^2}$.

The study of functions defined by infinite series will concern us in a later chapter of this book. Our immediate interest will be confined to functions of the type (1.4–7) defined by integration. We shall presently learn (see the end of §1.52) that if $f(x)$ is continuous on a given interval $a \leq x \leq b$, the general solution of the equation $dy/dx = f(x)$ on that interval is $y = F(x) + C$, where C is

an arbitrary constant and $F(x)$ is defined by (1.4–7). The integral is defined by a limiting process involving sums.

Every student who has come this far in his study of calculus knows that there is a very close connection between differentiation and integration. But elementary calculus books vary widely in their discussions of integration and in their treatment of the link between differentiation and integration. Thus we shall not assume any uniformity of knowledge on this subject by the readers of this book. The next sections will be devoted to the subject of integration and its relation to differentiation. Our aim will be to lay out a precise logical pattern of definitions and theorems which will provide a common understanding of the integration concept and of the connection between differentiation and integration.

EXERCISES

1. What is the logical objection to the following procedure? Let the symbol $\int_0^t e^{-s^2}\,ds$ be *defined* as that function $F(t)$ such that $F'(t) = e^{-t^2}$ and $F(0) = 0$. Then $\frac{d}{dt}(F(t) + C) = F'(t) = e^{-t^2}$, and therefore $v = \int_0^t e^{-s^2}\,ds + C$ is the general solution of the equation $dv/dt = e^{-t^2}$.

2. What is the logical objection to the following procedure? Let $f(x)$ be a given function defined when $a \leqq x \leqq b$. Let $F(x)$ be any differentiable function such that $F'(x) = f(x)$ for each x of the given interval. Then define $\int_a^b f(x)\,dx = F(b) - F(a)$.

3. Try to find a continuous $F(x)$ such that $F(0) = 0$ and $F'(x) = [x]$ when $0 \leqq x \leqq 3$ (where $[x]$ denotes the greatest integer $\leqq x$). Do you succeed fully? Where is the difficulty?

1.5 / DEFINITE INTEGRALS

We are going to define what we mean by the definite integral of a function. We start with a function f, which we suppose to be defined and continuous on a closed interval $a \leqq x \leqq b$. These things being given, the definite integral of f over the interval is a certain number, which we denote by

$$\int_a^b f(x)\,dx, \tag{1.5–1}$$

and which we arrive at by a defining process which we shall outline in four steps, as follows:

Step 1. Choose an integer $n \geqq 1$ and subdivide the interval $[a, b]$ into n subintervals by choosing points $x_0, x_1, \ldots, x_n$ such that $a = x_0 < x_1 < \cdots < x_n = b$. We adopt the notation

$$\Delta x_i = x_i - x_{i-1}, \quad i = 1, \ldots, n.$$

Step 2. In each subinterval choose an arbitrary point, denoting the point in the ith subinterval by x_i', so that

$$x_{i-1} \leqq x_i' \leqq x_i.$$

Step 3. Find the value of the function at each of the points chosen in Step 2, and form the sum

$$f(x_1') \Delta x_1 + f(x_2') \Delta x_2 + \cdots + f(x_n') \Delta x_n. \tag{1.5–2}$$

This is called an approximating sum.

Step 4. Find the limit of the sums (1.5–2) as n is increased and the maximum of the numbers $\Delta x_1, \ldots, \Delta x_n$ is made to approach zero. This limit is, by definition, the definite integral (1.5–1), so that

$$\int_a^b f(x)\, dx = \lim \sum_{i=1}^{n} f(x_i') \Delta x_i. \tag{1.5–3}$$

In Chapter 18, we shall study the theory of integration systematically. At that time we shall prove that the sums (1.5–2) do actually approach a limit in the case of any continuous function. For the present we take for granted the existence of this limit, and its uniqueness. A further discussion of the limit concept associated with the integral will be found in §1.63.

A definite integral is thus defined as the limit of a certain kind of sum associated with the function. A geometrical interpretation of the integral can be made in terms of the area under the curve $y = f(x)$ from $x = a$ to $x = b$. Each term in the approximating sum (1.5–2) is the area of one of the shaded rectangles in Fig. 14. The area under the curve is the limit of the sum of the areas of these

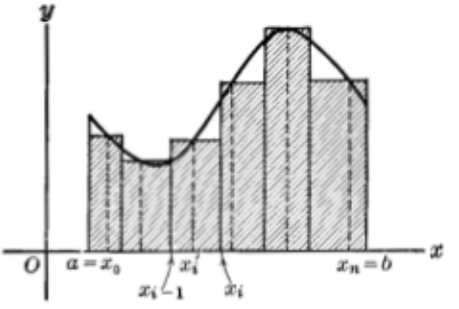


Fig. 14.

rectangles. We assume that the student is already familiar with this geometrical interpretation of the integral, and with the extension of the interpretation (by the concept of negative area) to those situations where the curve $y = f(x)$ goes below the x-axis. We emphasize, however, that we do not *define* the definite integral as the area under the curve. The area interpretation is merely a convenient method of bringing our intuition into play to aid us in grasping the nature of the definition (1.5–3). We "feel" that the area exists, and that a good approximation to it can be obtained by the sums (1.5–2), provided we take all the subintervals short enough. Actually, the area is defined as being equal to the limit in (1.5–3).

Area is only one of many geometrical and physical concepts whose exact quantitative measurement is furnished by a definite integral. The student is no doubt already familiar with such applications of integration as the finding of volumes (including volumes of solids of revolution) by slicing them into thin slabs or thin cylindrical shells, the calculation of arc-lengths of curves, the location of centroids and centers of gravity, and the reckoning of force due to water pressure on submerged plane surfaces. In every one of these applications the integral as the limit of a sum is the fundamental concept. Now we wish to emphasize that the concept is applicable to *any* continuous function, and that the concept itself does not depend on any geometrical or physical interpretation.

Suppose that m_i is the smallest value of $f(x)$ in the subinterval $x_{i-1} \leqq x \leqq x_i$. If we choose x_i' as a point at which $f(x)$ takes on the value m_i, the sum (1.5–2) becomes

$$m_1 \Delta x_1 + m_2 \Delta x_2 + \cdots + m_n \Delta x_n. \tag{1.5–4}$$

This particular approximating sum is called a *lower sum*. Likewise we define an *upper sum* by choosing x_i' to be a point of the subinterval $x_{i-1} \leqq x \leqq x_i$ at which $f(x)$ takes on its greatest value M_i for that subinterval. The upper sum is

$$M_1 \Delta x_1 + M_2 \Delta x_2 + \cdots + M_n \Delta x_n. \tag{1.5–5}$$

The upper and lower sums have the property that

$$s \leqq \int_a^b f(x)\, dx \leqq S, \tag{1.5–6}$$

where s represents the lower sum and S represents the upper sum. This system of inequalities makes it possible to obtain numerical estimates of the value of an integral. Upper and lower sums are discussed at greater length in §18.1.

Example 1. Estimate the value of $\int_2^6 (75x - x^3)\, dx$ by lower and upper sums, using four equal subintervals.

We take $f(x) = 75x - x^3$, $x_0 = 2$, $x_1 = 3$, $x_2 = 4$, $x_3 = 5$, $x_4 = 6$. See Fig. 15. Each $\Delta x_i = 1$. Since $f'(x) = 75 - 3x^2$, it appears that $f(x)$ increases from $x = 2$ to $x = 5$, and decreases from $x = 5$ to $x = 6$. From the accompanying table of values we

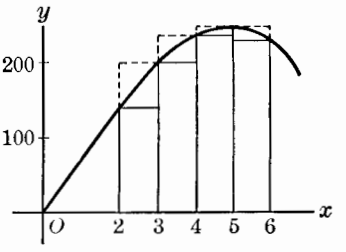


Fig. 15.

can then read off the values of m_i and M_i:

x	$f(x)$
2	142
3	198
4	236
5	250
6	234

i	m_i	M_i
1	142	198
2	198	236
3	236	250
4	234	250

Thus the lower and upper sums are

$$s = 142 + 198 + 236 + 234 = 810,$$

$$S = 198 + 236 + 250 + 250 = 934.$$

Therefore

$$810 < \int_2^6 (75x - x^3)\, dx < 934.$$

A better estimate could be obtained by using more and smaller subintervals.

One of the simple but very important facts about integrals is expressed by the formula

$$\int_a^c f(x)\, dx = \int_a^b f(x)\, dx + \int_b^c f(x)\, dx, \tag{1.5–7}$$

where $a < b < c$, and f is continuous on the interval $a \leqq x \leqq c$. The analytical proof of this formula runs as follows: Divide the whole interval $[a, c]$ into parts in such a way that $x = b$ is always one of the points of subdivision. Suppose that $[a, b]$ is divided into m parts and $[b, c]$ into n parts:

$$a = x_0 < x_1 < \cdots < x_m = b = \xi_0 < \xi_1 < \cdots < \xi_n = c.$$

Choose points x_i' and ξ_j' such that $x_{i-1} \leqq x_i' \leqq x_i$, $i = 1, \ldots, m$, and $\xi_{j-1} \leqq \xi_j' \leqq \xi_j$, $j = 1, \ldots, n$. Then, as $m \to \infty$ and $n \to \infty$, and as the greatest of the differences Δx_i, $\Delta \xi_j$ approaches zero, we have

$$\int_a^b f(x)\, dx = \lim \sum_{i=1}^{m} f(x_i')\, \Delta x_i,$$

$$\int_b^c f(x)\, dx = \lim \sum_{j=1}^{n} f(\xi_j')\, \Delta \xi_j.$$

But also,

$$\int_a^c f(x)\, dx = \lim \left[\sum_{i=1}^{m} f(x_i')\, \Delta x_i + \sum_{j=1}^{n} f(\xi_j')\, \Delta \xi_j \right].$$

Formula (1.5–7) then follows from the principle that the limit of a sum is equal to the sum of the limits. We apply the principle to the sum of the two expressions

$$\sum_{i=1}^{m} f(x_i')\,\Delta x_i \quad \text{and} \quad \sum_{j=1}^{n} f(\xi_j')\,\Delta \xi_j.$$

There are various limiting processes used in calculus: limits of functions, limits of sequences, and the type of limit (of approximating sums) used in defining an integral. For each of these limiting processes there is the principle that the limit of a sum is equal to the sum of the limits. For limits of functions this principle was formulated in the rule (1.1–4). A proof of the validity of the rule (1.1–4) is given in §1.64. With slight formal modifications, the idea of this proof applies equally well to sequences and to limits of approximating sums. Basically, all these various forms of the principle are covered by the theorem that the process of addition is a continuous function of the things which are added (see Theorem VI, §17.5).

Of course, the geometric interpretation of (1.5–7) is very simple and obvious: *The area between the curve* $y = f(x)$ *and the* x*-axis from* a *to* c *is the algebraic sum of the partial areas from* a *to be and from* b *to* c (see Fig. 16).

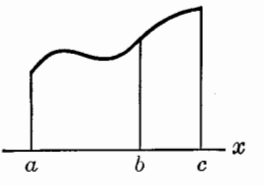


Fig. 16.

We do not in practice compute the precise values of definite integrals by direct application of the definition (1.5–3). In some simple cases, however, we may be able to find the precise value of the limit of the approximating sums by a direct examination of the sums. Usually these direct procedures involve subdivisions of the interval into equal parts.

Example 2. Find the value of $\int_a^b x^2\,dx$, assuming $0 < a < b$.
For convenience we write (by (1.5–7))

$$\int_a^b x^2\,dx = \int_0^b x^2\,dx - \int_0^a x^2\,dx.$$

We concentrate on finding the value of $\int_0^b x^2\,dx$.

Dividing $[0, b]$ into n equal parts, we write

$$x_0 = 0, x_1 = \frac{b}{n}, x_2 = \frac{2b}{n}, \ldots, x_n = \frac{nb}{n} = b.$$

Here each $\Delta x_i = \dfrac{b}{n}$. We choose $x_i' = x_i$. Then

$$\begin{aligned}\sum_{i=1}^{n} f(x_i')\,\Delta x_i &= \sum_{i=1}^{n} x_i^2\,\Delta x_i \\ &= \left[\left(\frac{b}{n}\right)^2 + \left(\frac{2b}{n}\right)^2 + \cdots + \left(\frac{nb}{n}\right)^2\right]\frac{b}{n} \\ &= \frac{b^3}{n^3}(1^2 + 2^2 + \cdots + n^2).\end{aligned}$$

There is a convenient formula for the sum of the squares of the integers from 1 to n (see Exercise 6):

$$1^2+2^2+\cdots+n^2=\frac{n(n+1)(2n+1)}{6}. \tag{1.5–8}$$

Combining the foregoing observations, we see that

$$\int_0^b x^2\,dx=\lim_{n\to\infty}\frac{n(n+1)(2n+1)b^3}{6n^3}. \tag{1.5–9}$$

Now

$$\frac{n(n+1)(2n+1)}{n^3}=2+\frac{3}{n}+\frac{1}{n^2}, \tag{1.5–10}$$

so that, as $n\to\infty$, the expression on the right in (1.5–10) approaches 2 as a limit. Hence, from (1.5–9),

$$\int_0^b x^2\,dx=\frac{2b^3}{6}=\frac{b^3}{3}.$$

Since b was arbitrary, it follows that

$$\int_0^a x^3\,dx=\frac{a^3}{3}.$$

Therefore

$$\int_a^b x^2\,dx=\frac{b^3-a^3}{3}.$$

EXERCISES

1. (a) Using four equal subintervals, calculate upper and lower sums for the integral $\int_{-1}^{3}(x^3-3x^2+3)\,dx$.

(b) Repeat (a), using eight equal subintervals.

(c) Calculate the value of the approximating sum (1.5–2), using four equal subintervals, and taking x_k' to be the midpoint of the kth subinterval.

2. (a) Calculate the value of the approximating sum (1.5–2) for the integral $\int_1^3\frac{dx}{x}$, using six equal subintervals and taking x_k' to be the midpoint of the kth subinterval.

(b) Calculate upper and lower sums for the integral in (a), using six equal subintervals. A table of reciprocals will be found convenient for this exercise.

3. Follow the instructions of Exercise 1 as applied to the integral $\int_0^4(4x^2-12x+10)\,dx$.

4. Apply the definition (1.5–3) to find the value of the integral $\int_a^b f(x)\,dx$ if $f(x)=c$, where c is a constant.

5. (a) Let $A_1(n)=1+2+\cdots+n$. Noting that $A_1(n)=n+(n-1)+\cdots+1$, show that $2A_1(n)=n(n+1)$.

(b) Using the formula for $A_1(n)$ found in (a), calculate $\int_a^b x\,dx$ by a method like that used in Example 2.

6. Let $A_2(n)=1^2+2^2+\cdots+n^2$. Obtain a formula for $A_2(n)$ as follows: Start with

$$(p+1)^3-p^3=3p^2+3p+1.$$

Write this out with $p = 0, 1, \ldots, n$, putting the results in order as shown.

$$1^3 - 0^3 = 3 \cdot 0^2 + 3 \cdot 0 + 1$$
$$2^3 - 1^3 = 3 \cdot 1^2 + 3 \cdot 1 + 1$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$(n+1)^3 - n^3 = 3 \cdot n^2 + 3 \cdot n + 1.$$

Now add, noting the cancellations on the left, to obtain

$$(n+1)^3 = 3A_2(n) + 3A_1(n) + (n+1).$$

Use the formula for $A_1(n)$ from Exercise 5(**a**) to solve for $A_2(n)$, obtaining $A_2(n) = \dfrac{n(n+1)(2n+1)}{6}$.

Imitate the foregoing procedure (**a**) to find the formula for $A_1(n)$ in this new way; (**b**) to obtain a formula for $A_3(n) = 1^3 + 2^3 + \cdots + n^3$. Then use your result to calculate $\int_a^b x^3\,dx$ by the method of Example 2.

7. Let $f(x) = e^x$ and assume $b > 0$. Using n equal subintervals and a method somewhat like that of Example 2, show that

$$\int_0^b e^x\,dx = \lim_{n\to\infty}(e^b - 1)\left(\frac{e^h - 1}{h}\right)^{-1},$$

where $h = b/n$. Use the definition of $f'(0)$ to calculate the value of the limit, and so find the value of the integral.

8. Taking $f(x) = x$, and choosing $x_i' = (x_{i-1} + x_i)/2$, show that the approximating sum (1.5–2) has a value which is independent of n and the choice of the points $x_1, x_2, \ldots, x_{n-1}$. From this result calculate $\int_a^b x\,dx$, using the definition (1.5–3).

9. The function $f(x)$ defined as $(\sin x)/x$ when $x \neq 0$ and $f(0) = 1$ is continuous for all values of x. Why? Show that the value of $f(x)$ decreases steadily as x increases from 0 to $\pi/2$. Below is given a table of values of $\sin x$. Use it to obtain high and low estimates of the value of $\int_0^{\pi/2} f(x)\,dx$.

n	1	2	3	4	5	6	7
$\sin\dfrac{n\pi}{16}$	0.1951	0.3827	0.5556	0.7071	0.8315	0.9239	0.9808

10. Calculate $\int_1^2 x^p\,dx$, where p is a positive integer, by the following procedure. Divide the interval $[1, 2]$ into n parts by the points $x_0 = 1$, $x_1 = h$, $x_2 = h^2, \ldots, x_n = h^n$, where $h = 2^{1/n}$, so that $h^n = 2$. This does not give subintervals of equal lengths, but of lengths in geometric progression. Choose $x_k' = x_{k-1}$, and show that (using (1.5–3))

$$\int_1^2 x^p\,dx = (2^{p+1} - 1)\lim_{n\to\infty}\frac{1 - 2^{1/n}}{1 - 2^{(p+1)/n}}.$$

Use the formula for the sum of a geometric progression on $(1-h)/(1-h^{p+1})$, and thus show that the limit in the above formula has the value $1/(p+1)$. Make use of the fact that $\lim_{n\to\infty} C^{1/n} = 1$ if $C > 0$. This is proved in §1.62, Exercise 16.

11. After careful study of Exercise 10, adapt the method to find the value of $\int_a^b x^p\,dx$ where $0 < a < b$ and p is a positive integer. Start with $x_0 = a$, $x_1 = ah$, $x_2 = ah^2$, etc., where $h = (b/a)^{1/n}$.

1.51 / THE MEAN VALUE THEOREM FOR INTEGRALS

The theorem which we prove in this section has an importance much like that of the law of the mean (Theorem IV, §1.2), in that it is valuable as a tool in systematic reasoning in calculus. We are introducing it here because of the use we shall have for it in our program of outlining the relation between differentiation and integration.

THEOREM VI. (*The mean value theorem for integrals.*) *Let f be continuous on the closed interval $[a, b]$. Then there is some number X such that $a \leqq X \leqq b$ and*

$$\int_a^b f(x)\,dx = (b-a)f(X). \tag{1.51–1}$$

Proof. Let m and M denote the minimum and maximum values of f on the interval. Consider the approximating sums (1.5–2) of Step 3 in the definition of the integral (§1.5).

We have

$$m \leqq f(x_i') \leqq M.$$

Therefore

$$m \sum_{i=1}^{n} \Delta x_i \leqq \sum_{i=1}^{n} f(x_i')\,\Delta x_i \leqq M \sum_{i=1}^{n} \Delta x_i.$$

But evidently

$$\sum_{i=1}^{n} \Delta x_i = b - a.$$

Hence the approximating sum (1.5–2) lies between $m(b-a)$ and $M(b-a)$. Consequently the definite integral, being the limit of the approximating sums, must also lie between these two numbers; that is,

$$m(b-a) \leqq \int_a^b f(x)\,dx \leqq M(b-a),$$

or

$$m \leqq \frac{1}{b-a}\int_a^b f(x)\,dx \leqq M. \tag{1.51–2}$$

Let us set

$$\mu = \frac{1}{b-a}\int_a^b f(x)\,dx. \tag{1.51–3}$$

We now reason that $f(x)$ must take on the value μ at some point $x = X$, $a \leqq X \leqq b$, since by (1.51–2) μ lies between the smallest and largest values of $f(x)$ on the interval. Once this argument is accepted, the proof is complete, for the equation $f(X) = \mu$ is equivalent to (1.51–1), in view of the definition of μ.

The existence of X such that $f(X) = \mu$ depends upon the hypothesis that f is

continuous. Such existence is made plausible by intuitive consideration of the variation in value of a continuous function. An indubitable proof of the existence of X must await our systematic consideration of the properties of continuous functions (in Chapter 3). The remarks which we made about the existence of m and M in connection with the proof of Rolle's theorem (§1.2) apply equally here to the existence of X.

The number μ defined by (1.51–3) is called the *average value* of the function $f(x)$ on the interval $[a, b]$. The sense in which this concept of average value is an extension of the simple notion of the arithmetic mean as an average value is indicated in Exercise 1.

EXERCISES

1. Let $[a, b]$ be divided into n equal parts, and let y_i be the value of $f(x)$ at the midpoint of the ith subinterval. The arithmetic mean of $y_1, \ldots, y_n$ is

$$A_n = \frac{y_1 + \cdots + y_n}{n}.$$

Show that $\mu = \lim_{n\to\infty} A_n$.

2. By interpreting the integral as an area, calculate the average value of $f(x) = \sqrt{a^2 - x^2}$ on the interval $-a \leqq x \leqq a$.

3. A right circular cone of altitude H and radius of base R has its axis along the x-axis. For a given value of x let $A(x)$ denote the area of cross section of the cone by a plane perpendicular to the x-axis at that point. What is the average value of $A(x)$, x ranging over all values for which the plane cuts the cone?

4. In Theorem VI it was asserted that an X can be found on the closed interval $[a, b]$ such that (1.51–1) holds. If $f(x)$ is constant on $[a, b]$, say $f(x) \equiv C$, then X may be taken as *any* point of the closed interval, for in that case $\int_a^b f(x)\,dx = C(b - a)$, and $C = f(X)$, no matter how we choose X. Hence certainly we can choose X so that $a < X < b$. Show that this can also be done if $f(x)$ is not constant on $[a, b]$. State precisely what you are taking for granted about continuous functions.

1.52 / VARIABLE LIMITS OF INTEGRATION

Before coming to the main subject of this section it will be well to consider a matter of notation. In the symbolic expression

$$\int_a^b f(x)\,dx$$

we refer to x as the *variable of integration*. The value of the integral does not depend upon the letter which is used for the variable of integration. For example,

$$\int_0^2 x^3\,dx = \int_0^2 t^3\,dt = \int_0^2 u^3\,du.$$

In cases where the limits of integration are literal symbols it is important to avoid using the same letter for a limit of integration and also for the variable of

integration. The danger of confusion is made apparent by asking: What is the value of the expression

$$\int_0^x x^3\,dx$$

when $x = 2$?

Let us now suppose that we have a function $f(t)$ which is continuous on the interval $a \leqq t \leqq b$. Regarding x as a variable, consider the function $F(x)$ defined by the integral

$$F(x) = \int_a^x f(t)\,dt. \tag{1.52–1}$$

It is clear that the integral does actually define a function of the upper limit x provided $a \leqq x \leqq b$, for once the function $f(t)$ and the lower limit $t = a$ are chosen and fixed, the integral has a definite numerical value which depends only on x. If we resort to the interpretation of the integral as an area, we may obtain a geometrical representation of the function $F(x)$ as the area under the curve $y = f(t)$ from $t = a$ to $t = x$ (see Fig. 17). It is natural to complete the definition (1.52–1) by setting $F(a) = 0$. As a matter of fact, it is usual to define

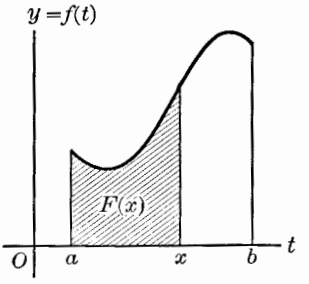


Fig. 17.

$$\int_a^a f(t)\,dt = 0 \tag{1.52–2}$$

and

$$\int_b^a f(t)\,dt = -\int_a^b f(t)\,dt, \quad a < b. \tag{1.52–3}$$

These formalities are convenient when dealing with integrals as functions of the limits of integration. With these two formulas available it is not difficult to see that formula (1.5–7) is valid for any positions of a, b, c on an interval where f is continuous.

A graph of $y = F(x)$, corresponding to a graph of $y = f(t)$ as shown in Fig. 17, would appear somewhat as in Fig. 18.

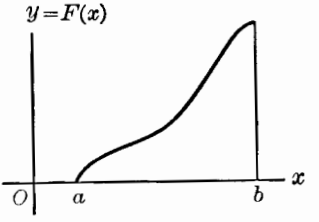


Fig. 18.

We are going to be primarily interested in the derivative of the function $F(x)$ defined by the integral (1.52–1).

THEOREM VII. *Let $f(t)$ be continuous, $a \leqq t \leqq b$, and define*

$$F(x) = \int_a^x f(t)\,dt, \quad a \leqq x \leqq b.$$

Then $F(x)$ is differentiable, with derivative

$$F'(x) = f(x). \tag{1.52–4}$$

The formula (1.52–4) *may be put verbally in the form: the derivative of the definite integral of a continuous function with respect to the upper limit of integration is equal to the value of the integrand function at this upper limit.*

Proof. We consider two points x, $x + \Delta x$ of the interval $[a, b]$. Then

$$F(x + \Delta x) - F(x) = \int_a^{x+\Delta x} f(t)\, dt - \int_a^x f(t)\, dt$$

$$= \int_x^{x+\Delta x} f(t)\, dt.$$

By the mean value theorem for integrals (Theorem VI, §1.51) we have

$$\int_x^{x+\Delta x} f(t)\, dt = \Delta x \cdot f(X),$$

where X is some number between x and $x + \Delta x$. Thus we see that

$$\frac{F(x + \Delta x) - F(x)}{\Delta x} = f(X).$$

In this equation we now hold x fixed and make Δx approach zero. Then X approaches x, and $f(X)$ approaches $f(x)$. Therefore

$$\lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = f(x).$$

The limit on the left is $F'(x)$, by definition. Thus (1.52–4) is established. It is clear from the proof that if x is at one end of the interval $[a, b]$, Δx must be restricted to have but one sign (e.g., $\Delta x > 0$ if $x = a$). In this case $F'(x)$ is a one-sided derivative.

Theorem VII furnished us with a complete solution to the problem raised in §1.4 in connection with (1.4–1). If $f(x)$ is a given function, continuous on the closed interval $[a, b]$, then the general solution of the equation

$$\frac{dy}{dx} = f(x)$$

on this interval is

$$y = \int_a^x f(t)\, dt + C,$$

where C is an arbitrary constant. This follows at once from Theorem VII and the italicized statement accompanying (1.4–6).

EXERCISES

1. The functions f, g, h are assumed to be continuous for all values of their independent variables. Complete each of the following equations:

(a) $\dfrac{d}{du}\displaystyle\int_2^u f(x)\,dx =$ (b) $\dfrac{d}{dx}\displaystyle\int_1^x g(s)\,ds =$ (c) $G(y) = \displaystyle\int_{y_0}^{y} h(t)\,dt$, $G'(y) =$

2. If $\phi(t) = \int_0^t (3+x^3)^{-1/2}\,dx$, find (a) $\phi'(0)$, (b) $\phi'(1)$, (c) $\phi'(s)$.

3. If $F(x) = \displaystyle\int_1^x \frac{\sin y}{y}\,dy$, find (a) $F'\left(\frac{\pi}{6}\right)$, (b) $F'\left(\frac{\pi}{2}\right)$, (c) $F'(\pi)$.

4. If $G(x) = \int_{-2}^{x} |s|\,ds$, find (a) $G'(-1)$, (b) $G'(0)$, (c) $G'(2)$, (d) $G'(a)$.

5. (a) If $F(x) = \int_0^x t(t-1)e^{-t^2}\,dt$, find the points of relative maxima and minima of $F(x)$. (b) What is the value of $F(0)$? (c) For what values of x is $F'(x) > 0$, and for what values of x is $F'(x) < 0$?

6. If $F(x) = \int_0^x t^3 e^{-t^3}\,dt$, find the absolute minimum value of $F(x)$.

1.53 / THE INTEGRAL OF A DERIVATIVE

The theorem which we shall prove in this section is fundamental, for it establishes the standard technique whereby definite integrals are calculated in practice. The four-step defining process of arriving at a definite integral, as set forth in §1.5, is difficult to apply. For a large and important class of integrands the following theorem provides a convenient method of finding the value of the integral.

THEOREM VIII. *Let f be a given function continuous on the closed interval $[a, b]$. Suppose that F is any differentiable function such that $F'(x) = f(x)$ when $a \leqq x \leqq b$. Then*

$$\int_a^b f(x)\,dx = F(b) - F(a). \tag{1.53–1}$$

Proof. We are by hypothesis given a function $F(x)$ whose derivative is $f(x)$. By Theorem VII (§1.52) we know another function with this same derivative, namely

$$\int_a^x f(t)\,dt.$$

Thus the function

$$G(x) = F(x) - \int_a^x f(t)\,dt$$

is constant, by Theorem V (§ 1.2), since its derivative is zero.

Now $$G(a) = F(a) - 0,$$

by (1.52–2). Also, $$G(b) = F(b) - \int_a^b f(t)\,dt.$$

Since $G(x)$ is constant we have $G(b) = G(a)$,

or

$$F(b) - \int_a^b f(t)\,dt = F(a).$$

This result is equivalent to (1.53–1), so the proof is complete.

Example 1. Find the value of the integral $\int_0^{\pi/2} \sin x\,dx$.

Applying Theorem VIII, we seek a function of x whose derivative is $\sin x$. Such a function is $-\cos x$. Therefore

$$\int_0^{\pi/2} \sin x\,dx = -\cos\frac{\pi}{2} + \cos 0 = 1.$$

We are now in a position to see clearly the connection between differentiation and integration. As concepts, by their definition, these processes are quite independent of each other. It turns out, however, that each process is in a certain sense inverse to the other. The two aspects of this mutual inverseness are displayed by Theorems VII and VIII. If we want a function defined when $a \leqq x \leqq b$ and having as its derivative a certain given continuous function $f(x)$, the class of all functions satisfying our want is the family $\int_a^x f(t)\,dt + C$. If, on the other hand, we wish to integrate a given continuous function $f(x)$, we can do so by the formula

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

provided we can find a function $F(x)$ having $f(x)$ as its derivative at all points of the interval $[a, b]$.

Example 2. Evaluate the integral $\int_{-10}^{-2} \frac{dx}{x}$.

We seek a function whose derivative is $1/x$ when $-10 \leqq x \leqq -2$. The familiar formula

$$\frac{d}{dx}\log x = \frac{1}{x} \tag{1.53–2}$$

will not quite do, for $\log x$ is not defined if $x < 0$. But, if $x < 0$, $\log(-x)$ is defined, and

$$\frac{d}{dx}\log(-x) = \frac{1}{-x}(-1) = \frac{1}{x}. \tag{1.53–3}$$

Hence, by Theorem VIII, with $f(x) = 1/x$, $F(x) = \log(-x)$, we have

$$\int_{-10}^{-2} \frac{dx}{x} = \log(-x)\Big|_{-10}^{-2} = \log 2 - \log 10 = \log \tfrac{1}{5}.$$

The formulas (1.53–2) and (1.53–3) can be combined in the single formula

$$\frac{d}{dx}\log|x| = \frac{1}{x} \text{ if } x \neq 0. \tag{1.53–4}$$

We now consider a theorem which is much used in transforming definite integrals by substitution.

THEOREM IX. *Suppose the following conditions are fulfilled:*

(1) $f(x)$ *is continuous,* $a \leqq x \leqq b$;

(2) $\phi(t)$ *and* $\phi'(t)$ *are continuous, and* $a \leqq \phi(t) \leqq b$ *when* $\alpha \leqq t \leqq \beta$;

(3) $\phi(\alpha) = a$, $\phi(\beta) = b$.

Let $g(t) = f(\phi(t))\phi'(t)$. *Then*

$$\int_a^b f(x)\,dx = \int_\alpha^\beta g(t)\,dt. \tag{1.53–5}$$

We note that the formula (1.53–5) is easily remembered in the following way: when $x = \phi(t)$, $dx = \phi'(t)\,dt$, so that $f(x)\,dx = g(t)\,dt$. The limits $x = a$, $x = b$ correspond to the limits $t = \alpha$, $t = \beta$ by (3). The proof of the theorem is left to the student (Exercise 10).

EXERCISES

1. What is wrong with the following equations?

(a) $\displaystyle\int_{-1}^{2} \frac{dx}{x^2} = -\frac{1}{x}\bigg|_{-1}^{2} = -\tfrac{1}{2} - 1 = -\tfrac{3}{2}$. **(b)** $\displaystyle\int_0^{\pi} \sec^2 x\,dx = \tan x\,\bigg|_0^{\pi} = 0$.

2. Show that $\dfrac{d}{dx}\log|x-c| = \dfrac{1}{x-c}$. Under what restrictions on a, b, c is the formula $\displaystyle\int_a^b \frac{dx}{x-c} = \log\left|\frac{b-c}{a-c}\right|$ correct?

3. Find the values of the following integrals:

(a) $\displaystyle\int_4^{10} \frac{dx}{x-2}$; **(b)** $\displaystyle\int_{-4}^{-10} \frac{dx}{x-2}$; **(c)** $\displaystyle\int_3^4 \frac{x\,dx}{x^2-25}$;

(d) $\displaystyle\int_{-4}^{-3} \frac{x\,dx}{x^2-25}$; **(e)** $\displaystyle\int_{-1}^{2} \frac{dx}{9-x^2}$; **(f)** $\displaystyle\int_4^{10} \frac{dx}{9-x^2}$.

4. Show that $\dfrac{d}{dx}\log\left|\dfrac{a+x}{a-x}\right| = \dfrac{2a}{a^2-x^2}$ if $x^2 \neq a^2$,

and hence that

$$\int_{x_1}^{x_2} \frac{dx}{a^2-x^2} = \frac{1}{2a}\log\left\{\left|\frac{a+x_2}{a-x_2}\right|\left|\frac{a-x_1}{a+x_1}\right|\right\}$$

provided that x_1 and x_2 are not separated by either of the points $x = a$, $x = -a$.

5. The student will need to recall that, by the standard conventions about principal values of the inverse sine and inverse tangent, $y = \sin^{-1} x$ is the unique y such that $x = \sin y$ and $-\pi/2 \leqq y \leqq \pi/2$, while $y = \tan^{-1} x$ is the unique y such that $x = \tan y$ and $-\pi/2 < y < \pi/2$. Find the values of the following integrals:

(a) $\displaystyle\int_{-1/2}^{1/2} \frac{dx}{\sqrt{1-x^2}}$ **(b)** $\displaystyle\int_{-1}^{\sqrt{3}} \frac{dx}{1+x^2}$ **(c)** $\displaystyle\int_{-\sqrt{3}}^{1} \frac{dx}{3+x^2}$ **(d)** $\displaystyle\int_{-2}^{2\sqrt{3}} \frac{dx}{\sqrt{16-x^2}}$.

6. Show that

$$\tan^{-1} x = \int_0^x \frac{dt}{1+t^2};$$

$$\sin^{-1} x = \int_0^x \frac{dt}{\sqrt{1-t^2}}, \quad -1 < x < 1.$$

7. Show that

$$\int_{x_1}^{x_2} \tan x \, dx = \log \left| \frac{\cos x_1}{\cos x_2} \right|, \quad -\frac{\pi}{2} < x_1 \leqq x_2 < \frac{\pi}{2}.$$

8. Standard tables of integrals list the formula

$$\int \frac{dx}{a + b\cos x} = \frac{2}{\sqrt{a^2-b^2}} \tan^{-1}\left[\frac{\sqrt{a^2-b^2}\tan\frac{x}{2}}{a+b}\right], \quad a^2 > b^2.$$

(a) Let

$$F(x) = \frac{2}{\sqrt{a^2-b^2}} \tan^{-1}\left[\frac{\sqrt{a^2-b^2}\tan\frac{x}{2}}{a+b}\right]$$

and verify that

$$F'(x) = \frac{1}{a + b\cos x}$$

whenever x is not an odd multiple of π. What can be said about $F(x)$ for these exceptional values of x?

(b) Assuming $0 < b < a$, find the limit of $F(x)$ as $x \to \pi$ from the left; as $x \to \pi$ from the right.

(c) Considering $F(x)$ as defined at π by its limit as $x \to \pi$ from the left, use Theorem VIII to show that

$$\int_0^{\pi} \frac{dx}{a + b\cos x} = \frac{\pi}{\sqrt{a^2-b^2}} \text{ if } a > b > 0.$$

(d) Find the value of $\displaystyle\int_0^{3\pi/2} \frac{dx}{5 - 3\cos x}$.

9. **(a)** Discuss critically the integration formula

$$\int \frac{dx}{a^2\cos^2 x + b^2 \sin^2 x} = \frac{1}{ab} \tan^{-1}\left(\frac{b}{a}\tan x\right),$$

assuming that a and b are positive. Compare with Exercise 8 **(a)**, **(b)**.

(b) Explain the reason for the apparent failure of Theorem VIII in the obviously false result

$$\int_0^{\pi} \frac{dx}{\cos^2 x + 4\sin^2 x} = \tfrac{1}{2}\tan^{-1}(2\tan x)\Big|_0^{\pi}$$
$$= \tfrac{1}{2}\tan^{-1}(0) - \tfrac{1}{2}\tan^{-1}(0) = 0.$$

(c) Show that

$$\int_0^{\pi/2} \frac{dx}{a^2\cos^2 x + b^2\sin^2 x} = \frac{\pi}{2ab}.$$

10. Prove Theorem IX with the aid of the following suggestions: Define $F(x) =$

$\int_a^x f(s)\,ds$, $H(t) = F(\phi(t))$. Show that $H'(t) = g(t)$ (by what theorem?). Then $H(\beta) - H(\alpha) = \int_a^\beta g(t)\,dt$ (by what theorem?). Now complete the proof of Theorem IX.

11. Show that $\int_0^{\pi/2} \sin^n x\,dx = \int_0^{\pi/2} \cos^n y\,dy$ by an appropriate substitution.

12. Show that $\int_0^1 x^m(1-x)^n\,dx = \int_0^1 x^n(1-x)^m\,dx$ by an appropriate substitution.

13. If f is continuous on $[a, b]$, show that $\int_a^b f(x)\,dx = \int_a^b f(a+b-x)\,dx$.

14. If ϕ is a continuous function on $[0, 1]$, show that $\int_0^{\pi/2} \phi(\sin x)\,dx = \int_{\pi/2}^{\pi} \phi(\sin x)\,dx$, and hence that $\int_0^{\pi} \phi(\sin x)\,dx = 2\int_0^{\pi/2} \phi(\sin x)\,dx$.

15. Show that the substitution $x = a\cos^2 t + b\sin^2 t$ fulfills condition (2) of Theorem IX for the integral $\int_a^b \sqrt{(x-a)(b-x)}\,dx$. Use the substitution to find the value of the integral.

16. Show that $\int_{-a}^{a} f(x)\,dx = \int_0^a [f(x) + f(-x)]\,dx$ if f is continuous on $[-a, a]$. What do you conclude about the value of the integral if f is an odd function? an even function? (We call f *even* if $f(-x) = f(x)$ for all x, and *odd* if $f(-x) = -f(x)$.)

17. Show that $\int_a^b xf''(x)\,dx = bf'(b) - f(b) + f(a) - af'(a)$.

18. Show that, if f is continuous on $[0, 1]$, $\int_0^{\pi} xf(\sin x)\,dx = (\pi/2)\int_0^{\pi} f(\sin x)\,dx$. Use this result to find the value of $\displaystyle\int_0^{\pi} \frac{x\sin x}{1+\cos^2 x}\,dx$.

1.6 / LIMITS

Most students of calculus get their first extensive experience with the limit concept in the course of learning about differentiation. In the process of working out formulas for the derivatives of x^n, $\sin x$, and other elementary functions, as well as in the establishment of the rules for differentiating sums, products, and quotients, students are taught to use some of the fundamental theorems about limits, such as:

> The limit of a sum is equal to the sum of the limits.
>
> The limit of a product is equal to the product of the limits.
>
> The limit of a quotient is equal to the quotient of the limits, provided the limit of the denominator is not zero.

These are not fully and precisely formulated theorems in the form here given; nevertheless, each statement conveys the central idea of an important theorem about limits. No doubt most students accept the truth of these three propositions as being intuitively evident. Probably the most that can be expected, perhaps all that is desirable, in an elementary course in calculus, is the cultivation in the student of an awareness that these propositions exist and that it is necessary to appeal to them in building up the structure of calculus. As we proceed to a more advanced level, however, it becomes more important for us to analyze the limit concept carefully, and to see how the whole theory of limiting processes is developed. As with any part of mathematics, we cannot build a clear and precise theory of limits unless we formulate our basic definitions in terms sharp enough for use in the giving of clean-cut proofs.

There are at least three recognizably distinct limit concepts in elementary calculus. We describe them briefly in turn.

(1) The limit of a function of a continuous variable: $\lim_{x\to x_0} f(x)$. As examples of this kind of limit we cite

$$(a)\ \lim_{x\to 1}\frac{6x}{x^2+2}=2,\quad (b)\ \lim_{x\to 0}\frac{\sin x}{x}=1,\quad (c)\ \lim_{x\to e}\log_e x=1.$$

The limits (1.11–2)–(1.11–4) defining a derivative are also of this type. The definition of this kind of limit was given in § 1.1 (see 1.1–2).

(2) The limit of a sequence of numbers: $\lim_{n\to\infty} s_n$.
As examples we may choose

$$(a)\ \lim_{n\to\infty}\frac{2n}{n+1}=2,\quad (b)\ \lim_{n\to\infty}\frac{\sin(n\pi/2)}{n}=0,$$

$$(c)\ \lim_{n\to\infty} 2^{1/n}=1,\quad (d)\ \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n=e.$$

Here the variable n is *discrete*, running through the natural numbers 1, 2, 3, ... We have not yet defined this kind of limit formally.

(3) The type of limit occurring in the definition of an integral:

$$\lim\sum_{i=1}^{n} f(x_i')\,\Delta x_i=\int_a^b f(x)\,dx$$

(in the notation of §1.5). Here the variable quantity is an approximating sum; it does not depend merely on n, nor does it depend on a single continuous variable x. This kind of limit is therefore different from either of the types (1) and (2).

We shall now discuss these three types of limit in more detail.

1.61 / LIMITS OF FUNCTIONS OF A CONTINUOUS VARIABLE

In speaking of $\lim_{x\to x_0} f(x)$ it is usually understood that there is some open interval (a, b) such that $a < x_0 < b$, and such that f is defined at each point of the interval except possibly at x_0 itself. We repeat the definition of a limit from §1.1:

Definition. *The function f has the limit A (a certain real number) as $x\to x_0$ provided that to each positive number ϵ there corresponds some positive number δ such that*

$$|f(x)-A|<\epsilon \quad \text{if} \quad 0<|x-x_0|<\delta \tag{1.61–1}$$

and if f is defined at x.

In certain cases f may be defined only on one side of x_0; then we speak of one-sided limits.

In all these cases we speak of x as a *continuous variable*, because it is free to assume all values on certain intervals of the real-number scale. The adjective "continuous" here contrasts with "discrete".

The foregoing definition of limit was first used systematically in the foun-

dations of calculus by the French mathematician Augustin-Louis Cauchy (1789–1857). Prior to Cauchy there had been various attacks on the problem of putting the fundamental concepts of the calculus on a sound basis. During the period from Newton and Leibniz to the work of Cauchy (roughly 1665–1821) a good part of the formal side of calculus and the technique of applying it to physics and geometry had been developed, but the reasoning was often hazy and dependent upon intuition rather than logic. The work of clarifying the fundamentals and establishing a satisfactory standard of logical rigor did not end with Cauchy, of course. A more adequate understanding of the real number system had yet to come.

We have put in this brief reference to mathematical history in order to draw a parallel. The student's understanding of calculus will normally pass through stages of development not unlike the historical ones, but with a difference. The student need not embrace all or even many of the misconceptions which have been nourished about calculus in the long evolution of the subject since the time of Newton and Leibniz, provided he will put aside preconceived notions of fundamental mathematical concepts and base his understanding on careful study of modern definitions and theorems. Intuition and experience must play their part in learning, of course.

The notation for a limit is frequently employed in conjunction with the use of the symbols $+\infty$, $-\infty$. These symbols are associated with the words "infinity" and "infinite," with which the student no doubt already has some familiarity. Our present object is to discuss the meanings of such symbolic assertions as

$$\lim_{x\to x_0} f(x) = +\infty, \quad \lim_{x\to x_0} f(x) = -\infty.$$

These statements are given various verbal renderings. The first one may be put in the form "$f(x)$ becomes positively infinite (or approaches plus infinity) as $x \to x_0$." For the second statement "positively" is replaced by "negatively," and "plus" by "minus." The meanings of statements are contained in the following definitions:

Definition. *We write $\lim_{x\to x_0} f(x) = +\infty$ if to every $M > 0$ there corresponds some $\delta > 0$ such that $M < f(x)$ whenever $0 < |x - x_0| < \delta$. We write $\lim_{x\to x_0} f(x) = -\infty$ if to every $M > 0$ there corresponds some $\delta > 0$ such that $f(x) < -M$ whenever $0 < |x - x_0| < \delta$.*

The definitions are modified in obvious ways for one-sided approach of x to x_0.

Example 1. $\lim_{x\to 0} x^{-2} = +\infty$.

To obtain this result from the definition, we consider the inequality $M < 1/x^2$, where M is any positive number. An equivalent inequality is $x^2 < 1/M$, or $|x| < M^{-1/2}$. Thus we may choose $\delta = M^{-1/2}$ and the assertion stated in Example 1 is seen to be true by definition.

A possible misconception of the definition may be allayed by the following example:

Example 2. Consider $f(x) = (1/x^2)\sin^2(\pi/x)$ as $x \to 0$.

In this case it is *not* true that $\lim_{x\to 0} f(x) = +\infty$, in spite of the fact that f takes on values as large as we please as near 0 as we please. For, considering the definition, choose $M = 1$. The inequality $(1/x^2)\sin^2(\pi/x) > 1$ is true for *some* values of x very close to 0, e.g., $x = \frac{2}{3}, \frac{2}{5}, \frac{2}{7}, \frac{2}{9}, \ldots$; but for certain other values of x near 0 it is not true, e.g., $x = \frac{1}{2}, \frac{1}{3}, \ldots, 1/n, \ldots$.

The symbols $+\infty$, $-\infty$ are also used in connection with the independent variable. We use $x \to +\infty$ as a symbolic equivalent of the phrase "x tends to plus infinity" (or "x becomes positively infinite"). This occurs in such contexts as $\lim_{x\to+\infty} f(x) = A$, which by definition means "to each $\epsilon > 0$ there corresponds an $M > 0$ such that $|f(x) - A| < \epsilon$ whenever $M < x$."

Example 3. $\lim_{x\to+\infty}(2x - 1)/(x - 3) = 2$.

To verify this assertion by direct application of the definition, we proceed as follows:

$$\frac{2x-1}{x-3} - 2 = \frac{2x - 1 - 2x + 6}{x-3} = \frac{5}{x-3}.$$

Therefore, if $x > 3$ we have

$$\left|\frac{2x-1}{x-3} - 2\right| = \frac{5}{x-3} < \epsilon$$

provided $(5/\epsilon) < x - 3$, or $3 + (5/\epsilon) < x$. This shows that, for given $\epsilon > 0$, the conditions of the definition are fulfilled with $M = 3 + (5/\epsilon)$.

The following theorem states explicitly an important principle which is frequently used in mathematical arguments.

THEOREM X. *Let $f(x)$ be defined at all points except $x = a$ of some open interval containing that point, and suppose that, as x tends to a, $f(x)$ approaches a limit which is positive. Then $f(x)$ is positive when x is sufficiently near the point a.*

Proof. Let the limit of $f(x)$ be A. Then, according to the definition, if any positive number ϵ is given, there is some corresponding positive number δ such that $|f(x) - A| < \epsilon$ when $0 < |x - a| < \delta$. Now $A > 0$, by hypothesis. Suppose that we take $A/2$ for the ϵ. Then there is a certain δ such that $|f(x) - A| < A/2$ if $0 < |x - a| < \delta$. Now $|f(x) - A| < A/2$ is equivalent to the double inequality $-A/2 < f(x) - A < A/2$, as the student will easily see. Using the left one of these last two inequalities, and transposing, we see that $-A/2 + A < f(x)$ when $0 < |x - a| < \delta$. Certainly then $f(x)$ is positive, for $A/2 > 0$. This completes the proof.

Example 4. As an illustration of the use of Theorem X, we shall prove the following statement: *Let f be defined in some open interval containing $x = a$, and*

suppose that $f'(a)$ exists and is positive. Then, for points x_1, x_2 sufficiently close to $x = a$, $x_1 < a < x_2$ implies

$$f(x_1) < f(a) < f(x_2). \tag{1.61–2}$$

To prove this assertion, consider the definition

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}.$$

We are assuming that $f'(a) > 0$; consequently, by Theorem X,

$$\frac{f(x) - f(a)}{x - a} > 0 \tag{1.61–3}$$

when x is sufficiently close to a. But from (1.61–3) we infer that $f(x) > f(a)$ if $x > a$, and $f(x) < f(a)$ if $x < a$. Thus the assertion about the inequalities (1.61–2) is seen to be true.

We shall state two further theorems which, like Theorem X, are frequently used in the kind of reasoning with limits which occurs regularly in calculus.

THEOREM XI. *Let f be defined at all points except $x = a$ of some open interval containing that point. Suppose that there is a number M such that $f(x) \leqq M$ when x is sufficiently near a. Further suppose that $\lim_{x \to a} f(x) = A$. Then $A \leqq M$.*

THEOREM XII. *Let f, g, h be functions of x defined at all points except $x = a$ of some open interval containing that point. Suppose that $f(x) \leqq g(x) \leqq h(x)$, and suppose that the limits $\lim_{x \to a} f(x)$, $\lim_{x \to a} h(x)$ exist and are equal. Then $\lim_{x \to a} g(x)$ exists also, and all three limits are equal.*

We leave the proofs of these theorems as exercises for the student, but we shall use the theorems whenever the need arises. These theorems have analogues for other kinds of limits (e.g., limits of sequences and the limits defining definite integrals). For example, the proof of (1.51–2) employed a principle similar to that of Theorem XI, as applied to the integral as the limit of approximating sums. One of the standard proofs that $\lim_{x \to 0} (\sin x)/x = 1$ uses the principle of Theorem XII, with $f(x) = \cos x$, $g(x) = (\sin x)/x$, $h(x) = 1/\cos x$.

EXERCISES

1. Show that $\lim_{x \to 0} (\cos x)/x^2 = +\infty$. SUGGESTION: Note that $\cos x \geqq \frac{1}{2}$ if $|x| \leqq \pi/3$. If $M > 0$ is given, choose δ as the lesser of the numbers $\pi/3$, $(2M)^{-1/2}$, and explain why $M < (\cos x)/x^2$ if $|x| < \delta$.

2. Show that $\lim_{x \to 0} (1 + \sin^2 x)/x^2 = +\infty$.

3. What is $\lim_{x \to 0} (\sin^2 x)/x^4$? Justify your answer.

4. Suppose $f(x) \geqq m > 0$ and $g(x) > 0$ when x is any point in an interval $a < x < b$

containing x_0. Also suppose $g(x) \to 0$ as $x \to x_0$. Show that

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = +\infty.$$

5. Show that $\lim_{x \to +\infty} (2 + \sin x)/e^{-x} = +\infty$.

6. Given $\epsilon > 0$, find M so that

$$\left| \frac{9x+4}{3x-1} - 3 \right| < \epsilon \text{ if } M < x.$$

What statement about a limit does this prove?

7. Suppose that, as $x \to +\infty$ (or as $x \to x_0$), $|f(x)| \leqq M$, where M is a constant, and that $|g(x)| \to +\infty$. Show that $\dfrac{f(x)}{g(x)} \to 0$.

8. Write $f(x) = \dfrac{4x^2 - x - 1}{x^2 + 2x - 3} = \dfrac{4 - (1/x) - (1/x^2)}{1 + (2/x) - (3/x^2)}$.

Then find $\lim_{x \to +\infty} f(x)$, using the rules for limits of sums and quotients.

9. If $P(x)$ and $Q(x)$ are polynomials of degrees m and n respectively, discuss $\lim_{x \to +\infty} \dfrac{P(x)}{Q(x)}$ according as $m > n$, $m = n$, $m < n$. Show that the results when $x \to -\infty$ are the same as when $x \to +\infty$, if $m \leqq n$ or if $m - n$ is even.

10. In the following cases discuss the behavior of the given function $f(x)$ as $x \to +\infty$. Does $f(x)$ approach a number as limit? Does $f(x)$ tend to $+\infty$ or to $-\infty$? Does it do none of these things?

(a) $\sin x$; **(b)** $\sin(1/x)$; **(c)** $(\sin x)/x$; **(d)** $x \sin x$; **(e)** $x^2 + x \sin x$; **(f)** $x^2 + x^2 \cos x$; **(g)** $x + x^2 \sin x$.

11. If $\lim_{h \to 0} \dfrac{f(x+h) - f(x)}{h} = +\infty$ we say that $f'(x) = +\infty$.

Likewise we define the meaning of $f'(x) = -\infty$. **(a)** If $f(x) = x^{1/3}$, what is $f'(0)$? **(b)** If $f(x) = x^{2/3}$, what about $f'(0)$? What about $f'_+(0)$ and $f'_-(0)$? **(c)** If $f(x) = 0$ when $x < 0$, $f(0) = 1$ and $f(x) = 2$ when $x > 0$, show that $f'(0) = +\infty$. This shows that $f'(x_0) = +\infty$ does not imply that f is continuous at x_0.

12. Prove that the law of the mean is true with the following hypotheses, which are weaker than those imposed in § 1.2; f is continuous when $a \leqq x \leqq b$, and, for each x of the open interval $a < x < b$, f is either differentiable (i.e., $f'(x)$ exists as a finite limit) or $f'(x)$ is $+\infty$ or $-\infty$ as defined in Exercise 11.

13. Prove Theorem XI and Theorem XII. For Theorem XI begin by supposing $M < A$, and show that this leads to a contradiction. For Theorem XII let $f(x)$ and $h(x)$ approach A as limit, and note that, if $\epsilon > 0$, the inequalities $A - \epsilon < f(x) < A + \epsilon$, $A - \epsilon < h(x) < A + \epsilon$ must hold when $|x - a|$ is sufficiently small.

14. State and prove a theorem similar to Theorem XI in which the inequalities are reversed.

1.62 / LIMITS OF SEQUENCES

A sequence is an ordered set of numbers in one-to-one correspondence with the positive integers. This correspondence may be shown by numbering the terms of

the sequence in order:

$$s_1, s_2, s_3, \ldots, s_n, \ldots.$$

The sequence is then denoted symbolically by $\{s_n\}$. As examples we cite:

(*a*) $2, 4, 6, 8, \ldots, 2n, \ldots$;

(*b*) $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots, 1/(2n), \ldots$;

(*c*) $\frac{1}{2}, \frac{2}{5}, \frac{3}{10}, \frac{4}{17}, \ldots, n/(n^2+1), \ldots$;

(*d*) $1, 1+\frac{1}{2}, 1+\frac{1}{2}+\frac{1}{3}, \ldots, 1+\frac{1}{2}+\cdots+(1/n), \ldots$;

(*e*) $1, 0, 1, 0, \ldots \left(s_n = \frac{1-(-1)^n}{2}, n = 1, 2, \ldots\right)$.

A sequence is in fact a particular kind of function, a function whose independent variable n ranges over the set of positive integers. We could (and ocassionally do) use the functional notation $f(n)$ for a sequence, but notations such as $\{x_n\}$, $\{s_n\}$, $\{a_n\}$ are more common. Observe that $\{s_n\}$ is the symbol for the sequence (function) as a whole, whereas s_n is the symbol for the nth term (the value of the function).

Sometimes it is convenient to have a notation in which the terms are numbered $0, 1, 2, \ldots$ instead of $1, 2, 3, \ldots$. For instance, if the sequence is $1, 2, 4, 8, 16, 32, \ldots$, it is convenient to denote it by $s_0, s_1, s_2, \ldots, s_n$, so that $s_n = 2^n$ (note that $s_0 = 2^0 = 1$).

The definition of the limit of a sequence is quite similar to the definition of the limit of a function of a continuous variable, as given in §1.61. Thinking of the sequence $\{s_n\}$ as a function, let us compare $\{s_n\}$ with a function $f(x)$ of the continuous variable x. The definition of $\lim_{n\to\infty} s_n$ is then very similar to the definition of $\lim_{x\to+\infty} f(x)$ (just preceding Example 3 in §1.61).

Definition. *We say that* $\lim_{n\to\infty} s_n = A$ *if for each positive* ϵ *there is some integer* N *depending on* ϵ *such that* $|s_n - A| < \epsilon$ *whenever* $N \leqq n$. *In this case we say that the sequence* $\{s_n\}$ *is convergent and that it has the limit* A.

Example 1. We shall show that $\lim_{n\to\infty}(\frac{2}{3})^n = 0$.

Suppose $\epsilon > 0$ is given. We wish to find N so that $N \leqq n$ will insure $(\frac{2}{3})^n < \epsilon$, or, what is equivalent, $1/\epsilon < (\frac{3}{2})^n$. Now, in the sequence $\{(\frac{3}{2})^n\}$, each term is half again as large as its predecessor, and the first term is $\frac{3}{2}$. Hence certainly $(\frac{3}{2})^n > n(\frac{1}{2})$. Thus, to get $1/\epsilon < (\frac{3}{2})^n$, it is amply sufficient to have $1/\epsilon < n(\frac{1}{2})$, or $2/\epsilon < n$. Therefore we take N as the first integer which is greater than $2/\epsilon$, and then certainly $(\frac{2}{3})^n < \epsilon$ if $N \leqq n$. This is proof that $\lim_{n\to\infty}(\frac{2}{3})^n = 0$.

Example 2. We now show that $\lim_{n\to\infty} a^n = 0$ if $0 < a < 1$. This includes Example 1 as a special case.

We can express a in the form

$$a = \frac{1}{1+h}, \quad h > 0.$$

This is because $0 < a < 1$; h is given by

$$h = \frac{1-a}{a}.$$

We now use (1.2–5), with $\alpha = n$;

$$a^{-n} = (1+h)^n \geqq 1 + nh.$$

This implies

$$a^n \leqq \frac{1}{1+nh}.$$

Suppose now that $\epsilon > 0$ is given. We wish to find an N such that $N \leqq n$ implies $a^n < \epsilon$. It is sufficient to have

$$\frac{1}{1+nh} < \epsilon, \quad \text{or} \quad \frac{1}{\epsilon} < 1 + nh, \quad \text{or} \quad \frac{1-\epsilon}{\epsilon} < nh.$$

Since $h > 0$, some multiple of h will exceed $(1-\epsilon)/\epsilon$. Let N be a positive integer such that $(1-\epsilon)/\epsilon < Nh$. Then $N \leqq n$ *implies* $a_n < \epsilon$, and the proof is complete.

Theorems X, XI, and XII have analogues for sequences. The student can easily formulate these analogues alone. The general theorems about sums, products, and quotients (see §1.6) also apply to sequences.

Example 3. Find

$$\lim_{n\to\infty} \frac{2n^3 + n^2 - 7n}{n^3 + 2n + 2}.$$

We write the general term of the sequence in the form

$$\frac{2 + (1/n) - (7/n^2)}{1 + (2/n^2) + (2/n^3)}.$$

By the theorem on sums,

$$\lim_{n\to\infty}\left(2 + \frac{1}{n} - \frac{7}{n^2}\right) = 2 \quad \text{and} \quad \lim_{n\to\infty}\left(1 + \frac{2}{n^2} + \frac{2}{n^3}\right) = 1.$$

By the theorem on quotients,

$$\lim_{n\to\infty} \frac{2 + (1/n) - (7/n^2)}{1 + (2/n^2) + (2/n^3)} = \frac{2}{1}.$$

Hence the required limit is 2.

Many important sequences have a property which is described by the word *monotonic*. A sequence $\{s_n\}$ is called *nondecreasing* if $s_1 \leqq s_2 \leqq s_3 \leqq \cdots$, i.e., if $s_n \leqq s_{n+1}$ for every positive integer n. It is called *strictly increasing* (or just *increasing*) if $s_1 < s_2 < s_3 < \cdots$. For example, the sequence

$$1, 1, 2, 2, 3, 3, \ldots$$

is nondecreasing. The nth term of this sequence is given by $\left[\frac{n+1}{2}\right]$ (the greatest

integer $\leqq \frac{n+1}{2}$). The sequence $\{s_n\}$, where $s_n = \frac{n}{n+1}$, is strictly increasing:

$$\tfrac{1}{2}, \tfrac{2}{3}, \tfrac{3}{4}, \ldots.$$

Likewise, we call a sequence *nonincreasing* if $s_n \geqq s_{n+1}$ for every n, and (*strictly*) *decreasing* if $s_n > s_{n+1}$ for every n. A sequence of any one of these four types is called *monotonic*.

One very important type of nondecreasing sequence is the sequence of decimal approximations to a positive number. For example, let s_n be the number obtained by writing the first n decimal places of the number $\frac{1}{3}$:

$$s_1 = 0.3,\ s_2 = 0.33,\ s_3 = 0.333, \ldots.$$

The general formula for s_n may be written

$$s_n = \frac{3}{10} + \frac{3}{10^2} + \cdots + \frac{3}{10^n}.$$

Using the formula

$$a + ar + ar^2 + \cdots + ar^{n-1} = a\left(\frac{1-r^n}{1-r}\right) \tag{1.62–1}$$

for the sum of a geometric progression, we find

$$s_n = \tfrac{3}{10}\,\frac{1-(\frac{1}{10})^n}{1-\frac{1}{10}} = \tfrac{1}{3}[1-(\tfrac{1}{10})^n].$$

We see from this that

$$\lim_{n\to\infty} s_n = \tfrac{1}{3}. \tag{1.62–2}$$

When we write

$$\tfrac{1}{3} = 0.333\ldots,$$

the meaning is exactly that expressed by (1.62–2).

Similar remarks apply to all nonterminating decimal representations, e.g.,

$$\tfrac{2}{3} = \lim_{n\to\infty} s_n, \quad s_n = 0.\underbrace{66\ldots6}_{n}, \quad \sqrt{3} = \lim_{n\to\infty} x_n,$$

where $x_1 = 1.7$, $x_2 = 1.73$, $x_3 = 1.732$, $x_4 = 1.7321$, etc.

Suppose that $\{s_n\}$ is a nondecreasing sequence. There are just two possibilities: either (1) there is some number M such that $s_n \leqq M$ for every n, or (2) $s_n \to +\infty$ (this means that no matter what M is chosen, $M < s_n$ for all sufficiently large values of n). In case (1) we say that the sequence is *bounded above*. Likewise, for a nonincreasing sequence $\{s_n\}$ there are just two possibilities: either (1) the sequence is *bounded below*, i.e., there is some number M such that $M \leqq s_n$ for every n, or (2) $s_n \to -\infty$ as $n \to \infty$.

A sequence is called *bounded* if it is bounded both above and below.

Observe that an increasing or nondecreasing sequence is always bounded *below*, since $s_1 \leqq s_n$ for every n. Likewise, a decreasing or nonincreasing sequence is bounded above, since $s_n \leqq s_1$.

There is a very important theorem which asserts that a monotonic sequence is convergent (i.e., has a limit) provided it is bounded.

THEOREM XIII. *Suppose that either*

$$(a)\ s_n \leqq s_{n+1} \quad \text{and} \quad s_n \leqq M,\ n = 1, 2, \ldots,$$

or

$$(b)\ s_{n+1} \leqq s_n \quad \text{and} \quad M \leqq s_n,\ n = 1, 2, \ldots,$$

where M is a constant. Then $\lim_{n\to\infty} s_n$ exists.

The great importance of this theorem lies in the fact that by using it we can be sure that certain sequences are convergent without knowing precisely what the limits are. The theorem is proved in Chapter 2 (Theorems III and IV). The proof is based upon a careful discussion of the nature of the real number system.

Example 4. Let $s_n = \dfrac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2 \cdot 4 \cdot 6 \cdots (2n)}$. Show that this sequence is monotonic and bounded, and therefore convergent.

In order to make sure the notation is understood, let us write a few terms of the sequence, by substituting successively $n = 1, 2, 3, \ldots$. We have

$$s_1 = \frac{1}{2},\ s_2 = \frac{1 \cdot 3}{2 \cdot 4} = \frac{3}{8},\ s_3 = \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6} = \frac{5}{16},\ \text{etc.}$$

The sequence is decreasing. For,

$$s_2 = \tfrac{3}{4}s_1,\ s_3 = \tfrac{5}{6}s_2,\ s_4 = \tfrac{7}{8}s_3,$$

and in general

$$s_{n+1} = \frac{2n+1}{2n+2}\, s_n < s_n.$$

All the terms are positive; so that $0 < s_n \leqq \frac{1}{2}$. Thus the sequence is bounded. Hence it must have a limit. This argument does not show what the limit may be.

Example 5. Consider the sequence defined by

$$s_n = \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!},$$

the first few terms of which are

$$s_1 = 1,\ s_2 = 1 + \frac{1}{1 \cdot 2} = \frac{3}{2},\ s_3 = 1 + \frac{1}{1 \cdot 2} + \frac{1}{1 \cdot 2 \cdot 3} = \frac{5}{3}.$$

This sequence is plainly monotonic, for

$$s_{n+1} = s_n + \frac{1}{(n+1)!},\ s_n < s_{n+1}.$$

We shall show that the sequence is bounded above. Now, if $n > 2$,

$$\frac{1}{n!} = \frac{1}{1 \cdot 2 \cdot 3 \cdots n} < \frac{1}{\underbrace{2 \cdot 2 \cdots 2}_{n-1}} = \frac{1}{2^{n-1}}.$$

Therefore, if $n > 2$,

$$s_n < 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{n-1}}.$$

But, by (1.62–1) taking $a = 1$, $r = \frac{1}{2}$, we have

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

Therefore $s_n < 2$ for all n. Consequently, by Theorem XIII, the sequence $\{s_n\}$ is convergent. The theorem does not tell us the exact value of the limit, though of course we can see that it is not larger than 2.

Example 6. Let $\{s_n\}$ be the sequence

$$s_n = \left(1 + \frac{1}{n}\right)^n.$$

We shall show that the sequence is increasing and bounded above. First we write the binomial expansion

$$(1 + a)^n = 1 + na + \frac{n(n-1)}{1 \cdot 2} a^2 + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} a^3 + \cdots + a^n.$$

The coefficient of $a^k (1 \leqq k \leqq n)$ is

$$\frac{n(n-1)(n-2) \cdots (n-k+1)}{k!}.$$

Putting $1/n$ in place of a, we have

$$\left(1 + \frac{1}{n}\right)^n = 1 + n\left(\frac{1}{n}\right) + \frac{n(n-1)}{1 \cdot 2}\left(\frac{1}{n}\right)^2 + \cdots + \left(\frac{1}{n}\right)^n.$$

The expression on the right has $n + 1$ terms, a typical one of which is

$$\frac{n(n-1)(n-2) \cdots (n-k+1)}{k! n^k} = \frac{1}{k!} \cdot 1 \cdot \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right).$$

Thus

$$\left(1 + \frac{1}{n}\right)^n = 1 + 1 + \frac{1 - \frac{1}{n}}{2!} + \frac{\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)}{3!} + \cdots + \frac{\left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{n-1}{n}\right)}{n!}. \qquad (1.62\text{–}3)$$

Now suppose that n is replaced by $n+1$, so as to form the corresponding formula for the expression

$$\left(1+\frac{1}{n+1}\right)^{n+1}.$$

We see from (1.62–3) that each of the numerators after the first two terms on the right in (1.62–3) is increased when n is replaced by $n+1$. Moreover, the total number of terms on the right is increased from $n+1$ to $n+2$. Hence it is certainly the case that

$$\left(1+\frac{1}{n}\right)^{n} < \left(1+\frac{1}{n+1}\right)^{n+1}.$$

This shows that our sequence is increasing. Moreover, from (1.62–3) it is clear that

$$\left(1+\frac{1}{n}\right)^{n} < 1+1+\frac{1}{2!}+\frac{1}{3!}+\cdots+\frac{1}{n!}.$$

In Example 5, we saw that

$$1+\frac{1}{2!}+\frac{1}{3!}+\cdots+\frac{1}{n!}<2.$$

Therefore

$$\left(1+\frac{1}{n}\right)^{n}<3. \tag{1.62–4}$$

This shows that our sequence is bounded above. It therefore has a limit. The limit is denoted by the letter e:

$$e=\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{n}. \tag{1.62–5}$$

This number e is taken as the base in defining natural logarithms, of which we assume the student already has a working knowledge.

The definition of the limit of a sequence enables us (theoretically, at least) to decide whether any specified number A is or is not the limit of a given sequence $\{s_n\}$. To make the decision in a given case we must work with inequalities to form an estimate of the magnitude of the difference $s_n - A$ as n becomes larger and larger. In practice we often find the limits of sequences by using the theorem on limits of sums, products, and quotients. In this procedure the given sequence is expressed in terms of other sequences whose limits we already know.

There are many cases, however, in which we cannot find the limit of a sequence by direct use of the definition or by using the theorem on limits of sums, products, and quotients. It is very important to be able to recognize, by intrinsic characteristics of the sequence itself, that the sequence is convergent. Then we can say with certainty: "By virtue of such and such a characteristic possessed by this sequence $\{s_n\}$, *there must exist* a number A such that

$\lim_{n\to\infty} s_n = A$." One such feature is the property of being monotonic and bounded. Not all convergent sequences possess this characteristic, of course, but it is nevertheless of great importance.

In higher analysis there are many situations in which we assert the *existence* of something having specified properties. A problem leading to such an assertion may be called an *existence problem.* Many important concepts are introduced via existence problems. Definite integrals and functions defined by infinite series are examples.

EXERCISES

1. Which of the following sequences is convergent? **(a)** $\{(-1)^n\}$, **(b)** $\left\{\frac{(-1)^n}{n}\right\}$, **(c)** $\{n(-1)^n\}$, **(d)** $\{n[1-(-1)^n]\}$, **(e)** $\{(-1)^{2n+1}\}$, **(f)** $\left\{\frac{\sin n}{n}\right\}$.

2. If $s_n = \frac{n^2+n-1}{3n^2+1}$ and $\epsilon > 0$, find N so that $|s_n - \frac{1}{3}| < \epsilon$ if $N \leqq n$.

3. If $s_n = \frac{6n^3+2n+1}{n^3+n^2}$, find A so that if $0 < \epsilon < 6$, $|s_n - A| < \epsilon$ provided $n > \frac{6-\epsilon}{\epsilon}$. What does this show about $\lim_{n\to\infty} s_n$?

4. If $s_n = \frac{10^n}{n!}$, show that $s_n \leqq \frac{10^{10}}{10!} \cdot \frac{10}{n}$ if $n > 10$. Hence, if $\epsilon > 0$, find N so that $|s_n| < \epsilon$ if $N \leqq n$. From what value of n onward is $s_{n+1} < s_n$?

5. Let $s_1 = \frac{2}{1} \cdot \frac{1}{1^2}$, $s_2 = \frac{2\cdot 4}{1\cdot 3} \cdot \frac{1}{2^2}$, $s_3 = \frac{2\cdot 4\cdot 6}{1\cdot 3\cdot 5} \cdot \frac{1}{3^2}$, etc. Write the general expression for s_n and show that $s_n \leqq \frac{2}{n}$. What do you conclude about $\lim_{n\to\infty} s_n$? Show that $s_{n+1} < s_n$.

6. Let $P(x) = a_0x^p + a_1x^{p-1} + \cdots$ and $Q(x) = b_0x^q + b_1x^{q-1} \cdots$ be polynomials, with $a_0b_0 \neq 0$. Discuss $\lim_{n\to\infty} P(n)/Q(n)$ according as **(a)** $p < q$, **(b)** $p = q$, **(c)** $p > q$. (Compare with Example 3.)

7. **(a)** If $P(x)$ is a polynomial of degree $r > 0$, show that $\lim_{n\to\infty} P(n+1)/P(n) = 1$. **(b)** What is $\lim_{n\to\infty} P(2n)/P(n)$?

8. Find

$$\lim_{n\to\infty} \frac{1-\left(1-\frac{1}{n}\right)^4}{1-\left(1-\frac{1}{n}\right)^3}.$$

9. Using rationalization techniques, find the limits of the following sequences: **(a)** $\{\sqrt{n+1}-\sqrt{n}\}$; **(b)** $\{\sqrt{n}(\sqrt{n+1}-\sqrt{n})\}$.

10. If $\lim_{n\to\infty} \frac{x_n - x}{x_n + x} = 0$, show that $\lim_{n\to\infty} x_n = x$. (Write $y_n = \frac{x_n - x}{x_n + x}$ and solve for x_n.)

11. Let $f(x)$ be defined as $f(x) = \lim_{n\to\infty} \frac{1}{x^n + x^{-n}}$, where $x > 0$. Find the value of $f(x)$ for each positive x.

12. Consider $\lim_{n\to\infty}\left(\frac{1-x^2}{1+x^2}\right)^n$. For what values of x does the limit exist? Classify the values of x according to the value of the limit.

13. Let $f(x) = x/|x|$ if $x \neq 0$, and define $f(0) = 0$. Show that $f(x) = \lim_{n\to\infty}(2/\pi)\tan^{-1}(nx)$.

14. Find each of the following limits:

(a) $\lim_{n\to\infty}\left(\frac{1}{n^2}+\frac{2}{n^2}+\cdots+\frac{n}{n^2}\right)$;

(b) $\lim_{n\to\infty}\left(\frac{1^2}{n^3}+\frac{2^2}{n^3}+\cdots+\frac{n^2}{n^3}\right)$.

15. If $s_n=\frac{1}{\sqrt{n^2+1}}+\frac{1}{\sqrt{n^2+2}}+\cdots+\frac{1}{\sqrt{n^2+n}}$, show that $\left(1+\frac{1}{n}\right)^{-1/2}<s_n<1$, and hence find $\lim_{n\to\infty} s_n$.

16. (a) Prove that, if $C>1$, $\lim_{n\to\infty} C^{1/n}=1$, using the following suggestions. First explain why $C^{1/n}>1$. Then let $x_n=C^{1/n}-1$ and use (1.2–5) to show that $C>1+nx_n$. It then follows that $\lim_{n\to\infty} x_n=0$. (Why?) Why then does $C^{1/n}\to 1$? (b) If $0<C<1$, show that $\lim_{n\to\infty} C^{1/n}=1$ by applying the *result* (not the method) of (a).

17. Let $a_n=n^{1/2n}$. For $n>1$ write $a_n=1+x_n$. Use (1.2–5) to prove that $x_n\leqq\frac{\sqrt{n}-1}{n}<\frac{1}{\sqrt{n}}$. Use this to deduce that $n^{1/n}<1+\frac{2}{\sqrt{n}}+\frac{1}{n}$, and hence that $\lim_{n\to\infty} n^{1/n}=1$.

18. Let $s_n=n/C^n$, where $C>1$. Write $\sqrt{C}=1+x$ and use (1.2–5) to show that $\sqrt{s_n}<1/(x\sqrt{n})$. What do you conclude about $\lim_{n\to\infty} s_n$?

19. Let $s_n=\left(1+\frac{1}{n}\right)^{n+1}$. Show that $\frac{s_{n-1}}{s_n}=\frac{n-1}{n}\left(1+\frac{1}{n^2-1}\right)^{n+1}$, and then use (1.2–5) to prove that $s_{n-1}>s_n$. Show that $\lim_{n\to\infty} s_n=e$ (see (1.62–5)).

20. (a) Show that $\lim_{n\to\infty}\left(1+\frac{1}{2n}\right)^n=\sqrt{e}$.

(b) Observe that $\left(1+\frac{2}{n}\right)=\left(1+\frac{1}{n+1}\right)\left(1+\frac{1}{n}\right)$. Now find $\lim_{n\to\infty}\left(1+\frac{2}{n}\right)^n$.

(c) Find $\lim_{n\to\infty}\left(1+\frac{3}{n}\right)^n$.

21. Suppose $0<s_n$ and $s_{n+1}<rs_n$, where r is a constant such that $0<r<1$. Show that $\lim_{n\to\infty} s_n=0$.

22. Suppose $0<s_n$ and $s_{n+1}\geqq rs_n$, where r is a constant such that $r>1$. Show that $\lim_{n\to\infty} s_n=+\infty$.

23. Suppose $C>1$ and $s_n=C^{1/n}$. Show that $1<s_n$ and $s_{n+1}<s_n$ (assume the contrary in each case, and deduce a contradiction). Hence, by Theorem XIII, $\lim_{n\to\infty} s_n$ exists. Denote the limit by r. Why is $r\geqq 1$? Prove that $r=1$ by showing that $r>1$ leads to the contradiction that $1/C\leqq 0$.

24. Let $a_n=1+\frac{1}{1!}+\frac{1}{2!}+\cdots+\frac{1}{n!}$, $b_n=\left(1+\frac{1}{n}\right)^n$. In the course of Example 6 it was proved that $b_n<a_n$. Let $a=\lim_{n\to\infty} a_n$ (the existence of the limit follows from Example 5). By definition, $e=\lim_{n\to\infty} b_n$ (see (1.62–5)). Show that $a=e$, using the following suggestions: First explain why you know that $e\leqq a$. Then refer to (1.62–3) and explain why

$$1+\frac{1}{1!}+\frac{\left(1-\frac{1}{n}\right)}{2!}+\frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)}{3!}+\cdots+\frac{\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{p-1}{n}\right)}{p!}<e$$

if $1<p<n$. In this result let $n\to\infty$ and obtain $a_p\leqq e$. Now complete the argument as to why $a=e$.

1.63 / THE LIMIT DEFINING A DEFINITE INTEGRAL

Here we have to do with a limit different in kind from the limit of a function of a continuous variable and the limit of a sequence. Our discussion will be couched in terms of the notation used in the definition of a definite integral in §1.5. Consider the approximating sums

$$\sum_{i=1}^{n} f(x_i') \, \Delta x_i \tag{1.63–1}$$

associated with a given fixed function f which is defined on the interval $a \leqq x \leqq b$. Although the index n appears here, these sums do not form a sequence. A particular sum depends not merely on n, but on the points of subdivision $x_1, x_2, \ldots, x_{n-1}$ and on the intermediate points $x_1', \ldots, x_n'$. We may say that, the function and the interval being fixed, the approximating sum (1.63–1) is a function of n and of the points $x_1, x_1, \ldots, x_{n-1}, x_1', \ldots, x_n'$. When we say that

$$\lim \sum_{i=1}^{n} f(x_i') \, \Delta x_i = \int_a^b f(x) \, dx, \tag{1.63–2}$$

we mean that to each positive number ϵ corresponds another positive number δ such that

$$\left| \sum_{i=1}^{n} f(x_i') \, \Delta x_i - \int_a^b f(x) \, dx \right| < \epsilon$$

for all choices of n and the points x_i, x_i' such that the greatest of the numbers $\Delta x_1, \ldots, \Delta x_n$ is less than δ.

We give this definition here for comparison with the definitions of the two kinds of limits already discussed. The problem of showing that the sums (1.63–1) do actually have a limit (when f is a continuous function) is another example of an "existence problem" for a real number. (See the two final paragraphs of §1.62.) As with most such problems, we cannot obtain a solution without having at our command a systematic knowledge of the fundamentals of the real number system.

1.64 / THE THEOREM ON LIMITS OF SUMS, PRODUCTS, AND QUOTIENTS

In § 1.6 we stated three fundamental theorems about limits [see also (1.1–4)–(1.1–6)]. The student has used these theorems from the very beginning of his study of calculus. We are now going to give formal statement and proof of the propositions as they apply to limits of functions of a continuous variable. First, however, it will be convenient to consider certain rules governing the use of absolute values. The absolute value $|A|$ of a number A was defined in Example 3, §1.1. Now for any two numbers A, B, it is always true that

$$|A + B| \leqq |A| + |B|. \tag{1.64–1}$$

There are four cases to consider: (1) A and B both positive, (2) A and B both negative, (3) $A = 0$ or $B = 0$, (4) A and B of opposite sign, and neither of them equal to 0. In the first three cases it is easily seen that $|A + B| = |A| + |B|$, while in case (4) $|A + B| < |A| + |B|$. Thus (1.64–1) is true in all cases.

We now turn to the limit theorem.

THEOREM XIV. *Let f and g be defined in an interval containing $x = x_0$, but not necessarily at the point $x = x_0$ itself. Suppose that the limits*

$$\lim_{x \to x_0} f(x), \quad \lim_{x \to x_0} g(x)$$

exist. Then the sum $f(x) + g(x)$ and the product $f(x)g(x)$ approach limits as x tends to x_0, given by

$$\lim_{x \to x_0} \{f(x) + g(x)\} = \lim_{x \to x_0} f(x) + \lim_{x \to x_0} g(x), \tag{1.64–2}$$

$$\lim_{x \to x_0} \{f(x)g(x)\} = \{\lim_{x \to x_0} f(x)\}\{\lim_{x \to x_0} g(x)\}. \tag{1.64–3}$$

Furthermore, if $\lim_{x \to x_0} g(x) \neq 0$, the quotient $\dfrac{f(x)}{g(x)}$ has a limit given by

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = \frac{\lim_{x \to x_0} f(x)}{\lim_{x \to x_0} g(x)}. \tag{1.64–4}$$

Proof. Let the limits of $f(x)$ and $g(x)$ be denoted by A and B, respectively. We shall prove (1.64–2) and (1.64–4), leaving the proof of (1.64–3) as an exercise for the student. The prove (1.64–2) we must show that, if a positive ϵ is given, we can choose a positive δ such that

$$|(f(x) + g(x)) - (A + B)| < \epsilon \text{ if } 0 < |x - x_0| < \delta. \tag{1.64–5}$$

Now

$$(f(x) + g(x)) - (A + B) = (f(x) - A) + (g(x) - B),$$

and therefore, by an application of (1.64–1),

$$|(f(x) + g(x)) - (A + B)| \leqq |f(x) - A| + |g(x) - B|. \tag{1.64–6}$$

Now by hypothesis we can make $|f(x) - A|$ and $|g(x) - B|$ as small as we like by restricting x to lie sufficiently near x_0. In particular, there are positive numbers δ_1 and δ_2 such that

$$|f(x) - A| < \frac{\epsilon}{2} \text{ if } 0 < |x - x_0| < \delta_1 \tag{1.64–7}$$

and

$$|g(x) - B| < \frac{\epsilon}{2} \text{ if } 0 < |x - x_0| < \delta_2. \tag{1.64–8}$$

Let δ be the smaller of the numbers δ_1, δ_2. It then follows from (1.64–6), (1.64–7), and (1.64–8) that (1.64–5) holds. This completes the proof of (1.64–2).

In proving (1.64–4) we first of all observe that, since $\lim_{x\to x_0} g(x) = B$ and $B \neq 0$, we can be sure that $|g(x)| > \frac{1}{2}|B|$ if we require x to be sufficiently near x_0. It suffices to choose δ_0 so that

$$|g(x) - B| < \tfrac{1}{2}|B| \text{ if } 0 < |x - x_0| < \delta_0,$$

for then

$$B = (B - g(x)) + g(x),$$

and by (1.64–1) we have

$$|B| \leqq |B - g(x)| + |g(x)| < \tfrac{1}{2}|B| + |g(x)|,$$

and consequently

$$\tfrac{1}{2}|B| < |g(x)| \text{ if } 0 < |x - x_0| < \delta_0. \tag{1.64–9}$$

Now we can write

$$\frac{f(x)}{g(x)} - \frac{A}{B} = \frac{Bf(x) - Ag(x)}{g(x)B} = \frac{B[f(x) - A] + A[B - g(x)]}{g(x)B},$$

and so

$$\left|\frac{f(x)}{g(x)} - \frac{A}{B}\right| \leqq \frac{|B||f(x) - A| + |A||B - g(x)|}{|g(x)||B|}. \tag{1.64–10}$$

(Here we have used the fact that the absolute value of a product is the product of the absolute values.) Now let a positive ϵ be assigned arbitrarily. We wish to choose a positive δ so that the left member of (1.64–10) is less than ϵ if $0 < |x - x_0| < \delta$. Let

$$\epsilon_1 = \frac{|B|^2\epsilon}{2(|B| + |A|)}.$$

Choose δ_1 and δ_2 so that

$$|f(x) - A| < \epsilon_1 \text{ if } 0 < |x - x_0| < \delta_1, \tag{1.64–11}$$

$$|g(x) - B| < \epsilon_1 \text{ if } 0 < |x - x_0| < \delta_2. \tag{1.64–12}$$

At the same time we make sure that $\delta_2 < \delta_0$ so that (1.64–9) will hold. Now, making use of (1.64–9), (1.64–11), and (1.64–12) we see that the right member of (1.64–10) is less than

$$\frac{|B|\epsilon_1 + |A|\epsilon_1}{\frac{1}{2}|B|^2} = \epsilon$$

if $0 < |x - x_0| < \delta$, where δ is the smaller of the numbers δ_1, δ_2. This completes the proof of (1.64–4).

EXERCISES

1. Prove part (1.64–3), using the following suggestions: Show that $|g(x)| < |B| + 1$ if $|g(x) - B| < 1$. Write

$$f(x)g(x) - AB = (f(x) - A)g(x) + (g(x) - B)A,$$

and use (1.64–1). If $\epsilon > 0$ is given, let

$$\epsilon_1 = \frac{\epsilon}{|A| + |B| + 1}.$$

Now show that $|f(x)g(x) - AB| < \epsilon$ provided $|f(x) - A| < \epsilon_1$, $|g(x) - B| < 1$ and $|g(x) - B| < \epsilon_1$. Explain how these last inequalities are to be guaranteed by appropriate restrictions on x. Write out the entire proof in full detail.

2. Formulate and prove the counterpart of Theorem XIV for limits of sequences.

MISCELLANEOUS EXERCISES

1. Suppose that f is continuous on $[a, b]$, that $f'(x)$ and $f''(x)$ exist when $a < x < b$, that $f(a) = f(b) = 0$, and that there is a number c such that $a < c < b$ and $f(c) > 0$. Prove that there is a number ξ between a and b such that $f''(\xi) < 0$.

2. Let $f(x) = \log_a x$ $(0 < a, a \neq 1)$. Deduce that $f'(x) = \dfrac{f'(1)}{x}$, using nothing about logarithms other than: (1) the assumption that $f'(x)$ exists when $x > 0$, and (2) the property $\log_a(xu) = \log_a x + \log_a u$ when $x > 0$ and $u > 0$.

3. Find $\lim\limits_{x \to \infty} \{\sqrt{(x + a)(x + b)} - x\}$.

4. In each of the following cases, investigate the one-sided limits at the point indicated, and decide whether or not $\lim_{x \to x_0} f(x)$ exists.

(a) $f(x) = [x] + [3 - x]$, $x_0 = 2$;

(b) $f(x) = \dfrac{e^{\tan x} - 1}{e^{\tan x} + 1}$, $x_0 = \dfrac{\pi}{2}$;

(c) $f(x) = 2^{-1/x} \sin \dfrac{1}{x}$, $x_0 = 0$.

5. Let $s_1 = \sqrt{2}$, $s_{n+1} = \sqrt{2s_n}$, $n = 1, 2, \ldots$; find $\lim_{n \to \infty} s_n$.

6. Suppose $0 < a_1 \leqq a_2 \leqq \cdots \leqq a_k$, where the a's are fixed numbers. Let $b_n = (a_1^n + a_2^n + \cdots + a_k^n)^{1/n}$. Show that $\lim_{n \to \infty} b_n = a_k$.

7. **(a)** Show that $\lim\limits_{n \to \infty} \sum\limits_{p=0}^{n-1} \dfrac{n}{n^2 + p^2} = \dfrac{\pi}{4}$.

$\left(\text{Bring in consideration of } \int_0^1 \frac{dx}{1 + x^2}.\right)$

(b) Find $\lim\limits_{n \to \infty} \dfrac{1}{n} \sum\limits_{p=1}^{n} \sin \dfrac{p\pi\alpha}{n}$, where $\alpha > 0$.

8. In (1.2–4) take $f(x) = x^3$, $a \neq 0$. Obtain a formula connecting a, θ, and h, and show that $|\theta - \frac{1}{2}| \leqq \dfrac{2|h|}{3|a|}$. Hence conclude that $\lim\limits_{h \to 0} \theta = \frac{1}{2}$.

9. Let (x_0, y_0) be a point of the ellipse $b^2x^2 + a^2y^2 = a^2b^2$. Let the tangent to the ellipse at (x_0, y_0) intersect the x-axis at A and the y-axis at B. Find the minimum possible value of the distance AB.

10. Define $f(x) = x \tan^{-1}(1/x)$ if $x \neq 0$, and $f(0) = 0$. Is f continuous at $x = 0$? Is it differentiable there?

11. Show, using the law of the mean, that

$$|x \log x| < |x \log \xi| + \xi$$

if $0 < x < \xi$. From this deduce that $\lim_{x \to 0+} x \log x = 0$. Specify carefully how to choose δ so that $0 < x < \delta$ will imply $|x \log x| < \epsilon$, where ϵ is given in advance. Begin by choosing ξ in terms of ϵ in a suitable manner.

12. Suppose $f(x) = x^2 + 2x$, and define g by $g(t) = t^2 \sin(\pi/t) + t$ if $t \neq 0$, $g(0) = 0$. Let $F(t) = f(g(t))$. Find $F'(0)$.

13. Find the absolute maximum and minimum of $f(x) = x^2(1-x)^3$ on the interval $-1 \leqq x \leqq 2$ *without using a graph.*

14. Consider the functions

$$\begin{aligned} f_0(x) &= \cos x - 1 \\ f_1(x) &= \sin x - x \\ f_2(x) &= \cos x - 1 + \tfrac{1}{2}x^2 \\ f_3(x) &= \sin x - x + \tfrac{1}{6}x^3 \\ f_4(x) &= \cos x - 1 + \tfrac{1}{2}x^2 - \tfrac{1}{24}x^4 \\ f_5(x) &= \sin x - x + \tfrac{1}{6}x^3 - \tfrac{1}{120}x^5. \end{aligned}$$

Note that $f_1'(x) = f_0(x)$ and that $f_0(x) \leqq 0$ if $x > 0$. Thus $f_1(x)$ decreases as x increases, and since $f_1(0) = 0$, we conclude that $f_1(x) < 0$ when $x > 0$. Next note that $f_2'(x) = -f_1(x)$. Explain how you conclude that $f_2(x) > 0$ when $x > 0$. Continuing in this way, show that for $x > 0$,

$$x - \tfrac{1}{6}x^3 < \sin x < x - \tfrac{1}{6}x^3 + \tfrac{1}{120}x^5$$

and

$$1 - \tfrac{1}{2}x^2 + \tfrac{1}{24}x^4 - \tfrac{1}{120}x^6 < \cos x < 1 - \tfrac{1}{2}x^2 + \tfrac{1}{24}x^4.$$

What is the generalization?

15. Suppose that f satisfies the hypotheses: f is defined and continuous when $a \leqq x < b$, differentiable when $a < x < b$, $f(a) = 0$, and $f(x) > 0$ if $a < x < b$. Prove that there cannot be a positive constant M such that $0 \leqq \dfrac{f'(x)}{f(x)} \leqq M$ when $a < x < b$.

16. Let $\{a_n\}$ be a sequence and suppose that $\{\sigma_n\}$ is the sequence whose nth term is the arithmetic mean of the first n terms of $\{a_n\}$, i.e. $\sigma_n = (1/n) \sum_{k=1}^{n} a_k$. Prove that if $\{a_n\}$ converges, then $\{\sigma_n\}$ converges and has the same limit. Construct an example which shows that $\{\sigma_n\}_1^\infty$ may converge even when $\{a_n\}_1^\infty$ does not.

2 / THE REAL NUMBER SYSTEM

2 / NUMBERS

Our experience with numbers begins with the positive integers (also called the whole numbers, or natural numbers). Next we become acquainted with zero, and in due course we become familiar with negative integers and with rational fractions (ratios of positive and negative integers). At some stage we learn the adjective *irrational* for numbers such as $\sqrt{2}$, $\sqrt[3]{5}$, π. In algebra we meet the equation $x^2 = -1$ and are told that it gives rise to a new number i. Numbers such as i, $2i$, $-7i$ are called *pure imaginary*, and numbers such as $3+5i$ are called *complex*. Our learning, with numbers as with everything else, proceeds mainly by particular cases and illustrative examples. But in due time it is possible to reduce our knowledge to order and to give it logical coherence by a systematic study of number and number systems. We make a beginning on such a systematic study in this chapter.

The integers, the rational fractions, and the irrational numbers compose the number system which lies at the foundation of calculus and of all analysis. The numbers of this system are called the *real* numbers, and the system of all such numbers is called the real number system.

The adjectives "real" and "imaginary," as applied to numbers, have entirely conventional technical meanings. They are not meant to convey, and the student should not let them convey, any implications whatsoever of a philosophical nature about existence or nonexistence as genuine entities.

We shall not attempt to define the concept of number or to build up the concepts of rational and irrational numbers from the concept of the integers. Instead, we shall take the real number system as something known (though somewhat imperfectly and unsystematically) by the student, and shall make specific the algebraic laws governing the real numbers. The student is already familiar with most of these laws, but it is now very important to know a complete set of properties of the real number system. From such a complete set we may deduce every property of the system.

2.1 / THE FIELD OF REAL NUMBERS

Addition and multiplication are the fundamental operations with ordinary numbers. These operations conform to the following laws:

	Addition	*Multiplication*
The commutative law:	$a + b = b + a$	$ab = ba$
The associative law:	$a + (b + c) = (a + b) + c$	$a(bc) = (ab)c$
The distributive law:	$a(b + c) = ab + ac$	

Throughout this section the word "number" will be understood to mean "real number," and symbols $a, b, c, x, \ldots$ will stand for real numbers.

There are two numbers with special properties, namely 0 and 1. The special properties are expressed by the laws

$$a + 0 = a \quad \text{and} \quad a \cdot 1 = a \tag{2.1–1}$$

for every number a.

The number 0 is special for addition, while 1 is special for multiplication. The operations of subtraction and division may be defined with the aid of these special numbers in the following way: To every number a corresponds its negative, $-a$, which is the "additive inverse" of a. By this we mean that $x = -a$ satisfies the equation

$$a + x = 0. \tag{2.1–2}$$

Likewise every number a except 0 has a multiplicative inverse, denoted by a^{-1}. That is, if $a \neq 0$, $x = a^{-1}$ satisfies the equation

$$ax = 1. \tag{2.1–3}$$

We then *define* the subtraction of b from a by the equation

$$a - b = a + (-b). \tag{2.1–4}$$

Similarly, we define the division of a by b as

$$\frac{a}{b} = a(b^{-1}). \tag{2.1–5}$$

The properties of the real numbers which we have just been discussing are summed up briefly in the language of modern algebra by saying that the real numbers form a *field*. The word "field" here has a special technical meaning. *When we say that a system of numbers F constitutes a field we mean the following:*

1. *If a and b are in F, then $a + b$ and ab are in F.*
2. *The commutative, associative, and distributive laws hold.*
3. *F contains distinct special numbers 0 and 1 with the properties* (2.1–1).
4. *Equation* (2.1–2) *has a solution in F for each a, and* (2.1–3) *has a solution in F for each $a \neq 0$.*

In abstract algebra it is shown how the other familar laws of elementary algebra are deducible from the laws governing a field. Among the important rules that can be proved are: $a \cdot 0 = 0$ and $(-a)(-b) = ab$. We shall not undertake any systematic deductions of this kind. We mention, however, the rule:

$$\text{If } ab = 0 \text{ and } b \neq 0, \text{ then } a = 0. \tag{2.1–6}$$

This is proved as follows: Since $b \neq 0$, there is a number b^{-1} such that $bb^{-1} = 1$. From $ab = 0$ we conclude that $a(bb^{-1}) = 0 \cdot b^{-1}$, or $a \cdot 1 = 0$, or $a = 0$.

Note that the system of integers (positive, negative, and zero) is *not* a field, although it fails to be one only through the fact that a^{-1} need not be an integer

when a is. The rational numbers *do* form a field. Thus there are fields other than the field of real numbers.

2.2 / INEQUALITIES. ABSOLUTE VALUE

One of the important properties of the real numbers is that they are *ordered*; that is, there is a notion "a is less than b" expressed by the inequality $a < b$; of any two numbers a, b one and only one of the following three things is true:

$$a < b \quad \text{or} \quad a = b \quad \text{or} \quad b < a.$$

The properties of order can be stated simply in terms of properties of positive numbers. We express the fact that a is positive by the symbols $0 < a$. The basic laws governing the positive numbers are three in number:

$$\text{If } 0 < a \text{ and } 0 < b, \text{ then } 0 < a + b. \tag{2.2–1}$$

$$\text{If } 0 < a \text{ and } b, \text{ then } 0 < ab. \tag{2.2–2}$$

For each a, one and only one of the following relations is true:

$$0 < a \text{ or } 0 = a \text{ or } 0 < -a. \tag{2.2–3}$$

In terms of positivity we lay down the following definitions:

If $b - a$ is positive, we say that a is less than b and write $a < b$. Under the same conditions we also say that b is greater than a and write $b > a$.

Other properties of order, or rules for manipulating inequalities, are deducible from the above definitions and the three basic laws. Among the important rules are the following:

$$0 < a^2 \text{ if } a \neq 0. \tag{2.2–4}$$

$$\text{If } a < b \text{ and } b < c, \text{ then } a < c. \tag{2.2–5}$$

$$\text{If } a < b, \text{ then for any } c,\ a + c < b + c. \tag{2.2–6}$$

$$\text{If } a < b \text{ and } 0 < c, \text{ then } ac < bc. \tag{2.2–7}$$

A field is said to be ordered if certain of its members are distinguished by calling them positive and if this notion of "being positive" satisfies the laws (2.2–1), (2.2–2), and (2.2–3).

For many purposes it is convenient to introduce the symbolism $a \leqq b$, meaning that either $a < b$ or $a = b$. The symbolism $b \geqq a$ has the same meaning.

The notion of the *absolute value* of a real number is defined as follows (the absolute value of a is denoted by $|a|$):

$$|a| = a \text{ if } 0 < a; \quad |a| = -a \text{ if } a < 0; \quad |0| = 0.$$

In actual calculations with absolute values we rely largely upon the two rules:

$$|ab| = |a||b|, \tag{2.2–8}$$

$$|a + b| \leqq |a| + |b|. \tag{2.2–9}$$

(See §1.64 for a brief discussion of (2.2–9)).

The further rule

$$\Big||a|-|b|\Big| \leqq |a-b| \tag{2.2-10}$$

is also convenient; it is deducible with the aid of (2.2–9).

We observe in passing that the inequality $|x| \leqq \epsilon$ (where $\epsilon > 0$) is equivalent to the inequalities $x \leqq \epsilon$ and $-\epsilon \leqq x$, which we write as a double inequality $-\epsilon \leqq x \leqq \epsilon$. By setting $x = b - a$ and doing a little transposing we see that $|a-b| \leqq \epsilon$ is equivalent to $b - \epsilon \leqq a \leqq b + \epsilon$.

EXERCISES

1. Prove (2.2–4). Note first that $a^2 = a \cdot a$ and $a^2 = (-a) \cdot (-a)$. Then use (2.2–2) and (2.2–3).

2. Prove (2.2–5). Appeal to (2.2–1) and the fact that $x < y$ is equivalent to $0 < y - x$.

3. Prove (2.2–6).

4. Prove (2.2–7).

5. Write $a = (a-b) + b$ and apply the rule (2.2–9) to obtain $|a| \leqq |a-b| + |b|$, whence $|a| - |b| \leqq |a-b|$. Why is it also true that $|b| - |a| \leqq |a-b|$? Explain now why (2.2–10) is correct.

2.3 / THE PRINCIPLE OF MATHEMATICAL INDUCTION

The natural numbers 1, 2, 3, . . . occupy a position of especial importance in the field of real numbers. To single out the class of all natural numbers from the rest of the real numbers we may make the following assertion: The totality of natural numbers forms the *smallest* class of real numbers possessing the following two properties:

1. The number 1 is a member of the class.
2. If x is a member of the class, so is $x + 1$.

The statement that the natural numbers form the smallest class having these properties means that any collection of real numbers having these two properties must include all the natural numbers.

The characteristic feature of the totality of natural numbers, as just described, is logically equivalent to the principle of mathematical induction (also called complete induction). We formulate the principle as follows:

Let $A(n)$ denote a proposition (e.g., a verbal statement, or a formula) associated with the natural number n. Suppose it is possible to show that the proposition is true if $n = 1$, and suppose also that, for each particular n, we can prove the truth of $A(n+1)$ if we assume the truth of $A(n)$. Then $A(n)$ is true for every natural number n.

To see the validity of the principle, let S be the class of natural numbers n such that $A(n)$ is true. The assumption that $A(1)$ is true means that 1 belongs to S. Also, the fact that assumption of the truth of $A(n)$ allows us to deduce the

truth of $A(n+1)$ means that if n is in S, so is $n+1$. Hence, by the remarks made in the first paragraph of this section, S must contain all the natural numbers. That is, $A(n)$ must be true for every n.

We shall not analyze in detail the logical position of the principle of mathematical induction in the description of the natural numbers. If the real numbers are thought of as having been built up from the numbers 1, 2, 3, . . ., we need the principle of mathematical induction as an *axiom* about the natural numbers. If, on the other hand, we start from the assumption that the real numbers are somehow presented to us as a particular sort of ordered field, we may *define* the natural numbers as the smallest class of real numbers which contains the number 1, and which contains $x+1$ if it contains x. The theory of classes or sets, into which we prefer not to venture at this stage, permits us to *prove* that there *is* such a smallest class. The principle of mathematical induction is, from this latter point of view, a *theorem* about the class of natural numbers.

By way of illustrating the principle of mathematical induction, we shall prove formally some things about natural numbers.

Example 1. Every natural number is positive.

In the first place, $0<1$. This follows from (2.2–4), since $1 \neq 0$ and $1 = 1^2$. To complete the proof of the general assertion, suppose $0<n$ for a particular n. Then $0<n$ and $0<1$ imply $0<n+1$, by (2.2–1). Thus $0<n$ for every natural number n, by the induction principle.

From now on we shall usually refer to "the natural numbers" as "the positive integers." The unmodified term "integers" refers to the class consisting of the positive integers 1, 2, 3, . . ., their negatives -1, -2, -3, . . ., and the number 0.

We shall take for granted without formal proof such familiar facts as that *there is no integer between n and n + 1, if n is an integer.*

Example 2. If S is a class of positive integers containing at least one member, it contains a smallest number.

Most students will feel inclined to accept this as true without demonstration. A proof is logically necessary, however. Observe that the assertion ceases to be true if the word "integers" is replaced by "real numbers"; for instance, there is no smallest member of the class of positive rational numbers.

The proof of the assertion in Example 2 uses the characteristic properties of the class of natural numbers in an interesting way. Observe, in the first place, that if 1 is in S, 1 *is* the smallest member of S, since there is no positive integer less than 1. Hence, we need consider only the case in which 1 does not belong to S. We now let T be the class of positive integers p such that $p<n$ for every n in S. This class T contains 1 since 1 is not in S. On the other hand, T does not contain *all* positive integers, since *some* positive integers belong to S. Hence, it cannot be that T contains $p+1$ whenever it contains p, for in that case T *would* contain all positive integers, by the characteristic properties of the class of all such integers. There must therefore be some integer p_0 in T such that p_0+1 is not in T. Then, by the way in which T is defined, there must be an integer n_0 in S

such that $n_0 \leqq p_0 + 1$. We assert that n_0 is the smallest integer in S. For, if n is any member of S, we have $p_0 < n$. Let $m = n - p_0$, or $p_0 + m = n$. Here m is a positive integer. Therefore, $n_0 \leqq p_0 + 1 \leqq p_0 + m = n$, or $n_0 \leqq n$. This completes the argument.

EXERCISES

1. Prove by induction that, for every natural number n, either $1 = n$ or $1 < n$.

2. Prove the validity of the following form of the principle of mathematical induction, resting your argument on the form enunciated in the text. *Let $B(n)$ denote a proposition associated with the integer n. Suppose $B(n)$ is known (or can be shown) to be true when $n = n_0$, and suppose the truth of $B(n+1)$ can be deduced if the truth of $B(n)$ is assumed. Then $B(n)$ is true for every integer n such that $n_0 \leqq n$.*
SUGGESTION: Let $A(n)$ be the proposition $B(n_0 + n - 1)$.

2.4 / THE AXIOM OF CONTINUITY

The facts expressed in the statement that the real numbers form an ordered field are quite familiar. We are now going to discuss a much less familiar property of the real number system. Most students beginning a course in advanced calculus will have had no experience in making use of this property, and quite possibly may never have heard of it. We call it the *axiom of continuity.*

The Axiom of Continuity. *Suppose that all real numbers are separated into two collections, which we denote by L and R, in such a way that*

1. *every number is either in L or in R.*
2. *each collection contains at least one number.*
3. *if a is in L and b is in R, then $a < b$.*

Then there is a number c such that all numbers less than c are in L and all numbers greater than c are in R. (The number c itself may belong either to L or to R, depending on the particular way in which L and R are formed.)

It is convenient to have a name for a separation of all real numbers into collections L and R meeting the specifications (1)–(3). We call such a separation a *cut*; the number c is then called the *cut number.* The cut number corresponding to a particular cut is unique. For suppose a given cut has the distinct cut numbers c_1 and c_2. One of them is the greater, say $c_1 < c_2$. Consider the number

$$b = \frac{c_1 + c_2}{2},$$

which lies halfway between c_1 and c_2: $c_1 < b < c_2$. Now $c_1 < b$ implies that b is in R, by one of the properties of the cut number c_1. Likewise $b < c_2$ implies that b is in L. Hence b is in both L and R. This is impossible, however, for by the specification (3) L and R cannot have any members in common. The assumption of distinct cut numbers has led to a contradiction. Therefore, we conclude that any cut has but one cut number.

It is clear that if c belongs to L then L consists of all numbers x such that $x \leqq c$, while R consists of all numbers x such that $c < x$. If c belongs to R, then L consists of all x such that $x < c$ and R consists of all x such that $c \leqq x$.

The idea of a cut was originated by the German mathematician Dedekind for use in his theory of the structure of the real number system (published in 1872 under the title *Continuity and Irrational Numbers*). If the number system is built up by stages from the integers, it is possible to arrange the exposition in such a way that what we have called the axiom of continuity is provable as a theorem about the real numbers. But since we are taking the point of view that the real numbers are somehow given to us to work with, we are listing our assumptions about them. The axiom of continuity is one of our assumptions. It is a far-reaching assumption, for with it we can deal satisfactorily with the existence problems for real numbers referred to near the end of each of §§1.62 and 1.63.

No further assumptions need be made about the real number system. The system is described with logical completeness by saying that it is an ordered field satisfying the axiom of continuity.

At this point it is convenient to introduce a theorem which expresses what is called the *Archimedean law* of real numbers. Its proof is a good illustration of arguments using the axiom of continuity.

THEOREM I. *Let a and b be positive real numbers. There exists a positive integer n such that $b < na$.*

Proof. Suppose the theorem false, so that $na \leqq b$ for every positive integer n. We shall define a cut as follows: Let L consist of all numbers x such that $x < na$ for *some* n, and let R consist of all numbers not in L, i.e., all numbers y such that $na \leqq y$ for every n. We must verify that the three specifications for a cut are fulfilled:

1. Every number is in either L or R, by definition.
2. b is in R by our initial supposition, and a is in L, since $a < na$ if $n = 2$. Thus neither L nor R is without members.
3. If x is in L and y is in R, we have $x < na \leqq y$ for some n, and hence $x < y$.

Now let c be the cut number. We observe that all the numbers na are in L, since $na < (n+1)a$. Therefore $na \leqq c$, for $c < na$ would mean that na is in R, by one of the properties of a cut number. Hence, also $(n+1)a \leqq c$. But this implies $na \leqq c - a$. This being true for all n, we conclude that $c - a$ is in R. But $c - a < c$, and hence $c - a$ is in L. We have now reached a contradiction, and the proof is complete.

2.5 / RATIONAL AND IRRATIONAL NUMBERS

Numbers of the form p/q, where p and q are integers, are called *rational*. Real numbers which are not rational are called *irrational*. The theory of the nature of irrational numbers began with the ancient Greeks. It has been known since the

time of Pythagoras that $\sqrt{2}$ is not a rational number. (With a little knowledge about expressing integers as products of their prime factors the student may easily prove this fact for himself.) The theory of incommensurable ratios, developed by Eudoxus, is in essence a geometrical treatment of irrational numbers. It was not, however, until the nineteenth century that mathematicians (Dedekind with his cuts important among them) arrived at the understanding we have today of the real number system and of the position of irrationals in it.

It can be shown that between any two real numbers there are both rational and irrational numbers. If a and b are given real numbers with $a < b$, we obtain a rational number r such that $a < r < b$ by the following argument: since $0 < 1$ and $b - a > 0$ there exists (by Theorem I, §2.4) a positive integer n such that $1 < n(b - a)$. Let m be the smallest integer such that $m > na$. Then $(m - 1) \leqq na$, and therefore

$$m \leqq na + 1 < na + n(b - a) = nb.$$

Consequently

$$a < \frac{m}{n} < b,$$

so that $r = m/n$ is a rational number of the required sort. Finding an irrational number between a and b is left as an exercise for the student (see Exercise 2).

It follows from the previous paragraph that if x is a real number, rational or irrational, we can find a rational number r as near it as we please. That is, if ϵ is any positive number, we can find r so that $x - \epsilon < r < x + \epsilon$. We do in fact, in any actual computations or measurements, use rational approximations to real numbers. For many purposes we even limit ourselves to special kinds of rational numbers, namely decimal fractions. A decimal fraction is simply a rational number of the form $m/10^n$, where m and n are integers and $n \geqq 0$. There is a number of this form between any two real numbers, as may be readily shown from the fact that if $a < b$, then $1 < 10^n(b - a)$ for a suitably chosen n.

EXERCISES

1. Assume that $\sqrt{2} = m/n$, where m and n are integers and the fraction is reduced to its lowest terms. Then $m^2 = 2n^2$. Now deduce a contradiction, and hence prove that $\sqrt{2}$ is irrational. Use the fact that if the product of two integers is even, then at least one of them is even.

2. Show that if $a < b$ there is an irrational number x between a and b.

SUGGESTION: Let y be a positive irrational number, e.g., $\sqrt{2}$. Then show that there are integers m, n with $m \neq 0$ and $n > 0$, such that $a < (m/n)y < b$. Why is $x = (m/n)y$ irrational?

3. Prove by induction that $n < 10^n$ if n is a positive integer. Then prove that if $a < b$, there are integers m, n, with $n > 0$, such that $a < m/10^n < b$.

2.6 / THE AXIS OF REALS

It is customary and convenient to use geometric language a good deal in speaking about the number system. On a given straight line we take an arbitrary

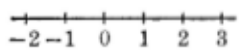


Fig. 19.

point as an origin, an arbitrary direction as positive, and an arbitrary unit of length. We then mark off segments of unit length on either side of the origin, thus obtaining the points which we label as shown in Fig. 19. There is a one-to-one correspondence between the real numbers and the points on the line. This entitles us, for brevity, to speak of "the point a" instead of "the point corresponding to the number a." The inequality $a < b$ has the geometrical interpretation that b lies in the positive direction from a along the line.

We call the line, thus regarded as a geometrical representation of the real number system, *the axis of reals, or the real number scale.*

2.7 / LEAST UPPER BOUNDS

By a *set* of real numbers we mean an aggregate or class of numbers. It may be formed according to any rule, and the number of its members may be finite or infinite. If the conditions laid down for determining the set are such that no number satisfies them, the set is said to be empty. It is very convenient to have a brief symbolism to indicate that a number belongs to a given set. The statement that the number s belongs to the set S is expressed symbolically in the form $s \in S$ (read s is a member of S). The symbolic form of the statement that s does *not* belong to S is $s \notin S$. Thus, if S is the set of prime positive integers, $3 \in S$ and $8 \notin S$.

If S is a set of numbers, and if M is a number such that $s \leqq M$ for each $s \in S$, we say that M is an *upper bound* of S. Evidently any number larger than M is also an upper bound of S. If A is an upper bound of S and if there is no number smaller than A which is also an upper bound for S, we call A the *least upper bound* of S. Obviously a set cannot have more than one least upper bound.

Example. The set S of numbers of the form $n/(n+1)$, for all positive integers n, consists of 1/2, 2/3, 3/4, 4/5 . . . etc. Evidently 1 is an upper bound of S. But more is true; 1 is the (unique) *least* upper bound of S. To verify this we must show that if $c < 1$, c cannot be an upper bound of S, i.e., that there is some n such that $c < n/(n+1)$. To get such an n we appeal to Theorem I in §2.4, which tells us that there exists an n such that $1 < n(1-c)$. But $c < 1$, and so $c < n(1-c) = n - nc$, or $nc + c < n$. But then $(n+1)c < n$, and so $c < n/(n+1)$. This completes the argument.

The following theorem is of fundamental importance:

THEOREM II. *If S is a set of real numbers which is not empty and which has an upper bound, then it has a least upper bound.*

Proof. We appeal to the axiom of continuity. Let L be the set of all numbers x such that $x < s$ for some s in S, and let R be the set of all numbers y such that

$s \leqq y$ for every s in S. Clearly L and R together comprise all real numbers. If $s \in S$, then $s - 1 \in L$; $A \in R$ if A is an upper bound of S. Thus neither L nor R is empty. If $x \in L$ and $y \in R$, we have $x < s$ for some s in S. But $s \leqq y$, and therefore $x < y$. We have therefore defined a cut. Let c be the cut number. We shall prove that c is the least upper bound of S. It certainly is *an* upper bound. For if we suppose $c < s$ for some s in S, we can choose a number z between c and s. Then $z \in R$ since $c < z$, and $z \in L$ since $z < s$; thus we have a contradiction, for a number cannot belong to both L and R. If b is any number smaller than c, the properties of the cut number insure that $b \in L$, and hence $b < s$ for some s in S. Thus b cannot be an upper bound of S. The proof that c is the least upper bound of S is now complete.

Theorem II expresses a property of the real number system which is a direct consequence of the axiom of continuity. It is easily demonstrated that if the statement of Theorem II is taken as an axiom concerning the real numbers, the truth of the axiom of continuity may be deduced (making it a theorem instead of an axiom). For a key to this demonstration see Exercise 1. Thus the axiom of continuity and the existence of least upper bounds as stated in Theorem II are equivalent propositions. Hereafter, in arguments where we could lean equally well either on the axiom of continuity or on Theorem II, we shall usually appeal to the latter.

As an immediate application we shall prove Theorem XIII of Chapter 1 (§1.62). We reword it slightly.

THEOREM III. *Let $\{x_n\}$ be a sequence such that $x_1 \leqq x_2 \leqq \cdots \leqq x_n \leqq x_{n+1} \leqq \ldots$, and suppose that the set of numbers x_n has an upper bound: $x_n \leqq M$ for every n. Then the sequence is convergent, its limit being the least upper bound of the numbers x_n.*

Proof. Let A be the least upper bound of the numbers x_n. Then if $\epsilon > 0$ we have $A - \epsilon < x_n$ for *some* n, say $n = N$, and $x_n \leqq A$ for every n. Since $x_N \leqq x_n$ for every $n \geqq N$ (by virtue of the assumption that $x_n \leqq x_{n+1}$), we see that $A - \epsilon < x_n \leqq A$ if $N \leqq n$. Thus by definition $\lim_{n\to\infty} x_n = A$. This proves the theorem.

The notion of *lower bound* of a set, and of the *greatest lower bound*, are defined in exactly the same way as upper bound and least upper bound, except that the notions of "less than" and "least" are replaced throughout by "greater than" and "greatest." We may summarize the defining properties of the least upper bound and greatest lower bound as follows:

The set S has the least upper bound A if $s \leqq A$ for every s in S and if, ϵ being any positive number, $A - \epsilon < s$ for at least one s in S.

The set S has the greatest lower bound B if $B \leqq s$ for every s in S and, ϵ being any positive number, $s < B + \epsilon$ for at least one s in S.

THEOREM IV. *If S is a set of real numbers which is not empty and which has a lower bound, then it has a greatest lower bound.*

Proof. We could give a proof similar to that of Theorem II. Instead, we base the proof on Theorem II itself. Define a set T as consisting of all numbers $t = -s$, where $s \in S$. If M is a lower bound of S, we have $M \leqq s$, or $-s \leqq -M$ for each s in S. Thus T has an upper bound $-M$. By Theorem II, T has a least upper bound, say A. We leave it for the student to show that $-A$ is the greatest lower bound of S, thus completing the proof.

Just as Theorem IV matches Theorem II, so there is a theorem which matches Theorem III.

THEOREM V. *Let $\{x_n\}$ be a sequence such that $x_1 \geqq x_2 \geqq \cdots$ (in general $x_n \geqq x_{n+1}$), and suppose that the set of numbers x_n has a lower bound: $x_n \geqq M$ for every n. Then the sequence is convergent, its limit being the greatest lower bound of the numbers x_n.*

We leave it for the student to base a proof of this theorem on Theorem IV or to deduce the proof from Theorem III by considering the sequence $\{y_n\} = \{-x_n\}$.

EXERCISES

1. Let L and R be sets of real numbers satisfying the conditions of the axiom of continuity. Assuming the truth of Theorem II, show that the set L has a least upper bound c and that this least upper bound has the properties asserted for the cut number in the axiom of continuity.

2. Prove Theorem IV by an argument similar to that used in proving Theorem II, using the axiom of continuity.

3. Carry out the proof of Theorem V in each of two ways, as suggested in the text.

4. Prove Theorem I (§2.4), starting as follows: Suppose the assertion in the theorem false, so that $na \leqq b$ for all positive integers n. Use Theorem II and the defining property of a least upper bound to arrive at a contradiction.

5. Let $x_1 = \sqrt{2}$, $x_{n+1} = \sqrt{2 + x_n}$. Use mathematical induction to show that $x_n < x_{n+1}$. Next show that if $x_n \geqq 2$ (where $n > 1$), then $x_{n-1} \geqq 2$ also. How do you conclude from this that $x_n < 2$ for all n? State why $\lim_{x \to \infty} x_n$ exists, and find the limit.

6. Suppose $c > 0$, and let $x_1 = \sqrt{c}$, $x_{n+1} = \sqrt{c + x_n}$. Show that $x_n < x_{n+1}$ and $x_n < 1 + \sqrt{c}$ (see Exercise 5). State why $\lim_{n \to \infty} x_n$ exists, and find the limit.

2.8 / NESTED INTERVALS

The theorem which we are going to discuss in this section is an immediate consequence of Theorems III and V of §2.7. We are going to use the language of geometry rather than the language of numbers; the theorem is about closed intervals on the real axis. It will be convenient to denote intervals by single letters, such as I, I_1, I_2, and so on. If I_1 and I_2 are closed intervals, and if the end points of I_2 lie in I_1, we say that I_2 is contained in I_1.

Now suppose that we are given a sequence of closed intervals, $I_1, I_2, I_3, \ldots, I_n, \ldots$ with the property that I_2 is contained in I_1, I_3 is contained in I_2, and so on; that is, we assume that each interval contains the one which follows it in the

sequence. Let us denote the length of the interval I_n by the symbol l_n, and let us assume that $\lim_{n\to\infty} l_n = 0$. In these circumstances we shall say that $\{I_n\}$ is a sequence of *nested* intervals, or that the sequence forms a *nest.*

THEOREM VI. *If $\{I_n\}$ is a nest of closed intervals, there is exactly one point which is common to all the intervals.*

Proof. Let the interval I_n be described by the inequalities $a_n \leqq x \leqq b_n$, $(a_n < b_n)$. The fact that $\{I_n\}$ is a nest is then described as follows. The inequalities

$$a_n \leqq a_{n+1},\ b_{n+1} \leqq b_n, \quad n = 1, 2, \ldots \tag{2.8–1}$$

correspond to the fact that I_n contains I_{n+1}. Also

$$\lim_{n\to\infty} (b_n - a_n) = 0, \tag{2.8–2}$$

since $l_n = b_n - a_n$. Now $a_n < b_n \leqq b_1$, and $a_1 \leqq a_n < b_n$. It now follows by Theorem III of §2.7 that the sequence $\{a_n\}$ has a limit; likewise, by Theorem V, $\{b_n\}$ has a limit. Let us write

$$\lim_{n\to\infty} a_n = a, \quad \lim_{n\to\infty} b_n = b.$$

These two limits must coincide, that is $a = b$, by virtue of (2.8–2). Now $a_n \leqq a = b$, for a is the least upper bound of the sequence $\{a_n\}$; likewise $a = b \leqq b_n$. Hence the point a belongs to every one of the intervals I_n. Obviously there cannot be more than one such point common to all the intervals. For if two such points existed, say at a positive distance h apart, we could choose n so large that $l_n = b_n - a_n < h$, and under this condition the two points could not both lie in I_n.

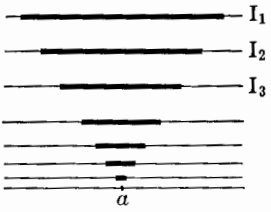


Fig. 20.

Theorem VI is illustrated in Fig. 20. Uses of Theorem VI will appear in the next chapter.

MISCELLANEOUS EXERCISES

1. Suppose f is defined as follows: $f(x) = 1$ if x is a rational number, $f(x) = 0$ if x is an irrational number. For what values of x (if any) is f continuous?

2. Show that $\lim_{n\to\infty} \dfrac{n^n}{n!e^n}$ exists without finding the limit.

3. Show that $\lim_{n\to\infty} \dfrac{1}{n}\left\{\dfrac{2\cdot 4\cdots(2n)}{1\cdot 3\cdots(2n-1)}\right\}^2$ exists without finding the limit.

4. If $x_n = (\log n)/n$, show that $x_n > x_{n+1}$ when $n \geqq N$, where N is a certain fixed integer. What is N? What do you conclude about the sequence $\{x_n\}$?

5. Let S be the set of all rational numbers r such that $r^2 < 2$. What is the least upper bound of S? What is the greatest lower bound of S?

6. Let S be the point set consisting of all the points $x_n = (-1)^n\left[2 - \frac{4}{2n}\right]$, $n = 1, 2, \ldots$. Find the least upper bound and greatest lower bound of S.

7. The same as Exercise 6, except that $x_n = (-1)^n + (1/n)$.

8. Let a sequence of numbers $a_1, a_2, \ldots$ be such that $(2 - a_n)\, a_{n+1} = 1$. **(a)** Show that $\lim_{n\to\infty} a_n$ exists. Consider two cases: either $a_n < 2$ for all n, or $a_n > 2$ for some n. **(b)** Find the limit of the sequence.

9. Let x_1 and y_1 be given with $x_1 > y_1 > 0$, and define

$$x_{n+1} = \frac{x_n + y_n}{2},\quad y_{n+1} = \sqrt{x_n y_n},\quad n = 1, 2, \ldots$$

Show **(a)** that $y_n < y_{n+1} < x_1$, **(b)** that $y_1 < x_{n+1} < x_n$, **(c)** that $0 < x_{n+1} - y_{n+1} < (x_1 - y_1)/2^n$. Explain why the sequences $\{x_n\}$ and $\{y_n\}$ are convergent and have the same limit.

10. The number $\sqrt{3}$ is defined as the positive number c such that $c^2 = 3$. That there *is* such a number may be proved as follows: Let S be the set of all positive numbers x such that $x^2 < 3$. This set is not empty, since $1 \in S$. It is a bounded set, for, if $x \in S$, either $x \leqq 1$ or $x > 1$, and in the latter case $x < x^2 < 3$. By Thoerem II, S has a least upper bound, which we shall denote by c. It remains to prove that $c^2 = 3$. (Why must c be positive?) If $\epsilon > 0$, there is some x in S such that $c - \epsilon < x$. Hence $(c - \epsilon)^2 < 3$, or $c^2 - 2c\epsilon + \epsilon^2 < 3$, since $x^2 < 3$. Now let $\epsilon \to 0$, and by Theorem XI, §1.61, we conclude that $c^2 \leqq 3$. Finally, we show that $c^2 < 3$ is impossible. For, if $c^2 < 3$, the fact that $\lim_{h\to 0}(c^2 + 2ch + h^2) = c^2$ shows that $(c + h)^2 < 3$ if h is sufficiently small. This means that $c + h$ is in S if h is any sufficiently small positive number. Since c is the least upper bound of S, this is a contradiction. Therefore $c^2 = 3$.

Now let the student develop a similar argument to show that if $A > 0$ and n is a positive integer, there is a positive number c such that $c^n = A$.

3 / CONTINUOUS FUNCTIONS

3 / CONTINUITY

In §1.12 and §1.2 we pointed out the need to know that if a function is continuous at all points of a finite closed interval it actually attains an absolute maximum and an absolute minimum at points of the interval. Again, in §1.51, we saw that another property of continuous functions occupies a key position in the proof of the mean-value theorem for integrals. After our study of the real number system in Chapter 2 we are prepared to prove that continuous functions do in fact possess the properties referred to above.

The definition of a continuous function was given in §1.1. We repeat the definition.

Definition. *Let f be a function which is defined in some interval containing the point x_0 either inside or at one end. We say that f is continuous at x_0 provided that $\lim_{x\to x_0} f(x) = f(x_0)$. If x_0 is at one end of the interval, x must approach x_0 from one side only. We say that f is continuous on an interval if it is continuous at each point of the interval.*

Often it is convenient to express the definition of continuity in an alternative but equivalent way, using inequalities: f is continuous at x_0 provided that to each positive number ϵ corresponds some positive number δ such that $|f(x)-f(x_0)|<\epsilon$ whenever $|x-x_0|<\delta$ and x is in the interval on which f is defined. Observe that $|f(x)-f(x_0)|<\epsilon$ is equivalent to the double inequality $f(x_0)-\epsilon<f(x)<f(x_0)+\epsilon$. The choice of the number δ will as a rule depend both on ϵ and on x_0 (and of course on the particular function f).

Among the important theorems about continuity is the following assertion about the continuity of sums, products, and quotients:

THEOREM I. *Let f and g be functions defined on the same interval. If $f(x)$ and $g(x)$ are continuous at a point $x=x_0$, so are $f(x)+g(x)$ and $f(x)\cdot g(x)$. If $g(x_0)\neq 0$, the quotient $\dfrac{f(x)}{g(x)}$ is also continuous at $x=x_0$.*

The proof stems directly from the fundamental limit theorem (Theorem XIV, §1.64). We have, for example,

$$\lim_{x\to x_0}\frac{f(x)}{g(x)}=\frac{\lim_{x\to x_0} f(x)}{\lim_{x\to x_0} g(x)}=\frac{f(x_0)}{g(x_0)},$$

provided $g(x_0) \neq 0$. This is precisely the statement that the quotient $\frac{f(x)}{g(x)}$ is continuous at $x = x_0$. The other parts of the theorem are proved in the same way.

EXERCISES

1. Assuming it is known that the functions $\sin x, \cos x$ are continuous for all values of x, discuss the continuity of $\tan x$ and $\operatorname{ctn} x$. Is either of these functions discontinuous at a point where it is defined?

2. If f is defined and continuous on $a \leqq x \leqq b$, is $1/f(x)$ defined on the same interval? What about continuity of $1/f(x)$? May it be discontinuous at a point where it is defined?

3. May a product $f(x) \cdot g(x)$ be continuous at a point x_0 where g is discontinuous? Support your answer. May the product be continuous at a point where both f and g are discontinuous?

4. May a sum $f(x)+g(x)$ be continuous at a point where f is continuous and g discontinuous? May the sum be continuous at a point where both f and g are discontinuous?

5. Suppose $f(x)$, $g(x)$, and $\frac{f(x)}{g(x)}$ are defined in some open interval containing x_0, and suppose $\frac{f(x)}{g(x)}$ and $g(x)$ are continuous at x_0. May f be discontinuous at x_0?

3.1 / BOUNDED FUNCTIONS

A set of real numbers is said to be *bounded* if the set has both an upper and a lower bound. Consider a function f, defined on a given interval. We say that the function is *bounded on the interval* if the set of all the values of the function is a bounded set. This means that there is some number A such that $|f(x)| \leqq A$ for all x on the interval; or, alternatively, there are two numbers m, M such that $m \leqq f(x) \leqq M$ for all x on the interval. We make this definition whether or not the interval on which the function is defined is closed.

Example 1. The function $f(x) = \sin x$ is bounded on the interval $0 \leqq x \leqq 2\pi$, because $-1 \leqq \sin x \leqq 1$. Actually, the function is bounded on *every* interval in this case.

Example 2. The function $f(x) = 1/x$ is not bounded on the interval $0 < x \leqq 1$, for there is no upper bound to its values. We can make $f(x)$ as large as we please by taking x sufficiently near zero.

Example 3. The function $f(x) = x \sin x$ is not bounded on the infinite interval $0 \leqq x$. There is neither an upper nor a lower bound to the values of the function, since $f(n\pi/2) = n\pi/2$ if $n = 1, 5, 9, 13, \ldots$ and $f(n\pi/2) = -n\pi/2$ if $n = 3, 7, 11, 15, \ldots$.

THEOREM II. *If a function is continuous on a finite closed interval, it is bounded on that interval.*

First Proof. Let the function f be continuous on the interval $a \leqq x \leqq b$. We observe in the first place that if x_0 is any point of the interval $[a, b]$ there is *some* subinterval containing x_0 in which f is bounded. For, if we take $\epsilon = 1$ in the criterion for continuity as stated in §3, and denote the corresponding δ by δ_1, we have $f(x_0) - 1 < f(x) < f(x_0) + 1$ provided that x is a point of the interval $[a, b]$ such that $x_0 - \delta_1 < x < x_0 + \delta_1$. Note that this subinterval extends on both sides of x_0 if $a < x_0 < b$, and on one side of x_0 if $x_0 = a$ or $x_0 = b$. To show the dependence of δ_1 on x_0 we shall write $\delta_1 = \delta_1(x_0)$. A second observation of importance is this: If f is bounded on each of two abutting or overlapping subintervals, then it is bounded on the single interval which consists of all points which belong to either one or both of the original subintervals. For example, if $|f(x)| \leqq 10$ when $0 \leqq x \leqq 3/2$ and $|f(x)| \leqq 15$ when $1 \leqq x \leqq 3$, then certainly $|f(x)| \leqq 15$ when $0 \leqq x \leqq 3$.

Now let us define T as the set of all numbers t for which $a < t \leqq b$ and such that f is bounded on the interval $a \leqq x \leqq t$. The set T is not empty, for by the previous paragraph it certainly contains t if $a < t < a + \delta_1(a)$ (see Fig. 21). Furthermore, T has the upper bound b. Therefore, by Theorem II, §2.7, T has a least upper bound, say c. To complete the proof we shall prove two things: (1) that c belongs to T, so that f is bounded on $[a, c]$, and (2) that $c = b$. By definition of c we know that either c belongs to T, or there are points of T as near c as we please on the left of c. Now, as we observed at the outset, f is bounded in *some* subinterval containing c and extending at least a distance $\delta_1(c)$ to the left of c. There will certainly be a point of T, say t_1, such that $c - \delta_1(c) < t_1 \leqq c$. If $t_1 = c$, (1) is clear; if $t_1 < c$, f is bounded on each of the overlapping intervals $[a, t_1]$, $[c - \delta_1(c), c]$, and hence on the single interval $[a, c]$. In either case we see that c belongs to T. This proves (1).

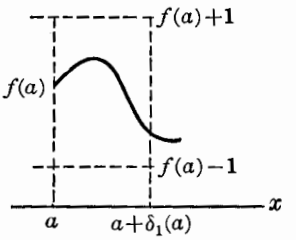


Fig. 21.

To prove (2) we assume $c < b$ and deduce a contradiction. Once more we use the fact that f is bounded in *some* subinterval containing c; since $c < b$ this subinterval will extend somewhat to the right of c, say as far as $c + \delta_1(c)$. But now we know that f is bounded on each of the abutting intervals $[a, c]$, $[c, c + \delta_1(c)]$, and hence on the single interval $[a, c + \delta_1(c)]$. This means, however, that $c + \delta_1(c)$ is in the set T, by definition. Here we have a contradiction, since $c + \delta_1(c)$ is greater than the least upper bound of T. We have thus proved (2), and thereby completed the proof of the theorem.

Second Proof. This proof uses the theorem of nested intervals (Theorem VI, §2.8). As in the first proof, we observe that if a function is bounded on each of two abutting closed intervals, it is bounded on the interval which is obtained by combining the two intervals into one. Now suppose the theorem to be proved were false; that is, suppose f is not bounded on $[a, b]$. Denote the interval $[a, b]$ by I_1. Consider the closed intervals $[a, (a + b)/2]$, $[(a + b)/2, b]$ obtained by

bisecting I_1. On at least one of these closed subintervals (denote such a one by I_2) f must fail to be bounded. We proceed to bisect I_2, obtaining a new subinterval I_3 on which f fails to be bounded. By repetition of this process we generate a sequence I_n of closed intervals on each of which f is not bounded. The length of I_n is $(b-a)/2^{n-1}$. Hence it is clear that I_n is a nest, as defined in §2.8. By Theorem VI of §2.8 there is a single point, say $x = c$, which is in each of the intervals I_n, and hence in the interval $[a, b]$. Now, as shown at the beginning of our first proof, f is bounded on *some* interval containing the point c. Denote such an interval by J. Since the length of I_n tends to zero as n increases, and since c is in I_n, it is clear that J must contain I_n when n is sufficiently large. But this involves a contradiction, for f is *not* bounded on I_n and it *is* bounded on J. Because of this contradiction, our initial assumption that the theorem is false must be rejected. We have thus completed the proof.

EXERCISES

1. Let $f(x) = 2x \sin(1/x) - \cos(1/x)$. Is this function bounded on the interval $0 < x \leqq 1$?

2. Consider the function $\tan^{-1} x$, defined for all values of x. Is it bounded?

3. Which of the following functions are bounded on the indicated intervals?

(a) $\dfrac{1-x}{1+x}, -1 < x < 1$; **(c)** $\dfrac{1}{\sqrt{5} - 2\sin x}, 0 \leqq x \leqq 2\pi$;

(b) $\dfrac{|x|}{x}, 0 < x < 1$; **(d)** $\dfrac{\sin x}{x}, 0 < x \leqq \dfrac{\pi}{2}$; **(e)** $\dfrac{1}{x} \sin \dfrac{\pi}{x}, 0 < x \leqq 1$.

4. Without attempting to find exact absolute maxima, find numbers M such that $|f(x)| \leqq M$ on the intervals indicated in each of the following cases:

(a) $f(x) = x^{17} - 6x^{13} + 5x^2 - 2, -1 \leqq x \leqq 1$;

(b) $f(x) = 3\sin^2 x - 2\cos x - \sin \dfrac{x}{2} \cos \dfrac{x}{3}, 0 \leqq x \leqq 2\pi$;

(c) $f(x) = \dfrac{x^3 - x^2 + 1}{1 + x^4}, -1 \leqq x \leqq 2$.

3.2 / THE ATTAINMENT OF EXTREME VALUES

Suppose we are given a function f, and suppose we know that the function is bounded on a certain given interval. Let m and M be the greatest lower bound and least upper bound, respectively, of the values of $f(x)$ on the given interval. Is it necessarily the case that $f(x)$ actually takes on the values m and M on the interval? Examples show that the answer to this question is negative.

Example 1. Suppose we define $f(x) = x^2$ if $0 \leqq x < 1$, $f(x) = 0$ if $x = 1$ (see Fig. 22). This function has $M = 1$ for the interval $0 \leqq x \leqq 1$, but there is no x on the interval such that $f(x) = 1$. Note, however, that the function is not continuous at $x = 1$.

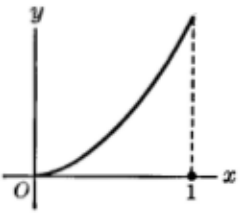


Fig. 22.

Example 2. Let $f(x) = \dfrac{1}{x+1}\sin\dfrac{1}{x}$, $0 < x$. This function is continuous for all positive values of x. We have

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

and hence $-1 < f(x) < 1$ when $x > 0$. Actually $m = -1$ and $M = 1$ for this function on any interval $0 < x \leqq a$, where $a > 0$. As x approaches 0, $f(x)$ oscillates in value between $1/(x+1)$ and $-1/(x+1)$ an infinite number of times (see Fig. 23). If N is any number such that $-1 < N < 1$, there are infinitely many values of x as near 0 as we please such that $f(x) = N$. But the values ± 1 are never attained.

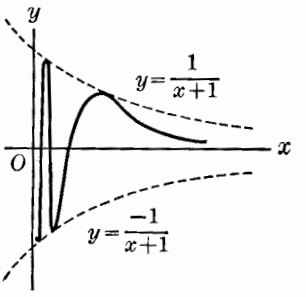


Fig. 23.

There is, however, a companion to Theorem II of the preceeding section, in which it was assumed that f is continuous on an interval that is both finite and closed. In that case the answer to the question raised in the opening paragraph of the present section is positive.

THEOREM III. *Let f be continuous when $a \leqq x \leqq b$, and let m and M be the greatest lower bound and least upper bound of the values $f(x)$ on this interval. Then $f(x)$ assumes each of the values m and M at least once in the interval.*

Proof. Suppose the value M is never attained, so that $M - f(x) > 0$ for all x in the interval. Consider the function

$$g(x) = \frac{1}{M - f(x)}.$$

It is a continuous function, since the denominator never vanishes (see Theorem I, §3). Hence, by Theorem II, §3.1, $g(x)$ is bounded. Let A be an upper bound of $g(x)$ (A is necessarily positive):

$$\frac{1}{M - f(x)} \leqq A.$$

This inequality may be transformed successively into

$$\frac{1}{A} \leqq M - f(x), f(x) \leqq M - \frac{1}{A}.$$

But we now have the result that $M - (1/A)$ is an upper bound for $f(x)$. This is a contradiction, however, for $M - (1/A) < M$, and M is the *least* upper bound of $f(x)$. We must then conclude that $M - f(x)$ vanishes for some value of x. This completes the proof so far as M is concerned. A similar proof can be given that the value m is assumed. Alternatively, one can consider the function $-f(x)$,

whose least upper bound is $-m$. Then, by what has already been proved, $-f(x) = -m$ must hold for some x. This is equivalent to $f(x) = m$, of course.

EXERCISES

1. Let $f(x) = 3x$ if $0 < x < 1$, and define $f(0) = 1$, $f(1) = 2$. Find m and M for f on the interval [0, 1]. Are these extreme values attained? What can you say about the continuity of f?

2. Let $f(x) = x - [x]$, where $[x]$ denotes the greatest integer less than or equal to x. Find m and M for f on $[0, 2]$. Are these extreme values attained? Graph the function.

3. Find m and M for $f(x) = \tan^{-1} x$, where x can range over all values. Are these extreme values attained?

4. Let $f(x) = \dfrac{x^2 + x^4}{2} + \dfrac{x^2 - x^4}{2} \sin \dfrac{\pi}{x}, 0 < x \leqq 1.$

Find m and M for this function on the indicated interval. Are these extreme values attained?

5. Given a point Q and a circle in the xy-plane, with Q not on the circle, explain with the aid of Theorem III how you know that there is a point P_1 on the circle which is at least as near to Q as any other point of the circle is; also explain why there is a point P_2 on the circle which is at least as far from Q as any other point on the circle is.

6. Given a point Q and a parabola in the xy-plane, with Q not on the parabola, explain why there is a point P on the parabola at least as close to Q as any other point of the parabola is? Why, in contrast to the case of the circle in Exercise 5, is there no point of the parabola at maximum distance from Q? If you wish, you may choose a co-ordinate system in which the parabola has a very simple equation.

3.3 / THE INTERMEDIATE-VALUE THEOREM

THEOREM IV. *Suppose that f is continuous on the closed interval $a \leqq x \leqq b$, and that $f(a) \neq f(b)$. Then, as x varies from a to b, $f(x)$ takes on every value between $f(a)$ and $f(b)$.*

This theorem expresses a property of continuous functions which has a simple geometric interpretation. Suppose, for example, that $f(a) < f(b)$, and that k is a number between $f(a)$ and $f(b)$. Consider the graph of $y = f(x)$. It is a continuous curve joining the points $(a, f(a))$ and $(b, f(b))$. These points are on opposite sides of the line $y = k$; the theorem asserts that the curve $y = f(x)$ intersects the line $y = k$ at some point $x = c$ between a and b (see Fig. 24).

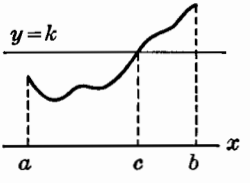


Fig. 24.

With this geometric interpretation, the student may be strongly tempted to say that the theorem is obviously true and requires no further proof. If the student does so, however, he or she is not relying upon our definition of continuity given in §3, but upon intuitive assumptions about the geometrical meaning of the term "continuous curve." We wish to show that the theorem can be proved on the basis of the definition.

Before proving Theorem IV we shall find it convenient to prove the following proposition:

THEOREM V. *Let f be defined on an interval containing the point c, and let f be continuous at $x = c$. Then, if $f(c) \neq 0$, there is a subinterval containing c throughout which $f(x)$ has the same sign as $f(c)$.*

Suppose, for example, that $f(x)$ is defined when $c - h < x < c + h$, where $h > 0$, and that $f(c) > 0$. Then if f is continuous at $x = c$ the theorem asserts that we can choose δ, $0 < \delta < h$, so that $f(x) > 0$ when $|x - c| < \delta$ (see Fig. 25). The proof is an immediate consequence of Theorem X, §1.61, since $\lim_{x \to c} f(x) = f(c) > 0$. The proof when $f(c) < 0$ is left to the student. If c is at one end of the interval on which f is defined, the subinterval on which f is of constant sign will extend on one side only of c.

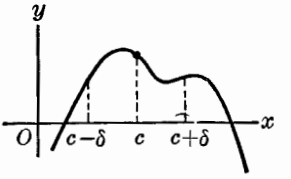


Fig. 25.

Proof of Theorem IV. Consider the function $g(x) = f(x) - k$, where k is between $f(a)$ and $f(b)$. Let us assume for definiteness that $f(a) < k < f(b)$. We shall assume the theorem false and deduce a contradiction, thus completing the proof. Thus we assume that $g(x)$ is never zero on the interval $[a, b]$.

Now $g(a) < 0$ and $g(b) > 0$. Bisect the interval $[a, b]$. At the midpoint $g(x)$ is not zero, and hence is either positive or negative. Choose that half interval for which $g(x)$ is positive at one end and negative at the other, denoting it by I_2. The interval $[a, b]$ we denote by I_1. We now repeat the bisection process, successively obtaining intervals $I_1, I_2, I_3, \ldots$ such that $g(x)$ is negative at the left end of I_n and positive at the right end. These intervals obviously form a nest, and so by Theorem VI, §2.8, close down on a unique point c. Now $g(c) \neq 0$, and so by Theorem V $g(x)$ is of one sign throughout some interval J containing c. But such an interval will contain I_n when n is sufficiently large, because the length of I_n approaches zero as n increases. Thus $g(x)$ takes on both positive and negative values in J. We have now reached a contradiction, and the proof is complete.

EXERCISES

1. Suppose f is defined as $f(x) = |x| \log |x|$ if $x \neq 0$, $f(0) = e$. Without investigating the limit of $f(x)$ as $x \to 0$, explain why it is certain that f is not continuous at $x = 0$.

2. A point Q is inside a circle C. The point of C nearest Q is a distance d from Q, and the point of C furthest from Q is a distance D from Q. Explain how you know that there is a point P on C whose distance from Q is $\frac{1}{2}(d + D)$.

3. If f is continuous and $f(x) \geqq 0$ on $[a, b]$, while $f(x) > 0$ for at least one x of the interval $[a, b]$, prove that $\int_a^b f(x)\,dx > 0$. Give a carefully reasoned proof, not merely an intuitive argument based on geometric plausibility.

4. Let P be a point on an ellipse. Consider rays (half-lines) emanating from P. Explain how you know that there is one such ray which divides the area enclosed by the ellipse into two equal parts.

5. Suppose f is continuous on $[a, b]$ and let λ be a number such that $0 < \lambda < 1$. Show that there is an x such that $a < x < b$ and $\int_a^x f(t)\,dt = \lambda \int_a^b f(t)\,dt$, provided that the condition $\int_a^b f(t)\,dt \neq 0$ is fulfilled.

MISCELLANEOUS EXERCISES

1. Six functions are defined below, each by a certain formula. For each function answer the questions: **(a)** For what values of x is the function defined? **(b)** For what values of x is the function continuous? **(c)** Are there any values of x where the function is not already defined, but may be defined so as to make the function continuous?

(i) $f(x) = \dfrac{x^3 - 1}{x^2 + x + 1}$, (iii) $h(x) = \dfrac{2^{1/x} + 3}{2^{1/x} + 1}$,

(ii) $g(x) = \dfrac{x - 1}{x^2 - 1}$, (iv) $F(x) = \dfrac{x^2 - 16}{x + 6 - 5\sqrt{x}}$,

(v) $G(x) = \dfrac{x^2 - 5x - 50}{(x^2 - 8x - 20)\sqrt{x^2 - 25}}$,

(vi) $H(x) = \dfrac{3x + 5}{\sqrt{2x - 3} - \sqrt{5x - 6} + \sqrt{3x - 5}}$.

2. Given that f is defined for all x, continuous at $x = 0$, and that, for all x and y, $f(x + y) = f(x)f(y)$, show that f is continuous for all values of x.

3. Given that f is defined for all x, continuous at $x = 0$, and that, for all x and y, $f(x + y) = f(x) + f(y)$, show that f is continuous for all values of x.

4. A function f is defined and continuous on the interval $a \leqq x \leqq b$, and $f(x) = 0$ when x is rational. Using Theorem V, explain why $f(x) = 0$ at *all* points of the interval.

5. Show that $|f(x) - f(x_0)| < |x - x_0|$ if $f(x) = \sqrt{4 + x^2}$ and $x \neq x_0$. What does this prove about f?

6. Let $P(x)$ be a polynomial of odd degree with real coefficients. Then the equation $P(x) = 0$ has at least one real root. Prove this by use of Theorem IV.

7. Let $P(x) = x^n + a_1 x^{n-1} + \cdots + a_n$, where n is an even positive integer, the a's are real, and $a_n < 0$. Show that the equation $P(x) = 0$ has at least two real roots. What more can you say about them?

8. Explain why $\dfrac{x^2 + 1}{x + 2} + \dfrac{x^4 + 1}{x - 3} = 0$ has at least one root between -2 and 3.

9. Let $f(x) = \dfrac{A}{a^2 + x} + \dfrac{B}{b^2 + x} + \dfrac{C}{c^2 + x} - 1$, where A, B, and C are positive and $a > b > c > 0$. Discuss the nature of the graph of $y = f(x)$ and explain why the equation $f(x) = 0$ has exactly three roots x_1, x_2, x_3 satisfying the inequalities $-a^2 < x_1 < -b^2 < x_2 < -c^2 < x_3$.

10. Suppose that f is continuous for all values of x, that $\lim_{x \to -\infty} f(x) = -1$, and $\lim_{x \to +\infty} f(x) = 10$. Explain how you use Theorem IV to show that there is at least one value of x such that $f(x) = 0$.

11. Suppose that f is given continuous for all x, with $f(x) < 0$ when $x < x_1$ and $f(x) > 0$ when $x > x_2$, where $x_1 < x_2$. **(a)** Define a cut (see §2.4) in such a way that the cut number c satisfies the conditions $f(c) = 0$, $f(x) < 0$ if $x < c$. **(b)** Define a cut in such a way that the cut number satisfies the conditions $f(c) = 0$, $f(x) > 0$ if $x > c$.

12. Prove Theorem III by using a repeated bisection method, as in the second proof of Theorem II. At each bisection, retain a half interval on which the least upper bound of f is M. Let $\{I_n\}$ be the resulting nest of intervals. Choose a point x_n in I_n such that $f(x_n) > M - (1/n)$. Why is this possible? What happens as $n \to \infty$? Write out the whole argument carefully, showing how you are led to a point c such that $f(c) = M$.

13. Consider the following theorem: *If f is continuous when $a \leqq x \leqq b$ and if $f(a)f(b) < 0$ (i.e., if $f(a)$ and $f(b)$ are of unlike signs), then there is some point c between a and b at which $f(c) = 0$.* As we saw in §3.3, the proof of Theorem IV may be made very easily once this theorem is proved. Give a proof, using the existence of a least upper bound, as guaranteed in Theorem II, §2.7. Start by defining S to be the set of all x such that $a \leqq x < b$ and $f(x)f(a) > 0$. Why does S have an upper bound? Let c be this upper bound and prove that $f(c) = 0$. Write out your whole argument clearly, with specific justification for each step in the reasoning.

14. A function f is defined and continuous when $0 \leqq x \leqq 1$, and $f(1) = 2$. The function has the further property that the value $f(x)$ is always a rational number. Find $f(0)$.

15. A function f is defined when $0 \leqq x \leqq 2$, with at most one point of discontinuity. Furthermore, the value $f(x)$ is rational if $0 \leqq x < 1$ and irrational if $1 < x \leqq 2$. Why must f have *exactly* one point of discontinuity? What is that point?

16. Let f be defined on $[0, 1]$ as follows: $f(x) = x$ if x is rational, $f(x) = 1 - x$ if x is irrational. **(a)** For what values of x is f continuous? **(b)** In spite of the fact that f does not satisfy the hypothesis of Theorem IV on $[0, 1]$, show that as x varies from 0 to 1, $f(x)$ takes on every value between $f(0)$ and $f(1)$. What is x, for example, if $f(x) = \sqrt{2}/2$?

17. Let f be defined for all x by the conditions: $f(x) = 0$ if x is irrational or if $x = 0$, $f(x) = 1/n$ if x is the rational number m/n, where $m \neq 0$, $n > 0$, and the fraction is reduced to its lowest terms. Explain carefully **(a)** why f is discontinuous at x_0 if x_0 is rational and not zero, **(b)** why f is continuous at x_0 if x_0 is irrational. **(c)** Is f continuous at $x = 0$?

18. Suppose that f is defined and has a continuous derivative when $a \leqq x \leqq b$, and suppose that $f(a) = f(b) = 0$, but that $f(x)$ is not 0 for every x. Let λ be a constant different from zero, and let $g(x) = f'(x) + \lambda f(x)$. Prove that there is some number ξ such that $a < \xi < b$ and $g(\xi) = 0$. This is fairly hard. There are two cases: Case 1, in which $f(x)$ assumes both positive and negative values, and Case 2, in which $f(x)$ takes on values of one sign only. The proof in Case 1 is easy. For Case 2 it is fairly easy to prove that $g(\xi) = 0$ for some ξ such that $a \leqq \xi \leqq b$, but it is much more difficult to prove the result as originally stated, with $a < \xi < b$.

19. A function f is continuous on $a \leqq x \leqq b$ and differentiable on $a < x < b$. Furthermore, $f(a) = a$ and $f(b) = b$. Show that there are two points x_1, x_2 such that $a < x_1 < x_2 < b$ and $\dfrac{1}{f'(x_1)} + \dfrac{1}{f'(x_2)} = 2$.

20. Suppose that f is defined and differentiable on the closed interval $[a, b]$, with $f'(a) > 0$, $f'(b) < 0$. Prove that $f'(\xi) = 0$ for some ξ such that $a < \xi < b$. Show also that the same conclusion may be reached if $f'(a) < 0$, $f'(b) > 0$.

21. There is a counterpart of Theorem IV for functions which need not be continuous, but are known to be derivatives of continuous functions. The theorem reads as follows: *Suppose that f is defined and differentiable on $[a, b]$, with $f'(a) \neq f'(b)$. Let k be a number between $f'(a)$ and $f'(b)$. Then $f'(\xi) = k$ at some point ξ, where $a < \xi < b$.* This theorem is known as Darboux's theorem. Prove it by setting $g(x) = f(x) - kx$ and using the results of Exercise 20.

22. Let f be a function which is defined for all x, continuous at $x = 0$, and such that $f(x + y) = f(x) + f(y)$ for all values of x and y. Show that $f(x) = Cx$, where $C = f(1)$. Begin by proving (**a**) that $f(m/n) = (m/n)f(1)$ if m and n are positive integers, (**b**) that $f(-x) = -f(x)$, and (**c**) that $f(0) = 0$. Then note Exercise 3 and apply Exercise 4 to the function $f(x) - xf(1)$.

4 / EXTENSIONS OF THE LAW OF THE MEAN

4 / INTRODUCTION

In this chapter we consider various generalizations of the law of the mean (Theorem IV, §1.2), and related topics. The most important extension of the law of the mean is Taylor's formula with remainder. In elementary calculus Taylor's formula is often closely associated with the study of expansions of functions in power series. There is indeed an important connection between Taylor's formula and expansions in power series, but the formula is not important solely because of that connection. The chapter closes with a discussion of l'Hospital's rule.

4.1 / CAUCHY'S GENERALIZED LAW OF THE MEAN

Cauchy's generalization of the law of the mean deals with two functions instead of with just one.

THEOREM I. *Let $F(x)$ and $G(x)$ be continuous on the closed interval $a \leqq x \leqq b$, and differentiable on the open interval $a < x < b$. Assume further that $G(a) \neq G(b)$, and that $F'(x)$ and $G'(x)$ never vanish simultaneously. Then for some value $x = X$ such that $a < X < b$ we have*

$$\frac{F(b) - F(a)}{G(b) - G(a)} = \frac{F'(X)}{G'(X)} . \tag{4.1-1}$$

As a special case we obtain the ordinary law of the mean (Theorem IV, §1.2) if we take $G(x) = x$.

Proof. As with the proof of the ordinary law of the mean, we appeal to Rolle's theorem (§1.2). Let us set

$$k = \frac{F(b) - F(a)}{G(b) - G(a)}, \tag{4.1-2}$$

and define

$$\phi(x) = F(x) - F(a) - k[G(x) - G(a)].$$

Observe that $\phi(a) = 0$. Also, because of the definition of k, we see that $\phi(b) = 0$. Now

$$\phi'(x) = F'(x) - kG'(x).$$

The function $\phi(x)$.satisfies the conditions of Rolle's theorem; therefore $\phi'(X) =$

0 for some X, $a < X < b$. That is,

$$0 = F'(X) - kG'(X).$$

In this formula $G'(X) \neq 0$. For if $G'(X) = 0$ then $F'(X) = 0$ also, whereas we assumed that these derivatives never vanish simultaneously. Thus we can write

$$k = \frac{F'(X)}{G'(X)},$$

a formula which is equivalent to (4.1–1) because of (4.1–2). This completes the proof.

A geometrical interpretation of (4.1–1) may be given as follows: Let a plane curve be represented parametrically by equations

$$y = F(t), x = G(t), a \leqq t \leqq b. \tag{4.1–3}$$

The slope of the curve for a given t is

$$\frac{dy}{dx} = \frac{F'(t)}{G'(t)}. \tag{4.1–4}$$

The constant k in (4.1–2) is the slope of the straight line joining the points on the curve corresponding to $t = a$ and $t = b$, respectively. The theorem says that the two slopes (4.1–2) and (4.1–4) are equal for at least one value of t between a and b (see Fig. 26).

The uses of Theorem I are largely in proving other theorems, notably Theorem III (§4.3) and Theorem VI (§4.5).

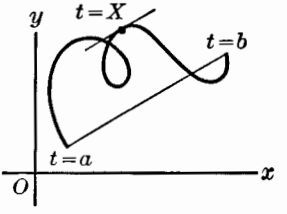


Fig. 26.

EXERCISE

Cauchy's generalized mean-value formula (4.1–1) may be written in the form

$$\begin{vmatrix} F(a) & G(a) & 1 \\ F(b) & G(b) & 1 \\ F'(X) & G'(X) & 0 \end{vmatrix} = 0.$$

This suggests the following more general theorem:

Suppose that F, G, H are continuous when $a \leqq x \leqq b$ and differentiable when $a < x < b$. Then there is a value X such that $a < X < b$ and

$$\begin{vmatrix} F(a) & G(a) & H(a) \\ F(b) & G(b) & H(b) \\ F'(X) & G'(X) & H'(X) \end{vmatrix} = 0.$$

Prove this theorem by considering

$$\phi(x) = \begin{vmatrix} F(a) & G(a) & H(a) \\ F(b) & G(b) & H(b) \\ F(x) & G(x) & H(x) \end{vmatrix}.$$

4.2 / TAYLOR'S FORMULA WITH INTEGRAL REMAINDER

Consider a polynomial $P(x)$ of degree n:

$$P(x) = b_0x^n + b_1x^{n-1} + \cdots + b_n, \quad n \geqq 0, \tag{4.2–1}$$

where $b_0 \neq 0$. If we choose any particular value of x, say $x = a$, it is possible to express $P(x)$ as a sum of powers of $(x - a)$, the highest power being n:

$$P(x) = c_0(x - a)^n + c_1(x - a)^{n-1} + \cdots + c_n. \tag{4.2–2}$$

That this is the case may be seen as follows: It is clearly true if $n = 0$, for in that case the two expressions for $P(x)$ are identical in form. For $n = 1$ we have a linear function $P(x) = b_0x + b_1$, and we wish to express it in the form $c_0(x - a) + c_1$. Choosing $c_0 = b_0$, we have

$$P(x) - b_0(x - a) = b_1 + ab_0,$$

so that $P(x) = b_0(x - a) + c_1$ with $c_1 = b_1 + ab_0$. In general, we proceed by induction, assuming that the desired type of representation is possible with polynomials of degree $\leqq n - 1$, where $n \geqq 1$. Then for the polynomial (4.2–1) we choose $c_0 = b_0$, so that $P(x) - b_0(x - a)^n$ is a polynomial of degree at most $n - 1$. Hence we can express this polynomial as a sum of powers of $(x - a)$:

$$P(x) - b_0(x - a)^n = c_1(x - a)^{n-1} + \cdots + c_n.$$

This is equivalent to (4.2–2), and completes the induction proof.

Once we know that the representation (4.2–2) is possible, it is very easy to find convenient formulas for $c_0, \ldots, c_n$. Let us differentiate (4.2–2) k times, where $0 \leqq k \leqq n$ [if $k = 0$, $P^{(k)}(x)$ means $P(x)$]. After doing this we set $x = a$. In this process the only term of $P(x)$ which leads to a non-zero result is $c_{n-k}(x - a)^k$. Therefore

$$P^{(k)}(a) = \left\{\frac{d^k}{dx^k}[c_{n-k}(x - a)^k]\right\}_{x=a} = k!c_{n-k},$$

where, according to the usual convention, $0! = 1$.
Consequently

$$c_0 = \frac{P^{(n)}(a)}{n!}, c_1 = \frac{P^{(n-1)}(a)}{(n-1)!}, \ldots, c_n = P(a),$$

so that (4.2–2) may be written in the form

$$P(x) = P(a) + P'(a)(x - a) + \frac{P''(a)}{2!}(x - a)^2 + \cdots + \frac{P^{(n)}(a)}{n!}(x - a)^n,$$

where we have reversed the order of the terms to suit our convenience.

Now let us ask whether there is any counterpart of this formula when the polynomial $P(x)$ is replaced by a function $f(x)$ which is not a polynomial. That is, let us ask what relation the expression

$$f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n$$

bears to $f(x)$. Naturally we assume that f can be differentiated n times at $x = a$. Let us assume more than that, however. A convenient assumption is that f and its first $n + 1$ derivatives are continuous in a closed interval containing $x = a$.

Now by Theorem VIII, §1.53, we know that

$$\int_a^x f'(t)\,dt = f(x) - f(a).$$

Let us write this in the form

$$f(x) = f(a) + \int_a^x f'(t)\,dt. \tag{4.2–4}$$

We transform the integral by integration by parts, taking

$$u = f'(t), \quad dv = dt,$$
$$du = f''(t)\,dt, \quad v = -(x - t).$$

Thus

$$\int_a^x f'(t)\,dt = -f'(t)(x - t)\Big|_a^x + \int_a^x f''(t)(x - t)\,dt,$$

and so

$$f(x) = f(a) + f'(a)(x - a) + \int_a^x f''(t)(x - t)\,dt. \tag{4.2–5}$$

We can now integrate by parts again, taking

$$u = f''(t), \qquad dv = (x - t)\,dt,$$
$$du = f^{(3)}(t)\,dt, \qquad v = \frac{-(x - t)^2}{2!}.$$

The integral in (4.2–5) now becomes

$$\int_a^x f''(t)(x - t)\,dt = -f''(t)\frac{(x - t)^2}{2!}\Big|_a^x + \int_a^x f^{(3)}(t)\frac{(x - t)^2}{2!}\,dt,$$

and so

$$f(x) = f(a) + f'(a)(x - a) + f''(a)\frac{(x - a)^2}{2!} + \frac{1}{2!}\int_a^x f^{(3)}(t)(x - t)^2\,dt.$$

It is now clear that repeated integration by parts will lead to the formula

$$f(x) = f(a) + f'(a)(x - a) + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n + \frac{1}{n!}\int_a^x (x - t)^n f^{(n+1)}(t)\,dt. \tag{4.2–6}$$

This is the generalization of (4.2–3) which we have been seeking. The function $f(x)$ has been expressed as a polynomial of degree n in $(x - a)$, *plus a remainder term.*

We state our findings in a formal theorem.

THEOREM II. *Let $f(x)$ and its first $n+1$ derivatives ($n \geqq 0$) be continuous in a closed interval containing $x = a$ (either inside or at one end). Let x be any point of this interval. Then*

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_{n+1}, \tag{4.2–7}$$

the remainder being given by

$$R_{n+1} = \frac{1}{n!}\int_a^x (x-t)^n f^{(n+1)}(t)\,dt. \tag{4.2–8}$$

We have indicated the procedure for proving the theorem by successive integration by parts, starting from (4.2–4). If one wishes, he may give the proof more formally by mathematical induction.

Formula (4.2–7) is called *Taylor's formula with remainder.* Various formulas for the remainder may be given, as we shall see in §4.3.

The size of the remainder may sometimes be estimated from (4.2–8). Thus, for example, if $|f^{(n+1)}(t)| \leqq M$ when $a \leqq t \leqq x$, we can see that

$$|R_{n+1}| \leqq \frac{M}{n!}\int_a^x (x-t)^n\,dt = \frac{M(x-a)^{n+1}}{(n+1)!}.$$

4.3 / OTHER FORMS OF THE REMAINDER

It is possible to obtain (4.2–7) with a different formula for R_{n+1}, under slightly less stringent assumptions. It will suffice to assume merely that $f^{(n+1)}(x)$ *exists* on an interval, without necessarily being continuous.

THEOREM III. *Let f and its first n derivatives be continuous when $a \leqq x \leqq b$ (where $a < b$), and let the $(n+1)$st derivative $f^{(n+1)}(x)$ exist when $a < x < b$. Then there is a value $x = X$, $a < X < b$, such that*

$$f(b) = f(a) + f'(a)(b-a) + \cdots + \frac{f^n(a)}{n!}(b-a)^n + \frac{f^{(n+1)}(X)}{(n+1)!}(b-a)^{n+1}. \tag{4.3–1}$$

The same formula holds in case $b < a$, all the inequalities then being reversed.

This is a generalization of the law of the mean, and actually coincides with the law of the mean in the special case $n = 0$.

It is difficult to give a proof of (4.3–1) which will seem well motivated and free from artifice. The following proof has been discovered as a result of careful study and a certain amount of trial and error.

We define two functions

$$F(x) = f(b) - f(x) - f'(x)(b-x) - \cdots - \frac{f^{(n)}(x)}{n!}(b-x)^n, \tag{4.3–2}$$

and

$$G(x) = \frac{(b-x)^{n+1}}{(n+1)!}. \tag{4.3–3}$$

Observe that $F(b) = G(b) = 0$. In calculating the derivative of $F(x)$ we find that a great deal of cancellation occurs between terms arising from the differentiation of the right member of (4.3–2). Thus

$$\frac{d}{dx}[-f'(x)(b-x)] = -f''(x)(b-x) + f'(x).$$

The $f'(x)$ here cancels the derivative of the previous term, $-f(x)$; the term $-f''(x)(b-x)$ is canceled by one of the terms coming from the differentiation of $-\frac{f''(x)}{2!}(b-x)^2$. The final result, which the student should verify, is

$$F'(x) = -\frac{f^{(n+1)}(x)}{n!}(b-x)^n. \tag{4.3–4}$$

We also have

$$G'(x) = -\frac{(b-x)^n}{n!}. \tag{4.3–5}$$

Let us now apply Cauchy's form of the law of the mean (4.1–1). Since $F(b)$ and $G(b)$ are zero, it reads

$$\frac{F(a)}{G(a)} = \frac{F'(X)}{G'(X)}, \quad \text{or} \quad F(a) = \frac{F'(X)}{G'(X)}G(a).$$

Taking account of (4.3–3), (4.3–4), and (4.3–5), this may be written

$$F(a) = f^{(n+1)}(X)\frac{(b-a)^{n+1}}{(n+1)!}. \tag{4.3–6}$$

If we now put $x = a$ in (4.3–2) and use (4.3–6), we obtain the desired formula (4.3–1). This proof is valid whether $a < b$ or $b < a$, because Cauchy's formula (4.1–1) is unaffected by an interchange of a and b.

The formula in Theorem III is written in a variety of different ways by changes in notation. One important form commonly occurring in the literature is obtained by putting $b = a + h$, where h may be either positive or negative. The number X between a and $a + h$ may then be written in the form $X = a + \theta h$, where θ is some number such that $0 < \theta < 1$. Thus we have

$$f(a+h) = f(a) + f'(a)h + \cdots + \frac{f^{(n)}(a)}{n!}h^n + \frac{f^{(n+1)}(a+\theta h)}{(n+1)!}h^{n+1}. \tag{4.3–7}$$

Another form results by putting $b = x$ in (4.3–1). In this form we may write

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_{n+1}, \tag{4.3–8}$$

with

$$R_{n+1} = \frac{f^{(n+1)}(X)}{(n+1)!}(x-a)^{n+1}, \tag{4.3–9}$$

where X lies between x and a.

Formula (4.3–9) is called *Lagrange's form* of the remainder.

Brook Taylor (English) published in 1715 the form of the infinite series (4.3–10) which bears his name today. The formulas (4.3–8) and (4.3–9) were derived by J. L. Lagrange (French) in 1797. Lagrange's proof utilized integrals. The first proof using the law of the mean appears to have been by Ampère in 1806.

The student should distinguish between (4.3–8) and Taylor's *series*, which is the infinite-series formula

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \cdots. \qquad (4.3\text{–}10)$$

This expansion of a function in powers of $(x-a)$ is valid under certain circumstances. Proofs of such validity in particular cases may be made with the aid of Taylor's formula and one or the other of the forms of the remainder. Power-series developments are considered systematically in Chapter 21.

***Example* 1.** Write Taylor's formula for $f(x) = 1/(2+x)$ with $a = -1$ and $n = 2$, using Lagrange's form of the remainder.

We have $f(-1) = 1$ and

$$f'(x) = \frac{-1}{(2+x)^2}, \qquad f'(-1) = -1,$$

$$f''(x) = \frac{2}{(2+x)^3}, \qquad f''(-1) = 2,$$

$$f^{(3)}(x) = \frac{-6}{(2+x)^4}, \qquad f^{(3)}(-1) = -6.$$

Thus (4.3–8) becomes

$$\frac{1}{2+x} = 1 - (x+1) + (x+1)^2 + R_3$$

with

$$R_3 = \frac{-6}{(2+X)^4}\frac{(x+1)^3}{3!} = -\frac{(x+1)^3}{(2+X)^4},$$

where X is between x and -1. If we wish to estimate R_3 we observe that $1/(2+X)$ lies between $1/(2+x)$ and 1, and hence $|R_3|$ lies between $\dfrac{|x+1|^3}{(2+x)^4}$ and $|x+1|^3$. For example, if $x = -0.9$, R_3 is negative, and in absolute value between $0.001/(1.1)^4$ and 0.001.

***Example* 2.** Write Taylor's formula for $f(x) = \log(1+x)$ with $a = 0$ and a general value of n. Estimate R_{n+1} if $0 \leqq x \leqq \frac{1}{2}$.

We have $f(0) = 0$ and

$$f'(x) = \frac{1}{1+x}, \qquad f'(0) = 1,$$

$$f''(x) = \frac{-1}{(1+x)^2}, \qquad f''(0) = -1,$$

$$\cdots\cdots\cdots\cdots \qquad \cdots\cdots\cdots$$

$$f^{(n)}(x) = \frac{(-1)^{n-1}(n-1)!}{(1+x)^n}, \qquad f^{(n)}(0) = (-1)^{n-1}(n-1)!$$

The general formula for $f^{(n)}(x)$ may be surmised after the first few instances and verified by induction. Thus

$$\log(1+x) = x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 - \cdots + (-1)^{n-1}\frac{1}{n}x^n + R_{n+1}, \tag{4.3–11}$$

$$R_{n+1} = (-1)^n \frac{1}{n+1}\left(\frac{x}{1+X}\right)^{n+1}, \tag{4.3–12}$$

where X is between 0 and x. Of course we must have $0 < 1 + x$, or $-1 < x$. If $0 \leqq x \leqq \frac{1}{2}$, then

$$0 \leqq \frac{x}{1+X} \leqq \frac{1}{2},$$

and so

$$|R_{n-1}| \leqq \frac{1}{n+1} \cdot \frac{1}{2^{n+1}}.$$

There are many other possible formulas for R_{n+1} besides that of Lagrange. Probably the most important of these other forms is that due to Cauchy. It appears in formula (4.3–13), which follows.

THEOREM IV. *An alternative form of the remainder in* (4.3–1), *with hypotheses as in Theorem III, is*

$$R_{n+1} = \frac{f^{(n+1)}(X)}{n!}(b-X)^n(b-a), \tag{4.3–13}$$

where X is some number between a and b (in general different from the X of (4.3–1)). *With $b = a + h$, $X = a + \theta h$ $(0 < \theta < 1)$, this takes the form*

$$R_{n+1} = \frac{f^{(n+1)}(a+\theta h)}{n!}h^{n+1}(1-\theta)^n. \tag{4.3–14}$$

This last form is an alternative to the last term in (4.3–7), *but the θ's in the two forms are usually different.*

Proof. We start with $F(x)$ as defined by (4.3–2), but instead of using (4.3–3) we define $G(x) = b - x$. Then, following the method of proof of Theorem III, we have (4.3–4) and the formula $G'(x) = -1$ in place of (4.3–5).

Just as before we have, for some X between a and b,

$$F(a) = \frac{F'(X)}{G'(X)}G(a),$$

or

$$F(a) = \frac{f^{(n+1)}(X)}{n!}(b-X)^n(b-a).$$

This is equivalent to (4.3–13), for comparison of (4.3–2) and (4.3–1) shows that the remainder term is $R_{n+1} = F(a)$.

Other forms of the remainder, as well as other proofs of Theorems III and IV, are indicated in the exercises.

Example 3. Find Cauchy's form of R_{n+1} for $f(x) = \log(1+x)$ with $a = 0$.

Using the work of Example 2, and putting $h = x$, $a = 0$ in (4.3–14), we find

$$R_{n+1} = (-1)^n \frac{x^{n+1}(1-\theta)^n}{(1+\theta x)^{n+1}}. \tag{4.3–15}$$

Example 4. Show that, in Taylor's formula (4.3–11) for $\log(1+x)$, the following estimates of the remainder hold:

$$|R_{n+1}| < \frac{|x|^{n+1}}{1+x} \text{ if } -1 < x < 0, \tag{4.3–16}$$

$$|R_{n+1}| < \frac{x^{n+1}}{n+1} \text{ if } 0 < x \leqq 1. \tag{4.3–17}$$

We use Cauchy's form of the remainder to get (4.3–16), and Lagrange's form to get (4.3–17). If $-1 < x < 0$ we write (4.3–15) in the form

$$R_{n+1} = (-1)^n \frac{x^{n+1}}{1+\theta x}\left(\frac{1-\theta}{1+\theta x}\right)^n,$$

and observe that $1 + \theta x > 1 + x$ and

$$0 < \frac{1-\theta}{1+\theta x} < 1.$$

The inequality (4.3–16) is then seen to be correct. On the other hand, if $0 < x \leqq 1$, we observe that $0 < X < x$ and hence $1 + X > 1$ in (4.3–12). The inequality (4.3–17) is then seen to be correct.

We observe from the foregoing inequalities that $R_{n+1} \to 0$ as $n \to \infty$ if $-1 < x < 0$ or if $0 < x \leqq 1$. Of course $R_{n+1} = 0$ if $x = 0$. It follows that Taylor's formula, neglecting the remainder, gives a better and better approximation to $\log(1+x)$ when n is increased, provided that $-1 < x \leqq 1$. Accordingly, we have the Taylor's series expansion, valid if $-1 < x \leqq 1$:

$$\log(1+x) = x - \tfrac{1}{2}x^2 + \cdots + (-1)^{n-1}\frac{x^n}{n} + \cdots. \tag{4.3–18}$$

EXERCISES

1. Arrange x^4 in powers of $(x-3)$.

2. Write Taylor's formula with Lagrange's remainder in the case of $f(x) = \dfrac{1}{x^2+3}$ with $a = 1$, $n = 2$.

3. Write Taylor's formula with Lagrange's remainder in each of the following cases:

(a) $f(x) = \sin^2 x$, $a = 0$, $n = 3$;

(b) $f(x) = \tan x$, $a = 0$, $n = 3$;

(c) $f(x) = e^{-x^2}$, $a = 0$, $n = 3$;

(d) $f(x) = \log(1 + e^x)$, $a = 0$, $n = 2$;

(e) $f(x) = \dfrac{\log(1-x)}{1-x}$, $a = 0$, $n = 1$.

4. Find R_5 in (4.3–9) with $f(x) = \sin x$, $a = \pi/2$.

5. Write Taylor's formula for $f(x) = \sin x$ with $a = 0$. Show that

$$|R_{n+1}| \leqq \frac{|x|^{n+1}}{(n+1)!}.$$

Show that the same inequality holds for $f(x) = \cos x$, $a = 0$.

6. Show that

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + R_{n+1},$$

with

$$0 < R_{n+1} < e^x \frac{x^{n+1}}{(n+1)!} \text{ if } 0 < x$$

and

$$|R_{n+1}| < \frac{|x|^{n+1}}{(n+1)!} \text{ if } x < 0.$$

7. Show that

$$(1-x)^{-1/2} = 1 + \tfrac{1}{2}x + \frac{1 \cdot 3}{2 \cdot 4}x^2 + \cdots + \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots (2n)}x^n + R_{n+1},$$

and write both Lagrange's and Cauchy's form of the remainder. Show that

$$|R_{n+1}| < \frac{1 \cdot 3 \cdots (2n+1)}{2 \cdot 4 \cdots (2n+2)}|x|^{n+1} \text{ if } -1 < x < 0$$

and

$$|R_{n+1}| < \frac{1 \cdot 3 \cdots (2n+1)}{2^{n+1}n!}\frac{x^{n+1}}{(1-x)^{3/2}} \text{ if } 0 < x < 1.$$

8. Observe that the curves $y = \sin x$, $y = \lambda x$ intersect near $x = \pi$ if λ is small. Let $f(x) = \sin x - \lambda x$ and apply Taylor's formula with $a = \pi$, $n = 2$, assuming that x is near π and neglecting R_3. Use this result to show that an approximate solution of $\sin x = \lambda x$ is $x = \pi/(1 + \lambda)$.

9. Proceed as in Exercise 8 to find an approximate formula for the solution of $\operatorname{ctn} x = \lambda x$ near $x = \pi/2$ (assuming that λ is small).

10. Suppose that f is twice differentiable in the interval $a < x < b$ and that $f''(x) \geqq 0$ at each point. If $a < x_0 < b$ and $y_0 = f(x_0)$, show that in the given interval no point of the curve $y = f(x)$ is below the line which is tangent to the curve at (x_0, y_0). In particular, if $f'(x_0) = 0$, the function has a minimum value at x_0.

11. In the method of proof for Theorem III let the function $G(x)$ in (4.3–3) be replaced by $G(x) = (x - b)^p$. Then carry on the method and show that the expression for the last term in (4.3–7) is replaced by

$$R_{n+1} = \frac{f^{(n+1)}(a + \theta h)}{n!p}h^{n+1}(1-\theta)^{n+1-p}, \qquad 0 < \theta < 1.$$

This is called Schlömilch's form of the remainder. Lagrange's form is the special case $p = n + 1$, and Cauchy's form is the special case $p = 1$.

12. The following suggestions provide a method of proving Taylor's formula with remainder without appeal to Cauchy's generalized law of the mean. Suppose that g is continuous on the closed interval $[a, a + h]$ and differentiable on the open interval

$(a, a+h)$. Assume further that g has the properties (i) $g(a)=1$, (ii) $g(a+h)=0$, (iii) $g'(x)\neq 0$ if $a<x<a+h$. Define $F(x)$ by (4.3–2) with $a+h$ in place of b. Define $\phi(x)=F(x)-g(x)F(a)$. Apply Rolle's theorem to ϕ and hence obtain a formula for $R_{n+1}=F(a)$. The result is

$$f(a+h)=f(a)+f'(a)h+\cdots+\frac{f^{(n)}(a)}{n!}h^n-\frac{f^{(n+1)}(a+\theta h)h^n(1-\theta)^n}{n!g'(a+\theta h)}.$$

If we choose

$$g(x)=\left(\frac{a+h-x}{h}\right)^{n+1}$$

we obtain (4.3–7) (Lagrange's form). The choice $g(x)=\dfrac{a+h-x}{h}$ leads to Cauchy's form (4.3–14).

4.4 / AN EXTENSION OF THE MEAN VALUE THEOREM FOR INTEGRALS

Theorem VI of §1.51 may be generalized somewhat as follows:

THEOREM V. *Let $f(x)$ and $p(x)$ be continuous functions defined on the closed interval $a\leqq x\leqq b$, and suppose that $p(x)\geqq 0$ for every x of the interval. Then there is some number X such that $a\leqq X\leqq b$ and*

$$\int_a^b f(x)p(x)\,dx=f(X)\int_a^b p(x)\,dx. \tag{4.4–1}$$

Proof. Let m and M be the minimum and maximum values of $f(x)$ on the interval $[a, b]$. Then

$$mp(x)\leqq f(x)p(x)\leqq Mp(x)$$

and so

$$m\int_a^b p(x)\,dx\leqq\int_a^b f(x)p(x)\,dx\leqq M\int_a^b p(x)\,dx. \tag{4.4–2}$$

There are two cases to consider:

$$(1)\ \int_a^b p(x)\,dx>0;\ (2)\ \int_a^b p(x)\,dx=0.$$

In case (1) we have

$$m\leqq\frac{\displaystyle\int_a^b f(x)p(x)\,dx}{\displaystyle\int_a^b p(x)\,dx}\leq M.$$

Since the quotient of the integrals lies between the extreme values attained by $f(x)$ on the interval, it follows by Theorem IV of §3.3 that there must be some

point $x = X$ on the interval such that

$$f(X) = \frac{\int_a^b f(x)p(x)\,dx}{\int_a^b p(x)\,dx}.$$

This is equivalent to (4.4–1).

In case (2) we see by (4.4–2) that the left member of (4.4–1) is zero. The right member is also zero, no matter how we choose X. Thus the proof is complete.

As one application of Theorem V we shall show how Lagrange's form of the remainder in Taylor's formula (see (4.3–9)) may be deduced from the integral form (4.2–8) of the remainder. Suppose first that $a < x$. Then the function $(x-t)^n$ is $\geqq 0$ when $a \leqq t \leqq x$. Hence we can apply Theorem V to the integral in (4.2–8), taking $p(t) = (x-t)^n$. Thus

$$\int_a^x (x-t)^n f^{(n+1)}(t)\,dt = f^{(n+1)}(X)\int_a^x (x-t)^n\,dt$$
$$= f^{(n+1)}(X)\left[\frac{-(x-t)^{n+1}}{n+1}\right]_a^x = f^{(n+1)}(X)\frac{(x-a)^{n+1}}{n+1},$$

where X is some number such that $a \leqq X \leqq x$. When this result is put back into (4.2–8) we obtain the Lagrange form (4.3–9). If $x < a$ we write (4.2–8) in the form

$$R_{n+1} = \frac{(-1)^{n+1}}{n!}\int_x^a (t-x)^n f^{(n+1)}(t)\,dt.$$

If we take $p(t) = (t-x)^n$ in this integral, then $p(t) \geqq 0$, and we can apply Theorem V. We leave it for the student to carry out the details of showing that we are once more led to formula (4.3–9).

It should be pointed out that in the present section we are assuming the *continuity* of $f^{(n+1)}(x)$, whereas in §4.3 we merely assumed the existence of $f^{(n+1)}(x)$.

4.5 L'HOSPITAL'S RULE

Theorem XIV of §1.64 states that

$$\lim_{x\to c}\frac{f(x)}{g(x)} = \frac{\lim\limits_{x\to c} f(x)}{\lim\limits_{x\to c} g(x)},$$

provided that $\lim_{x\to c} f(x)$ and $\lim_{x\to c} g(x)$ both exist and $\lim_{x\to c} g(x) \neq 0$. There are many instances in which these conditions are not fulfilled, however, e.g.,

$$\lim_{x\to 0}\frac{e^x - \cos x}{\tan x} = ? \qquad \lim_{x\to+\infty}\frac{\log x}{x} = ?$$

$$\lim_{x\to+\infty}\frac{e^x}{x^{10}} = ? \qquad \lim_{x\to 0}\frac{e^{-1/x^2}}{x^5} = ?$$

The important cases not covered by the theorem of §1.64, and not otherwise easily disposed of, are of two kinds. The first kind is that in which $f(x)$ and $g(x)$ both approach 0 as $x \to c$. The second kind is that in which $|g(x)| \to \infty$ as $x \to c$. In both cases it may be that $x \to c$ from one side only; it may also happen that $x \to c$ is replaced by $x \to +\infty$ or $x \to -\infty$.

There is a useful rule for finding the limit of a quotient in the most commonly occurring instances of the cases just mentioned. This rule was popularized by the Marquis de l'Hospital in his textbook on calculus published in 1696, and is generally known as l'Hospital's rule. The exact statement of the rule actually amounts to the statement of two theorems, or of one theorem with two alternative assumptions leading to the same conclusion. Before coming to the formal statement of the theorem, we need to make a few preliminary remarks. We shall have two functions $f(x)$ and $g(x)$ to consider. Let c denote either a real number or one of the symbols $+\infty$, $-\infty$. We assume that f and g are defined on a portion of the x-axis which we denote by I, and we assume that, if c is a real number, I is a finite open interval with c as one of its end-points, while if c is $+\infty$ or $-\infty$, I is a semi-infinite open interval, extending indefinitely in the positive direction if $c = +\infty$, and indefinitely in the negative direction if $c = -\infty$. We then talk about limits as $x \to c$, it being understood that x is to range over I, and to approach c from one side only. We furthermore assume that the derivatives $f'(x)$, $g'(x)$ exist at each point of I, and that $g(x)$ and $g'(x)$ are never equal to zero:

THEOREM VI. *(L'Hospital's rule.) Suppose either that*

Case 1. $f(x) \to 0$ *and* $g(x) \to 0$ *as* $x \to c$,

or that

Case 2. $|g(x)| \to \infty$ *as* $x \to c$.

Let A denote either a real number or one of the symbols $+\infty$, $-\infty$. Suppose that

$$\lim_{x \to c} \frac{f'(x)}{g'(x)} = A. \tag{4.5–1}$$

Then it is also true that

$$\lim_{x \to c} \frac{f(x)}{g(x)} = A. \tag{4.5–2}$$

Proof. The proof of Theorem VI is easiest when I is a finite interval and Case 1 is assumed. Indications as to the procedure to be followed are given in Exercise 8. The situation is less simple if c is $+\infty$ or $-\infty$, or if we assume Case 2. We shall give the proof in such a way that the argument does not depend on whether or not I is a finite interval, and also so that the arguments for Case 1 and Case 2 are very much alike.

We take x and y to be any distinct points of I, with y between x and c. Then,

by Cauchy's generalized law of the mean (Theorem I, §4.1), there is some point X between x and y such that

$$\frac{f(x)-f(y)}{g(x)-g(y)}=\frac{f'(X)}{g'(X)}. \tag{4.5–3}$$

Observe that X also is between x and c. We are assured that $g(x)\neq g(y)$ by the ordinary law of the mean, since $g'(x)$ is never zero, by hypothesis.

Now consider Case 1, and rewrite (4.5–3) in the form

$$\frac{\dfrac{f(x)}{g(x)}-\dfrac{f(y)}{g(x)}}{1-\dfrac{g(y)}{g(x)}}=\frac{f'(X)}{g'(X)}.$$

Now suppose, for definiteness, that A is not $+\infty$ or $-\infty$. Suppose $\epsilon>0$. The meaning of (4.5–1) is that there is some number x_0 such that, if x is between x_0 and c,

$$A-\epsilon<\frac{f'(x)}{g'(x)}<A+\epsilon. \tag{4.5–5}$$

If x is such a number, so is X, and hence, no matter how y is chosen, it follows from (4.5–4) that

$$A-\epsilon<\frac{\dfrac{f(x)}{g(x)}-\dfrac{f(y)}{g(x)}}{1-\dfrac{g(y)}{g(x)}}<A+\epsilon.$$

Now make $y\to c$. Then, since we are assuming Case 1, we conclude that

$$A-\epsilon\leqq\frac{f(x)}{g(x)}\leqq A+\epsilon.$$

This holds if x is between x_0 and c, where x_0 may depend on ϵ. But this means that (4.5–2) is true.

The argument is essentially the same if A is $+\infty$ or $-\infty$. For example, if $A=+\infty$, we take any number M, and choose x_0 so that

$$M<\frac{f'(x)}{g'(x)}$$

if x is between x_0 and c. Then we find that, for such values of x,

$$M\leqq\frac{f(x)}{g(x)}.$$

So we get (4.5–2) from (4.5–1) in this case also. The student may wish to review the definitions of limits involving the symbols $+\infty$, $-\infty$; these definitions are found in §1.61.

To treat Case 2, we return to (4.5–3) and write it in the form

$$\frac{\dfrac{f(y)}{g(y)}-\dfrac{f(x)}{g(y)}}{1-\dfrac{g(x)}{g(y)}}=\frac{f'(X)}{g'(X)}. \tag{4.5–6}$$

As in the previous argument, we may suppose the assumption (4.5–1) expressed in the form (4.5–5), with the consequence from (4.5–6) that

$$A-\epsilon<\frac{\dfrac{f(y)}{g(y)}-\dfrac{f(x)}{g(y)}}{1-\dfrac{g(x)}{g(y)}}<A+\epsilon$$

for all x between x_0 and c and all y between x and c. We shall now let $y\to c$, keeping x fixed. Since $|g(y)|\to\infty$, we may safely assume that

$$1-\frac{g(x)}{g(y)}>0.$$

We then see that

$$(A-\epsilon)\left[1-\frac{g(x)}{g(y)}\right]+\frac{f(x)}{g(y)}<\frac{f(y)}{g(y)}<\frac{f(x)}{g(y)}+(A+\epsilon)\left[1-\frac{g(x)}{g(y)}\right].$$

As $y\to c$, the left and right members of this set of inequalities approach $A-\epsilon$ and $A+\epsilon$, respectively. Certainly then it will be true that, for y beyond a certain point,

$$A-2\epsilon<\frac{f(y)}{g(y)}<A+2\epsilon.$$

This, however, is equivalent to saying that

$$\lim_{y\to c}\frac{f(y)}{g(y)}=A,$$

which is what we set out to prove. As in Case 1, the argument is not very different if A is $+\infty$ or $-\infty$.

Example 1. Take $f(x)=e^x-\cos x$ and $g(x)=\tan x$ with $c=0$. The conditions of Theorem VI, Case 1, are fulfilled with x approaching zero from either side. Thus

$$\lim_{x\to 0}\frac{e^x-\cos x}{\tan x}=\lim_{x\to 0}\frac{e^x+\sin x}{\sec^2 x}=\frac{1+0}{1}=1.$$

If, in attempting to apply Theorem VI, it should turn out that we have $\lim_{x\to c} f'(x)=0$ and $\lim_{x\to c} g'(x)=0$, or that $\lim_{x\to c}|g'(x)|\to\infty$, it may be that the rule can be applied a second time.

Example 2. Find $\lim_{x\to 0}\dfrac{e^x+e^{-x}-2}{3x^2}$.

Here one application of Theorem VI gives

$$\lim_{x\to 0}\frac{e^x+e^{-x}-2}{3x^2}=\lim_{x\to 0}\frac{e^x-e^{-x}}{6x}.$$

A second application gives

$$\lim_{x\to 0}\frac{e^x-e^{-x}}{6x}=\lim_{x\to 0}\frac{e^x+e^{-x}}{6}=\tfrac{2}{6}=\tfrac{1}{3}.$$

Hence the original limit is equal to $\frac{1}{3}$.

The effect of Theorem VI is to replace the original problem of finding $\lim_{x\to c}\frac{f(x)}{g(x)}$ by the new problem of finding $\lim_{x\to c}\frac{f'(x)}{g'(x)}$, provided this latter limit exists or is $+\infty$ or $-\infty$. Any legitimate methods may be brought to bear on the new problem, including algebraic or trigonometric transformations of the problem, breaking the quotient up into a product of simpler quotients, repeated use of Theorem VI, etc.

Example 3. Find $\lim_{x\to 0}\frac{(e^x-1)\sin x}{\cos x-\cos^2 x}$.

We first apply Theorem VI:

$$\lim_{x\to 0}\frac{(e^x-1)\sin x}{\cos x-\cos^2 x}=\lim_{x\to 0}\frac{(e^x-1)\cos x+e^x\sin x}{-\sin x+2\cos x\sin x.}$$

Next we simplify the new quotient:

$$\frac{(e^x-1)\cos x+e^x\sin x}{-\sin x+2\cos x\sin x}=\frac{\cos x}{2\cos x-1}\cdot\frac{e^x-1}{\sin x}+\frac{e^x}{2\cos x-1}.$$

Now

$$\lim_{x\to 0}\frac{\cos x}{2\cos x-1}=\frac{1}{2-1}=1,\quad \lim_{x\to 0}\frac{e^x}{2\cos x-1}=\frac{1}{2-1}=1,$$

and

$$\lim_{x\to 0}\frac{e^x-1}{\sin x}=\lim_{x\to 0}\frac{e^x}{\cos x}=\frac{1}{1}=1.$$

Hence the original limit is equal to 2.

The next example illustrates Case 1 with $c=+\infty$.

Example 4. Find $\lim_{x\to+\infty} x\log\frac{x+1}{x-1}$.

Here we write

$$x\log\frac{x+1}{x-1}=\frac{\log\dfrac{x+1}{x-1}}{\dfrac{1}{x}},$$

taking $f(x) = \log\dfrac{x+1}{x-1}$, $g(x) = \dfrac{1}{x}$. It is clear that Theorem VI is applicable. Thus

$$\lim_{x\to+\infty} \frac{\log\dfrac{x+1}{x-1}}{1/x} = \lim_{x\to+\infty} \frac{\log(x+1) - \log(x-1)}{1/x}$$

$$= \lim_{x\to+\infty} \frac{\dfrac{1}{x+1} - \dfrac{1}{x-1}}{-1/x^2}.$$

On simplification, the new limit takes the form

$$\lim_{x\to+\infty} \frac{2x^2}{x^2-1} = 2,$$

so that the original limit has the value 2.

Next we illustrate Case 2 of Theorem VI.

Example 5. Find $\lim\limits_{x\to+\infty} \dfrac{e^x}{x^{10}}$.

Case 2 of Theorem VI is applicable, but it must be applied ten times before we reach a result. We have

$$\lim_{x\to+\infty} \frac{e^x}{x^{10}} = \lim_{x\to+\infty} \frac{e^x}{10x^9} = \lim_{x\to+\infty} \frac{e^x}{10\cdot 9x^8} = \cdots = \lim_{x\to+\infty} \frac{e^x}{10!} = +\infty.$$

It might at first sight appear as though we could solve the problem of Example 5 by use of Case 1. For we can write

$$\frac{e^x}{x^{10}} = \frac{x^{-10}}{e^{-x}},$$

and if $f(x) = x^{-10}$, $g(x) = e^{-x}$, the conditions of Case 1 are satisfied. But when we differentiate, things get worse instead of better:

$$\lim_{x\to+\infty} \frac{x^{-10}}{e^{-x}} = \lim_{x\to+\infty} \frac{-10x^{-11}}{-e^{-x}} = \lim_{x\to+\infty} \frac{11\cdot 10x^{-12}}{e^{-x}} = \cdots.$$

As one of the very interesting applications of l'Hospital's rule, let us discuss the function e^{-1/x^2}.

Example 6. Let us define $f(x) = e^{-1/x^2}$ if $x \neq 0$, and $f(0) = 0$. Then all the derivatives of $f(x)$ have the value 0 at $x = 0$, as we shall now show.

In the first place, f is continuous at $x = 0$, for $f(x) \to 0$ as $x \to 0$ (see Example 6, §1.1). Now, by definition,

$$f'(0) = \lim_{x\to 0} \frac{f(x) - f(0)}{x - 0} = \lim_{x\to 0} \frac{e^{-1/x^2}}{x}.$$

This appears to be a place to use Case 1 of Theorem VI, but the application of

the method gives

$$\lim_{x\to 0}\frac{e^{-1/x^2}}{x}=\lim_{x\to 0}\frac{2x^{-3}e^{-1/x^2}}{1}=\lim_{x\to 0}\frac{2e^{-1/x^2}}{x^3},$$

and the new limit is harder to deal with, rather than easier. We may, however, use Case 2 of Theorem VI, as follows:

$$\lim_{x\to 0}\frac{e^{-1/x^2}}{x}=\lim_{x\to 0}\frac{1/x}{e^{1/x^2}}=\lim_{x\to 0}\frac{-1/x^2}{-\frac{2}{x^3}e^{1/x^2}}=\lim_{x\to 0}\tfrac{1}{2}xe^{-1/x^2}=0.$$

This shows that $f'(0)=0$.

To deal with higher derivatives we observe that, if $x\neq 0$,

$$f'(x)=2x^{-3}e^{-1/x^2},$$
$$f''(x)=4x^{-6}e^{-1/x^2}-6x^{-4}e^{-1/x^2}.$$

It is easy to see, by induction, that $f^{(n)}(x)$ is a linear combination of terms of the form

$$e^{-1/x^2}/x^m$$

with $0<m\leqq 3n$. Consequently, to see that $f^{(n)}(0)=0$, as well as to see that $f^{(n)}(x)$ is continuous at $x=0$, it will be sufficient to show that

$$\lim_{x\to 0}\frac{e^{-1/x^2}}{x^m}=0 \tag{4.5–7}$$

for all positive integers m. We prove (4.5–7) by repeated use of Case 2 of Theorem VI:

$$\lim_{x\to 0}\frac{x^{-m}}{e^{1/x^2}}=\lim_{x\to 0}\frac{-mx^{-m-1}}{-2x^{-3}e^{1/x^2}}=\frac{m}{2}\lim_{x\to 0}\frac{x^{-m+2}}{e^{1/x^2}};$$

after a finite number of steps the exponent in the numerator will be positive, and then the limit is seen to be 0.

The graph of $y=e^{-1/x^2}$ is indicated in Fig. 27. It is very flat (but of course not perfectly straight) in the neighborhood of the origin.

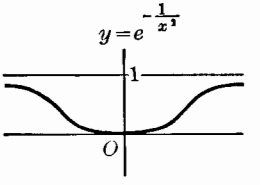


Fig. 27.

There are various kinds of functions whose limiting values may be found by using suitable devices to bring the problem to a form where l'Hospital's rule is applicable. The principal types of problems and the appropriate devices are indicated in Exercises 3 and 4.

EXERCISES

1. Evaluate the following limits:

(a) $\lim_{x\to 0+}\dfrac{\sin 3x}{1-\cos 4x}$. **(c)** $\lim_{x\to 0}\dfrac{1}{x}\log\dfrac{1+x}{1-x}$. **(e)** $\lim_{x\to 0}\dfrac{\tan^{-1}x-x}{x^3}$.

(b) $\lim_{x\to 0}\dfrac{\tan x-\sin x}{x^3}$. **(d)** $\lim_{x\to 0}\dfrac{x^3\sin x}{(1-\cos x)^2}$. **(f)** $\lim_{x\to+\infty}\dfrac{\log(1+e^{-x})}{e^{-x}}$.

2. Evaluate the following limits:

(a) $\lim_{x\to+\infty} \frac{x^n}{e^x}$, n a positive integer.

(b) $\lim_{x\to+\infty} \frac{(\log x)^n}{x}$, n a positive integer.

(c) $\lim_{x\to+\infty} \frac{\log x}{\sqrt{1+x^3}}$.

(d) $\lim_{x\to+\infty} \frac{x^3(\log x)^2}{e^x}$.

(e) $\lim_{x\to+\infty} \frac{\log(1+e^{2x})}{x}$.

(f) $\lim_{x\to+\infty} \frac{\log(1+xe^{2x})}{x^2}$.

3. One sometimes wants to find the limit of a function of the form $y = [F(x)]^{G(x)}$ as x approaches some limit x_0, or as $x \to \infty$. In such problems $F(x)$ is positive. If $F(x)$ and $G(x)$ both approach nonzero limits, the problem presents no unusual difficulty. There are, however, three cases in which the limit of y is not apparent without some investigation:

CASE 1. $F(z) \to 1$ and $G(x) \to \infty$;
CASE 2. $F(x) \to \infty$ and $G(x) \to 0$;
CASE 3. $F(x) \to 0$ and $G(x) \to 0$.

In these cases it is usually appropriate to investigate the limit of the logarithm of y:

$$\log y = G(x) \log F(x) = \frac{\log F(x)}{1/G(x)}.$$

The rule of l'Hospital may then be applicable. If $\log y \to b$, then $y \to e^b$.

Use this procedure to evaluate the following limits:

(a) $\lim_{x\to 0+} x^x$.

(b) $\lim_{x\to+\infty} x^{1/x}$.

(c) $\lim_{x\to 0} \left(\frac{\sin x}{x}\right)^{1/x^2}$.

(d) $\lim_{x\to+\infty} \left(1+\frac{1}{2x}\right)^{x^2}$.

(e) $\lim_{x\to+\infty} (e^x + e^{2x})^{1/x}$.

(f) $\lim_{x\to 0+} \left(\log \frac{1}{x}\right)^x$.

4. Sometimes a limit is not easy to determine because it is of the type $\lim_{x\to x_0}(f(x) - g(x))$, where $f(x)$ and $g(x)$ both become infinite as $x \to x_0$ (or as $x \to \infty$). In practice the best plan for such cases is usually to bring the expression $f(x) - g(x)$ to the form of a single quotient, e.g.,

$$\lim_{x\to 0}\left(\frac{1}{x} - \frac{1}{\sin x}\right) = \lim_{x\to 0} \frac{\sin x - x}{x \sin x}.$$

The limit may then be treated by l'Hospital's rule, or by devices of algebraic or trigonometric reduction. In some cases an algebraic device alone will be sufficient, e.g.,

$$\lim_{x\to+\infty} (x - \sqrt{x^2+1}) = \lim_{x\to+\infty} \frac{(x-\sqrt{x^2+1})(x+\sqrt{x^2+1})}{x+\sqrt{x^2+1}}$$
$$= \lim_{x\to+\infty} \frac{-1}{x+\sqrt{x^2+1}} = 0.$$

Evaluate the following limits:

(a) $\lim_{x\to 0}\left(\frac{1}{x} - \frac{1}{\sin x}\right)$.

(b) $\lim_{x\to 0}\left(\frac{1}{x \sin x} - \frac{1}{x^2}\right)$.

(c) $\lim_{x\to 0+}\left(\frac{1}{x} - \log\frac{1}{x}\right)$.

(d) $\lim_{x\to 0}\left(\frac{1}{\log(1+x)} - \frac{1}{x}\right)$.

(e) $\lim_{x\to 0}\left(\frac{1}{x^2} - \frac{1}{\tan^2 x}\right)$.

(f) $\lim_{x\to+\infty} x(\sqrt{x^2+a^2} - x)$.

5. Suppose that f and g have continuous derivatives of the first n orders in a closed interval $a \leqq x \leqq b$. Furthermore, assume that $f(a) = f'(a) = \cdots = f^{(n-1)}(a) = 0$, $g(a) = g'(a) = \cdots = g^{(n-1)}(a) = 0$, and that $g^{(n)}(a) \neq 0$. Without using l'Hospital's rule, show that

$$\lim_{x \to a+} \frac{f(x)}{g(x)} = \frac{f^{(n)}(a)}{g^{(n)}(a)}.$$

Use Taylor's formula with remainder.

6. What can you conclude in Theorem VI, if $\lim \dfrac{f'(x)}{g'(x)}$ fails to exist either as a definite numerical limit or as $+\infty$ or $-\infty$? Give your answer after a careful examination of the limits

$$\lim_{x \to 0} \frac{x^2 \sin \dfrac{1}{x}}{\sin x} \quad \text{and} \quad \lim_{x \to \infty} \frac{x - \sin x}{x}.$$

7. Evaluate each of the following limits:

(a) $\lim\limits_{x \to \infty} x \sin \dfrac{1}{x}$. **(c)** $\lim\limits_{x \to +\infty} \dfrac{1}{x^2} \displaystyle\int_1^x \sin^2 t \, dt$.

(b) $\lim\limits_{x \to +\infty} \dfrac{1}{x} \displaystyle\int_1^x \frac{\log t}{1+t} \, dt$. **(d)** $\lim\limits_{x \to 0} \dfrac{e - (1+x)^{1/x}}{x}$.

8. Give a proof of Case 1 of Theorem VI, assuming that c is a real number, and utilizing the following suggestions: Let b be a point of the interval, I, and consider the functions f, g on the *closed* interval with end-points b and c, after defining $f(c) = g(c) = 0$. Now apply Theorem I, §4.1, and make the deduction of (4.5–2) from (4.5–1). Explain the argument carefully. Why do we define $f(c) = g(c) = 0$?

MISCELLANEOUS EXERCISES

1. Suppose that f is defined and differentiable in an interval containing $x = a$ (the point $x = a$ may be at one end of the interval, in which case derivatives at $x = a$ are to be considered as one-sided limits). Suppose also that $f''(a)$ exists, but assume nothing else about second derivatives. Show that

$$f''(a) = \lim_{x \to a} \frac{f(x) - f(a) - (x-a)f'(a)}{\dfrac{(x-a)^2}{2!}}.$$

2. Suppose that f satisfies the conditions of Exercise 1, and that $x = a$ is an interior point of the interval in question. Show that, if f has a relative minimum at $x = a$, then $f''(a) \geqq 0$, while $f''(a) \leqq 0$ if f has a relative maximum at $x = a$. These are *necessary* conditions for a relative extreme at $x = a$. Now assume that $f'(a) = 0$ and $f''(a) > 0$, and prove that f must have a relative minimum at $x = a$. These are *sufficient* conditions for a relative minimum. State a set of sufficient conditions for a relative maximum at $x = a$, and prove the sufficiency of the conditions.

3. Generalize the result of Exercise 1, obtaining

$$f^{(n)}(a) = \lim_{x \to a} \frac{f(x) - f(a) - (x-a)f'(a) - \cdots - \dfrac{(x-a)^{n-1}}{(n-1)!} f^{(n-1)}(a)}{\dfrac{(x-a)^n}{n!}},$$

with the assumptions that f is defined and has derivatives of orders $1, \ldots, n-1$ in an interval including $x = a$, while $f^{(n)}(a)$ exists. Suggestion: Apply Theorem VI $(n-1)$ times. Use this result to show that there is a function $E_n(x)$ defined in the interval, except at $x = a$, such that

$$f(x) = f(a) + (x-a)f'(a) + \cdots + \frac{(x-a)^{n-1}}{(n-1)!} f^{(n-1)}(a) + \frac{(x-a)^n}{n!}\{f^{(n)}(a) + E_n(x)\}$$

and $\lim_{x\to a} E_n(x) = 0$. The function $E_n(x)$ depends upon the function $f(x)$, of course. Find $E_2(x)$ explicitly if $f(x)$ is defined as $x^4 \sin(1/x)$ when $x \neq 0$, and $f(0) = 0$; take $a = 0$.

4. With the conditions of Exercise 3 on f suppose that $f'(a) = \cdots = f^{(n-1)}(a) = 0$ and that f has a relative minimum at $x = a$. Also suppose that n is even. Show that $f^{(n)}(a) \geqq 0$. What if there is a relative maximum at $x = a$? What about sufficient conditions for a relative maximum or minimum when $f'(a) = \cdots = f^{(n-1)}(a) = 0$? What can you say if n is odd, $f'(a) = \cdots = f^{(n-1)}(a) = 0$ and $f^{(n)}(a) \neq 0$? (Assume that $x = a$ is an interior point of the interval.)

5. Use the result of Exercise 3 to prove the following theorem:

Let f and g be defined and have derivatives of orders $1, \ldots, n-1$ in an interval including $x = a$, while $f^{(n)}(a)$ and $g^{(n)}(a)$ exist and $g^{(n)}(a) \neq 0$. Suppose that both f and g, as well as their first $n-1$ derivatives, have the value 0 at $x = a$. Then

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{f^{(n)}(a)}{g^{(n)}(a)}.$$

This theorem holds for $n \geqq 1$. It may give a result when Theorem VI is inapplicable, e.g., in the case $f(x) = x^2 \sin(1/x)$, $f(0) = 0$, $g(x) = x$.

6. Calculate the following limits:

(a) $\lim_{x\to 0+} x(\log x)^n$, n a positive integer. **(b)** $\lim_{x\to 1} x^{1/(1-x)}$.

7. Discuss $\lim_{x\to+\infty}(a^x + b^x)^{1/x}$, assuming $0 < a \leqq b$. Generalize to the case of $\lim_{x\to+\infty}(a_1^x + a_2^x + \cdots + a_n^x)^{1/x}$, where $0 < a_1 \leqq a_2 \leqq \cdots \leqq a_n$.

8. Find $\lim_{x\to 0} \dfrac{x\int_0^x e^{-t^2}\,dt}{1 - e^{-x^2}}$.

9. If $f(x) = e^{-1/x^2}$, show that $f^{(n)}(x) = e^{-1/x^2} P_n(1/x)$, where $P_n(t)$ is a polynomial of degree $3n$ in t, with leading term $2^n t^{3n}$. Show also that $P_{n+1}(t) = 2t^3 P_n(t) - t^2 P_n'(t)$.

10. Show that to each positive integer n corresponds a number θ_n, $0 < \theta_n < 1$, such that $\log\left(1 + \dfrac{1}{n}\right) = \dfrac{1}{n} - \dfrac{\theta_n}{2n^2}$.

11. Apply (4.3–7) to $f(x) = 1/(1-x)$ with $a = 0$, $h = x$, and find the form of the remainder term. Then compare with the algebraic indentity

$$\frac{1}{1-x} = 1 + x + \cdots + x^n + \frac{x^{n+1}}{1-x},$$

and so find that $\theta = \dfrac{1 - (1-x)^{1/(n+2)}}{x}$ (where $x < 1$). Now find $\lim_{x\to 0} \theta$.

12. Suppose that f is differentiable in an interval of which $x = a$ is an interior point, and assume that $f''(a)$ exists. Show that

$$f''(a) = \lim_{h\to 0} \frac{f(a+2h) - 2f(a+h) + f(a)}{h^2}.$$

This may be proved using the result of Exercise 1. If one assumes that the second derivative exists on the whole interval and is continuous at $x = a$, the proof may be given by using (4.3–7) in a suitable way.

5 / FUNCTIONS OF SEVERAL VARIABLES

5 / FUNCTIONS AND THEIR REGIONS OF DEFINITION

Thus far in this book we have dealt with functions of a single independent variable. But we do not go far in either pure or applied mathematics until we have occasion to consider functions of two or more variables. We assume that the student has some familiarity with the concept of a function of several independent variables.

One of the first things that claims our attention when we begin to study functions of several variables is the nature of the region of definition of such a function. The functions of one variable which we study in calculus are usually defined on intervals of the real axis. There are only a few different types of intervals. If the interval is finite, it may contain both its end-points, or just one, or neither. If the interval is infinite, but is not the entire axis, it has just one end-point, and this may or may not be counted as belonging to the interval. There is much more variety in the case of functions of several variables. We shall give some illustrative examples, taking the number of independent variables to be two.

Example 1. $f(x, y) = \log(1 - x^2 - y^2)$.

The function is defined only when $x^2 + y^2 < 1$, since otherwise the logarithm is undefined. The region of definition is the interior of the unit circle with center at the origin. In Fig. 28 the circle is dashed to indicate that the boundary of the circular area does not belong to the region of definition.

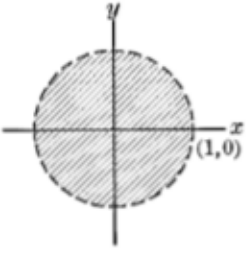


Fig. 28.

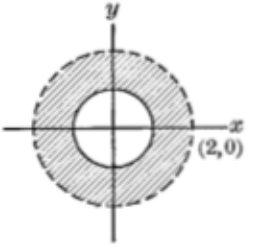


Fig. 29.

Example 2. $F(x, y) = \sqrt{x^2 + y^2 - 1} + \log(4 - x^2 - y^2)$.

Here we must have $x^2 + y^2 \geqq 1$ in order for the square root to be real, while we must have $x^2 + y^2 < 4$ for the logarithm to be defined. The region of definition

of $F(x, y)$ is the annular region between the circles $x^2+y^2=1$ and $x^2+y^2=4$. The inner circumference is part of the region of definition, while the outer circumference is not (see Fig. 29).

Example 3. $g(x, y)=\dfrac{x}{y^2-4x}$.

The function is defined except when the denominator is zero, that is, everywhere except at the points of the parabola $y^2=4x$ (see Fig. 30).

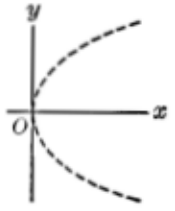


Fig. 30.

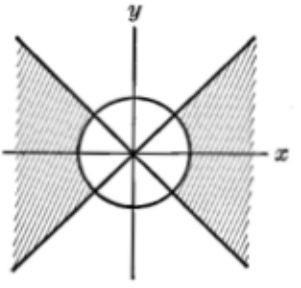


Fig. 31.

Example 4. $G(x, y)=\sqrt{x^2-y^2}+\sqrt{x^2+y^2-1}$.

The region of definition here is defined by the inequalities $x^2 \geqq y^2$, $x^2+y^2 \geqq 1$. The lines $x-y=0$, $x+y=0$ divide the plane into four quadrants. The inequality $x^2 \geqq y^2$ states that the point (x, y) lies in (or on the edge of) one of those two of the four quadrants which contain the x-axis. The other inequality states that (x, y) lies outside or on the circle $x^2+y^2=1$. Hence the region of definition of $G(x, y)$ is that part of the xy-plane which is shaded in Fig. 31.

Similar examples might be given for functions of three independent variables. The region of definition might be the interior of a cube, the interior and boundary of an ellipsoid, the space between two concentric spheres, or the interior of a surface formed like the inner tube of a bicycle tire.

Because of the great variety of possible regions of definition of a function of two or more variables, it is desirable to devote some attention to matters of terminology about configurations of points in the plane. Not only will this make it easier for us to state things clearly, but it will eventually become absolutely indispensable in developing parts of our subject. We shall sometimes use the word "domain" for "region of definition," and by the range of a function we shall mean the set of values which the function takes on.

5.1 / POINT SETS

In §2.7 we explained the meaning of the phrase "a set of real numbers." Since we identify real numbers with points on the axis of reals, we may equally well

speak of sets of points on a line. We are now going to talk about sets of points, or point sets, in the plane or in space of three dimensions. For the sake of simplicity and definiteness we shall speak principally of point sets in the plane.

In dealing with intervals we found it necessary to introduce the notions of closed and open intervals. Likewise, in our discussion of regions of definition of a function of several variables, we find it necessary to introduce and use the concepts *open set* and *closed set.* Furthermore, it is well to give a precise definition of the *boundary* of a set. It is much more difficult to define these concepts carefully than one might at first suppose, because of the great variety of configurations of points which are available for consideration. We proceed as follows: First let us define the phrase *a circular neighborhood* of the point (x_0, y_0) to mean the set of all points (x, y) lying inside some circle with center at (x_0, y_0). That is, if $\delta > 0$, the set of all (x, y) such that

$$(x - x_0)^2 + (y - y_0)^2 < \delta^2$$

constitutes a circular neighborhood of (x_0, y_0) (see Fig. 32). A neighborhood may have any positive number as its radius. The further concepts which we are now going to introduce are built upon the concept of a neighborhood. It is convenient to abandon the coordinate notation for the time being, and denote points by such symbols as P, Q, P_1, P_2, ...

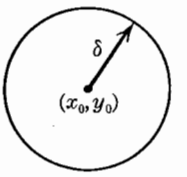
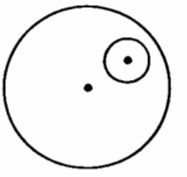


Fig. 32.

Definition. *A set S is called open if each point P of S has some circular neighborhood which belongs entirely to the set S.*

Example 1. A circular neighborhood, for instance the set of all points inside the circle $x^2 + y^2 = 1$, is an open set. For, if P is inside the unit circle, say at a distance r from the center (where $r < 1$), the circular neighborhood of P of radius δ also lies inside the unit circle provided δ is chosen so that $0 < \delta < 1 - r$ (see Fig. 33).


Fig. 33.

Example 2. The set of all points in the plane *not* on the parabola $y^2 = 4x$ is an open set. The student should verify this for himself.

Example 3. Let S consist of all points on or inside the circle $x^2 + y^2 = 1$. This set is not open, for if P is on the circle there is no neighborhood of P which belongs entirely to S.

Definition. *If S is any point set, the set of all points of the plane which are not in S is called the complement of S. On occasion it is convenient to use the notation $C(S)$ for the complement of S.*

To illustrate this notion, consider the sets defined in the foregoing examples. For the S of Example 1, $C(S)$ is the set of all points outside or on the circle $x^2 + y^2 = 1$, i.e., all points for which $x^2 + y^2 \geqq 1$. In Example 2, $C(S)$ consists of

the points on the parabola $y^2 = 4x$. In Example 3, $C(S)$ is the exterior of the circle, i.e., all points such that $x^2 + y^2 > 1$.

Definition. *A set S is called closed if its complement is open.*

The set of Example 3 is closed (the student should verify this to his own satisfaction); the sets of Examples 1 and 2 are not closed. A set consisting of any finite number of points is closed. A set may be neither open nor closed, as we see in the next example.

Example 4. Let S be the set of all points for which $1 \leqq x^2 + y^2 < 4$. This set is the region of definition of the function $F(x, y)$ of Example 2, §5. It is not open, because a point of the circle $x^2 + y^2 = 1$ has no circular neighborhood which belongs entirely to S (see Fig. 29). The complement of S has two parts: the set of all points for which $x^2 + y^2 < 1$, and the set of all points for which $x^2 + y^2 \geqq 4$. It is easily seen that $C(S)$ is not open, for a point on the circle $x^2 + y^2 = 4$ has no circular neighborhood which belongs entirely to $C(S)$. Therefore S is not closed. The set $C(S)$ is shown in Fig. 34.

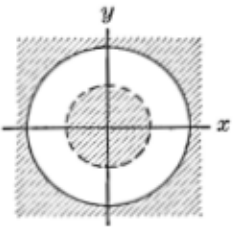


Fig. 34.

If S is a set, the complement of $C(S)$ is S itself. Hence, by definition, $C(S)$ is closed if S is open. Thus, if one of the two sets S, $C(S)$ is open, the other is closed.

Definition. *If S is a point set, a point P is called a boundary point of S if every neighborhood of P contains at least one point of S and one point of the complement $C(S)$. The collection of all boundary points of S is called the boundary of S. We denote it by $B(S)$.*

The sets introduced in Examples 1–4 have the following boundaries:

Example 1. $B(S)$ is the circle $x^2 + y^2 = 1$.

Example 2. $B(S)$ is the parabola $y^2 = 4x$.

Example 3. $B(S)$ is the circle $x^2 + y^2 = 1$.

Example 4. $B(S)$ consists of the two circles $x^2 + y^2 = 1$ and $x^2 + y^2 = 4$.

It is clear from the definition of boundary that a set S and its complement $C(S)$ have the same boundary. If a set S is open, no boundary point of S is actually in S. If S is closed, $B(S)$ is part of S. These statements may be verified by referring to the definitions. In Example 4, $B(S)$ is partly in S and partly in $C(S)$.

Definition. *A point P of a set S is called an interior point of S if there is some circular neighborhood of P which belongs entirely to S. The interior of a set S is the set consisting of all interior points of S.*

An open set consists entirely of interior points. A set may have no interior points.

We are going to be most interested in sets of points which are of one of the following kinds:

(*a*) Open.

(*b*) A closed set formed by an open set together with its boundary.

(*c*) The boundary of an open set.

Furthermore, the sets which we consider will usually have boundaries which are composed of one or more curves or segments of curves (straight lines included), or possibly of isolated points.

Definition. *By a region we shall mean a set of points which is either a nonempty open set or such a set together with some or all of the points forming its boundary.*

All the sets occurring as domains of definition of the functions in Examples 1–4 of §5 are regions in the sense just defined. Of these regions, those of examples 1 and 3 are open; that of Example 4 is closed (i.e., the entire boundary is part of the region); that of Example 2 is neither open nor closed (it contains only one of the two circles forming the boundary).

If we wish to deal with point sets in three-dimensional space we define *spherical* neighborhoods instead of *circular* neighborhoods. Once this is done we can define open set, closed set, and boundary just as in the case of point sets in the plane. These concepts also apply to point sets on a line. In that case we define a neighborhood of x_0 as an open interval $x_0 - \delta < x < x_0 + \delta$, where δ is any positive number.

Finally, we remark that we could use square neighborhoods instead of circular neighborhoods without affecting the whole development of the concepts of open sets, closed sets, boundary of a set, and so forth. A square neighborhood of (x_0, y_0) is the set of all points inside a square with center (x_0, y_0) (see Fig. 35). If the square has sides of length 2δ the neighborhood is defined by the inequalities

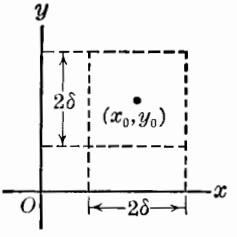


Fig. 35.

$$|x - x_0| < \delta, |y - y_0| < \delta.$$

In the future we shall often speak of neighborhoods without the adjectives *circular* or *square.* It will usually be immaterial which is meant or which the student chooses to think of on any particular occasion.

Recall that the *union* of any collection of sets consists of all points belonging to at least one of them, and the *intersection* consists of all points belonging to each of them. Two sets are said to be *disjoint* in case they have no points in common.

Example 5. An open rectangle is a rectangle without its boundary. It consists

entirely of interior points. For a rectangle with each side parallel to a coordinate axis, to say that R is open means that there are number pairs a, b with $a < b$ and c, d with $c < d$ such that R is the set of all points (x, y) for which $a < x < b$ and $c < y < d$. The same thing is frequently said more briefly as follows:

$$\{R = (x, y) : a < x < b \quad \text{and} \quad c < y < d\}.$$

A property of open rectangles which we shall find useful later in proving the inverse function theorem (in Chapter 12) is brought out in the following: *Assertion: No open rectangle R is the union of two nonempty disjoint open subsets.*

When we undertake to justify this very plausible assertion, we see that the key to a proof is a good understanding of the property of openness. Let us reason by contradiction and begin by supposing that R is the union of two nonempty disjoint open subsets, A and B. Now consider the line segment $[PQ]$ connecting a point P in A to a point Q in B. Since R is an open rectangle, the line segment $[PQ]$ obviously lies entirely in R. Since the subset A to which P belongs is open, there is some positive number r such that the disc of radius r centered at P is contained in A. Therefore there is some interval extending along $[PQ]$, from P toward Q, which lies in A. Since B also is open, the same argument shows that all points of the line segment $[PQ]$ which are sufficiently close to Q must, like Q itself, lie in B.

Now let D be the set of all numbers which are distances from P of points on $[PQ]$ which belong to A. Then D is a set of nonnegative numbers which is bounded above, since the distance from P to Q is obviously one upper bound. By the least upper bound property of the real numbers (§2.7), D must have a least upper bound; call it d. By the preceding paragraph, we know that d is a positive number less than the distance from P to Q. From now on, we shall concentrate our attention on that point C of $[PQ]$ which is at the distance d from P. Obviously C must belong to R, yet we shall soon see that it cannot belong to either A or B. This contradiction will prove the *Assertion.*

Suppose first that C belongs to A. Then C must be farther than any other point of A, that is also in $[PQ]$, from the point P. Since A is open there must be some disc centered at C and contained in A. But this implies that there are points of A on the segment $[PA]$ which are farther than C is from P, which is a contradiction. Now try assuming that C belongs to B. Clearly no point between C and Q can belong to A. Since B is open, there is some disc of positive radius μ centered at C and contained in B. But this would imply that $d - \mu$ is an upper bound for D, contradicting the fact that d is the least upper bound. This completes the proof.

EXERCISES

1. The set S consists of all points (x, y) such that $x^2 + y^2 < 1$ and $x < 0$ if $y = 0$. Describe S in geometrical language, with the aid of a figure. Is S open, closed, or neither? What is the boundary of S?

2. The set S consists of all points (x, y) such that either $x^2 + y^2 = 1$ or $y = 0$ and $0 \leqq x \leqq 1$. Does this set have any interior points? Is it closed?

3. The set S consists of all points (x, y) such that $y \geqq x^2$ and $y \leqq 1$. Draw a figure, and describe S in geometrical language. Is S open, closed, or neither? What is the boundary of S?

4. The set S consists of all points (x, y) such that $0 < xy \leqq 1$ and $x > 0$. Is this set open, closed, or neither? What is $B(S)$?

5. The set S consists of all points (x, y) such that $y = \sin(1/x)$ and $x > 0$. Does this set have any interior points? It is closed? What is $B(S)$?

6. The set S consists of all points (x, y) for which $x^2 + y^2 < 4$ and $y > 0$ except for the points with $0 < y \leqq 1$ and $x = 1/n$, $n = 1, 2, \ldots$, i.e., except for the points of a certain infinite sequence of line segments each one unit long. Is this set S open? What is the boundary of S? Are there any points of $B(S)$ which are also in S? Is S a region? Is its complementary set $C(S)$ a region?

7. Let $f(x, y) = \left(y - \sin\dfrac{1}{x}\right)^{-1}$, the function being defined whenever this expression has a meaning, but not otherwise. Is the set of points where f is defined a region? What is the boundary of the set?

8. Let $f(x, y) = \log \sin x + y^{-1/2}$, the function being defined whenever this expression has a meaning (real numbers only are to be considered). Describe, with the aid of a diagram, the set of points (x, y) where f is defined. Is the set open, closed, or neither? Of what does its boundary consist?

5.2 / LIMITS

We wish to define what is meant by the statement "$f(x, y)$ approaches A as a limit when the point (x, y) approaches (x_0, y_0)." The statement is written in the form

$$\lim_{(x, y) \to (x_0, y_0)} f(x, y) = A. \tag{5.2–1}$$

In giving the definition we shall assume that the function f is defined in a region R and that (x_0, y_0) is either an interior point of R or on the boundary of R. The point (x_0, y_0) *may*, but *need not*, belong to R. If it is in R, the meaning of (5.2–1) has nothing whatever to do with the value $f(x_0, y_0)$ at the point (x_0, y_0). The statement (5.2–1) is now defined to mean that if ϵ is any positive number, there is some neighborhood of (x_0, y_0) such that if (x, y) is in the neighborhood, in R, and different from (x_0, y_0), then $|f(x, y) - A| < \epsilon$. This definition may be compared with that for functions of one variable in §§1.1, 1.61. The limit notion can be expressed verbally as follows: The meaning of (5.2–1) is that $f(x, y)$ is in a prescribed neighborhood of A on the real axis provided (x, y) is any point other than (x_0, y_0) in a suitably chosen (sufficiently small) neighborhood of (x_0, y_0) in the plane.

With the adoption of the term *neighborhood* we obtain a unification of the limit concept for functions of one, two, or three independent variables. The extension to more than three variables causes no trouble and involves no new

principle. We continue to use geometric language; the meaning of a "spherical neighborhood" in a space of four variables is made clear by the inequality

$$(x-x_0)^2+(y-y_0)^2+(z-z_0)^2+(w-w_0)^2<\delta^2.$$

The fundamental theorems about limits carry over to functions of several variables. We cite particularly Theorems X (§1.61) and XIV (§1.64).

When we say that $f(x, y)\to A$ as $(x, y)\to(x_0, y_0)$, it must be stressed that *the limit must exist and be the same, no matter how* (x, y) *approaches* (x_0, y_0). The student will recall that, for a function of one variable, $f(x)\to A$ as $x\to x_0$ means that $f(x)\to A$ as $x\to x_0+$ and also as $x\to x_0-$. But in the case of two variables, (x, y) can approach (x_0, y_0) in a infinite number of ways. If it is possible to find two different modes of approach to (x_0, y_0) such that $f(x, y)$ approaches different limits in the two cases, or no limit at all in at least one of the cases, then $\lim_{(x, y)\to(x_0, y_0)} f(x, y)$ does not exist.

Example 1. Let $f(x, y)=\dfrac{x^2-y^2}{x^2+y^2}$. This function is defined except at the origin. Let us show that the limit of $f(x, y)$ as $(x, y)\to(0, 0)$ does not exist. If $(x, y)\to(0, 0)$ along the x-axis, we have $f(x, 0)=1 (x\neq 0)$. If $(x, y)\to(0, 0)$ along the y-axis, we have $f(0, y)=-1$ $(y\neq 0)$. Thus the limits for the two modes of approach are 1 and -1 respectively. This shows that $f(x, y)$ has no limit as $(x, y)\to(0, 0)$.

To prove directly that a certain function approaches a certain limit as $(x, y)\to(x_0, y_0)$, we have to work with inequalities. The following example will illustrate the technique. It is not our intent, at this stage of a student's training, to have him cultivate extensively the technique of working exercises of the type represented by the example. The purpose is merely to make clearer the essential content of the definition of a limit.

Example 2. Show that

$$\lim_{(x, y)\to(0, 0)} \frac{2x^3-y^3}{x^2+y^2}=0. \tag{5.2–2}$$

In terms of inequalities, this means that if ϵ is any positive number, we have to show that another positive number δ (depending on ϵ) can be found, such that

$$\left|\frac{2x^3-y^3}{x^2+y^2}\right|<\epsilon \text{ if } 0<(x^2+y^2)^{1/2}<\delta; \tag{5.2–3}$$

in other words, denoting the function under consideration by $f(x, y)$, we have to show that, if $\epsilon>0$, there is *some* circular neighborhood of the origin (whose radius we denote by δ) such that $|f(x, y)-0|<\epsilon$ if (x, y) is in the specified neighborhood of the origin but not actually at the origin. We proceed to find such a number δ, considering ϵ as given.

Now $$|2x^3-y^3|\leqq 2|x|^3+|y|^3=2|x|x^2+|y|y^2.$$

Also, $$|x|\leqq(x^2+y^2)^{1/2} \quad \text{and} \quad |y|\leqq(x^2+y^2)^{1/2}.$$

Therefore $$|2x^3 - y^3| \leqq (x^2 + y^2)^{1/2}(2x^2 + y^2) \leqq 2(x^2 + y^2)^{3/2},$$

and $$\left|\frac{2x^3 - y^3}{x^2 + y^2}\right| \leqq 2(x^2 + y^2)^{1/2} \text{ if } 0 < x^2 + y^2.$$

It is now clear that (5.2–3) will hold if δ is chosen in any manner such that $0 < \delta \leqq \epsilon/2$. Thus (5.2–2) is proved.

In work of this kind the student will find the simple inequalities

$$2|ab| \leqq a^2 + b^2, \tag{5.2–4}$$

$$|a| + |b| \leqq \sqrt{2}(a^2 + b^2)^{1/2} \tag{5.2–5}$$

quite useful. See Exercise 6 for remarks about these inequalities.

EXERCISES

1. Find the limit of $\frac{x^2 - y^2}{x^2 + y^2}$ as (x, y) approaches $(0, 0)$ along the line $y = x$; along the line $y = mx$.

2. Does $\lim_{(x, y) \to (0, 0)} \frac{xy}{x^2 + y^2}$ exist? Give reasons.

3. Examine the behavior of $\frac{x^4 y^4}{(x^2 + y^4)^3}$ as (x, y) approaches $(0, 0)$ along various straight lines. Then consider what happens for approach to the origin along the curve $y^2 = x$. Is there a limit as $(x, y) \to (0, 0)$ without restriction?

4. Show in each case that the given function does not approach a limit as $(x, y) \to (0, 0)$, by examining the behavior of the function for at least two modes of approach.

(a) $\frac{x - y}{x^2 + y^2}$. (c) $\frac{x^2 + y}{(x^2 + y^2)^{1/2}}$.

(b) $\frac{xy^2}{x^2 + y^4}$. (d) $\frac{x^4 + 3x^2y^2 + 2xy^3}{(x^2 + y^2)^2}$.

5. Define a function by setting $f(x, y) = 0$ if $y \leqq 0$ or if $y \geqq x^2$, and $f(x, y) = 1$ if $0 < y < x^2$. Show that $f(x, y) \to 0$ as $(x, y) \to (0, 0)$ along any straight line through the origin. Find a curve through the origin along which $f(x, y) = 1$ (except at the origin).

6. If A and B are nonnegative numbers, the inequality $A \leqq B$ is equivalent to $A^2 \leqq B^2$. Use this fact to prove the correctness of (5.2–4); then show that (5.2–5) is correct.

7. Let $f(x, y) = xy\left(\frac{x^2 - y^2}{x^2 + y^2}\right)$. Show that $|f(x, y)| \leqq \frac{1}{2}(x^2 + y^2)$, and hence prove that $f(x, y)$ approaches a limit as $(x, y) \to (0, 0)$.

8. Let $f(x, y) = \frac{x^2y^2}{x^2 + y^2}$. If $\epsilon > 0$, find δ so that $0 < (x^2 + y^2)^{1/2} < \delta$ implies $|f(x, y)| < \epsilon$.

9. If $\epsilon > 0$, show that $|2x^2 - 6xy + 5y^2| < \epsilon$ when $(x^2 + y^2)^{1/2} < (\epsilon/13)^{1/2}$.

10. Show that $\frac{x^4 + y^4}{x^2 + y^2} < \epsilon$ if $0 < x^2 + y^2 < \delta^2$, for a suitably chosen δ depending on ϵ.

11. Show that $|x^3 - y^3| \leqq (x^2 + y^2)^{3/2}$.

12. Does $\dfrac{2x^5 + 2y^3(2x^2 - y^2)}{(x^2 + y^2)^2}$ approach a limit as $(x, y) \to (0, 0)$?

5.3 / CONTINUITY

The notion of continuity depends on the notion of limit, as was pointed out at the beginning of Chapter 3.

Definition. *Let $f(x, y)$ be defined in a region R, and let (x_0, y_0) be a point of R. We say that f is continuous at this point if*

$$\lim_{(x, y) \to (0, 0)} f(x, y) = f(x_0, y_0).$$

If (x_0, y_0) is an interior point of R, the mode of approach of (x, y) to (x_0, y_0) is unrestricted in this definition. But, if (x_0, y_0) is a boundary point of R, there is the restriction that (x, y) must remain in R. We say that f is continuous in R if it is continuous at each point of R.

If f and g are defined in the same region R, and each is continuous at a point (x_0, y_0) of R, then the sum and product functions

$$f(x, y) + g(x, y), \quad f(x, y)g(x, y)$$

are also continuous at (x_0, y_0). The quotient function

$$\frac{f(x, y)}{g(x, y)}$$

is continuous at (x_0, y_0) provided $g(x_0, y_0) \neq 0$. These assertions are direct generalizations of Theorem I, §3. They may be extended to functions of more than two variables.

The theorem of Chapter 3 all have important analogues for functions of several independent variables. We do not wish at this point to prove all these analogous theorems, but we shall discuss the statements of certain theorems which will be used in the chapters immediately following.

In dealing with the analogues of Theorems II (§3.1) and III (§3.2) of Chapter 3 it is necessary to introduce the concept of a *bounded point set.*

Definition. *A point set S in the plane is called bounded if all its points are inside some sufficiently large circle. For a point set in space the definition is similar; we write "sphere" instead of "circle."*

Examples. The interior and boundary of a triangle form a bounded point set. The set of all points between the lines $y = 0$, $y = 1$ is not a bounded point set.

We now state two important theorems.

THEOREM I. *If a function is continuous at each point of a closed and bounded*

region R, the function is bounded on the region (i.e., the values of the function form a bounded set of real numbers).

THEOREM II. *Let f be continuous on a closed and bounded region R. Let m and M be the greatest lower bound and least upper bound of the values of f on R. Then f takes on each of the values m, M at least once in R.*

Proofs of these two theorems will be considered later, in §§17.2, 17.3.

Theorem V of Chapter 3 (§3.3) has the following analogue. We state it for the case of two independent variables.

THEOREM III. *Let f be defined in an open set containing the point (x_0, y_0). Suppose that f is continuous at the point and that $f(x_0, y_0) \neq 0$. Then there is a neighborhood of (x_0, y_0) throughout which $f(x, y)$ has the same sign as at (x_0, y_0).*

The proof is left to the student.

There is a feature of the continuity of a function $f(x, y)$ which deserves notice. If we fix y, say $y = y_0$, $f(x, y_0)$ is a function of x alone. Likewise $f(x_0, y)$ is a function of y alone. It can happen that each of these functions of a single variable is continuous, and yet that $f(x, y)$ is not continuous. An illustration of this possibility is given in Exercise 3.

There is another theorem which will be needed later. It deals with composite functions, and may be roughly stated in the form: *A continuous function of continuous functions is continuous.* The number of variables is immaterial.

Examples. $F(z) = \sin z$ is a continuous function of z, and $f(x, y) = (1 + xy)^2$ is a continuous function of x, y. Therefore

$$F(f(x, y)) = \sin(1 + xy)^2$$

is a continuous function of x, y. Or, again,

$$F(x, y, z) = x^2 + y^2 + z^2 \quad \text{and} \quad f(x, y) = x(1 + x^2 + y^2)^{-3/2}$$

are continuous functions of x, y, z and x, y, respectively. Therefore

$$F(x, y, f(x, y)) = x^2 + y^2 + \frac{x^2}{(1 + x^2 + y^2)^3}$$

is a continuous function of x, y.

We formalize one such theorem about composite functions.

THEOREM IV. *Let $F(x, y, z)$ be continuous in an open set B of space. Let $f(x, y)$ be continuous in an open set R of the xy-plane. Writing $z = f(x, y)$, suppose that the point (x, y, z) is in B when (x, y) is in R. Then the composite function $F(x, y, f(x, y))$ is continuous in R.*

$F(f(x,y))$ is cts where $F(z)$ is cts on D and $f(x,y) \in D$ if $(x,y) \in R$

e.g. $F(z) = \frac{1}{2-z}$, $f(x,y) = \sin(x^2+y^2)$

F is cts on $(-\infty, 2)$, $(2, \infty)$ and

We shall not give a proof here. This theorem is a special case of the Theorem II which is proved in §11.7.

EXERCISES

1. If $f(x, y) = \dfrac{\sin(x^2 + y^2)}{x^2 + y^2}$ when $x^2 + y^2 \neq 0$, how must $f(0, 0)$ be defined so as to make f continuous at $(0, 0)$?

2. Let us define $f(x, y) = \dfrac{\sin(xy)}{x}$ if $x \neq 0$, and $f(x, y) = y$ if $x = 0$. Does f have any points of discontinuity?

3. If we define $f(x, y) = xy/(x^2 + y^2)$ when $x^2 + y^2 \neq 0$, and $f(0, 0) = 0$, show that f is discontinuous at $(0, 0)$, but also that $f(x, 0)$ and $f(0, y)$ are continuous functions of x and y, respectively, with no exceptions.

4. Let $f(x, y) = (x^2 + y^2)\tan^{-1}(y/x)$ if $x \neq 0$, and define $f(0, 0) = 0$, but do not consider f defined if $x = 0$ and $y \neq 0$. **(a)** Is f continuous at $(0, 0)$ according to the definition in the text? **(b)** Is it possible to define f at the one additional point $(0, 1)$ so as to make it continuous there?

5. Let $f(x, y) = xy \log(xy)$ if $xy > 0$, and define $f(x, y) = 0$ if $xy = 0$. Where, if at all, is f discontinuous?

6. Let $f(x, y) = (5x + y)/(x - y)$. Show directly by the definition that f is continuous at $(4, 1)$ by proving that, if $\epsilon > 0$, $|f(x, y) - f(4, 1)| < \epsilon$ provided (x, y) is in a sufficiently small neighborhood of $(4, 1)$. Start by showing that

$$|f(x, y) - f(4, 1)| \leqq 2|x - 4| + 8|y - 1|$$

at the points of the square $3 < x < 5, 0 < y < 2$.

7. If $f(x, y) = e^{-1/|x-y|}$ when $x \neq y$, how must f be defined when $x = y$ so as to make it continuous at all points of the plane?

8. Let us define $f(x, y) = 0$ if $y \leqq 0$ or if $x^2 \leqq y$, and $f(x, y) = 4y(x^2 - y)/x^4$ if $0 < y < x^2$.

(a) Is f continuous at $(0, 0)$? **(b)** Discuss possible discontinuity at other points on the line $y = 0$ or the curve $y = x^2$.

9. Let $f(x, y) = x(1 - x^2 - y^2)^{-1/2}$, the region R of definition being defined by $x^2 + y^2 < 1$. It is possible to add the single point $(0, 1)$ to R and define f so as to make it continuous at that point? Consider values of f on the circle $x^2 + y^2 - y = 0$, and also at other points in R near $(0, 1)$.

10. If $f(x, y, z) = xyz/(x^2 + y^2 + z^2)$ when $x^2 + y^2 + z^2 \neq 0$, is it possible to define $f(0, 0, 0)$ so as to make f continuous at the origin?

5.4 / MODES OF REPRESENTING A FUNCTION

The standard method of representing a function of one variable is by graphing in rectangular co-ordinates. We write $y = f(x)$ and plot the points (x, y). If f is continuous on an interval, the graph will be a curve in the plane.

The corresponding procedure for the case of two independent variables is familiar. We write $z = f(x, y)$, and plot the points (x, y, z). If f is continuous in a region R of the xy-plane we obtain a surface in space (see Fig. 36).

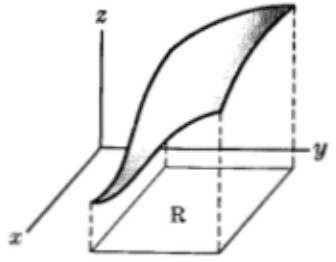


Fig. 36.

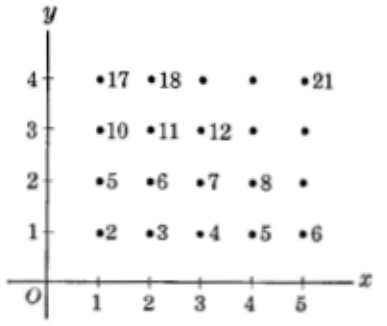


Fig. 37.

When we go to three independent variables there is no satisfactory analogue of the foregoing methods of graphical representation, for we cannot draw upon any familiar geometric intuition to visualize $w = f(x, y, z)$ as defining a configuration in space of four dimensions. There is, however, another mode of representation which is helpful. It is available as well in the case of two independent variables, and since the figures are easier to draw, we begin with that case.

When f is defined in a region R, we can think of each point of R as being given a label, namely, the value $f(x, y)$ at that point. A good example is obtained by thinking of the xy-plane as a map on which elevations above sea level are marked at various locations, $f(x, y)$ being the elevation in feet at (x, y) (see Fig. 37). To carry this example further, imagine that the map is a topographic map with *contour lines* drawn in, showing lines of equal elevation. Each line is labeled; there is a line for 500 feet above sea level, others for 400, 600, and so on. In the aggregate, the configuration of these lines, together with their numbering, gives us a good visual representation of the elevation as a function of x and y.

This "topographic map" idea can be carried over to any function $f(x, y)$. Instead of contour lines we consider curves along which $f(x, y)$ is constant in value. Such a curve is called a *level curve* of the function. If the constant value is C, the equation of the level curve is $f(x, y) = C$. See Fig 38 and Fig. 39. Isobars

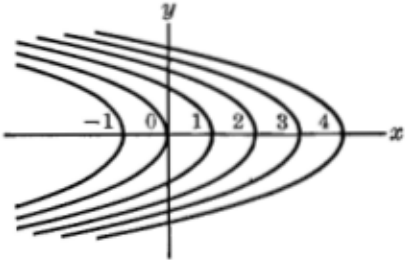


Fig. 38. Level curves of $f(x, y) = x + y^2$.

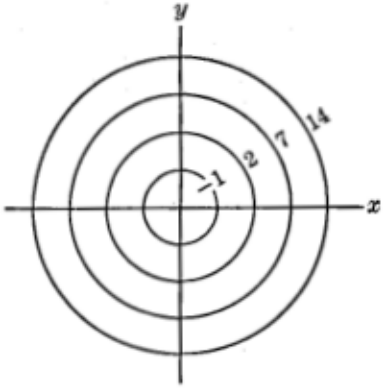


Fig. 39. Level of curves of $f(x, y) = x^2 + y^2 - 2$.

(curves of equal atmospheric pressure) on a meteorological chart furnish another good example of level curves of a function.

The three-dimensional analogue of this mode of representation is now easily grasped. Instead of level curves we shall have *level surfaces* $f(x, y, z) = C$. Common physical examples of functions of three variables which are conveniently visualized in this way are density and temperature in a gas or other medium.

6 / THE ELEMENTS OF PARTIAL DIFFERENTIATION

6 / PARTIAL DERIVATIVES

In this chapter the main objective is the exposition of the fundamentals of differential calculus for functions of two or more variables, with emphasis on the formal procedures which are used in dealing with various kinds of particular problems. The subject is developed from the beginning. Many students will already know some things about partial differentiation from courses in elementary calculus. For them the earlier parts of this chapter will serve as a review. By concentrating on procedures and techniques in this chapter, and putting most of the theoretical considerations in later chapters, the authors have hoped to make the treatment of partial differentiation adaptable to the needs of students of varying degrees of preparation. For further expression of the authors' intentions in the organization of Chapters 6, 7, and 8, see the introductory sections of Chapters 7 and 8.

Let $f(x, y)$ be defined in a region R of the xy-plane. If we think of y as fixed and x as variable, the derivative of $f(x, y)$ with respect to x is called the *partial derivative* with respect to x. This partial derivative is denoted by $\frac{\partial f}{\partial x}$. If we write $u = f(x, y)$, the partial derivative is also denoted by $\frac{\partial u}{\partial x}$. Likewise, the partial derivative with respect to y, $\frac{\partial f}{\partial y}$ or $\frac{\partial u}{\partial y}$, is the derivative of $f(x, y)$ with respect to y when x is regarded as a constant.

Example 1. If

$$u = x^2y + e^{-xy^3},$$

$$\frac{\partial u}{\partial x} = 2xy - y^3e^{-xy^3},$$

$$\frac{\partial u}{\partial y} = x^2 - 3xy^2e^{-xy^3}.$$

Similar definitions and notations apply in dealing with functions of three or more independent variables.

The partial derivatives of $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ are called the second partial derivatives of $f(x, y)$. There are in all four second derivatives of $f(x, y)$. The notations for

these derivatives, if we write $u = f(x, y)$, are the following:

$$\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) = \frac{\partial^2 u}{\partial x^2}, \qquad \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial x}\right) = \frac{\partial^2 u}{\partial y\,\partial x},$$

$$\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial y}\right) = \frac{\partial^2 u}{\partial x\,\partial y}, \qquad \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial y}\right) = \frac{\partial^2 u}{\partial y^2}.$$

Example 2. For the function of Example 1 we have

$$\frac{\partial^2 u}{\partial x^2} = 2y + y^6 e^{-xy^3},$$

$$\frac{\partial^2 u}{\partial y\,\partial x} = 2x + 3xy^5 e^{-xy^3} - 3y^2 e^{-xy^3},$$

$$\frac{\partial^2 u}{\partial x\,\partial y} = 2x + 3xy^5 e^{-xy^3} - 3y^2 e^{-xy^3},$$

$$\frac{\partial^2 u}{\partial y^2} = 9x^2 y^4 e^{-xy^3} - 6xy e^{-xy^3}.$$

We observe that

$$\frac{\partial^2 u}{\partial y\,\partial x} = \frac{\partial^2 u}{\partial x\,\partial y} \tag{6–1}$$

in this example. We shall ordinarily find that the relation (6–1) holds true for the functions we meet in practice, for the relation is valid at a point provided both the second derivatives are defined in a neighborhood of the point and continuous at the point. This will be proved in §7.2 (Theorem III).

A partial derivative of $f(x, y)$ is again a function of x, y. To denote the value of $\dfrac{\partial f}{\partial x}$ at the point (x_0, y_0) we may use one of the expressions

$$\left(\frac{\partial f}{\partial x}\right)_{(x_0, y_0)}, \qquad \left.\frac{\partial f}{\partial x}\right|_{(x_0, y_0)}.$$

These notations are rather awkward, however; it is desirable to have a standard functional notation for partial derivatives. For a function $f(x, y)$ of two independent variables we shall write

$$f_1(x, y) = \frac{\partial f}{\partial x}, \qquad f_2(x, y) = \frac{\partial f}{\partial y}.$$

For the value of a partial derivative at a point we then have expressions such as

$$f_1(x_0, y_0) = \left(\frac{\partial f}{\partial x}\right)_{(x_0, y_0)}, \qquad f_2(a, b) = \left(\frac{\partial f}{\partial y}\right)_{(a, b)}.$$

For second derivatives we use the notations

$$f_{11}(x, y) = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right), \qquad f_{12}(x, y) = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right).$$

$$f_{21}(x, y) = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right), \qquad f_{22}(x, y) = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right).$$

Observe the ordering of the numerical subscripts in relation to the order of carrying out the differentiations; 1 refers to x and 2 refers to y.

The notation is extended in an obvious way to derivatives of order higher than the second, and also to functions of more than two independent variables. Thus for example,

$$f_{122}(x, y) = \frac{\partial^2}{\partial y^2}\left(\frac{\partial f}{\partial x}\right),$$

and

$$g_3(x, y, z) = \frac{\partial g}{\partial z}, \qquad g_{123}(x, y, z) = \frac{\partial}{\partial z}\left[\frac{\partial}{\partial y}\left(\frac{\partial g}{\partial x}\right)\right].$$

6.1 / IMPLICIT FUNCTIONS

We often deal with functions which are defined implicitly as the solution of certain equations. In ordinary practice we can find the partial derivatives of such a function by the same procedures which we learn in elementary calculus.

Example 1. Find $\dfrac{\partial z}{\partial x}$ from the equation

$$\frac{x^2}{16} + \frac{y^2}{12} + \frac{z^2}{9} = 1, \tag{6.1–1}$$

on the understanding that z is dependent and x, y are independent.

We have

$$\frac{2x}{16} + \frac{2z}{9}\frac{\partial z}{\partial x} = 0,$$

and so

$$\frac{\partial z}{\partial x} = -\frac{9x}{16z}. \tag{6.1–2}$$

The equation (6.1–1) actually defines two functions of (x, y), corresponding to the two choices of sign in

$$z = \pm 3\left(1 - \frac{x^2}{16} - \frac{y^2}{9}\right)^{1/2}. \tag{6.1–3}$$

By substituting (6.1–3) in (6.1–2) we obtain the partial derivative for each of these two functions:

$$\frac{\partial z}{\partial x} = \mp \frac{3x}{16}\left(1 - \frac{x^2}{16} - \frac{y^2}{9}\right)^{-1/2}. \tag{6.1–4}$$

The result (6.1–4) could also have been obtained by differentiating (6.1–3) directly.

The procedure can also be applied in the case of functions defined by simultaneous equations.

Example 2. If u and v are defined as functions of x, y by the equations

$$\begin{aligned} u\cos v - x &= 0 \\ u\sin v - y &= 0, \end{aligned} \tag{6.1–5}$$

find the partial derivatives $\dfrac{\partial u}{\partial x}, \dfrac{\partial v}{\partial x}$.

Method I. One method of procedure is to attempt to solve for u, v in terms of x, y. If this can be accomplished, we can then calculate the required partial derivatives directly.

From (6.1–5) we have

$$u^2\cos^2 v = x^2,\ u^2\sin^2 v = y^2.$$

Now add these equations and use a familiar trigonometric identity. The result is

$$u^2 = x^2 + y^2, \text{ or } u = \pm\sqrt{x^2+y^2}. \tag{6.1–6}$$

Next, going back to (6.1–5), we substitute the value just found for u. We find

$$\cos v = \frac{x}{\pm\sqrt{x^2+y^2}},\quad \sin v = \frac{y}{\pm\sqrt{x^2+y^2}}, \tag{6.1–7}$$

the same sign being taken before the radical in both cases. We might also write

$$\tan v = \frac{y}{x} \tag{6.1–8}$$

in cases $x \neq 0$. We see that there are in general two possible determinations of u from (6.1–6); for v there are an infinite number of possible determinations from (6.1–7), differing by multiples of 2π. The derivatives of u may be found from (6.1–6):

$$\frac{\partial u}{\partial x} = \pm\frac{x}{\sqrt{x^2+y^2}}.$$

In finding the derivatives of v it is easier to work from (6.1–8). We have

$$\sec^2 v\,\frac{\partial v}{\partial x} = \frac{-y}{x^2},\ \sec^2 v = 1+\tan^2 v = 1+\frac{y^2}{x^2},$$

$$\frac{\partial v}{\partial x} = -\frac{y}{x^2\sec^2 v} = \frac{-y}{x^2+y^2}.$$

This result could have been obtained by expressing v as an inverse tangent and then differentiating. It should, however, be noted that v is not necessarily the *principal value* of the inverse tangent of y/x.

Method II. As an alternative to the first method we may proceed directly to differentiate the equations (6.1–5). In doing so we must bear in mind that x, y are independent variables and that u, v are dependent variables. Since we are seeking partial derivatives with respect to x, we think of y as a constant. Then

$$\begin{aligned} \cos v \frac{\partial u}{\partial x} - u \sin v \frac{\partial v}{\partial x} - 1 = 0, \\ \sin v \frac{\partial u}{\partial x} + u \cos v \frac{\partial u}{\partial x} - 0 = 0. \end{aligned} \tag{6.1–9}$$

These simultaneous equations are now solved for $\frac{\partial u}{\partial x}$ and $\frac{\partial v}{\partial x}$. The solution may be achieved by elimination, or by determinants and Cramer's rule. We leave it for the student to carry out the solution and find

$$\frac{\partial u}{\partial x} = \cos v, \qquad \frac{\partial v}{\partial x} = \frac{-\sin v}{u}. \tag{6.1–10}$$

To reconcile these answers with the answers as found by Method I, observe that, by (6.1–5) and (6.1–6),

$$\cos v = \frac{x}{u} = \frac{x}{\pm\sqrt{x^2+y^2}},$$

$$\frac{-\sin v}{u} = -\frac{y}{u^2} = \frac{-y}{x^2+y^2}.$$

One of the things to be noted in connection with problems like that of Example 2 is that the solution by Method II can be carried out even if the explicit solution for u and v in terms of x, y is impracticable.

In all the procedures which have been illustrated here it has been taken for granted that the given equations *do* implicitly define certain functions, and that these functions *do have* partial derivatives. The deeper theoretical questions on these matters are considered fully in Chapter 8.

EXERCISES

1. Find $\frac{\partial u}{\partial y}$ and $\frac{\partial v}{\partial y}$ from Example 2. Use both of the methods given in the text, and reconcile your answers.

2. Suppose that u and v are defined in terms of x, y by the equations $u^2 - v^2 = xy$, $uv = x^2 + y^2$. Find the first partial derivatives of u and v with respect to x and y, respectively. Use the implicit function procedure (Method II). See how far you can go with Method I in this case.

3. Assuming that z is defined as a function of x and y by the equation $4\sin^2 x + 2\cos(y+z) = 2$, find $\partial z/\partial x$ when $\partial z/\partial y$ when $x = \pi/6$, $y = \pi/3$, $z = 0$.

4. Find $\partial z/\partial x$ and $\partial z/\partial y$ from the equation $x^3 + y^3 + z^3 - 3xyz = 0$.

6.2 / GEOMETRICAL SIGNIFICANCE OF PARTIAL DERIVATIVES

Just as the ordinary derivative of a function of one variable has its geometric realization in the slope of a line which is tangent to a curve, so the partial derivatives of a function of two variables have a geometrical significance in connection with a plane which is tangent to a surface. Our purpose in this section is to show how the partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$, when $x = a$ and $y = b$, are related to the plane which is tangent to the surface $z = f(x, y)$ at the point (a, b, c). In this section we shall not give a formal definition of the tangent plane. We reserve full discussion of this matter to §6.4, because the concept of the tangent plane is the geometrical counterpart of the concept of the differential of a function of two variables.

Let S be the surface $z = f(x, y)$, and let (a, b, c) be a point on S. Then $c = f(a, b)$. Consider the line through the point (a, b, c) parallel to the z-axis. Let us visualize various planes containing this line, each such plane cutting the surface S in a curve. One such plane, cutting the y-axis perpendicularly at $y = b$, is shown in Fig. 40. In the diagram, the curve of the intersection of S and this plane, $y = b$, is represented as having a tangent line L at the point (a, b, c). One can also imagine a plane $x = a$, passing through (a, b, c) and intersecting the x-axis perpendicularly at $(a, 0, 0)$. More generally, one can imagine a plane different from either of these two, but so placed that it passes through (a, b, c) and is parallel to the z-axis. The student should construct for himself a diagram similar to Fig. 40, with a plane through (a, b, c) parallel to the z-axis but cutting neither the x-axis nor the y-axis at right angles. This exercise in geometrical visualization will be helpful for the following discussion.

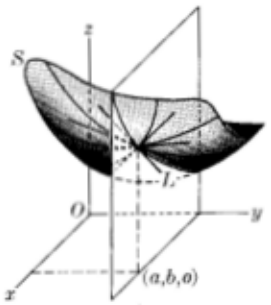


Fig. 40.

Suppose that each plane through (a, b, c) and parallel to the z-axis cuts the surface S in a curve which has at the point (a, b, c) a tangent line which is not parallel to the z-axis. Suppose further that all of these tangent lines, corresponding to the different planes of the type described, lie in a single plane. Then this single plane must surely be the tangent plane to the surface S at (a, b, c), if indeed there is such a tangent plane. Fig. 40 shows four different curves on S, together with their tangent lines, all intersecting at (a, b, c).

Assuming now that there is a tangent plane to S at (a, b, c) not parallel to the z-axis, let us see how to find its equation. If $\cos\alpha$, $\cos\beta$, $\cos\gamma$ are the direction cosines of a line which is normal (perpendicular) to this plane, the equation of the plane can be written

$$(\cos\alpha)(x - a) + (\cos\beta)(y - b) + (\cos\gamma)(z - c) = 0.$$

Since the plane is not parallel to the z-axis, we know that $\cos\gamma \neq 0$; we can therefore solve for $z - c$ by dividing by $\cos\gamma$, thus obtaining an equation of the

form

$$z - c = A(x - a) + B(y - b). \tag{6.2–1}$$

Our problem is to find A and B. If we put $y = b$ in this equation, we find the relation between x and z on the line in which the tangent plane (6.2–1) cuts the plane $y = b$. This line of intersection must be the line L which is tangent to the curve in which the surface S is cut by the plane $y = b$ (see Fig. 40). The relation between x and z on L is obtained by putting $y = b$ in (6.2–1); that is,

$$z - c = A(x - a).$$

The relation between x and z along the curve in which S is cut by the plane $y = b$ is

$$z = f(x, b).$$

Hence, according to the usual relationship between slopes and derivatives, A must be the derivative of z with respect to x when y is held constantly equal to b:

$$A = \frac{\partial}{\partial x} f(x, b)\bigg|_{x=a} = f_1(a, b). \tag{6.2–2}$$

In the same way, putting $x = a$ in (6.2–1) and interpreting B as a slope in the relation between y and z, with x held constantly equal to a, we see that

$$B = \frac{\partial}{\partial y} f(a, y)\bigg|_{y=b} = f_2(a, b). \tag{6.2–3}$$

Consequently, we see that the equation of the plane tangent to S at (a, b, c) can be written in the form

$$z - c = f_1(a, b)(x - a) + f_2(a, b)(y - b), \tag{6.2–4}$$

or

$$z = f(a, b) + f_1(a, b)(x - a) + f_2(a, b)(y - b). \tag{6.2–5}$$

The foregoing discussion of the tangent plane is not complete because of the fact that no full and formal definition of the tangent plane has been given. However, the discussion here is wholly consistent with the definition to be given in §6.4. It is shown there that when S has a tangent plane at (a, b, c), not parallel to the z-axis, then f has partial derivatives when $x = a$ and $y = b$. The arguments we have just given can then be used to derive the equation (6.2–4) for the tangent plane. Observe that the assumption that f has partial derivatives at $x = a$, $y = b$ rules out the possibility that the tangent plane might be parallel to the z-axis.

The line through (a, b, c) perpendicular to the plane (6.2–5) is called the *normal* to the surface S at that point. As we see from the equation (6.2–5), the direction of the normal is determined by the ratios

$$f_1(a, b) : f_2(a, b) : -1,$$

for the direction cosines of the normal to a plane are proportional to the coefficients of x, y, and z respectively, in the equation of the plane. Here is a result to be remembered:

The line normal to the surface $z = f(x, y)$ at a given point has direction ratios $\frac{\partial z}{\partial x} : \frac{\partial z}{\partial y} : -1$, the partial derivatives being evaluated at the point in question.

Example 1. (*a*) Find the equation of the plane tangent to the paraboloid $48z = 2x^2 + 3y^2$ at the point $(3, 2, \frac{5}{8})$. (*b*) Find the direction cosines of the normal to the surface at the point.

(*a*) We have $$48\frac{\partial z}{\partial x} = 4x, \qquad 48\frac{\partial z}{\partial y} = 6y;$$

at the point in question, therefore, $\frac{\partial z}{\partial x} = \frac{1}{4}$, $\frac{\partial z}{\partial y} = \frac{1}{4}$, and the equation of the tangent plane is

$$z - \tfrac{5}{8} = \tfrac{1}{4}(x - 3) + \tfrac{1}{4}(y - 2), \text{ or } 2x + 2y - 8z = 5.$$

(*b*) To obtain the direction cosines from the ratios $\frac{1}{4} : \frac{1}{4} : -1$, we first compute

$$[(\tfrac{1}{4})^2 + (\tfrac{1}{4})^2 + (-1)^2]^{1/2} = \frac{3\sqrt{2}}{4}.$$

The direction cosines are found by dividing $\frac{1}{4}$, $\frac{1}{4}$, -1 by $\frac{3\sqrt{2}}{4}$. They are, accordingly, $\frac{1}{3\sqrt{2}}, \frac{1}{3\sqrt{2}}, \frac{-4}{3\sqrt{2}}$.

Example 2. Show that at every point of intersection of the two surfaces $z = 2(x^2 + y^2)$, $8z = 17 - (x^2 + y^2)$, the normals to the two surfaces are perpendicular. (Because of this we say that the surfaces intersect *orthogonally*.)

The student will readily find that the surfaces are paraboloids of revolution intersecting each other all along the circle $x^2 + y^2 = 1$ in the plane $z = 2$. There are no other intersections. Now, at a point of the first paraboloid the direction ratios of the normal to the surface are

$$4x : 4y : -1;$$

for the second paraboloid the direction ratios of the normal are found to be

$$-\frac{x}{4} : -\frac{y}{4} : -1.$$

The condition for perpendicularity of these two normals at a point which is common to the two surfaces is therefore

$$4x\left(\frac{-x}{4}\right) + 4y\left(\frac{-y}{4}\right) + 1 = 0,$$

tangent planes orthogonal
also normals meet orthogonally

or $-(x^2+y^2)+1=0$. Since this equation is satisfied along the intersection of the surfaces, the demonstration of perpendicularity is complete.

EXERCISES

1. Find the equation of the tangent plane to the surface $z = e^{-x}\sin y$ **(a)** at $x=0$, $y=\pi/2$, **(b)** at $x=0$, $y=\pi$, **(c)** at $x=0$, $y=0$. **(d)** Make as good a diagram as you can of the surface for $0 \leqq y \leqq \pi$, $x>0$.

2. Find the equation of the plane tangent to the surface $x^3+2xy^2-7z^3+3y+1=0$ at $(1,1,1)$.

3. Prove that the plane tangent to the surface $z=x^2-y^2$ at the point (a,b,c) is pierced by the z-axis at the point for which $z=-c$.

4. Find the points of the praraboloid $z=x^2+y^2-1$ at which the normal to the surface coincides with the line joining the origin to the point. What is the acute angle between the normal and the z-axis at these points?

5. If $a^2 \neq b^2$, prove that no normal to the surface $z=(x^2/a^2)+(y^2/b^2)-c$, at a point for which $x\neq 0$ and $y\neq 0$, can pass through the origin.

6. Prove that the spheres $x^2+y^2+z^2=16$, $x^2+(y-5)^2+z^2=9$ intersect orthogonally, using the method of Example 2.

6.3 / MAXIMA AND MINIMA

We sometimes have occasion to inquire about the largest or smallest value attained by a function under specified circumstances. In speaking about maximum (or minimum) values it is very important to distinguish between a *relative* maximum and an *absolute* maximum. Suppose we are dealing with a function $f(x, y)$ defined in a region R of the xy-plane.

Definition. *We say that the function f has a relative maximum at the point (a, b) if there is some neighborhood of (a, b) such that $f(x, y) \leqq f(a, b)$ for all points (x, y) of R which are in this neighborhood. We may express the definition otherwise by saying that the value of f at (a, b) is at least as big as at any of the points (x, y) around (a, b) and not too far away.*

Thus for instance, within a given range of mountains, the elevation of the land surface above sea level attains a relative maximum at the summit of any particular peak in the range.

Definition. *Let f be defined in a region R, and let S be any part of R (i.e., any point-set in R). In particular, S might be all of R. Suppose there is in S a point (a, b) such that $f(x, y) \leqq f(a, b)$ for all points (x, y) in S. We then say that on the set S the function f has an absolute maximum at (a, b).*

Observe that, on a given set S, f can have a relative maximum which is not an absolute maximum. Observe also that a function may fail to have an absolute maximum on a given set (think of the function $1/(xy)$ in the first quadrant).

Similar definitions are made for relative and absolute minima of a function.

In problems where we have to find the absolute maximum of a function on a given set we usually find that it is convenient to begin by looking for relative maxima. If there are only a few of the latter we may be able easily to select one which furnishes an absolute maximum. Hence it is useful to have criteria for locating relative extrema.

THEOREM I. *Let f be defined on a region R, and let the function have a relative extreme (maximum or minimum) at the point (a, b) of R. Suppose further that (a, b) is an interior point of R (not on the boundary), and that f has first partial derivatives at (a, b). Then these derivatives are zero at that point:*

$$f_1(a, b) = 0, \qquad f_2(a, b) = 0. \tag{6.3-1}$$

Proof. This theorem should be compared with Theorem III of §1.12. The proof is based on this earlier theorem. Consider $f(x, b)$; this is a function of the single variable x, its values being those of the function $f(x, y)$ along the line $y = b$. As a function of x, $f(x, b)$ has a relative extreme at $x = a$. Moreover, the derivative of $f(x, b)$ at $x = a$ is $f_1(a, b)$. Therefore, by Theorem III, §1.12, we conclude that $f_1(a, b) = 0$. In the same way, applying this earlier theorem to the function $f(a, y)$ of the single variable y, we conclude that $f_2(a, b) = 0$.

The hypothesis that (a, b) is an interior point of R is essential. A relative extreme can occur at a boundary point of R, and in that case equations (6.3–1) may not hold.

It is important to realize that, under the conditions stated in Theorem I, the vanishing of the first partial derivatives is a necessary, but not sufficient, condition for a relative extreme. If the surface $z = f(x, y)$ has at the point $x = a$, $y = b$ a tangent plane which is parallel to the xy-plane, then equations (6.3–1) hold; but z need not be a relative extreme at such a point. A "saddle-point" of a surface is an illustration of such a situation.

The foregoing definitions and Theorem I extend to functions of three or more variables in an obvious manner.

As in elementary calculus, sufficient conditions for a relative maximum or minimum can be formulated by adding to (6.3–1) certain conditions on the second derivatives of f at the point (a, b). We discuss such conditions in §7.6. For the present, however, we proceed to illustrate some uses of Theorem I.

Example 1. Find the point of the plane $2x - 3y - 4z = 25$ which is nearest to the point $(3, 2, 1)$.

If D is the distance from the point (x, y, z) of the plane to $(3, 2, 1)$, we have $D^2 = (x - 3)^2 + (y - 2)^2 + (z - 1)^2$ and $z = \frac{1}{4}(2x - 3y - 25)$. Hence, eliminating z,

$$D^2 = (x - 3)^2 + (y - 2)^2 + (\tfrac{1}{2}x - \tfrac{3}{4}y - \tfrac{29}{4})^2.$$

We seek the minimum value of D^2 as x, y range through all possible values. In this case *all* points are interior points of the region (namely the whole xy-plane), and D^2 has partial derivatives at all points. We therefore look for points at which

$\frac{\partial(D^2)}{\partial x} = \frac{\partial(D^2)}{\partial y} = 0$. The equations to be considered are

$$2(x-3) + 2(\tfrac{1}{2}x - \tfrac{3}{4}y - \tfrac{29}{4}) \cdot \tfrac{1}{2} = 0,$$
$$2(y-2) + 2(\tfrac{1}{2}x - \tfrac{3}{4}y - \tfrac{29}{4}) \cdot (\tfrac{-3}{4}) = 0.$$

On simplifying, we obtain

$$10x - 3y = 53,$$
$$-6x + 25y = -55.$$

The solution is found to be $x = 5$, $y = -1$. Substituting in the equation of the plane, we find $z = -3$. We now argue as follows: The function D^2 certainly has an absolute minimum (from the geometrical nature of the problem). This absolute minimum is also a relative minimum, and the conditions of Theorem I apply. But we obtain a unique point at which the two first partial derivatives vanish. Hence, this point must furnish the desired absolute minimum.

Example 2. Locate the points which might furnish relative maxima and minima of the function

$$f(x, y) = 2xy - (1 - x^2 - y^2)^{3/2}$$

in the closed region $x^2 + y^2 \leqq 1$ (which is the region of definition of the function). Hence, find the absolute maximum and minimum values of the function.

We first apply the criterion of Theorem I. We have

$$\frac{\partial f}{\partial x} = 2y + 3x(1 - x^2 - y^2)^{1/2},$$

$$\frac{\partial f}{\partial y} = 2x + 3y(1 - x^2 - y^2)^{1/2}.$$

The interior points of the region are those for which $x^2 + y^2 < 1$. The interior points which might furnish a relative maximum or minimum are among those which we find by solving the equations

$$\begin{aligned} 3x(1 - x^2 - y^2)^{1/2} &= -2y, \\ 3y(1 - x^2 - y^2)^{1/2} &= -2x. \end{aligned} \qquad (6.3\text{–}2)$$

An obvious solution of these equations is $x = 0$, $y = 0$. If neither x nor y is zero we may divide one equation by the other and obtain the result

$$\frac{x}{y} = \frac{y}{x}, \text{ or } x^2 = y^2.$$

Hence, substituting back in the first equation of (6.3–2) after squaring both sides, we obtain

$$9x^2(1 - 2x^2) = 4x^2, \text{ or } 9 - 18x^2 = 4.$$

Thus we find $x^2 = y^2 = \frac{5}{18}$. Going back again to (6.3–2) we have $(1 - x^2 - y^2)^{1/2} = \frac{2}{3}$,

and hence $3x(\frac{2}{3}) = -2y$, or $x = -y$. Note that $x = y$ is ruled out. There are therefore three points in all which satisfy (6.3–2). They are

$$P_0 = (0, 0),\ P_1 = (\tfrac{1}{3}\sqrt{\tfrac{3}{2}}, -\tfrac{1}{3}\sqrt{\tfrac{3}{2}}),\ P_2 = (-\tfrac{1}{3}\sqrt{\tfrac{3}{2}}, \tfrac{1}{3}\sqrt{\tfrac{3}{2}}).$$

The accompanying table of values may now be constructed:

Point	Value of f
P_0	-1
P_1 and P_2	$-23/27$

We emphasize that Theorem I does not assert that the function has relative extrema at all three of these points; it only states that if any relative extrema occur at interior points, such extrema are found among these three points. Before drawing any conclusions about absolute extrema we must investigate the behavior of the function on the boundary of the region. On the boundary we have $x^2 + y^2 = 1$ and therefore $f(x, y) = 2xy$. To look for extreme values of f on the boundary we might solve for y: $y = \pm\sqrt{1 - x^2}$, and look for the extremes of $\pm 2x\sqrt{1 - x^2}$ by the methods of elementary calculus. We find such extremes when $x^2 = \frac{1}{2}$. A more elegant procedure is to introduce the parametric equations $x = \cos\theta$, $y = \sin\theta$ for the boundary circle (here θ is the usual angle of polar co-ordinates). Then $2xy = 2\cos\theta\sin\theta = \sin 2\theta$, and we see that the values range between

$$-1\left(\text{at } \theta = \frac{3\pi}{4} \text{ or } \frac{7\pi}{4}\right) \text{ and } +1\left(\text{at } \theta = \frac{\pi}{4} \text{ or } \frac{5\pi}{4}\right).$$

We have now found four more points which must be considered along with the original three when we look for the absolute minimum and maximum values of the function in the closed region. When we compare the values ± 1 with the values at the points listed in the table, we see that the function has the absolute maximum value $+1$, and the absolute minimum value -1. The maximum occurs at the two boundary points $(\sqrt{2}/2, \sqrt{2}/2)$, $(-\sqrt{2}/2, -\sqrt{2}/2)$. The minimum occurs at the interior point P_0 and at the two boundary points $(\sqrt{2}/2, -\sqrt{2}/2)$, $(-\sqrt{2}/2, \sqrt{2}/2)$. Our work has not settled the questions as to whether the interior points P_1, P_2 are points of relative extrema or saddle points. They are in fact saddle points, as may be shown by an examination of the function in polar co-ordinates.

Example 3. A shelter for use at the beach is to be built in the form of a box-like space with canvas covering on the top, back, and ends. If 96 square feet of canvas are available, what should be the dimensions of the shelter to give it maximum cubic content?

Let the shelter be y feet between ends, x feet from front to back, and z feet high. Its volume is $V = xyz$. The area to be covered by canvas is

$$A = 2xz + xy + yz.$$

Since $A = 96$, we can use this last equation to eliminate one variable, say y:

$$y = \frac{96 - 2xz}{x + z},\quad V = 2xz\,\frac{48 - xz}{x + z}.$$

Here V is expressed in terms of the independent variables x, z; we can set $\frac{\partial V}{\partial x} = \frac{\partial V}{\partial z} = 0$ and solve for x and z. An alternative procedure which is in some ways preferable is the following: Differentiate both of the equations

$$V = xyz, A = 2xz + xy + yz$$

with respect to x and z, regarding y as a function of x and z. Differentiation with respect to x gives

$$0 = \frac{\partial V}{\partial x} = yz + xz\frac{\partial y}{\partial x}, 0 = 2z + x\frac{\partial y}{\partial x} + y + z\frac{\partial y}{\partial x}.$$

We now eliminate $\frac{\partial y}{\partial x}$ between these two equations

$$\frac{\partial y}{\partial x} = -\frac{y}{x}, 0 = 2z + x\left(-\frac{y}{x}\right) + y + z\left(-\frac{y}{x}\right),$$

$$2z - y + y - \frac{yz}{x} = 0, 2x = y.$$

Since x and z enter symmetrically, we infer that $2z = y$ also, and hence that $x = z$. To get the values of x, y, z we return to the formula for A. We now have

$$96 = 2x(x) + x(2x) + (2x)x = 6x^2.$$

Hence $x^2 = 16$, $x = z = 4$, $y = 8$. The volume of the shelter of these dimensions is 128 cubic feet.

One logical issue still remains to be settled in the foregoing "solution" of the problem posed in Example 3. How do we know that we really found the dimensions which yield maximum volume? Our method was based on two assumptions: First, that there is a shelter of maximum volume under the given conditions, and second, that when V is expressed as a function of the independent variables x, z, the equations $\frac{\partial V}{\partial x} = \frac{\partial V}{\partial z} = 0$ are satisfied when V attains its maximum. If we can justify these two assumptions, our solution will be fully established. Let us then consider V as a function of x and z. The formula is

$$V = 2xz\frac{48 - xz}{x + z}; \tag{6.3–3}$$

we do not consider all values of x and z, however, but only those which have a meaning for the problem under consideration. Thus we must have $x \geqq 0$, $z \geqq 0$. We must have $xz \leqq 48$ also, since a negative volume would have no meaning. The region R in which we consider the values of V is thus composed of all points of the xz-plane for which $x \geqq 0$, $z \geqq 0$, $xz \leqq 48$, except the one point $x = 0$, $z = 0$. This point is ruled out, since V is not defined there. The region R is shown in Fig. 41.

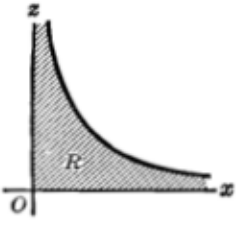


Fig. 41.

Let us now establish the fact that among all possible values of V in the region R, there is an absolute maximum value, and that this maximum occurs at an interior point of R. Observe that V is positive in the interior of R and that $V = 0$ at all points of the boundary of R except the origin (where V is not defined). Now, from the fact that $xz \leqq 48$ in R we see that

$$V \leqq 96\frac{48}{x+z},$$

and hence $V \to 0$ as a variable point (x, z) of R moves in such a way that either $x \to \infty$ or $z \to \infty$. Finally, $V \to 0$ as $(x, z) \to (0, 0)$. For, $2xz \leqq x^2 + z^2$ and $x + z \geqq \sqrt{x^2 + z^2}$ are true inequalities (the latter when x and z are non-negative), and therefore

$$0 \leqq V \leqq \frac{96(x^2+z^2)}{\sqrt{x^2+z^2}} = 96\sqrt{x^2+z^2},$$

so that $V \to 0$ as $(x, z) \to (0, 0)$. From the foregoing arguments it is now clear that if we form a new region R_0 by excluding from R the points for which $x^2 + z^2 < \delta^2$ and $x^2 + z^2 > 1/\delta^2$, where δ is sufficiently small, the values of V at these excluded points will all be smaller than *some* of the values of V in the remaining region R_0. Since R_0 is a bounded and closed region in which V is continuous, V must have a maximum value in R_0 (Theorem II, §5.3). The maximum value of V in R_0 will also be a maximum value of V in relation to all points of the larger region R. Since this maximum is positive, it must occur at an interior point of R.

We now apply Theorem I to draw the conclusion that $\frac{\partial V}{\partial x} = \frac{\partial V}{\partial z} = 0$ at the point where V is a maximum. Since there turned out to be only one point in R, namely $x = z = 4$, at which these conditions are satisfied, this point must be the point where V is a maximum.

EXERCISES

1. Find the point of the plane $x + 4y + 4z = 39$ nearest the point $(2, 0, 1)$.

2. Find the greatest value of the function $xy(c - x - y)$ in the closed triangular region with vertices $(0, 0)$, $(c, 0)$, and $(0, c)$. Assume $c > 0$.

3. Find the absolute maximum of $144x^3y^2(1 - x - y)$ in the first quadrant of the xy-plane.

4. Does $f(x, y) = x^2 + \frac{2}{3}xy + y^2 + (576/x) + (576/y)$ have any absolute extrema in the region $x > 0$, $y > 0$? If so, find where such extrema occur, and the type (maximum or minimum). Give all the supporting details of your argument.

5. Find the absolute minimum value of

$$f(x, y) = x^2 + y^2 + \left(\frac{2A - ax - by}{c}\right)^2$$

where A, a, b, and c are positive constants. All values of x and y are admitted. How do you know that a minimum exists?

6. Find the absolute extreme values of the function $f(x, y) = 2xy + (1 - x^2 - y^2)^{1/2}$ in the region $x^2 + y^2 \leqq 1$.

7. Find the absolute extreme values of the function $f(x, y) = xy - (1 - x^2 - y^2)^{3/2}$ in the region $x^2 + y^2 \leqq 1$.

8. **(a)** Introducing polar co-ordinates, show that the function of Example 2 becomes $f(x, y) = F(r, \theta) = r^2 \sin 2\theta - (1 - r^2)^{3/2}$. **(b)** Find the extreme values of F for $-1 \leqq r \leqq 1$, θ unrestricted, considering F as defined in a region of the $r\theta$ plane. **(c)** At the points of the $r\theta$-plane which correspond to the points P_1, P_2 of Example 2, show $\frac{\partial F}{\partial r} = 0$, $\frac{\partial F}{\partial \theta} = 0$, $\frac{\partial^2 F}{\partial r^2} < 0$, $\frac{\partial^2 F}{\partial \theta^2} > 0$. From these facts explain why these points are saddle points and not points of relative extrema.

9. Solve Exercises 6 and 7 by the introduction of polar co-ordinates as independent variables.

10. Find the greatest value of the function $\sin x \sin y \sin(x + y)$ in the closed triangular region with vertices $(0, 0)$, $(\pi, 0)$, $(0, \pi)$.

11. Find the absolute maximum value of the function $(x^2 + 2y^2)e^{-(x^2+y^2)}$, considering all possible values of x and y.

12. Find the absolute extrema of the function $3x^2 - 8xy - 4y^2 + 2x + 16y$ in the square $0 \leqq x \leqq 2$, $0 \leqq y \leqq 2$.

13. Find the absolute extrema of the function $x^3 + y^3 + 3xy^2 - 15x - 15y$ in the square $0 \leqq x \leqq 3$, $0 \leqq y \leqq 3$.

14. Find the minimum value of the function $(12/x) + (18/y) + xy$ in the first quadrant. How do you know there is a minimum?

15. Find the maximum value of the function $(xy - 4y - 8x)/x^2y^2$ in the first quadrant. How do you know there is a maximum?

16. A rectangular box without a top has length x, width y, and depth z. The combined area of the sides and bottom is fixed as S square feet. Express the volume V of the box as a function of x, y, and show that V is greatest when $x = y = (S/3)^{1/2}$, $z = x/2$. Justify your solution completely.

17. Consider the function $\sqrt{x^2 + y^2} + \sqrt{(x - 1)^2 + y^2}$. **(a)** Explain carefully why this function must have an absolute minimum at some point of the plane. **(b)** What is the minimum and where does it occur? **(c)** Are all the minimum points found by setting the partial derivatives equal to zero?

18. Consider the function $f(x, y) = |y| + \sqrt{x^2 + (y - 1)^2}$.
(a) At what points do one or both of the first partial derivatives of f fail to exist? **(b)** Find all points where they both exist and are equal to zero. **(c)** What is the absolute minimum value of f, and at what points does it occur?

19. For what position of the point (x, y) is the sum of the distance from (x, y) to the x-axis and twice the distance from (x, y) to the point $(0, 1)$ a minimum?

20. Let $f(x, y, z)$ be the sum of the three distances: from (x, y, z) to the y-axis, from (x, y, z) to the z-axis, and from (x, y, z) to the point $(1, 0, 0)$. Find the absolute minimum value of f, and where it occurs.

6.4 / DIFFERENTIALS

Our purpose in this section is to define what is meant by saying that a real-valued function of several real variables is differentiable and to define the

differential of such a function. These definitions are needed in order that we may derive the chain rules for differentiating composite functions (Theorem III, §6.5 and Theorem V, §7.3). We stress the case of two variables, but the ideas apply to functions of three or more variables, as we shall see.

In §1.3 we defined the differential of a function f of the independent variable x as the function of x and an independent variable dx whose value is $f'(x)\,dx$. The differential of f is defined at each point x where f has a derivative. The independent variable dx can have any value. For a fixed value of x the value of the differential is a multiple of dx, the multiplier being the value $f'(x)$ of the derivative of f at x. It is useful to restate the definition of differentiability as follows: the function f is called differentiable at x if it is defined at x and all points near x and if there exists a number C such that

$$\lim_{\Delta x \to 0} \frac{|f(x+\Delta x) - f(x) - C\Delta x|}{|\Delta x|} = 0. \tag{6.4–1}$$

We can rewrite (6.4–1) in the equivalent form

$$\lim_{\Delta x \to 0} \left| \frac{f(x+\Delta x) - f(x)}{\Delta x} - C \right| = 0,$$

and this new form is in turn equivalent to the assertion

$$\lim_{x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} = C. \tag{6.4–2}$$

But (6.4–2) is the same as the assertion that f has a derivative at x, with value $f'(x) = C$.

Let the functional symbol for the differential be df, and denote by $df(x; dx)$ the value of df as a function of x and dx. We use a semicolon rather than a comma to separate x and dx in order to emphasize the fact that dependence of df on dx is in general different from its dependence on x; df is a *linear* function of dx. We can replace dx by other symbols, such as Δx or h. We can write (6.4–1) in the form

$$\lim_{h \to 0} \frac{|f(x+h) - f(x) - df(x; h)|}{|h|} = 0. \tag{6.4–3}$$

This is because $C\Delta x = f'(x)\Delta x = df(x; \Delta x)$; we simply replace Δx by h. The characteristic features of the differential are: (1) that $df(x; h)$ is (for fixed x) a multiple of h and (2) that the limit relation (6.4–3) is valid. This limit relation can be described by saying that $df(x; h)$ is a good approximation to $f(x+h) - f(x)$ in the sense that the difference

$$f(x+h) - f(x) - df(x; h)$$

is small in comparison with $|h|$ as $|h| \to 0$.

The foregoing discussion of the one-variable case provides us with a motivation for the definitions of differentiability and the differential for the case of a function of two variables. For a function f of x and y we want the

differential $df(x, y; dx, dy)$ to be a linear combination $A\,dx + B\,dy$, where A and B are numbers determined by f and (x, y), and we want a limit relation analogous to (6.4–3). In speaking of points near (x, y) we represent them by $(x + h, y + k)$, where the distance from $(x + h, y + k)$ to (x, y) is $\sqrt{h^2 + k^2}$.

Definition. *The function f of (x, y) is called differentiable at (x, y) if it is defined for all points near (x, y) [that is, in a neighborhood of (x, y)] and if there exists numbers A, B [depending on f and (x, y)] such that*

$$\lim_{(h,k)\to(0,0)} \frac{|f(x+h, y+k) - f(x, y) - (Ah + Bk)|}{\sqrt{h^2 + k^2}} = 0. \tag{6.4–4}$$

The differential of f at (x, y) is then defined to be the function df of (x, y) and (dx, dy) with the value

$$df(x, y; dx, dy) = A\,dx + B\,dy. \tag{6.4–5}$$

Observe that the differential is a linear function of dx and dy, that is, a linear combination of them. The variables dx and dy can be assigned any values whatsoever. If we set $z = f(x, y)$, the value of the differential is often denoted by dz as given by the formula

$$dz = \frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy. \tag{6.4–6}$$

As an immediate consequence of the definitions we can see that when f is differentiable at (x, y) the partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist at (x, y) and are given by the formulas

$$\frac{\partial f}{\partial x} = f_1(x, y) = A, \qquad \frac{\partial f}{\partial y} = f_2(x, y) = B. \tag{6.4–7}$$

Take $k = 0$, $h \neq 0$ in (6.4–4) and let $h \to 0$. The result is

$$\lim_{h\to 0} \left| \frac{f(x+h, y) - f(x, y) - Ah}{h} \right| = 0,$$

which is equivalent to

$$\lim_{h\to 0} \frac{f(x+h, y) - f(x, y)}{h} = A.$$

Therefore, by definition, A is the partial derivative of f at (x, y). The result for B is obtained in the same way. It follows that when f is differentiable at (x, y) the pair of numbers A, B satisfying (6.4–4) is unique.

It is important to observe that the requirement on f of being differentiable at a point is stronger than the requirement that f have partial derivatives with respect to x and y at the point. This fact is illustrated by the functions in Exercises 2 and 7. In each case there the function is not differentiable at $(0, 0)$ and yet has first partial derivatives with respect to x and y, respectively, there.

It will be proved later that if the first partial derivatives of f exist throughout a neighborhood of a point and are continuous at that point, then f is differentiable at the point. (See Theorem II, §7.1.) This simple criterion assures us that most of the functions we ordinarily encounter are differentiable at most points.

The following theorem is useful (in the proof of Theorem V, §7.3, for example).

THEOREM II. *If a function is differentiable at a point, it is continuous there.*

Proof. Let us define a function u of (h, k) by the formula

$$u(h, k) = \frac{f(x+h, y+k) - f(x, y) - (Ah + Bk)}{\sqrt{h^2 + k^2}} \qquad (6.4\text{–}8)$$

when $(h, k) \neq (0, 0)$. Then we see from (6.4–4) that the limit of $u(h, k)$ as $(h, k) \to (0, 0)$ is 0. From the definition of u it follows that

$$f(x+h, y+k) - f(x, y) = Ah + Bk + u(h, k)\sqrt{h^2 + k^2}$$

and from this we see at once that

$$\lim_{(h,k)\to(0,0)} [f(x+h, y+k) - f(x, y)] = 0.$$

But by §5.3 this means that f is continuous at (x, y).

It follows from the theorem that if f is *not* continuous at (x, y) it cannot be differentiable there.

A function can be continuous without being differentiable, as the following example shows.

Example 1. Let $f(x, y) = \sqrt{x^2 + y^2}$. This function is continuous at all points, $(0, 0)$ included. But it is not differentiable at $(0, 0)$. In fact, it does not even have first partial derivatives at $(0, 0)$. To verify this consider the ratio (for $h \neq 0$)

$$\frac{f(h, 0) - f(0, 0)}{h} = \frac{\sqrt{h^2} - 0}{h} = \frac{|h|}{h}.$$

This ratio is 1 if $h > 0$ and -1 if $h < 0$. Hence it has no limit as $h \to 0$, and the partial derivative $\dfrac{\partial f}{\partial x}$ does not exist at $(0, 0)$. The argument is the same as regards $\dfrac{\partial f}{\partial y}$, by symmetry.

We shall presently give an example to illustrate the property of the differential expressed in (6.4–4). For the details of the example and other problems it is useful to know the following inequalities:

$$2|ab| \leqq a^2 + b^2; \qquad (6.4\text{–}9)$$

$$\sqrt{a^2 + b^2} \leqq |a| + |b| \leqq \sqrt{2}\sqrt{a^2 + b^2}. \qquad (6.4\text{–}10)$$

The proof of (6.4–9) is simple:

$$0 \leq (|a| - |b|)^2 = |a|^2 - 2|a||b| + |b|^2 = a^2 - 2|ab| + b^2.$$

On transposing $2|ab|$ we obtain (6.4–9). The first part of (6.4–10) follows from the obvious fact that

$$a^2 + b^2 \leqq a^2 + 2|a||b| + b^2 = (|a| + |b|)^2.$$

The second part of (6.4–10) is obtained with the aid of (6.4–9):

$$(|a| + |b|)^2 = a^2 + 2|ab| + b^2 \leqq a^2 + (a^2 + b^2) + b^2 = 2(a^2 + b^2).$$

Now extract square roots at each of the ends of the foregoing inequality.

Example 2. Let $f(x, y) = 3x^2y + 2xy^2 + 1$. Find the differential at $x = 1$, $y = 2$ after verifying that the limit relation (6.4–4) holds for this case.

We use (6.4–7) to find A and B. We have

$$\frac{\partial f}{\partial x} = 6xy + 2y^2, \qquad \frac{\partial f}{\partial y} = 3x^2 + 4xy.$$

Evaluating at (1, 2), we find $f(1, 2) = 15$, $f_1(1, 2) = 20$, $f_2(1, 2) = 11$, so we form the expression

$$f(1 + h, 2 + k) - f(1, 2) - (20h + 11k)$$
$$= 3(1 + h)^2(2 + k) + 2(1 + h)(2 + k)^2 + 1 - 15 - (20h + 11k).$$

Upon expansion and simplification, we find that the expression reduces to

$$2(3h^2 + 7hk + k^2) + hk(3h + 2k).$$

Thus, for this case, the expression on the left side of (6.4–4) becomes

$$\lim_{(h,k)\to(0,0)} \frac{|2(3h^2 + 7hk + k^2) + hk(3h + 2k)|}{\sqrt{h^2 + k^2}}.$$

Now, clearly, $3h^2 + k^2 \leqq 3(h^2 + k^2)$. Also, by (6.4–9), $2|hk| \leqq h^2 + k^2$. Therefore, the fraction whose limit we are considering is not larger than

$$\frac{(h^2 + k^2)(13 + \frac{1}{2}|3h + 2k|)}{\sqrt{h^2 + k^2}}$$

which equals

$$\sqrt{h^2 + k^2}(13 + \tfrac{1}{2}|3h + 2k|).$$

This clearly approaches 0 as $(h, k) \to (0, 0)$, so we are through with the verification. The differential at (1, 2) is

$$df(1, 2; h, k) = 20h + 11k.$$

We saw in Fig. 12 of §1.3 that for the function f, when we consider the differential of f at (x, y), the relationship between dx and dy is this: as dx varies,

the point $(x + dx, y + dy)$ moves along the line tangent to the graph at (x, y). There is a similar relationship in the case of a function of two variables. Let us assume that f is continuous, and consider the surface S defined by $z = f(x, y)$. To say that f is differentiable at a particular point (x_0, y_0) can be interpreted geometrically. The implication of differentiability is that at the point (x_0, y_0, z_0), where $z_0 = f(x_0, y_0)$, the surface S has a tangent plane not parallel to the z-axis. The equation of this plane is

$$z - z_0 = A(x - x_0) + B(y - y_0) \tag{6.4–11}$$

where A and B are the values of $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$, respectively, at (x_0, y_0). To show that this plane is indeed the tangent plane, we need a definition of what is meant by saying that a plane is tangent to a surface at a certain point. Let P_0 be the point (x_0, y_0, z_0) and let M be a plane containing P_0. We define M to be the tangent plane to S at P_0 if, when P is a point of S different from P_0, the angle between the line P_0P and the plane M approaches 0 as P approaches P_0. We shall show that this condition is fulfilled by the plane (6.4–11) as a consequence of the differentiability of f at (x_0, y_0).

For this purpose we anticipate a result from Chapter 10 that may already be familiar to the reader—that the cosine of the angle θ between two vectors is equal to the dot product of the two vectors divided by the product of their lengths. See (10.2–3). We take θ to be the nonobtuse angle between the line P_0P and the normal to the plane (6.4–11) at P_0. Because θ is the complement of the angle between P_0P and the plane, we have to show that $\theta \to \frac{\pi}{2}$, or equivalently, that $\cos\theta \to 0$. Now, let P be the point on S determined by $x = x_0 + h$, $y = y_0 + k$, where h and k are small and not both zero. The vector P_0P has components h, k, Δz, where

$$\Delta z = f(x_0 + h, y_0 + k) - f(x_0, y_0). \tag{6.4–12}$$

A unit vector normal to the plane (6.4–11) has components A/d, B/d, $-1/d$, where

$$d = \pm(A^2 + B^2 + 1)^{1/2},$$

and the sign of d is chosen so that the angle between P_0P and the normal is nonobtuse. Then (using formula (10.21–3) for the dot product),

$$\cos\theta = \frac{|Ah + Bk - \Delta z|}{(A^2 + B^2 + 1)^{1/2}[h^2 + k^2 + (\Delta z)^2]^{1/2}}. \tag{6.4–13}$$

Because A and B are constants we see from (6.4–13) that we have to prove that

$$\lim_{(h,k)\to(0,0)} \frac{|Ah + Bk - \Delta z|}{[h^2 + k^2 + (\Delta z)^2]^{1/2}} = 0. \tag{6.4–14}$$

Now, considering (6.4–12), we see that the fraction in (6.4–14) is not larger than

$$\frac{|f(x_0 + h, y_0 + k) - f(x_0, y_0) - (Ah + Bk)|}{(h^2 + k^2)^{1/2}}. \tag{6.4–15}$$

But the limit of the expression in (6.4–15) is 0 as a consequence of the differentiability of f at (x_0, y_0). Therefore (6.4–14) must be true.

It can be shown, conversely, that if (6.4–14) is true, then the limit of the expression in (6.4–15) is 0, and hence f is differentiable at (x_0, y_0) if and only if the plane (6.4–11) is tangent to S at P_0. We forego the details of the proof of this converse.

Now that we know the relationship between differentiability and the tangent plane, it is easy to see how a function f can have both partial derivatives $f_1(a, b), f_2(a, b)$ existing and yet fail to be differentiable at (a, b). All that is necessary is to have a surface $z = f(x, y)$ such that it has no tangent plane at $x = a$, $y = b$, and yet such that the surface is cut by the two planes $x = a$, $y = b$ in curves which have tangents at the point in question. The cone $z = \sqrt{x^2 + y^2}$ does not have a tangent plane at its vertex (the origin). The paraboloid $z = x^2 + y^2$ has the tangent plane $z = 0$ at the origin. We can readily imagine a surface which coincides with the paraboloid where the latter intersects the planes $x = 0$, $y = 0$, and which coincides with the cone where the cone intersects the planes $y = \pm x$. Such a surface $z = f(x, y)$ will have $f_1(0, 0) = 0, f_2(0, 0) = 0$, but it will not have a tangent plane at the origin (see Fig. 42). One such surface is defined by the equation

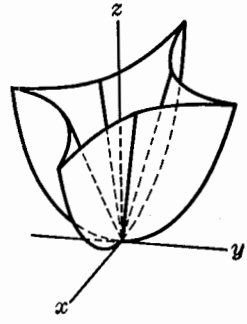


Fig. 42.

$$z = (|x| - |y|)^2 + \frac{2|xy|}{\sqrt{x^2 + y^2}} \qquad \text{(with } z = 0 \text{ when } x = y = 0\text{)}.$$

For functions of more than two variables the definitions of differentiability and differentials follow the model of the two-variable case. For a function f of the n variables $x_1, \ldots, x_n$ we want the differential of f at $(x_1, \ldots, x_n)$ to have as its value a linear combination

$$df(x_1, \ldots, x_n; h_1, \ldots, h_n) = A_1 h_1 + \cdots + A_n h_n \qquad (6.4\text{–}16)$$

of the independent variables $h_1, \ldots, h_n$, and we want this linear combination to be a suitable approximation to the difference

$$f(x_1 + h_1, \ldots, x_n + h_n) - f(x_1, \ldots, x_n)$$

when all the h_i's are sufficiently small. To express this precisely let us for convenience write

$$\|h\| = (h_1^2 + \cdots + h_n^2)^{1/2}.$$

Then $\|h\| \to 0$ means that each of the h_i's approaches 0. We now define f to be differentiable at $(x_1, \ldots, x_n)$ if there exist n numbers $A_1, \ldots, A_n$, depending on f and on the x_i's, such that

$$\lim_{\|h\| \to 0} \frac{|f(x_1 + h_1, \ldots, x_n + h_n) - f(x_1, \ldots, x_n) - (A_1 h_1 + \cdots + A_n h_n)|}{\|h\|} = 0. \qquad (6.4\text{–}17)$$

When this condition is satisfied we define the differential of f by (6.4–16).

Just as in the two-variable case, we see that, when f is differentiable at $(x_1, \ldots, x_n)$, f has first partial derivatives given by

$$\frac{\partial f}{\partial x_i} = A_i, \qquad i = 1, 2, \ldots, n. \tag{6.4-18}$$

If we use a dependent variable u to denote the value of f, then du is defined as a function of $x_1, \ldots, x_n$ and $dx_1, \ldots, dx_n$ by the formula

$$du = \frac{\partial f}{\partial u_1}\, dx_1 + \cdots + \frac{\partial f}{\partial x_n}\, dx_n. \tag{6.4-19}$$

As a matter of technique in using differentials, it is important to know the standard formulas

$$d(u + v) = du + dv,$$

$$d(uv) = u\, dv + v\, du,$$

$$d\left(\frac{u}{v}\right) = \frac{v\, du - u\, dv}{v^2}.$$

Here u and v may represent differentiable functions of several independent variables. Suppose, for example, that

$$u = f(x, y), v = g(x, y).$$

Then, by definition,

$$du = \frac{\partial u}{\partial x}\, dx + \frac{\partial u}{\partial y}\, dy,$$

$$dv = \frac{\partial v}{\partial x}\, dx + \frac{\partial v}{\partial y}\, dy,$$

$$d(uv) = \frac{\partial (uv)}{\partial x}\, dx + \frac{\partial (uv)}{\partial y}\, dy.$$

But

$$\frac{\partial (uv)}{\partial x} = u\, \frac{\partial v}{\partial x} + v\, \frac{\partial u}{\partial x},$$

with a similar formula for $\dfrac{\partial (uv)}{\partial y}$. With these relations before him the student can readily write out the work necessary to verify the relation $d(uv) = u\, dv + v\, du$.

We likewise have such formulas as

$$de^u = e^u\, du,$$

$$d \sin u = \cos u\, du,$$

$$d \tan^{-1} u = \frac{du}{1 + u^2},$$

where u is any differentiable function of several variables. Since these are all

exactly the same in appearance as the formulas of elementary calculus, we refrain from presenting a formal list.

All the formulas just considered afford us special instances of the following principle: *A differentiable function of differentiable functions is differentiable.* A fully precise statement and proof of the theorem embodying this principle will be given later (see §7.3). For the present we observe that it covers such statements as the following:

Let u = $f(x, y)$ and $v = g(x, y)$ be differentiable functions of x, y. Let $F(u, v)$ be a differentiable function of u, v, and let $w = F(f(x, y), g(x, y))$. The w is a differentiable function of x, y, and

$$dw = \frac{\partial F}{\partial u}\,du + \frac{\partial F}{\partial v}\,dv$$

when the right side of this last equation is expressed entirely in terms of x, y, dx, and dy.

A particular consequence of the principle enunciated in the previous paragraph is that the form of the relation

$$df(x, y) = \frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy$$

is valid even when x and y are not independent variables, but are differentiable functions of other variables. This invariance of form is one of the very important properties of differentials.

Example 2. If $z = \log(x^2 + y^2)$, find dz.

We use the formula

$$d \log u = \frac{du}{u}$$

with $u = x^2 + y^2$. The result is

$$dz = \frac{d(x^2 + y^2)}{x^2 + y^2} = \frac{2x\,dx + 2y\,dy}{x^2 + y^2}.$$

The coefficient of dx here is $\frac{\partial z}{\partial x}$, and that of dy is $\frac{\partial z}{\partial y}$, as the student should verify.

Example 3. Let a, b, c be the sides of a triangle, and let θ be the angle opposite the side c. Regarding c as a function of a, b, and θ, find the differential dc. Use the result to find c approximately when $a = 6.20$, $b = 5.90$, and $\theta = 58°$.

By the law of cosines we have

$$c^2 = a^2 + b^2 - 2ab \cos \theta.$$

Hence (using radian measure for θ),

$$2c\,dc = 2a\,da + 2b\,db + 2ab \sin \theta\,d\theta - 2a \cos \theta\,db - 2b \cos \theta\,da. \tag{6.4-20}$$

This equation permits us to calculate dc when a, b, θ, da, db, $d\theta$ are known. To solve the numerical problem proposed in the example we start from a triangle with $a = 6$, $b = 6$, $\theta = \pi/3$ (60°). Then $\cos\theta = \frac{1}{2}$ and $c = 6$. We are interested in Δc when $\Delta a = 0.20$, $\Delta b = -0.10$, and $\Delta\theta = -\pi/90$ radians (the equivalent of −2°). Hence we set $da = 0.20$, $db = -0.10$, $d\theta = -\pi/90$, and use dc as an approximation for Δc. We obtain from (6.4–20), after dividing by 2,

$$6dc = 1.20 - 0.60 + 36\left(\frac{\sqrt{3}}{2}\right)\left(-\frac{\pi}{90}\right) + 0.30 - 0.60,$$

$$6dc = -0.79,\ dc = -0.13.$$

Hence the new value of c is approximately $6 - 0.13 = 5.87$. This result does in fact agree with the exact value of c to two decimal places.

EXERCISES

1. Find the differential of $f(x, y)$ in each of the following cases.

(a) $x^4 + y^4 + 6xy$ **(c)** $(x^2 + y^2)/xy$

(b) $e^x y$ **(d)** $(x^2 + y^2)^{3/2}$

2. Suppose $f(x, y) = \dfrac{xy}{x^2 + y^2}$ for all (x, y) except $(0, 0)$, and let $f(0, 0) = 0$. Show that $f_1(0, 0)$ and $f_2(0, 0)$ both exist with values equal to 0 but that f is discontinuous at $(0, 0)$.

3. (**a**) Show that

$$x\,d\left(\frac{x}{\sqrt{x^2 + y^2}}\right) + y\,d\left(\frac{y}{\sqrt{x^2 + y^2}}\right) = 0$$

for all values of x and y such that $x^2 + y^2 > 0$, and for all values of dx and dy. (**b**) Generalize (**a**) by proving that $\Sigma_{k=1}^{n} x_k d(x_k/r) = 0$ if $r = (x_1^2 + \cdots + x_n^2)^{1/2}$ and $x_1, \ldots, x_n$ are independent variables.

4. Find dz at $x = 1$, $y = \pi/2$ if $z = x^2 y + e^x \sin y$. What is the resulting value of dz if $dx = \pi - e$, $dy = e^2$?

5. If $f(x, y) = (50 - x^2 - y^2)^{1/2}$, find an approximate value of the difference $f(3, 4) - f(2.9, 4.1)$ by use of differentials.

6. If u and v are differentiable functions of x and y, prove the formula

$$d\left(\frac{u}{v}\right) = \frac{v\,du - u\,dv}{v^2},$$

assuming that $v \neq 0$.

7. Let $f(x, y) = \sqrt{|xy|}$. (**a**) Verify that $f_1(0, 0) = f_2(0, 0) = 0$. (**b**) Does the surface $z = f(x, y)$ have a tangent plane at $x = 0$, $y = 0$? Consider the section of the surface made by the plane $x = y$. (**c**) Verify that for this case the requirement (6.4–4), with $x = y = 0$, cannot be satisfied, no matter how A and B are chosen.

8. Assuming that f_1 and f_2 both exist at (x, y), prove that if

$$\lim_{(h,k)\to(0,0)} \frac{f(x + h, y + k) - f(x, y) - f_1(x, y)h - f_2(x, y)k}{\sqrt{h^2 + k^2}}$$

exists, the limit is 0.

6.5 / COMPOSITE FUNCTIONS AND THE CHAIN RULE

We use the term *composite function* for a function which is obtained from a given function of one or more variables by substituting other functions in place of the variables in the first function. As an example, consider the function

$$F(x, y) = \sqrt{x^2 + y^2}.$$

It is a composite function which may be built up as follows: Let

$$f(u) = u^{1/2}, \phi(x, y) = x^2 + y^2.$$

On replacing u by $\phi(x, y)$ we get

$$f(\phi(x, y)) = F(x, y).$$

As another example, consider the function

$$f(x, y) = xy + \log \frac{y}{x}.$$

In this function let us replace x and y by certain functions of new variables r, θ, as follows:

$$x = r \cos \theta, y = r \sin \theta.$$

The result is a function of r and θ:

$$F(r, \theta) = r^2 \sin \theta \cos \theta + \log \tan \theta.$$

Note that we may also write

$$F(r, \theta) = f(r \cos \theta, r \sin \theta).$$

Most of the functions we deal with are built up as composite functions. We meet the concept of a composite function very early in the study of elementary differential calculus; it is there that we learn the very important rule of differentiation embodied in Theorem II, §1.11. We are now going to be concerned with extensions of this rule for functions of several variables.

Let us first consider a function of three variables, say

$$u = F(x, y, z). \tag{6.5-1}$$

We are going to suppose that the variables x, y, z are made to depend upon a variable t; let the notation for this dependence be

$$x = f(t), y = g(t), z = h(t). \tag{6.5-2}$$

On substituting these functions for x, y, z in the function $u = F(x, y, z)$, we obtain u as a composite function of t; if we denote this function by $G(t)$, we have

$$u = F(f(t), g(t), h(t)) = G(t). \tag{6.5-3}$$

The formula of differentiation for this composite function is

$$\frac{du}{dt} = \frac{\partial u}{\partial x}\frac{dx}{dt} + \frac{\partial u}{\partial y}\frac{dy}{dt} + \frac{\partial u}{\partial z}\frac{dz}{dt}. \tag{6.5-4}$$

We shall presently prove this formula under suitable hypotheses. It is very important for the student to learn the structure of this formula; as a part of such learning, he must grasp clearly the role of the variables in the notation of each term in (6.5–4). In the term $\frac{\partial u}{\partial x}$, u and x are related by (6.5–1); u is dependent, and x is one of the three independent variables x, y, z. In the term $\frac{dx}{dt}$, x and t are related by the first of equations (6.5–2); x is dependent and t is independent. In the term $\frac{du}{dt}$, u and t are related by (6.5–3); u is dependent and t is independent. An alternative notation for (6.5–4) is

$$\frac{dG}{dt} = \frac{\partial F}{\partial x}\frac{df}{dt} + \frac{\partial F}{\partial y}\frac{dg}{dt} + \frac{\partial F}{\partial z}\frac{dh}{dt}. \tag{6.5–5}$$

This notation is clearer and less subject to misunderstanding. However, both methods of writing the formula are widely used, and the student should become familiar with both of them.

There are other varieties of composite functions in addition to the type presented in (6.5–3). The function F may depend on a different number of variables (e.g., 2, or 4, 5, . . .). Also, the variables x, y, z may depend upon more than one variable. Suppose, for example, that we have

$$\begin{aligned} u &= F(x, y), \\ x &= f(s, t),\ y = g(s, t). \end{aligned} \tag{6.5–6}$$

Then, under suitable differentiability assumptions, if we write $G(s, t) = F(f(s, t), g(s, t))$, we have the differentiation formulas

$$\begin{aligned} \frac{\partial G}{\partial s} &= \frac{\partial F}{\partial x}\frac{\partial f}{\partial s} + \frac{\partial F}{\partial y}\frac{\partial g}{\partial s}, \\ \frac{\partial G}{\partial t} &= \frac{\partial F}{\partial x}\frac{\partial f}{\partial t} + \frac{\partial F}{\partial y}\frac{\partial g}{\partial t}. \end{aligned} \tag{6.5–7}$$

The general rule covering all formulas such as (6.5–5) and (6.5–7) is often called the *chain rule*, or the *composite-function rule*.

To describe the situation generally, let us use the term *first-class variables* for the independent variables on which F depends, and the term *second-class variables* for the variables on which G depends. Observe that G is formed by replacing each first-class variable in F by a function of the second-class variables. In the differentiation formulas such as (6.5–7) we have as many different formulas as there are variables of the *second* class; each formula has as many terms as there are variables of the *first* class.

We shall now formulate and prove a theorem about formulas such as (6.5–5) or (6.5–7). For simplicity we assume the situation is that represented in (6.5–6).

THEOREM III. *Let $F(x, y)$ be defined in some region R of the xy-plane having*

the point (a, b) as an interior point, and let F be differentiable at (a, b). Let $f(s, t)$, $g(s, t)$ be defined in some neighborhood of the point (s_0, t_0) in the st-plane. Let these functions admit first partial derivatives at (s_0, t_0), and suppose further that

$$a = f(s_0, t_0),\ b = g(s_0, t_0). \tag{6.5–8}$$

Then the composite function $G(s, t) = F(f(s, t), g(s, t))$ has first partial derivatives at (s_0, t_0), given by formulas (6.5–7), where $\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}$ are evaluated at (a, b), and the partial derivatives with respect to s, t are evaluated at s_0, t_0.

Proof. It is to be emphasized that, although we are assuming that F is differentiable at (a, b), we are *not* drawing a conclusion about differentiability of G. Nor are we assuming differentiability of f and g. We deal merely with partial derivatives of f, g, and G. This form of the chain rule is therefore a little different from the chain rule for differentials given in Theorem V, §7.3.

The proof begins with use of the fact that F is differentiable at (a, b). This means that the ratio

$$\frac{F(a + \Delta x, b + \Delta y) - F(a, b) - F_1(a, b)\Delta x - F_2(a, b)\Delta y}{\sqrt{(\Delta x)^2 + (\Delta y)^2}} \tag{6.5–9}$$

approaches 0 as $\Delta x \to 0$ and $\Delta y \to 0$. Here we assume that Δx and Δy are small, but not both 0. Let us define ϵ (depending on Δx and Δy) to be the value of the fraction in (6.5–9) if the denominator is not 0; if $\Delta x = \Delta y = 0$ we define $\epsilon = 0$. Then we can write

$$F(a + \Delta x, b + \Delta y) - F(a, b) = F_1(a, b)\Delta x + F_2(a, b)\Delta y + \epsilon\sqrt{(\Delta x)^2 + (\Delta y)^2}. \tag{6.5–10}$$

This is true even if $\Delta x = \Delta y = 0$. It is clear from the definition of ϵ that $\epsilon \to 0$ as $\Delta x \to 0$ and $\Delta y \to 0$.

Now we undertake to prove the first formula in (6.5–7). The second formula is proved in the same way. For small $\Delta s (\neq 0)$ let us write

$$\begin{aligned} \Delta x &= f(s_0 + \Delta s, t_0) - f(s_0, t_0), \\ \Delta y &= g(s_0 + \Delta s, t_0) - g(s_0, t_0), \\ \Delta u &= F(a + \Delta x, b + \Delta y) - F(a, b). \end{aligned} \tag{6.5–11}$$

Then it follows from (6.5–8) and the definition of G that

$$\Delta u = G(s_0 + \Delta s, t_0) - G(s_0, t_0),$$

so that, by definition,

$$\lim_{\Delta s \to 0} \frac{\Delta u}{\Delta s} = G_1(s_0, t_0)$$

if the limit exists. Now, from (6.5–10) and the third formula in (6.5–11) we see

that

$$\frac{\Delta u}{\Delta s} = F_1(a, b)\frac{\Delta x}{\Delta s} + F_2(a, b)\frac{\Delta y}{\Delta s} + \epsilon \frac{\sqrt{(\Delta x)^2 + (\Delta y)^2}}{\Delta s}.$$

We are now ready to take limits on both sides of this equation as $\Delta s \to 0$. On the right we have

$$\frac{\Delta x}{\Delta s} \to f_1(s_0, t_0), \frac{\Delta y}{\Delta s} \to g_1(s_0, t_0).$$

Also $\epsilon \to 0$ because $\Delta x \to 0$ and $\Delta y \to 0$. [The fact that $\Delta x \to 0$ is a consequence of the continuity of $f(s, t_0)$ as a function of s alone, at $s = s_0$; see Theorem I, §1.11. The reasoning for $\Delta y \to 0$ is the same.] Finally,

$$\left|\frac{\sqrt{(\Delta x)^2 + (\Delta y)^2}}{\Delta s}\right| = \left[\left(\frac{\Delta x}{\Delta s}\right)^2 + \left(\frac{\Delta y}{\Delta s}\right)^2\right]^{1/2}$$

has a definite limit. Thus we see that

$$G_1(s_0, t_0) = F_1(a, b)f_1(s_0, t_0) + F_2(a, b)g_1(s_0, t_0).$$

This is the same as the first formula in (6.5–7) when the partial derivatives are expressed in the appropriate notations to show the values at (a, b) and (s_0, t_0).

Example 1. If $F(x, y)$ is a given differentiable function of x, y, and if we introduce polar co-ordinates r, θ by writing $x = r\cos\theta$, $y = r\sin\theta$, then F is transformed into a function G of r and θ:

$$G(r, \theta) = F(r\cos\theta, r\sin\theta). \tag{6.5–12}$$

Use the chain rule to find $\frac{\partial G}{\partial r}$ and $\frac{\partial G}{\partial \theta}$ in terms of $\frac{\partial F}{\partial x}$ and $\frac{\partial F}{\partial y}$.

Here the first-class variables are x, y and the second-class variables are r, θ. We have

$$\frac{\partial x}{\partial r} = \cos\theta, \qquad \frac{\partial y}{\partial r} = \sin\theta,$$

$$\frac{\partial x}{\partial \theta} = -r\sin\theta, \quad \frac{\partial y}{\partial \theta} = r\cos\theta.$$

The chain rule, in the form (6.5–7) with r, θ replacing s, t, then gives

$$\begin{aligned} \frac{\partial G}{\partial r} &= \frac{\partial F}{\partial x}\cos\theta + \frac{\partial F}{\partial y}\sin\theta, \\ \frac{\partial G}{\partial \theta} &= -\frac{\partial F}{\partial x} r\sin\theta + \frac{\partial F}{\partial y} r\cos\theta. \end{aligned} \tag{6.5–13}$$

To emphasize the meaning of these formulas, let us evaluate each part of them for $r = 1$, $\theta = \pi/6$. The corresponding values of x, y are $x = \sqrt{3}/2$, $y = \frac{1}{2}$. Then,

using the subscript notation for partial derivatives, equations (6.5–13) become

$$G_1\left(1, \frac{\pi}{6}\right) = F_1\left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)\frac{\sqrt{3}}{2} + F_2\left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)\frac{1}{2},$$
$$G_2\left(1, \frac{\pi}{6}\right) = -F_1\left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)\frac{1}{2} + F_2\left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)\frac{\sqrt{3}}{2}. \tag{6.5–14}$$

Thorough familiarity with the chain rule is of great importance to any student who wants to read books and journals in which mathematical methods are used extensively. Adequate training in the use of the chain rule must include, among other things, stress on the idea of transforming a given function of one set of variables into a new function of another set of variables, with a resulting set of formulas relating the two sets of partial derivatives of these functions. Ideas of this sort are constantly used in the theory of partial differential equations, and such ideas are at the very root of tensor analysis, which is an important subject beyond the scope of this book. Tensor analysis is an outgrowth and generalization of vector analysis. A brief discussion of some of the fundamental concepts of Euclidean vector analysis is given in Chapter 10. The chain rule plays an important part in the discussion of such concepts as *gradient*, *divergence*, and *curl* in vector analysis.

For an example of a physical problem where the chain rule is involved see §6.51.

Matters of notation play a considerable role in connection with the chain rule. Wide varieties of usage exist in mathematical writing where the chain rule is concerned. Consider the situation expressed in (6.5–6). Instead of writing the chain rule here in the form (6.5–7), it is frequently written

$$\frac{\partial u}{\partial s} = \frac{\partial u}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial s}, \tag{6.5–15}$$

with a similar formula in which t replaces s. The difference between (6.5–7) and (6.5–15) is that in (6.5–15) we have used *dependent variables* in place of *functional symbols*. The student must bear in mind some of the subtleties of distinction between a dependent variable and a function. If we write $u = F(x, y)$, u denotes the value of F at the point (x, y). If $x = f(s, t)$ and $y = g(s, t)$, and if $F(x, y)$ is thereby transformed into $G(s, t)$, we may also write $u = G(s, t)$, since the value of G at (s, t) is the same as the value of F at (x, y) when x and y are the values of f and g, respectively, at (s, t). But though we may write $u = F(x, y) = G(s, t)$, this does not mean that F and G are the same function; in general they are not.

Example 2. If F is a differentiable function of two variables, and

$$u = F(s^2 - t^2, t^2 - s^2),$$

show that

$$t\frac{\partial u}{\partial s} + s\frac{\partial u}{\partial t} = 0. \tag{6.5–16}$$

One method of handling this problem is to introduce new variables

$$x = s^2 - t^2,\ y = t^2 - s^2,$$

so that $u = F(x, y)$. Then, by the chain rule,

$$\frac{\partial u}{\partial s} = \frac{\partial u}{\partial x}(2s) + \frac{\partial u}{\partial y}(-2s),$$
$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x}(-2t) + \frac{\partial u}{\partial y}(2t).$$

The correctness of (6.5–16) is now readily deduced.

It sometimes happens that a problem involves several variables and several relations between them. Consider, for example, the four variables V, S, r, h connected by the two equations

$$V = \pi r^2 h,\ S = 2\pi rh + 2\pi r^2. \tag{6.5–17}$$

These formulas arise if we consider a right circular cylinder of height h, radius of base r, volume V, and total surface area S. Of the four variables, just two are independent. Ordinarily we most naturally think of r and h as independent, but other choices are legitimate. We may choose r and S as independent. Then

$$h = \frac{S}{2\pi r} - r,\ V = \tfrac{1}{2}rS - \pi r^3. \tag{6.5–18}$$

In view of what has been said, it is evident that a notation such as $\frac{\partial V}{\partial r}$ is ambiguous, for if we calculate from (6.5–17) we have $\frac{\partial V}{\partial r} = 2\pi rh$, while if we calculate from (6.5–18) we have $\frac{\partial V}{\partial r} = \frac{1}{2}S - 3\pi r^2$, and these two results are not in agreement. What is needed is a notation that makes clear the choice of the independent variables. A customary notation employs subscripts. According to this practice,

$$\left(\frac{\partial V}{\partial r}\right)_h$$

indicates that we are regarding V as a function of the independent variables r, h, with h held constant in the differentiation. With this notation we have

$$\left(\frac{\partial V}{\partial h}\right)_r = \pi r^2,\ \left(\frac{\partial V}{\partial S}\right)_r = \tfrac{1}{2}r,\ \left(\frac{\partial S}{\partial h}\right)_r = 2\pi r,$$

and so on.

Situations like this arise in thermodynamics, for example, with different variables, of course, and different relations between them

EXERCISES

In all these exercises it is assumed without explicit reference that all the functions introduced are differentiable.

1. If w is a function of p, q, r and each of these latter variables is a function of s, write the chain-rule formula for $\dfrac{dw}{ds}$.

2. If $u = x^2 - y^2$ and $v = 2xy$ transform $F(u, v)$ into $G(x, y)$, find $\dfrac{\partial G}{\partial x}$ and $\dfrac{\partial G}{\partial y}$ in terms of $\dfrac{\partial F}{\partial u}$ and $\dfrac{\partial F}{\partial v}$.

3. Define the composite function G and write the formula or formulas of the chain rule in each of the following situations. Identify the first- and second-class variables in each case.

(a) $u = F(x, y, z), x = f(p, q), y = g(p, q), z = h(p, q)$.

(b) $u = F(w), w = \phi(x, y)$.

(c) $u = F(x, y, z, \theta), x = f(t), y = g(t), z = h(t), \theta = t$.

4. If $x = u$, $y = u + v$ and $z = u + v + w$ transform $F(x, y, z)$ into $G(u, v, w)$, find each of the first partial derivatives of G in terms of the first partial derivatives of F.

5. Suppose u depends on x_1, x_2, x_3 and the x's depend on ξ_1, ξ_2 as follows:

$$x_1 = a_{11}\xi_1 + a_{12}\xi_2$$
$$x_2 = a_{21}\xi_1 + a_{22}\xi_2$$
$$x_3 = a_{31}\xi_1 + a_{32}\xi_2$$

where the a's are constants. Write the formulas connecting $\dfrac{\partial u}{\partial \xi_1}$ and $\dfrac{\partial u}{\partial \xi_2}$ with $\dfrac{\partial u}{\partial x_1}$, $\dfrac{\partial u}{\partial x_2}$, and $\dfrac{\partial u}{\partial x_3}$. Generalize for the case of nx's and $m\xi$'s.

6. If $w = F(a + \alpha t, b + \beta t, c + \gamma t)$, where a, b, c and α, β, γ are constants, write a formula for $\dfrac{dw}{dt}$ involving $F_1(a + \alpha t, b + \beta t, c + \gamma t)$ and other similar expressions.

7. If $u = f(x - y, y - x)$, prove that $\dfrac{\partial u}{\partial x} + \dfrac{\partial u}{\partial y} = 0$.

8. If $u = F\left(\dfrac{y - x}{xy}, \dfrac{z - x}{xz}\right)$, prove that $x^2 \dfrac{\partial u}{\partial x} + y^2 \dfrac{\partial u}{\partial y} + z^2 \dfrac{\partial u}{\partial z} = 0$.

9. If $u = x^3 F\left(\dfrac{y}{x}, \dfrac{z}{x}\right)$, prove that $x \dfrac{\partial u}{\partial x} + y \dfrac{\partial u}{\partial y} + z \dfrac{\partial u}{\partial z} = 3u$.

10. Let $f(x, y)$ be transformed into $g(u, v)$ by $x = u \cos \theta - v \sin \theta$, $y = u \sin \theta + v \cos \theta$, where θ is constant. Show that

$$\left(\frac{\partial g}{\partial u}\right)^2 + \left(\frac{\partial g}{\partial v}\right)^2 = \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2$$

is an identity in u and v.

11. If $w = f(x, y)$ and $x = r \cosh \theta$, $y = r \sinh \theta$, find $\left(\dfrac{\partial w}{\partial r}\right)^2 - \dfrac{1}{r^2}\left(\dfrac{\partial w}{\partial \theta}\right)^2$ in terms of $\dfrac{\partial w}{\partial x}$ and $\dfrac{\partial w}{\partial y}$.

12. Let $F(x, t) = f(x + 2t) + f(3x - 2t)$. Set $u = x + 2t$, $v = 3x - 2t$, and show by the chain rule that

$$F_1(x, t) = f'(u) + 3f'(v),$$
$$F_2(x, t) = 2f'(u) - 2f'(v).$$

Hence show that $F_1(0, 0) = 4f'(0)$, $F_2(0, 0) = 0$.

13. Let $z = yf(x^2 - y^2)$. Show that $y\dfrac{\partial z}{\partial x} + x\dfrac{\partial z}{\partial y} = \dfrac{xz}{y}$.

14. If $z = xy + xF\left(\dfrac{y}{x}\right)$, show that $x\dfrac{\partial z}{\partial x} + y\dfrac{\partial z}{\partial y} = xy + z$.

15. Suppose u is a function of r and $r = (x^2 + y^2 + z^2)^{1/2}$. Show that $\left(\dfrac{\partial u}{\partial x}\right)^2 + \left(\dfrac{\partial u}{\partial y}\right)^2 + \left(\dfrac{\partial u}{\partial z}\right)^2 = \left(\dfrac{du}{dr}\right)^2$.

16. Consider the determinant

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

Let A_{ij} be the cofactor (appropriately signed minor) of a_{ij} in D. If the nine elements of D are regarded as independent variables, show that $\dfrac{\partial D}{\partial a_{ij}} = A_{ij}$. Now suppose that the elements a_{ij} are differentiable functions of x, and let $a'_{ij} = \dfrac{d}{dx}(a_{ij})$. Show by the chain rule that

$$\frac{dD}{dx} = \begin{vmatrix} a'_{11} & a'_{12} & a'_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a'_{21} & a'_{22} & a'_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a'_{31} & a'_{32} & a'_{33} \end{vmatrix}.$$

17. Generalize the results of Exercise 16 for determinants of order n. State the final result as a verbal rule.

18. If

$$F(x, y, z) = \begin{vmatrix} f_1(x) & f_2(x) & f_3(x) \\ f_1(y) & f_2(y) & f_3(y) \\ f_1(z) & f_2(z) & f_3(z) \end{vmatrix},$$

calculate $\dfrac{\partial F}{\partial x}$, $\dfrac{\partial F}{\partial y}$, and $\dfrac{\partial F}{\partial z}$ by the results of Exercise 16.

19. Write out in detail and prove the version of Theorem III that corresponds to (6.5–1), (6.5–2), and (6.5–5).

20. In Example 2 let $G(s, t) = F(s^2 - t^2, t^2 - s^2)$. Formula (6.5–16) now becomes $tG_1(s, t) + sG_2(s, t) = 0$. Rewrite the chain-rule formulas in the solution of Example 2 without using the letters x, y, u.

21. From the situation expressed by (6.5–17) find $\left(\dfrac{\partial S}{\partial V}\right)_h$, $\left(\dfrac{\partial h}{\partial V}\right)_r$, $\left(\dfrac{\partial S}{\partial V}\right)_r$, $\left(\dfrac{\partial h}{\partial r}\right)_V$ and $\left(\dfrac{\partial r}{\partial S}\right)_h$.

22. **(a)** If $u = x^2 + y^2 + z^2$ and $z = xyv$, how many meanings are there for $\dfrac{\partial u}{\partial y}$? **(b)** Find $\dfrac{\partial u}{\partial y}$ in each of its meanings, indicating the meaning in each case by proper subscripts.

23. **(a)** If $V = xyz$ and $S = xy + 2xz + 2yz$, find $\dfrac{\partial V}{\partial x}$ in each of its possible meanings,

indicating the meaning in each case by proper subscripts. (b) Find

$$\left(\frac{\partial S}{\partial x}\right)_{y,V}, \left(\frac{\partial S}{\partial V}\right)_{x,y} \text{ and } \left(\frac{\partial V}{\partial S}\right)_{x,y}.$$

24. If $u = F(x, y)$ and $y = f(x)$, show that $\dfrac{du}{dx} = \dfrac{\partial u}{\partial x} + \dfrac{\partial u}{\partial y}\dfrac{dy}{dx}$. Explain carefully the difference between $\dfrac{\partial u}{\partial x}$ and $\dfrac{du}{dx}$.

25. If $G(x, y) = F(x, y, f(x, y))$, show that $G_1(x, y) = F_1(x, y, z) + F_3(x, y, z)f_1(x, y)$, where $z = f(x, y)$. Classify the variables and show that this is an instance of the chain rule. Write the corresponding formula for $G_2(x, y)$.

6.51 / AN APPLICATION IN FLUID KINEMATICS

A very good illustration of the occurrence of formulas such as (6.5–4) or (6.5–5) in physics is furnished in the very beginning of the study of the flow of a fluid.

To build up a mathematical model of an arbitrary fluid in motion, let us fix our attention upon a certain portion of the fluid at a certain instant of time, say the fluid occupying a region R_0 at the time $t = 0$. Then, selecting an arbitrary particle of the fluid in R_0, we follow its motion as t increases. Each particle of the fluid will occupy a certain point of space at a certain time. At a general instant t the portion of the fluid which occupied R_0 at $t = 0$ will occupy a new region R; the particle which was at the point (x_0, y_0, z_0) of R_0 will have moved to a point (x, y, z) of R (see Fig. 43). We assume that the law of motion of the fluid is a definite (but perhaps very complicated) law, so that the co-ordinates (x, y, z) of the particle in R are determined by (i.e., are functions of) x_0, y_0, z_0, t, say

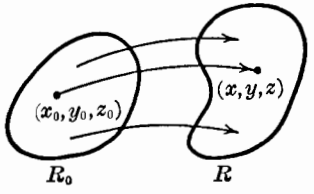


Fig. 43.

$$\begin{aligned} x &= f(x_0, y_0, z_0, t), \\ y &= g(x_0, y_0, z_0, t), \\ z &= h(x_0, y_0, z_0, t). \end{aligned} \tag{6.51–1}$$

These equations define the flow of the fluid for time subsequent to $t = 0$. In what follows we shall regard the functions f, g, h as known without specifying them in any particular manner.

The velocity of an individual particle at a given instant has components

$$\frac{\partial x}{\partial t}, \frac{\partial y}{\partial t}, \frac{\partial z}{\partial t}.$$

The velocity is a vector quantity; it is tangent to the path which is being followed by the particle under consideration.

Now let us consider the density of the fluid. If the mass of a substance is not distributed uniformly throughout its volume, it has variable density, and we must speak of the density *at a point*. Let us then consider the fluid in the region R at

time t; let ρ be the density at the point x, y, z. Then ρ is a function of x, y, z. Conceivably ρ might be constant, but it need not be so. (The fluid might be the atmosphere, and R might be a region extending from sea level up to a height of 30,000 feet.) It is important to note that ρ does not depend merely on x, y, z. It is also a function of t, for the fluid which occupies a given region at time t need not have the same mass distribution as the fluid which occupies that region at some other time. Let us write

$$\rho = F(x, y, z, t). \tag{6.51–2}$$

If in this equation we keep t fixed and vary the point (x, y, z), we obtain the density of the fluid in R at time t. On the other hand, if we fix the point (x, y, z) and vary t, then we obtain the density at (x, y, z) of the fluid occupying R at different times.

There is still another important way of examining the variability of ρ. Imagine an observer moving along with a particle of the fluid. If he were capable of measuring ρ at his own position at any time, by what equation would he describe his observations? For him ρ is the composite function obtained by substituting x, y, z from (6.51–1) in (6.51–2). The result is that ρ is a function of x_0, y_0, z_0, t. Since the observer stays with the same particle always, x_0, y_0, z_0 are to be considered as constants, and ρ becomes a function of t alone, say

$$\rho = G(t). \tag{6.51–3}$$

Now let us ask the question: From the point of view of the observer moving with the particle, what is the rate of change of ρ? The answer is, of course, the value of the derivative $G'(t)$. Since G is a composite function we may express $G'(t)$ by the chain rule. We see in (6.51–2) that F depends on four variables, x, y, z, t. These are *first-class* variables. By equations (6.51–1) the variables x, y, z, t. depend on other variables x_0, y_0, z_0, t. These are *second-class* variables. Note that t is both a first-class variable and a second-class variable. One way out of the perplexity which the student may at first experience in this situation may be found by an artifice of notation. Let us temporarily introduce a different letter, say θ, to stand for t in the set of first-class variables. We then write $\rho = F(x, y, z, \theta)$ and add the equation $\theta = t$ to the set (6.51–1). The chain rule then gives

$$\frac{dG}{dt} = \frac{\partial F}{\partial x}\frac{\partial f}{\partial t} + \frac{\partial F}{\partial y}\frac{\partial g}{\partial t} + \frac{\partial F}{\partial z}\frac{\partial h}{\partial t} + \frac{\partial F}{\partial \theta}\frac{\partial \theta}{\partial t}.$$

Having obtained the formula, we may forget the artifice, and write t again in place of θ. Note that $\frac{\partial \theta}{\partial t} = 1$. It is more usual, in the literature of this subject, to write the letters ρ, x, y, z instead of the functional symbols for these variables. When this is done we have

$$\frac{d\rho}{dt} = \frac{\partial \rho}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial \rho}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial \rho}{\partial z}\frac{\partial z}{\partial t} + \frac{\partial \rho}{\partial t}. \tag{6.51–4}$$

The derivative of ρ with respect to t occurs here in two different forms, with different meanings.

EXERCISES

1. Let the fluid flow be the movement of the atmosphere. From the point of view of an observer standing on a street corner with the wind rushing by, what is the expression for the rate of change of density of air on the street corner? If the observer were to ride in the gondola of a freely-drifting balloon, what would be the expression for the rate of change of density of air at the gondola?

2. Suppose the atmosphere over a great plain is in a state such that at any given time ρ is a function only of altitude above the plain. Suppose that a free balloon is drifting along at a constant elevation of 500 feet. Explain why an observer carried by the balloon finds that $\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t}$.

3. Suppose that, in (6.51–1), z is independent of t, and that

$$\rho = H(x^2 + y^2 + z^2, t).$$

Show that $\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t}$ at all points on the z-axis.

6.52 / SECOND DERIVATIVES BY THE CHAIN RULE

The purpose of this section is to illustrate by examples the use of the chain rule in dealing with second derivatives.

Example 1. If $F(x, y)$ is transformed into $G(r, \theta)$ by the equations $x = r\cos\theta$, $y = r\sin\theta$, find $\frac{\partial^2 G}{\partial r\,\partial\theta}$ in terms of derivatives of F with respect to x and y.

We start from the work already done in Example 1, §6.5. From (6.5–13) we have

$$\frac{\partial G}{\partial \theta} = -\frac{\partial F}{\partial x} r\sin\theta + \frac{\partial F}{\partial y} r\cos\theta.$$

We now differentiate both sides of this equation with respect to r, bearing in mind that $\frac{\partial F}{\partial x}$ and $\frac{\partial F}{\partial y}$ depend on x and y, and are thus to be regarded as composite functions of r and θ. We have

$$\begin{aligned}\frac{\partial^2 G}{\partial r\,\partial\theta} = &-\frac{\partial F}{\partial x}\frac{\partial}{\partial r}(r\sin\theta) - \frac{\partial}{\partial r}\left(\frac{\partial F}{\partial x}\right) r\sin\theta \\ &+ \frac{\partial F}{\partial y}\frac{\partial}{\partial r}(r\cos\theta) + \frac{\partial}{\partial r}\left(\frac{\partial F}{\partial y}\right) r\cos\theta.\end{aligned} \tag{6.52–1}$$

The problem now facing us is that of doing something further with expressions like $\frac{\partial}{\partial r}\left(\frac{\partial F}{\partial x}\right)$. Now $\frac{\partial F}{\partial x} = F_1(x, y)$. What we really have before us is the problem

of finding $\frac{\partial H}{\partial r}$, where $H(r, \theta) = F_1(r \cos \theta, r \sin \theta)$. This is a problem exactly like that of finding $\frac{\partial G}{\partial r}$, except that we have F_1 in place of F, and H in place of G. Thus, just as in (6.5–13),

$$\frac{\partial H}{\partial r} = \frac{\partial F_1}{\partial x} \cos \theta + \frac{\partial F_1}{\partial y} \sin \theta.$$

This may be written

$$\frac{\partial}{\partial r}\left(\frac{\partial F}{\partial x}\right) = \frac{\partial^2 F}{\partial x^2} \cos \theta + \frac{\partial^2 F}{\partial y\, \partial x} \sin \theta. \tag{6.52–2}$$

In a similar way we find

$$\frac{\partial}{\partial r}\left(\frac{\partial F}{\partial y}\right) = \frac{\partial^2 F}{\partial x\, \partial y} \cos \theta + \frac{\partial^2 F}{\partial y^2} \sin \theta. \tag{6.52–3}$$

If we now use (6.52–2) and (6.52–3) in (6.52–1), we obtain

$$\begin{aligned} \frac{\partial^2 G}{\partial r\, \partial \theta} = &- \frac{\partial F}{\partial x} \sin \theta - \left(\frac{\partial^2 F}{\partial x^2} \cos \theta + \frac{\partial^2 F}{\partial y\, \partial x} \sin \theta\right) r \sin \theta \\ &+ \frac{\partial F}{\partial y} \cos \theta + \left(\frac{\partial^2 F}{\partial x\, \partial y} \cos \theta + \frac{\partial^2 F}{\partial y^2} \sin \theta\right) r \cos \theta. \end{aligned} \tag{6.52–4}$$

In order to get this far we naturally assume that F and its partial derivatives F_1, F_2 are differentiable functions of x, y. Under this assumption it is true that $\frac{\partial^2 F}{\partial y\, \partial x} = \frac{\partial^2 F}{\partial x\, \partial y}$ (see Theorem IV, §7.2). Therefore (6.52–4) may be written in a slightly more compact form. If we write $u = F(x, y) = G(r, \theta)$, (6.52–4) becomes

$$\begin{aligned} \frac{\partial^2 u}{\partial r\, \partial \theta} = &- r \sin \theta \cos \theta \frac{\partial^2 u}{\partial x^2} + r(\cos^2 \theta - \sin^2 \theta) \frac{\partial^2 u}{\partial x\, \partial y} \\ &+ r \sin \theta \cos \theta \frac{\partial^2 u}{\partial y^2} - \sin \theta \frac{\partial u}{\partial x} + \cos \theta \frac{\partial u}{\partial y}. \end{aligned} \tag{6.52–5}$$

The work of finding formulas for $\frac{\partial^2 u}{\partial r^2}$ and $\frac{\partial^2 u}{\partial \theta^2}$ is similar.

Example 2. If $u = F(x, y)$ becomes $G(s, t)$ when we set

$$x = s + t,\ y = s - t,$$

find what $\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2}$ becomes in terms of derivatives with respect to s and t.

For this purpose we express s and t in terms of x and y; we regard s, t as first-class variables and x, y as second-class variables. We have

$$s = \tfrac{1}{2}(x + y),\ t = \tfrac{1}{2}(x - y).$$

Then

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial s}\frac{\partial s}{\partial x} + \frac{\partial u}{\partial t}\frac{\partial t}{\partial x} = \frac{1}{2}\frac{\partial u}{\partial s} + \frac{1}{2}\frac{\partial u}{\partial t}. \tag{6.52–6}$$

Likewise,

$$\frac{\partial u}{\partial y} = \frac{1}{2}\frac{\partial u}{\partial s} - \frac{1}{2}\frac{\partial u}{\partial t}. \tag{6.52–7}$$

Differentiating again,

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{2}\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial s}\right) + \frac{1}{2}\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial t}\right).$$

Now, by (6.52–6), with $\frac{\partial u}{\partial s}$ in place of u,

$$\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial s}\right) = \frac{1}{2}\frac{\partial}{\partial s}\left(\frac{\partial u}{\partial s}\right) + \frac{1}{2}\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial s}\right).$$

We find $\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial t}\right)$ in the same way. Thus

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{4}\frac{\partial^2 u}{\partial s^2} + \frac{1}{4}\frac{\partial^2 u}{\partial t\,\partial s} + \frac{1}{4}\frac{\partial^2 u}{\partial s\,\partial t} + \frac{1}{4}\frac{\partial^2 u}{\partial t^2}. \tag{6.52–8}$$

We shall assume enough about u to insure that $\frac{\partial^2 u}{\partial t\,\partial s} = \frac{\partial^2 u}{\partial s\,\partial t}$. The student should show for himself that

$$\frac{\partial^2 u}{\partial y^2} = \frac{1}{4}\frac{\partial^2 u}{\partial s^2} - \frac{1}{2}\frac{\partial^2 u}{\partial t\,\partial s} + \frac{1}{4}\frac{\partial^2 u}{\partial t^2}. \tag{6.52–9}$$

Subtraction of (6.52–9) from (6.52–8) now gives the result

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = \frac{\partial^2 u}{\partial t\,\partial s}. \tag{6.52–10}$$

The basic point to be noted in using the chain rule in connection with second derivatives is this: Suppose u is a function of first-class variables $x, y, \ldots$ and of second-class variables $s, t, \ldots$. Then, if we have written down a chain-rule formula for one of the derivatives $\frac{\partial u}{\partial s}, \frac{\partial u}{\partial t}, \ldots$, this same formula is valid if we replace u throughout by any one of the derivatives $\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \ldots$. We are thus able to express symbols like $\frac{\partial}{\partial s}\left(\frac{\partial u}{\partial x}\right)$ entirely in terms of second derivatives of u with respect to the first-class variables $x, y, \ldots$.

EXERCISES

1. Find formulas comparable to (6.52–4) for $\frac{\partial^2 G}{\partial r^2}$ and $\frac{\partial^2 G}{\partial \theta^2}$ in the problem of Example 1. Hence show that

$$\frac{\partial^2 G}{\partial r^2} + \frac{1}{r}\frac{\partial G}{\partial r} + \frac{1}{r^2}\frac{\partial^2 G}{\partial \theta^2} = \frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial y^2}.$$

2. If $u = F(x, y)$ becomes $G(r, \theta)$ when $x = r\cosh\theta$, $y = r\sinh\theta$, show that

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} - \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} = \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2}.$$

3. If $G(s, t) = F(e^s \cos t,\ e^s \sin t)$, show that $G_{11} + G_{22} = e^{2s}(F_{11} + F_{22})$, where G_{11} and G_{22} are evaluated at (s, t) and F_{11} and F_{22} are evaluated at $(e^s \cos t,\ e^s \sin t)$.

4. Give a complete proof of (6.52–9).

5. If $\xi = x + ct$, $\eta = x - ct$, and $u = F(x, t) = G(\xi, \eta)$, find $\dfrac{\partial^2 u}{\partial t^2} - c^2\dfrac{\partial^2 u}{\partial x^2}$ in terms of derivatives with respect to ξ and η.

6. Show that $5\dfrac{\partial^2 u}{\partial x^2} + 2\dfrac{\partial^2 u}{\partial x\,\partial y} + 2\dfrac{\partial^2 u}{\partial y^2}$ becomes $\dfrac{\partial^2 u}{\partial \xi^2} + \dfrac{\partial^2 u}{\partial \eta^2}$ if we set $\xi = \frac{1}{3}(x + y)$, $\eta = \frac{1}{3}(x - 2y)$.

7. If $u = \phi(x - ct) + \psi(x + ct)$, show that $\dfrac{\partial^2 u}{\partial t^2} = c^2\dfrac{\partial^2 u}{\partial x^2}$.

8. **(a)** If $u = F(x^2 - 2xy)$, let $w = x^2 - 2xy$ and show that $\dfrac{\partial^2 u}{\partial x^2} = 2F'(w) + (2x - 2y)^2 F''(w)$.

(b) Find similar expressions for $\dfrac{\partial^2 u}{\partial x\,\partial y}$ and $\dfrac{\partial^2 u}{\partial y^2}$.

(c) Verify that $x\dfrac{\partial u}{\partial x} + (x - y)\dfrac{\partial u}{\partial y} = 0$.

9. **(a)** If $u = F(r)$ and $r = (x^2 + y^2 + z^2)^{1/2}$, show that $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} + \dfrac{\partial^2 u}{\partial z^2} = \dfrac{d^2 u}{dr^2} + \dfrac{2}{r}\dfrac{du}{dr}$. Hence find the form of $F(r)$ if this last expression is equal to zero when $r > 0$.

(b) Generalize the results of **(a)** for $u = F(r)$, $r = (x_1^2 + x_2^2 + \cdots + x_n^2)^{1/2}$.

10. If $V = \dfrac{1}{r}g\left(t - \dfrac{r}{c}\right)$, where c is constant and $r = (x^2 + y^2 + z^2)^{1/2}$, show that

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 V}{\partial t^2}.$$

11. If $u = F[x + f(y)]$, show that $\dfrac{\partial u}{\partial x}\dfrac{\partial^2 u}{\partial x\,\partial y} = \dfrac{\partial u}{\partial y}\dfrac{\partial^2 u}{\partial x^2}$.

12. Show that $y\dfrac{\partial^2 u}{\partial x^2} + (x + y)\dfrac{\partial^2 u}{\partial x\,\partial y} + x\dfrac{\partial^2 u}{\partial y^2}$ becomes $-2t^2\dfrac{\partial^2 u}{\partial s\,\partial t} - 2t\dfrac{\partial u}{\partial s}$ when we set $s = y^2 - x^2$, $t = y - x$.

13. If $u = x^3 F(y/x, z/x)$, show that

$$x^2\frac{\partial^2 u}{\partial x^2} - y^2\frac{\partial^2 u}{\partial y^2} - z^2\frac{\partial^2 u}{\partial z^2} - 2yz\frac{\partial^2 u}{\partial y\,\partial z} + 4y\frac{\partial u}{\partial y} + 4z\frac{\partial u}{\partial z} = 6u.$$

14. If $u = F(x, y) = G(s, t)$, where $s = xy$, $t = \dfrac{1}{y}$, find $\dfrac{\partial^2 u}{\partial x^2} + 2xy^2\dfrac{\partial u}{\partial x} + (y - y^3)\dfrac{\partial u}{\partial y} + x^2y^2u$ in terms of s, t, u, and derivatives of u with respect to s and t.

15. Prove that setting $x = e^s$, $y = e^t$ changes $x^2\dfrac{\partial^2 u}{\partial x^2} + y^2\dfrac{\partial^2 u}{\partial y^2} + x\dfrac{\partial u}{\partial x} + y\dfrac{\partial u}{\partial y}$ into $\dfrac{\partial^2 u}{\partial s^2} + \dfrac{\partial^2 u}{\partial t^2}$.

16. Suppose that $x = f(u, v)$, $y = g(u, v)$ transform $F(x, y)$ into $G(u, v)$. Suppose also that $\frac{\partial f}{\partial u} = \frac{\partial g}{\partial v}$ and $\frac{\partial f}{\partial v} = -\frac{\partial g}{\partial u}$.

Prove that $\frac{\partial^2 G}{\partial u^2} + \frac{\partial^2 G}{\partial v^2} = \left(\frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial y^2}\right)\left[\left(\frac{\partial f}{\partial u}\right)^2 + \left(\frac{\partial f}{\partial v}\right)^2\right]$ is an identity in u and v.

17. If $u = x^n f(y/x) + x^{-n} g(y/x)$, show that

$$x^2 \frac{\partial^2 u}{\partial x^2} + 2xy \frac{\partial^2 u}{\partial x\, \partial y} + y^2 \frac{\partial^2 u}{\partial y^2} + x \frac{\partial u}{\partial x} + y \frac{\partial u}{\partial y} = n^2 u.$$

18. (a) If $u = F(x, y) = G(s, t)$ and $x = f(s, t)$, $y = g(s, t)$, show that $G_{11} = F_{11} f_1^2 + F_{22} g_1^2 + 2F_{12} f_1 g_1 + F_1 f_{11} + F_2 g_{11}$, where for convenience all the variables have been omitted from the functional symbols. (b) Write this same formula using the variable u and not using any of the functional symbols F, G, f, g. (c) Write analogous formulas for G_{12} and G_{22}.

19. Let

$$\xi_1 = a_{11} x_1 + a_{12} x_2$$
$$\xi_2 = a_{21} x_1 + a_{22} x_2$$

where the coefficients $a_{11}, \ldots$ are constants. With this change of variable, show that $a \frac{\partial^2 u}{\partial x_1^2} + 2b \frac{\partial^2 u}{\partial x_1\, \partial x_2} + c \frac{\partial^2 u}{\partial x_2^2} = \alpha \frac{\partial^2 u}{\partial \xi_1^2} + 2\beta \frac{\partial^2 u}{\partial \xi_1\, \partial \xi_2} + \gamma \frac{\partial^2 u}{\partial \xi_2^2}$, where the coefficients a, b, c, and α, β, γ are related in such a way that

$$\alpha\gamma - \beta^2 = (ac - b^2)(a_{11} a_{22} - a_{21} a_{12})^2.$$

20. If the change of variable (with constant coefficients)

$$\xi_i = \sum_{j=1}^{n} c_{ij} x_j$$

changes

$$\sum_{i,j=1}^{n} a_{ij} \frac{\partial^2 u}{\partial x_i\, \partial x_j} \quad \text{into} \quad \sum_{k,l=1}^{n} b_{kl} \frac{\partial^2 u}{\partial \xi_k\, \partial \xi_l},$$

show that

$$b_{kl} = \sum_{i,j=1}^{n} c_{ki} a_{ij} c_{lj}.$$

6.53 / HOMOGENEOUS FUNCTIONS. EULER'S THEOREM

Consider the functions

$$x^2 + y^2, \frac{xy}{x + y}, x^2 y \log \frac{y}{x}.$$

Each of these functions has the interesting property that, if the variables x, y are replaced by tx, ty respectively, where t is a parameter, we obtain the original function multiplied by a power of t:

$$(tx)^2 + (ty)^2 = t^2(x^2 + y^2),$$

$$\frac{(tx)(ty)}{(tx) + (ty)} = t \frac{xy}{x + y},$$

$$(tx)^2 (ty) \log \frac{ty}{tx} = t^3 \left(x^2 y \log \frac{y}{x}\right).$$

Functions of this type are called *homogeneous*. In general, we say that $F(x, y)$ is *homogeneous of degree n* if

$$F(tx, ty) = t^n F(x, y) \tag{6.53–1}$$

for all values of x, y, and t for which $F(x, y)$ and $F(tx, ty)$ are defined. The degree n is a constant; it need not be an integer, and it may be negative. It may also be zero.

Example 1. **(a)** $\dfrac{x}{x^2+y^2}$ is homogeneous of degree -1. **(b)** $x^{1/3}+xy^{-2/3}$ is homogeneous of degree $\frac{1}{3}$. **(c)** $\dfrac{x^2-y^2}{x^2+y^2}$ is homogeneous of degree 0.

If the fundamental relation (6.53–1) holds true only when t is restricted to positive (or nonnegative) values, we say that F is *positively homogeneous* of degree n. An example in which the limitation to nonnegative t is essential is furnished by the function $\sqrt{x^2+y^2}$, which is positively homogeneous of degree 1. Since by definition the radical sign calls for the nonnegative square root of the radicand,

$$\sqrt{(tx)^2+(ty)^2} = t\sqrt{x^2+y^2}$$

holds only if $t \geqq 0$.

The definition of homogeneity is extended in an obvious way to functions of any number of variables.

The meaning of homogeneity can be interpreted geometrically. We refer back to §5.4, where we discussed modes of representing a function. If (x, y) is a point distinct from the origin in the plane, the set of all points (tx, ty), as t varies, fills out the line determined by (x, y) and the origin (see Fig. 44). The fundamental relation (6.53–1) states that the value of the function F at the point (tx, ty) is t^n times the value of the function at the point (x, y). Thus, once we know the value of F at a point other than $(0, 0)$ on the line, we can compute its value at all other points on the line where it is defined. We must of course know the degree n. If the function is merely *positively* homogeneous, the requirement $t \geqq 0$ limits us to points (tx, ty) on the same side of the origin as (x, y) itself. It is clear that if F is positively homogeneous, and if we know its values at all points of the circle $x^2+y^2=1$, we can compute its value at any other point where it is defined.

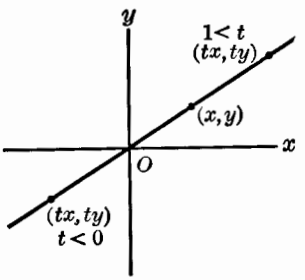


Fig. 44.

Many simple or important functions arising in applied mathematics are homogeneous or positively homogeneous. For example, if F is such that its values are directly proportional to the nth power of the distance from the origin to (x, y), then F is positively homogeneous of degree n. In fact, we can write

$$F(x, y) = Ar^n,\ r = (x^2+y^2)^{1/2},\ \text{whence}\ F(tx, ty) = t^n F(x, y)\ \text{ if }\ t \geqq 0.$$

One of the useful pieces of information about homogeneous functions is that

furnished by a theorem named after the 18th century Swiss mathematician Leonard Euler.

EULER'S THEOREM (THEOREM IV). *Let $F(x, y)$ be positively homogeneous of degree n. Then at any point where F is differentiable we have*

$$x\frac{\partial F}{\partial x} + y\frac{\partial F}{\partial y} = nF(x, y). \tag{6.53–2}$$

Proof. Write $u = tx$, $v = ty$. Consider u, v as first-class variables, and t, x, y as second-class variables. The partial derivative of $F(u, v) = F(tx, ty)$ with respect to t is

$$F_1(u, v)\frac{\partial u}{\partial t} + F_2(u, v)\frac{\partial v}{\partial t},$$

or

$$xF_1(tx, ty) + yF_2(tx, ty),$$

provided that F is differentiable at (tx, ty). On the other hand, for positive t, $F(tx, ty) = t^nF(x, y)$, and $\frac{\partial}{\partial t}(t^nF(x, y)) = nt^{n-1}F(x, y)$. Thus

$$xF_1(tx, ty) + yF_2(tx, ty) = nt^{n-1}F(x, y). \tag{6.53–3}$$

If we put $t = 1$ here we obtain the desired relation (6.53–2).

The theorem and the proof extend to functions of three or more independent variables. There is also a converse theorem; see Exercise 6.

Example 2. If $F(x, y)$ is positively homogeneous of degree 2, and $u = r^mF(x, y)$, where $r^2 = x^2 + y^2$, show that

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = r^m\left(\frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial y^2}\right) + m(m+4)r^{m-2}F.$$

In working this problem it is convenient to observe that

$$2r\frac{\partial r}{\partial x} = 2x, \quad \text{or} \quad \frac{\partial r}{\partial x} = \frac{x}{r}.$$

We have, therefore,

$$\frac{\partial u}{\partial x} = r^m\frac{\partial F}{\partial x} + mr^{m-1}\frac{\partial r}{\partial x}F = r^m\frac{\partial F}{\partial x} + mr^{m-2}xF.$$

Differentiating again,

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2} &= r^m\frac{\partial^2 F}{\partial x^2} + mr^{m-1}\frac{\partial r}{\partial x}\frac{\partial F}{\partial x} + mr^{m-2}\frac{\partial F}{\partial x} \\ &\quad + mr^{m-2}F + m(m-2)r^{m-3}\frac{\partial r}{\partial x}xF \\ &= r^m\frac{\partial^2 F}{\partial x^2} + 2mr^{m-2}x\frac{\partial F}{\partial x} + mr^{m-2}F + m(m-2)r^{m-4}x^2F.\end{aligned}$$

Because of the symmetrical occurrence of x and y in r, we can write down at once

$$\frac{\partial^2 u}{\partial y^2} = r^m \frac{\partial^2 F}{\partial y^2} + 2mr^{m-2}y \frac{\partial F}{\partial y} + mr^{m-2}F + m(m-2)r^{m-4}y^2 F.$$

If we now add, and take note of the relations

$$x \frac{\partial F}{\partial x} + y \frac{\partial F}{\partial y} = 2F, \; x^2 + y^2 = r^2,$$

we find

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = r^m \left(\frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial y^2} \right) + 4mr^{m-2}F + 2mr^{m-2}F + m(m-2)r^{m-2}F$$

$$= r^m \left(\frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial y^2} \right) + m(m+4)r^{m-2}F.$$

This is the required result. Question for the student: Where was use made of the homogeneity of F?

EXERCISES

1. (a) If F is positively homogeneous of degree n, and if F, F_1, and F_2 are differentiable, the equation

$$x_2 \frac{\partial^2 F}{\partial x^2} + 2xy \frac{\partial^2 F}{\partial x \, \partial y} + y^2 \frac{\partial^2 F}{\partial y^2} = n(n-1)F(x, y)$$

is valid. Prove this by differentiating (6.53–3) partially with respect to t. (b) State and prove the corresponding result for third derivatives.

2. If F is positively homogeneous of degree n, and differentiable, its first partial derivatives are positively homogeneous of degree $n-1$. Prove this.

3. Let $H(x, y)$ be positively homogeneous of degree p, and let $u = r^m H(x, y)$, where $r^2 = x^2 + y^2$. Show that

$$\Delta u = r^m \, \Delta H + m(2p + m)r^{m-2}H,$$

where, for any function $F(x, y)$, the notation ΔF means

$$\Delta F = \frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial y^2}.$$

4. The equation $\Delta F = 0$ (see the explanation of notation at the end of Exercise 3) is called Laplace's equation in two dimensions. Here it is understood that (x, y) are rectangular co-ordinates of a point in the plane. If $H(x, y)$ is positively homogeneous of degree p, and $\Delta H = 0$, show that $\Delta(r^{-2p}H) = 0$ also. Use Exercise 3. Examples are furnished by taking $H(x, y)$ to be one of the functions x, y, $x^2 - y^2$, $2xy$, $3x^2y - y^3$, with appropriate values of p in each case.

5. (a) If x, y, z are rectangular co-ordinates in space, we write

$$\Delta F(x, y, z) = \frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial y^2} + \frac{\partial^2 F}{\partial z^2}.$$

If $u = r^m H(x, y, z)$, where $r^2 = x^2 + y^2 + z^2$, and if H is positively homogeneous of degree

p, work out the result analogous to that of Exercise 3. (**b**) What is the result for three dimensions comparable to that of Exercise 4? (**c**) Develop generalizations of the results of (**a**) and (**b**) for the case of n dimensions, taking $r^2 = x_1^2 + \cdots + x_n^2$.

6. Suppose that $F(x, y)$ is defined and differentiable in an open region R, and suppose that $x\,\partial F/\partial x + y\,\partial F/\partial y = nF(x, y)$ at each point of the region. Then, if (x, y) is in R, the relation $F(tx, ty) = t^n F(x, y)$ holds in any interval $t_0 < t < t_1$ (where $t_0 \geq 0$) provided that $t = 1$ is in this interval and provided that, for all such t, the points (tx, ty) are in R. To prove this converse of Euler's theorem, let $f(t) = F(tx, ty)$, where (x, y) is a fixed point of R. Use the hypothesis on F to prove that $tf'(t) = nf(t)$. From this, infer that $f(t)t^{-n}$ is a constant (depending on x, y). Then complete the proof.

6.6 / DERIVATIVES OF IMPLICIT FUNCTIONS

In §6.1 we dealt with the differentiation of implicit functions in a variety of particular situations. We did not attempt to deal with general cases in which the functions were merely indicated by some functional symbol. In practice it is necessary to have formulas to deal with implicit functions in terms of general notation.

A simple but typical case is that arising when z is defined as a function of x, y by an equation of the form

$$F(x, y, z) = 0. \tag{6.6–1}$$

Suppose, for instance, that the equation is

$$x^2 + 2xz + z^2 - yz - 1 = 0,$$

so that in this case

$$F(x, y, z) = x^2 + 2xz + z^2 - yz - 1. \tag{6.6–2}$$

Proceeding as in §6.1, we have

$$\begin{aligned} 2x + 2x\frac{\partial z}{\partial x} + 2z + 2z\frac{\partial z}{\partial x} - y\frac{\partial z}{\partial x} &= 0, \\ 2x\frac{\partial z}{\partial y} + 2z\frac{\partial z}{\partial y} - y\frac{\partial z}{\partial y} - z &= 0, \\ \frac{\partial z}{\partial x} &= -\frac{2x + 2z}{2x + 2z - y}, \\ \frac{\partial z}{\partial y} &= -\frac{-z}{2x + 2z - y}. \end{aligned} \tag{6.6–3}$$

Now let us observe that if we regard x, y, z as independent in (6.6–2), then

$$\frac{\partial F}{\partial x} = 2x + 2z, \frac{\partial F}{\partial y} = -z, \frac{\partial F}{\partial z} = 2x + 2z - y.$$

Accordingly, equations (6.6–3) take on the form

$$\frac{\partial z}{\partial x} = -\frac{\dfrac{\partial F}{\partial x}}{\dfrac{\partial F}{\partial z}}, \quad \frac{\partial z}{\partial y} = -\frac{\dfrac{\partial F}{\partial y}}{\dfrac{\partial F}{\partial z}}. \tag{6.6–4}$$

We are going to show that these formulas are general; that is, they hold when z is defined by (6.6–1), subject to certain general assumptions on the otherwise arbitrary function F.

We are not just now concerned with knowing what assumptions to make about F in order to guarantee that equation (6.6–1) does in fact define z as a function of x, y. Questions of this kind will be dealt with in Chapter 8. We assume that z is a well-defined function of x, y, and that equation (6.6–1) is an identity in x, y when z is replaced by its functional value:

$$z = f(x, y), \tag{6.6–5}$$

$$F(x, y, f(x, y)) \equiv 0. \tag{6.6–6}$$

A simple instance of these last relations would be afforded by

$$z = \sqrt{1 - x^2 - y^2}, \; F(x, y, z) = x^2 + y^2 + z^2 - 1.$$

Another instance is

$$z = \log xy, \; F(x, y, z) = xy - e^z.$$

We also assume that F is a differentiable function of the *three independent* variables and that $f(x, y)$ has first partial derivatives. Under these conditions we shall prove formulas (6.6–4). Naturally we must assume that the denominator $\dfrac{\partial F}{\partial z}$ is not zero.

Under the foregoing hypotheses we look upon $G(x, y) = F(x, y, f(x, y))$ as a composite function of x, y. The first-class variables are x, y, z, and the second-class variables are x, y. The relations between the variables of the two classes are

$$x = x, \; y = y, \; z = f(x, y).$$

Therefore,

$$\frac{\partial x}{\partial x} = 1, \frac{\partial x}{\partial y} = 0, \frac{\partial y}{\partial x} = 0, \frac{\partial y}{\partial y} = 1.$$

The chain rule gives

$$\frac{\partial G}{\partial x} = \frac{\partial F}{\partial x}\frac{\partial x}{\partial x} + \frac{\partial F}{\partial y}\frac{\partial y}{\partial x} + \frac{\partial F}{\partial z}\frac{\partial z}{\partial x},$$

or

$$\frac{\partial G}{\partial x} = F_1 + F_3\frac{\partial z}{\partial x}.$$

Likewise

$$\frac{\partial G}{\partial y} = F_2 + F_3 \frac{\partial z}{\partial y}.$$

Now by (6.6–6), $G(x, y) \equiv 0$, and $\frac{\partial G}{\partial x} = \frac{\partial G}{\partial y} \equiv 0$. Therefore, assuming that $F_3 \neq 0$, we have

$$\frac{\partial z}{\partial x} = -\frac{F_1}{F_3}, \quad \frac{\partial z}{\partial y} = -\frac{F_2}{F_3}.$$

These are the same equations as (6.6–4) except for the difference in notation. It is understood that the derivatives F_1, F_2, F_3 are evaluated with $z = f(x, y)$.

It is easily seen that the foregoing considerations may be generalized as follows. Suppose a function of n variables, say $u = f(x_1, \ldots, x_n)$ is determined implicitly as a solution of an equation $F(x_1, \ldots, x_n, u) = 0$. Then, under suitable assumptions of differentiability, we have

$$\frac{\partial f}{\partial x_i} = -\frac{F_i(x_1, \ldots, x_n, f(x_1, \ldots, x_n))}{F_{n+1}(x_1, \ldots, x_n, f(x_1, \ldots, x_n))}, \; i = 1, \ldots, n. \tag{6.6–7}$$

Example 1. Show that if $z = f(x, y)$ is a solution of the equation $F(x, y, z) = 0$, then the line normal to the surface $z = f(x, y)$ at a given point has direction ratios $F_1 : F_2 : F_3$, the partial derivatives being evaluated at the point in question.

It was shown in §6.2 that the line normal to the surface has direction ratios $\frac{\partial z}{\partial x} : \frac{\partial z}{\partial y} : -1$. By (6.6–4) an equivalent set of ratios is

$$-\frac{F_1}{F_3} : -\frac{F_2}{F_3} : -1.$$

The ratios will not be altered if we multiply by $-F_3$. In this way we obtain the ratios $F_1 : F_2 : F_3$ for the direction of the normal line.

Next let us consider the case of two functions which arise as solutions of a pair of simultaneous equations. Suppose, for instance, that

$$u = f(x, y, z) \quad \text{and} \quad v = g(x, y, z) \tag{6.6–8}$$

are solutions of

$$F(x, y, z, u, v) = 0, \quad G(x, y, z, u, v) = 0. \tag{6.6–9}$$

That is, suppose that

$$F(x, y, z, f(x, y, z), g(x, y, z)) \equiv 0 \tag{6.6–10}$$

is true for all values of x, y, z in some region, with a similar relation for G. We assume that the functions F, G, f, g are all differentiable. The problem which we pose is that of expressing the partial derivatives of f and g in terms of those of F and G. As usual, we write

$$F_1 = \frac{\partial F}{\partial x}, \; F_2 = \frac{\partial F}{\partial y}, \ldots, \; F_5 = \frac{\partial F}{\partial v}, \ldots.$$

The required formulas are then of the form

$$\frac{\partial u}{\partial x} = -\frac{\begin{vmatrix} F_1 & F_5 \\ G_1 & G_5 \end{vmatrix}}{\begin{vmatrix} F_4 & F_5 \\ G_4 & G_5 \end{vmatrix}}, \quad \ldots, \quad \frac{\partial u}{\partial z} = -\frac{\begin{vmatrix} F_3 & F_5 \\ G_3 & G_5 \end{vmatrix}}{\begin{vmatrix} F_4 & F_5 \\ G_4 & G_5 \end{vmatrix}},$$

$$\frac{\partial v}{\partial x} = -\frac{\begin{vmatrix} F_4 & F_1 \\ G_4 & G_1 \end{vmatrix}}{\begin{vmatrix} F_4 & F_5 \\ G_4 & G_5 \end{vmatrix}}, \quad \ldots, \quad \frac{\partial v}{\partial z} = -\frac{\begin{vmatrix} F_4 & F_3 \\ G_4 & G_3 \end{vmatrix}}{\begin{vmatrix} F_4 & F_5 \\ G_4 & G_5 \end{vmatrix}}. \tag{6.6–11}$$

This is on the assumption that the determinant appearing in the denominators is not equal to zero. The expressions $F_1, \ldots, G_5$ are evaluated with u and v given by (6.6–8).

We shall indicate briefly the manner of deriving formulas (6.6–11). We regard x, y, z, u, v as first-class variables, and x, y, z as second-class variables. Then, from the identity (6.6–10), we have (differentiating with respect to x)

$$F_1 + F_4 \frac{\partial u}{\partial x} + F_5 \frac{\partial v}{\partial x} = 0.$$

There is a similar equation involving G:

$$G_1 + G_4 \frac{\partial u}{\partial x} + G_5 \frac{\partial v}{\partial x} = 0.$$

These two linear equations are now solved for $\frac{\partial u}{\partial x}$ and $\frac{\partial v}{\partial x}$, with the results indicated in (6.6–11). The procedure is entirely similar for finding the derivatives of u and v with respect to y or z.

The student need not memorize the formulas (6.6–11), though it is useful to have ready reference to such formulas. It is the procedure for deriving the formulas which is important, and the student should be able to carry out the derivation himself.

Determinants such as those occurring in the formulas (6.6–11) are called *Jacobians*, after Carl Jacobi, a prominent German mathematician of the nineteenth century. There is a more compact notation which is often used for a Jacobian:

$$\frac{\partial(F, G)}{\partial(u, v)} = \begin{vmatrix} \dfrac{\partial F}{\partial u} & \dfrac{\partial F}{\partial v} \\ \dfrac{\partial G}{\partial u} & \dfrac{\partial G}{\partial v} \end{vmatrix}.$$

This is called a Jacobian of second order. The form of a general nth-order

Jacobian involves n functions, each of n variables:

$$\frac{\partial(F_1, \ldots, F_n)}{\partial(u_1, \ldots, u_n)} = \begin{vmatrix} \frac{\partial F_1}{\partial u_1} & \frac{\partial F_1}{\partial u_2} & \cdots & \frac{\partial F_1}{\partial u_n} \\ \frac{\partial F_2}{\partial u_1} & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \frac{\partial F_n}{\partial u_1} & \cdot & \cdots & \frac{\partial F_n}{\partial u_n} \end{vmatrix}.$$

In concluding this section we emphasize that here we are concerned with the *structure* of general formulas for the derivatives of implicit functions. Questions of existence and differentiability of implicit functions are taken up in Chapter 8.

EXERCISES

1. If the equations

$$F(x, y, u, v, w) = 0, G(x, y, u, v, w) = 0, H(x, y, u, v, w) = 0$$

have solutions for u, v, and w as functions of x and y, show, taking for granted certain general conditions, that

$$\frac{\partial u}{\partial x} = -\frac{\frac{\partial(F, G, H)}{\partial(x, v, w)}}{\frac{\partial(F, G, H)}{\partial(u, v, w)}}.$$

Write five other similar formulas.

2. Suppose $y = f(x)$ is a solution of $G(x, y) = 0$, where $\frac{\partial G}{\partial y} \neq 0$, and let $g(x) = F(x, f(x))$. Show that $g'(x) = (F_1G_2 - F_2G_1)/G_2$, the right side being evaluated with $y = f(x)$.

3. If $z = f(x, y)$ is a solution of $F(x, y, z) = 0$ (with $F_3 \neq 0$), and if $H(x, y) = G(x, y, f(x, y))$, show that $\frac{\partial F}{\partial z}\frac{\partial H}{\partial y} = -\frac{\partial(F, G)}{\partial(y, z)}$. Write a similar formula involving $\frac{\partial H}{\partial x}$.

4. Suppose $u = f(x, y, z)$, $v = g(x, y, z)$ are solutions of $F(x, y, z, u, v) = 0$, $G(x, y, z, u, v) = 0$. Let $K(x, y, z) = H(x, y, z, f(x, y, z), g(x, y, z))$. Show that

$$\frac{\partial K}{\partial z} = \frac{\frac{\partial(F, G, H)}{\partial(z, u, v)}}{\frac{\partial(F, G)}{\partial(u, v)}}$$

under suitable conditions.

5. As an instance of (6.6–8) and (6.6–9) consider $u = x \sin xyz$, $v = y \cos xyz$ as solutions of $yu - xv \tan xyz = 0$, $y^2u^2 + x^2v^2 - x^2y^2 = 0$. Verify all six of the formulas (6.6–11) in this case.

6. If the equation $F(x, y, z) = 0$ can be solved for each one of the variables in terms of the other two, show, taking for granted certain general conditions, that

$$\left(\frac{\partial x}{\partial y}\right)_z \left(\frac{\partial y}{\partial z}\right)_x \left(\frac{\partial z}{\partial x}\right)_y = -1.$$

7. Let $G(x, y, z, v) = 0$ have solutions $x = f(y, z, v)$, $y = g(x, z, v)$, $z = h(x, y, v)$, and by means of one of these equations at a time let $u = F(x, y, z)$ become a function of (y, z, v), (x, z, v), and (x, y, v) respectively. Show that, subject to certain conditions,

$$\left(\frac{\partial u}{\partial y}\right)_{x,v} - \left(\frac{\partial u}{\partial y}\right)_{z,v} = \left(\frac{\partial u}{\partial z}\right)_{x,y} \cdot \left(\frac{\partial z}{\partial y}\right)_{x,v} - \left(\frac{\partial u}{\partial x}\right)_{y,z} \cdot \left(\frac{\partial x}{\partial y}\right)_{z,v}.$$

An example is furnished by $xyv - z = 0$, $u = x^2 + y^2 + z^2$, and the student should check the meaning of the problem in terms of this special case if he feels an illustration to be necessary.

8. Suppose $u = f(x, y)$ is a solution of $F(x, y, u) = 0$, and that $y = g(x, z)$ is a solution of $G(x, y, z) = 0$. Let $H(x, z) = f(x, g(x, z))$. Show that $F_3G_2H_2 = F_2G_3$ and $F_3G_2H_1 = F_2G_1 - F_1G_2$ are identities in x and z.

9. (a) Starting from (6.6–4), show that

$$\frac{\partial^2 z}{\partial x^2} = \frac{F_3^2F_{11} - 2F_1F_3F_{13} + F_1^2F_{33}}{-F_3^3}.$$

(b) Derive analogous formulas for $\dfrac{\partial^2 z}{\partial y\,\partial x}$ and $\dfrac{\partial^2 z}{\partial y^2}$.

10. If $z = f(x, y)$ satisfies an equation of the form $z = F(ax + by + cz)$, where a, b, and c are constants, show that $b\dfrac{\partial z}{\partial x} = a\dfrac{\partial z}{\partial y}$.

11. If $z = f(x, y)$ satisfies an equation of the form $F(x + y + z, x^2 + y^2 + z^2) = 0$, show that $(y - x) + (y - z)\dfrac{\partial z}{\partial x} + (z - x)\dfrac{\partial z}{\partial y} = 0$.

12. Suppose that the function $z = f(x, y)$ satisfies an equation of the form $F(ax + by + cz,\ Ax^2 + By^2 + Cz^2) = 0$, where a, b, c and A, B, C are constants. Show that $\dfrac{\partial z}{\partial x} = -\dfrac{aF_1 + 2AxF_2}{cF_1 + 2CzF_2}$.

13. If $z = \phi(x, y)$ satisfies the equation $F(f(x, y, z), g(x, y, z)) = 0$, show that

$$\frac{\partial z}{\partial y} = -\frac{F_1f_2 + F_2g_2}{F_1f_3 + F_2g_3}.$$

14. If $G_1(x_1, x_2, y)$, $G_2(x_1, x_2, y)$ and $f(x_1, x_2)$ are given, and if $g_i(x_1, x_2) = G_i(x_1, x_2, f(x_1, x_2))$ $(i = 1, 2)$, show that

$$\frac{\partial(g_1, g_2)}{\partial(x_1, x_2)} = \frac{\partial(G_1, G_2)}{\partial(x_1, x_2)} + \frac{\partial f}{\partial x_1}\frac{\partial(G_1, G_2)}{\partial(y, x_2)} + \frac{\partial f}{\partial x_2}\frac{\partial(G_1, G_2)}{\partial(x_1, y)},$$

with y replaced by $f(x_1, x_2)$ after the differentiations. This formula is used in the theory of first-order partial differential equations.

6.7 / EXTREMAL PROBLEMS WITH CONSTRAINTS

Many interesting maximum or minimum problems arise in such a form that we are required to find an extremal value of a function, say $F(x, y, z)$, where the variables x, y, z are not independent of each other, but are restricted by some relation existing between them, this relation being expressed by an equation $G(x, y, z) = 0$.

Example 1. Find the minimum value of

$$F(x, y, z) = (x-3)^2 + (y-2)^2 + (z-1)^2$$

subject to the condition

$$G(x, y, z) = 2x - 3y - 4z - 25 = 0.$$

This problem has already been solved; see Example 1, §6.3.

The equation $G(x, y, z) = 0$ is called a *constraint* on the variables x, y, z. It is immaterial whether the equation of constraint has the form $G(x, y, z) = 0$ or $G(x, y, z) = k$, where k is a specified constant, for the latter form of constraint can be written $G(x, y, z) - k = 0$.

An extremal problem with constraint may occur with any number of variables, and there may be more than one equation of constraint.

Example 2. Find maximum and minimum values of $x^2 + y^2 + z^2$ subject to the two conditions

$$x + y - z = 0, \frac{x^2}{16} + y^2 + z^2 = 1.$$

In this problem there are two constraints on the three variables, so that there is actually only one independent variable.

Example 3. Find the minimum value of $(x-u)^2 + (y-v)^2 + z^2$ subject to the condition $3x^2 + y^2 - 6x - 4y - 12z + 43 = 0$.

In this problem there are five variables and one constraint, which happens to involve only three of the five variables.

There are various methods for dealing with extremal problems with constraints. One method is to use the equation or equations of constraint to express certain of the variables in terms of the remaining variables. These latter variables are chosen as the independent variables, and the function whose extremal value is sought is then expressed in terms of the independent variables only. The solution is then carried out by standard methods. We shall call this general method the *method of direct elimination.* It is illustrated, in the case of the problem of Example 1 in the present section, by the solution given in Example 1, §6.3.

A second method may conveniently be called the *method of implicit functions.* Suppose the problem is to find the point or points at which $F(x, y, z)$ has extreme values subject to the conditions $G(x, y, z) = k$. Here we assume that F and G are given functions with continuous first partial derivatives, and that k is a given constant. Let us assume, for definiteness, that $\frac{\partial G}{\partial z} \neq 0$ and that the equation

$$G(x, y, z) - k = 0 \tag{6.7–1}$$

has a solution

$$z = f(x, y). \tag{6.7–2}$$

We then seek to make the quantity

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

a maximum or minimum. Accordingly we want to solve the equations

$$\frac{\partial u}{\partial x} = 0, \frac{\partial u}{\partial y} = 0. \tag{6.7–3}$$

Now

$$\frac{\partial u}{\partial x} = \frac{\partial F}{\partial x} + \frac{\partial F}{\partial z}\frac{\partial f}{\partial x}, \qquad \frac{\partial u}{\partial y} = \frac{\partial F}{\partial y} + \frac{\partial F}{\partial z}\frac{\partial f}{\partial y}, \tag{6.7–4}$$

where z is replaced by $f(x, y)$ after the differentiations are performed.

We also have the identity

$$G(x, y, f(x, y)) - k = 0,$$

from which it follows by differentiation that

$$\frac{\partial G}{\partial x} + \frac{\partial G}{\partial z}\frac{\partial f}{\partial x} = 0, \frac{\partial G}{\partial y} + \frac{\partial G}{\partial z}\frac{\partial f}{\partial y} = 0. \tag{6.7–5}$$

If we solve these equations for $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ and substitute in (6.7–4), we obtain the equation

$$\frac{\partial u}{\partial x} = \frac{\frac{\partial F}{\partial x}\frac{\partial G}{\partial z} - \frac{\partial F}{\partial z}\frac{\partial G}{\partial x}}{\frac{\partial G}{\partial z}},$$

and a similar equation for $\frac{\partial u}{\partial y}$. Equations (6.7–3) now take the form

$$\frac{\partial F}{\partial x}\frac{\partial G}{\partial z} - \frac{\partial F}{\partial z}\frac{\partial G}{\partial x} = 0, \frac{\partial F}{\partial y}\frac{\partial G}{\partial z} - \frac{\partial F}{\partial z}\frac{\partial G}{\partial y} = 0, \tag{6.7–6}$$

in which z is replaced by $f(x, y)$ after the differentiations are performed. Now the method of implicit functions for this extremal problem with constraint may be described as follows: We do not *actually* solve for z at the outset. Instead, we carry the work along and arrive at equations (6.7–6) as equations in *all three* variables. These two equations, together with the constraint (6.7–1), give us three equations which we solve as simultaneous equations in x, y, z. The required points of extreme value will be among the points found in this way. This general assertion is subject to some qualifications to rule out exceptional cases. We could, of course, think of y as a function of x and z, or of x as a function of y and z, the functional relation in each case being determined by the constraint. These alternatives would give us pairs of equations different from (6.7–6), but equivalent to them.

The implicit-function method was used in the solution of the problem of

Example 3, §6.3, with y regarded as a function of x and z. Our purpose just now is not to illustrate the method in particular cases, but to obtain the equations (6.7–6) in preparation for study of other aspects of extremal problems with constraints.

It is often useful to regard an extremal problem with constraint from a geometrical point of view. Consider, for instance, the problem of extremal value for the function $F(x, y, z)$ with the constraint $G(x, y, z) = k$ (a given constant). We shall suppose that the equation $G(x, y, z) = k$ defines a surface S possessing a tangent plane at each of its points with which we are concerned. We then consider the values of the function F at points of the surface S; we wish to find the points of S where F has maximum or minimum values. Now consider the family of surfaces

$$F(x, y, z) = C, \tag{6.7–7}$$

where C is a parameter. Suppose that M is the absolute maximum value attained by F at points of S. Then, if $C > M$, the surface (6.7–7) and the surface S will have no points in common, whereas if $C = M$ the two surfaces will have at least one point (x_0, y_0, z_0) in common, and this will be a point where F attains its maximum value on S. We shall prove that the two surfaces $G(x, y, z) = k$ and $F(x, y, z) = M$ are tangent at this point (x_0, y_0, z_0). The direction ratios of the normals to the two surfaces are, respectively,

$$G_1 : G_2 : G_3 \quad \text{and} \quad F_1 : F_2 : F_3,$$

with the partial derivatives evaluated at (x_0, y_0, z_0) (see Example 1, §6.6). The condition of tangency is therefore that these two sets of ratios be the same, i.e., that G_1, G_2, G_3 and F_1, F_2, F_3 be proportional. Now we saw earlier that equations (6.7–6) must hold at a point of extreme value, provided $G_3 \neq 0$. These equations may be written

$$F_1G_3 - F_3G_1 = 0, \; F_2G_3 - F_3G_2 = 0,$$

or

$$F_1 = \frac{F_3}{G_3} G_1, \; F_2 = \frac{F_3}{G_3} G_2,$$

whence it follows that F_1, F_2, F_3 and G_1, G_2, G_3 are proportional at (x_0, y_0, z_0). If it should happen that $G_3 = 0$ at the point, the same conclusion of proportionality can be reached by assuming that $G_2 \neq 0$ or $G_1 \neq 0$.

The argument and the conclusion are essentially the same in the case of a minimum rather than a maximum. Also, the surfaces will be tangent at a point of *relative* extreme value of F on the surface S. There may also be tangencies at points where F is neither a maximum nor a minimum. We sum up our findings in a formal statement.

THEOREM V. *Suppose the functions F and G have continuous first partial derivatives throughout a certain region of space. Let the equation $G(x, y, z) = k$ define a surface S, every point of which is in the interior of the*

region, and suppose that the three partial derivatives G_1, G_2, G_3 are never simultaneously zero at a point of S. Then a necessary condition for the values of $F(x, y, z)$ on S to attain an extreme value (either relative or absolute) at a point of S is that F_1, F_2, F_3 be proportional to G_1, G_2, G_3 at that point. If C is the value of F at the point, and if the constant of proportionality is not zero, the geometrical meaning of the proportionality is that the surface S and the surface $F(x, y, z) = C$ are tangent at the point in question.

A fully detailed analytic proof of this theorem depends upon the theory of implicit functions, as developed in Chapter 8. The geometric arguments which precede the statement of the theorem do not constitute a rigorous proof, but they give a good basis for understanding the theorem; with adequate knowledge of implicit-function theory, the argument given here can be developed into a complete proof.

Example 4. Consider the maximum and minimum values of $F(x, y, z) = x^2 + y^2 + z^2$ on the surface of the ellipsoid

$$G(x, y, z) = \frac{x^2}{64} + \frac{y^2}{36} + \frac{z^2}{25} = 1.$$

Since $F(x, y, z)$ is the square of the distance from (x, y, z) to the origin, it is clear that we are looking for the points on the ellipsoid at maximum and minimum distances from the center of the ellipsoid. The maximum occurs at the ends of the longest principal axis, namely at $(\pm 8, 0, 0)$. The minimum occurs at the ends of the shortest principal axis, namely at $(0, 0, \pm 5)$. Consider the maximum point $(8, 0, 0)$. The value of F at this point is 64, and the surface $F(x, y, z) = 64$ is a sphere. The sphere and the ellipsoid are tangent at $(8, 0, 0)$, as asserted in the general theory. In this case the ratios $G_1 : G_2 : G_3$ and $F_1 : F_2 : F_3$ at $(8, 0, 0)$ are $\frac{1}{4} : 0 : 0$ and $16 : 0 : 0$, respectively.

This example brings out the fact that the tangency of the surfaces, or the proportionality of the two sets of ratios, is a necessary but not a sufficient condition for a maximum or minimum value of F; for the condition of proportionality exists at the points $(0, \pm 6, 0)$, which are the ends of the principal axis of intermediate length. But the value of F is neither a maximum nor a minimum at this point, as the student can readily see by considering the geometrical situation.

A similar geometrical interpretation can be given to the problem of extremal values for $F(x, y)$ subject to a constraint $G(x, y) = k$. Here we have a curve defined by the constraint, and a one-parameter family of curves $F(x, y) = C$. At a point of extremal value of F the curve $F(x, y) = C$ through the point will be tangent to the curve defined by the constraint. Suitable hypotheses must be made to rule out singular cases.

Example 5. Find the maximum value of $F(x, y) = xy$ subject to the constraint $x^2 + y^2 = 8$.

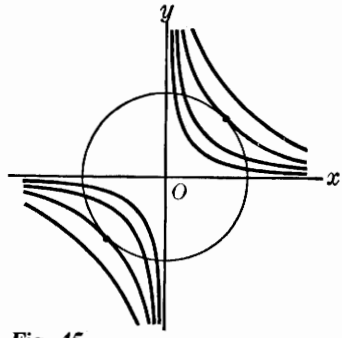


Fig. 45.

Here we have a fixed circle; the curves $xy = C$ are hyperbolas. The hyperbola $xy = 4$ is tangent to the circle at the points $(2, 2)$ and $(-2, -2)$ (see Fig. 45). These are the points at which the maximum value is attained.

An advantage of the geometrical point of view is that we can often use it to conclude with certainty that F actually attains an absolute maximum or minimum subject to the given constraint. Suppose, for example, that the surface S forms a closed and bounded set of points in space. Then, since F is continuous at the points of S, we can infer that F actually attains an absolute maximum and an absolute minimum on S. This inference is based on a theorem analogous to Theorem III, §3.2, and Theorem II, §5.3. The general theorem covering this situation is Theorem IV, §17.3.

Once we know that a maximum or minimum exists, we apply one of the methods which leads us to solve certain equations. If these equations have just a few solutions, as they often do in practice, it is an easy matter to tell which solutions give us the absolute maximum and which the absolute minimum

6.8 / LAGRANGE'S METHOD

In §6.7 we discussed extremal problems with constraints, and described two methods for handling such problems. We referred to these methods as (1) the method of direct elimination, and (2) the method of implicit functions. In this section we shall explain and illustrate a method known as Lagrange's method; it was devised by the great 18th-century mathematician Joseph Louis Lagrange. Our explanations depend upon Theorem V and the discussions leading up to it in §6.7.

In the case of the problem of extremal values of $F(x, y, z)$ subject to a constraint $G(x, y, z) = k$, Lagrange's method directs us to proceed as follows:

Form the function

$$u = F(x, y, z) + \lambda G(x, y, z), \tag{6.8–1}$$

where λ is a constant, as yet undetermined in value. Treat x, y, z as independent variables, and write down the conditions

$$\frac{\partial u}{\partial x} = 0, \frac{\partial u}{\partial y} = 0, \frac{\partial u}{\partial z} = 0. \tag{6.8–2}$$

Solve these three equations along with the equation of constraint

$$G(x, y, z) = k \tag{6.8–3}$$

to find the values of the four quantities x, y, z, λ. *More than one point* (x, y, z) *may be found in this way, but among the points so found will be the points of extremal values of* F.

To understand the reason for the validity of Lagrange's method, observe that the equations (6.8–2) are precisely

$$F_1 + \lambda G_1 = 0, F_2 + \lambda G_2 = 0, F_3 + \lambda G_3 = 0. \tag{6.8–4}$$

Here λ is a certain constant. These equations state, therefore, that at a point where they are satisfied, F_1, F_2, and F_3 are proportional to G_1, G_2, and G_2. But we know from Theorem V that such proportionality occurs at the points of the surface $G(x, y, z) = k$ where F has an extreme value. Thus the points of extreme value will be among those found by solving the four simultaneous equations (6.8–3) and (6.8–4). Thus Lagrange's method is justified in this type of problem.

The parameter λ occurring in Lagrange's method is called *Lagrange's multiplier.*

One of the great advantages of Lagrange's method over the method of implicit functions or the method of direct elimination is that it enables us to avoid making a choice of independent variables. This is sometimes very important; it permits the retention of symmetry in a problem where the variables enter symmetrically at the outset.

Example 1. Find the dimensions of the box of largest volume which can be fitted inside the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1, \tag{6.8–5}$$

assuming that each edge of the box is parallel to a co-ordinate axis.

Each of the eight corners of the box will lie on the ellipsoid. Let the corner in the first octant have co-ordinates (x, y, z); then the dimensions of the box are $2x$, $2y$, $2z$, and its volume is $V = 8xyz$. We wish to find the absolute maximum of V subject to the constraint (6.8–5). By the remarks at the end of §6.7 we know that an absolute maximum exists. Following Lagrange's method, we set

$$u = 8xyz + \lambda\left(\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2}\right).$$

The equations (6.8–2) in this case are

$$\begin{aligned} 8yz + 2\lambda \frac{x}{a^2} &= 0, \\ 8zx + 2\lambda \frac{y}{b^2} &= 0, \\ 8xy + 2\lambda \frac{z}{c^2} &= 0. \end{aligned} \tag{6.8–6}$$

After dividing by 2, let us multiply these three equations by x, y, z respectively, and add. In view of equation (6.8–5) the result is

$$12xyz + \lambda = 0, \text{ or } \lambda = -12xyz.$$

We now put this result back into the first of the equations (6.8–6), and obtain, after a slight simplification,

$$yz(a^2 - 3x^2) = 0.$$

By symmetry we obtain the two further equations

$$zx(b^2 - 3y^2) = 0, xy(c^2 - 3z^2) = 0.$$

For maximum V it is clear that we want *positive* values of x, y, and z. The only possible solutions of (6.8–6) meeting these requirements are thus seen to be

$$x = \frac{a}{\sqrt{3}}, y = \frac{b}{\sqrt{3}}, z = \frac{c}{\sqrt{3}},$$

and

$$\lambda = -12xyz = -\frac{4}{\sqrt{3}}abc.$$

The box of maximum volume therefore has dimensions

$$\frac{2a}{\sqrt{3}} \times \frac{2b}{\sqrt{3}} \times \frac{2c}{\sqrt{3}}.$$

Lagrange's method may be extended to the case of several constraints. The number of variables is immaterial. An undetermined multiplier is introduced corresponding to each constraint. We shall illustrate the extended method without formal proof.

Example 2. Find the semiaxes of the ellipse in which the plane

$$lx + my + nz = 0 \tag{6.8–7}$$

cuts the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1. \tag{6.8–8}$$

Since the plane passes through the origin, it is clear that the ends of the semiaxes of the ellipse are at the points (x, y, z) where the function

$$F(x, y, z) = x^2 + y^2 + z^2$$

is a maximum or a minimum subject to the side-conditions (6.8–7) and (6.8–8). We set up the expression

$$u = x^2 + y^2 + z^2 + \lambda_1(lx + my + nz) + \lambda_2\left(\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2}\right),$$

using two Lagrange multipliers, λ_1, λ_2. The method of Lagrange leads us to write

the equations $\frac{\partial u}{\partial x} = \frac{\partial u}{\partial y} = \frac{\partial u}{\partial z} = 0$, or

$$2x + \lambda_1 l + 2\lambda_2 \frac{x}{a^2} = 0,$$

$$2y + \lambda_1 m + 2\lambda_2 \frac{y}{b^2} = 0, \tag{6.8–9}$$

$$2z + \lambda_1 n + 2\lambda_2 \frac{z}{c^2} = 0.$$

Theoretically we may solve these three equations together with the two equations of constraint to find the values of x, y, z, λ_1, λ_2. Actually, the solution presents great difficulties. However, since we are merely interested in the lengths of the semiaxes of the ellipse, it will suffice to solve for the value of $F(x, y, z)$. Let us write $p^2 = F(x, y, z)$, so that p is the length of the semimajor or semiminor axis according as F has a maximum or a minimum.

If we multiply equations (6.8–9) by x, y, z respectively, and add, we obtain

$$2(x^2 + y^2 + z^2) + \lambda_1(lx + my + nz) + 2\lambda_2\left(\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2}\right) = 0.$$

In view of (6.8–7) and (6.8–8), this becomes

$$2p^2 + 2\lambda_2 = 0, \text{ or } \lambda_2 = -p^2.$$

Returning to (6.8–9) with this result, we have

$$2(a^2 - p^2)x + \lambda_1 a^2 l = 0, \tag{6.8–10}$$

and similar equations involving y and z. Except in special cases the situation will be such that x, y, z are all different from zero at the end of a semiaxis; and furthermore, p^2 will be different from a^2, b^2, and c^2. Thus in equation (6.8–10) we may assume $a^2 - p^2 \neq 0$ and $\lambda_1 \neq 0$ except in special cases. Thus

$$x = \lambda_1 \frac{a^2 l}{2(p^2 - a^2)}, \; y = \lambda_1 \frac{b^2 m}{2(p^2 - b^2)}, \; z = \lambda_1 \frac{c^2 n}{2(p^2 - c^2)}.$$

If we substitute these values in (6.8–7) and cancel the factor λ_1, we obtain the equation.

$$\frac{a^2 l^2}{p^2 - a^2} + \frac{b^2 m^2}{p^2 - b^2} + \frac{c^2 n^2}{p^2 - c^2} = 0. \tag{6.8–11}$$

When cleared of fractions this equation is a quadratic in p^2 as an unknown. The two roots of the equation thus determine the lengths of the semiaxes of the ellipse. For some consideration of the exceptional special cases see Exercise 18.

The foregoing problem illustrates very clearly the merits of Lagrange's method in the preservation of symmetry.

EXERCISES

1. A rectangular box lies in the first octant, with one corner at the origin and the diagonally opposite corner on the plane $(x/a)+(y/b)+(z/c)=1 (a, b, c>0)$. Find the maximum possible volume of the box.

2. Apply Lagrange's method in finding the extreme values of $x^2+y^2+z^2$ subject to the constraint $(x^2/a^2)+(y^2/b^2)+(z^2/c^2)=1$, where $a>b>c>0$.

3. A triangle is such that the product of the sines of its angles is a maximum. Show that the triangle is equilateral.

4. Find the maximum value of $xyz/(a^3x+b^3y+c^3z)$ subject to the conditions $xyz=A^3$, $x, y, z>0$ (a, b, c, A all >0).

5. Find the minimum of $x+y+z$ subject to the conditions $(a/x)+(b/y)+(c/z)=1$, $x, y, z>0$ (a, b, c and x, y, z all >0).

6. The perimeter of a triangle has a prescribed value $2s$. Determine the sides of the triangle so as to maximize the area.

7. Let $D=\begin{vmatrix} x & y \\ u & v \end{vmatrix}$. Find the maximum value of D^2 subject to the conditions $x^2+y^2=a^2$, $u^2+v^2=b^2$, where $a>0$, $b>0$. Solve the problem in two ways: (1) by Lagrange's method, and (2) by setting $x=a\cos\theta$, $y=a\sin\theta$, $u=b\cos\phi$, $v=b\sin\phi$, and using θ, ϕ as independent variables.

8. Find the minimum value of $x^3+y^3+z^3$ for positive x, y, and z, if it is required that $ax+by+cz=1$, where a, b, c are positive constants.

9. Suppose a, b, c are positive constants. If x, y, z are positive and $ayz+bzx+cxy=3abc$, show that $xyz \geqq abc$.

10. Solve the following problems by Lagrange's method: **(a)** Example 1, §6.3; **(b)** Example 3, §6.3; **(c)** Example 2, §6.7; **(d)** Example 3, §6.7.

11. Let the lengths of the sides of a fixed triangle of area A be a, b, c. From an interior point O draw the perpendiculars to the sides of the triangle, and let their lengths be x, y, z corresponding to a, b, c. If now a parallelepiped is constructed with edges x, y, z and volume V, show that V is a maximum when the lines from O to the vertices of the triangle divide it into three equal areas. What is the maximum value of V?

12. For the situation described in Exercise 11, show that

$$(x^2+y^2+z^2)(a^2+b^2+c^2) \geqq 4A^2.$$

13. Find the minimum distance from $(0, 0, c)$ to the cone $z^2=(x^2/a^2)+(y^2/b^2)$. Assume $c>0$ and $0<b<a$.

14. Given the ellipse $b^2x^2+a^2y^2=a^2b^2$, find the points (x, y) on the ellipse so that the line normal to the ellipse at (x, y) passes as far as possible from the origin. What is this greatest distance from the origin to a normal?

15. Find, by Lagrange's method, an extreme value of xyz subject to the conditions $(1/x)+(1/y)+(1/z)=c$, x, y, z all >0 (c a positive constant). Is this value an absolute maximum, an absolute minimum, or neither?

16. A particle is to travel from A to P and thence to B, by a broken line as indicated in Fig. 46. Velocity from A to P is v_1, and from P to B is v_2. Show by Lagrange's method that when the time of travel is least, $(\sin\theta_1)/(\sin\theta_2)=v_1/v_2$.

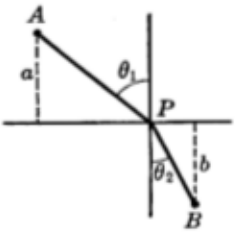


Fig. 46.

17. Let $p^2 = x^2 + y^2 + z^2$. Consider the extreme values of p^2, subject to the conditions $lx + my + nz = 0$, $(x^2 + y^2 + z^2)^2 = a^2x^2 + b^2y^2 + c^2z^2$, where $a > b > c > 0$. The minimum of p^2 is obviously zero. Show that, barring exceptional cases, the maximum is given by one of the roots of the equation

$$\frac{l^2}{p^2 - a^2} + \frac{m^2}{p^2 - b^2} + \frac{n^2}{p^2 - c^2} = 0.$$

Observe from the second constraint that the maximum of p^2 is less than a^2.

18. This exercise is devoted to one of the exceptional cases which can arise in connection with the derivation of (6.8–11) in Example 2. Suppose $a > b > 0$ and $b = c$, so that the ellipsoid has circular cross sections in planes perpendicular to the x-axis. Suppose also that $l \neq 0$. Show that, in this case, the minimum p^2 is b^2 and that the maximum p^2 is the sole root of (6.8–11).

19. Find the maximum and minimum values of the squared distance from the origin to the first octant portion of the curve in which the plane $x + y + z = 12$ meets the surface $xyz = 54$.

20. Find the minimum of $x_1^2 + \cdots + x_n^2$ subject to the constraint $a_1x_1 + \cdots + a_nx_n = 1$, where $a_1^2 + \cdots + a_n^2 > 0$.

21. Find the maximum of $(\Sigma_{i=1}^n a_ix_i)^2$ subject to the condition $\Sigma_{i=1}^n x_i^2 = 1$. Assume that at least one of the a's $\neq 0$.

22. If $P(x, y)$ is a point on the ellipse $256x^2 + 81y^2 = 2304$ and $Q(u, v)$ is a point on the line $4x + 3y = 24$, find the minimum value of the distance PQ.

23. Find the minimum distance between the curves $x^2 + y^2 = 1$, $x^2y = 16$.

24. Let $y = f(x)$ be the equation of a smooth curve, and let $P(a, b)$ be a point not on the curve. Let D be the distance from P to a variable point Q of the curve. Prove by Lagrange's method that PQ is normal to the curve when D is a minimum or a maximum.

25. **(a)** Use the result of Exercise 24 to give a simple solution of Exercise 23. **(b)** Proceed as directed in **(a)** to solve Exercise 22.

26. If $P(x_0, y_0, z_0)$ is a fixed point outside the ellipsoid $(x^2/a^2) + (y^2/b^2) + (z^2/c^2) = 1$, and Q is on the ellipsoid, show that the line PQ is normal to the ellipsoid when the distance PQ is a minimum.

27. Let $G(x, y, z) = 0$ define a surface S with a tangent plane at each point, and let $P(a, b, c)$ be a point not on S. Show that, if Q is a point on S, the line PQ is normal to S whenever the distance PQ attains a relative extreme. From this result prove that, if A and B are variable points on two smooth nonintersecting surfaces, then, when A and B are located so as to make the distance between them an absolute minimum, the line AB is normal to both surfaces.

28. Show that the maximum value of $x^2y^2z^2$ subject to $x^2 + y^2 + z^2 = R^2$ is $(R^2/3)^3$. Deduce from this that

$$(x^2y^2z^2)^{1/3} \leqq \frac{x^2 + y^2 + z^2}{3}$$

for all values of x, y, z. Generalize this result to n variables, and so obtain a proof of the inequality

$$(a_1a_2 \cdots a_n)^{1/n} \leqq \frac{a_1a_2 + \cdots + a_n}{n}$$

between the geometric mean and arithmetic mean of n positive numbers. What condition is both necessary and sufficient for equality?

29. In the case of two positive numbers, Exercise 28 gives

$$u^{1/2}v^{1/2} \leqq \tfrac{1}{2}u + \tfrac{1}{2}v.$$

Derive from this the fact that

$$\left|\sum_{i=1}^{n} a_i b_i\right| \leqq \left(\sum_{i=1}^{n} a_i^2\right)^{1/2}\left(\sum_{i=1}^{n} b_i^2\right)^{1/2}$$

if the a's and b's are any real numbers. This is known as *Cauchy's inequality*. Hint: Let

$$u = \frac{a_i^2}{A} \quad \text{and} \quad v = \frac{b_i^2}{B}, \quad \text{where}$$

$$A = \sum_{i=1}^{n} a_i^2 \quad \text{and} \quad B = \sum_{i=1}^{n} b_i^2.$$

What condition is both necessary and sufficient for equality?

30. If $\alpha_1, \alpha_2, \ldots, \alpha_n$ are nonnegative numbers whose sum is 1, then $\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$ is called a *weighted arithmetic mean* of the x's. If the x's are positive, then one can define the corresponding *weighted* geometric mean to be $x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$. Show that if all the x's are positive, the weighted geometric mean is always less than or equal to the corresponding weighted arithmetic mean. Hint: Maximize $x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$ subject to the constraint $\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = C$. This will show that for all positive x's having a given weighted arithmetic mean C, the weighted geometric mean is less than or equal to C.

Notice that this result contains that of Exercise 28 as a special case.

31. If u, v, p, and q are positive and $(1/p) + (1/q) = 1$, then Exercise 30 gives

$$u^{1/q}v^{1/q} \leqq \frac{1}{p}u + \frac{1}{q}v.$$

Deduce from this that under the above restrictions on p and q,

$$\left|\sum_{i=1}^{n} a_i b_i\right| \leqq \left(\sum_{i=1}^{n} |a_i|^p\right)^{1/p}\left(\sum_{i=1}^{n} |b_i|^q\right)^{1/q},$$

where the a's and b's are any real numbers. This is known as *Hölder's inequality*. Notice that it is a generalization of Cauchy's inequality.

Hint: Let

$$u = \frac{|a_i|^p}{\sum_{i=1}^{n} |a_i|^p} \quad \text{and} \quad v = \frac{|b_i|^q}{\sum_{i=1}^{n} |b_i|^q}.$$

32. Use Hölder's inequality (Exercise 30) to deduce the inequality of Minkowski:

$$\left(\sum_{i=1}^{n} |x_i + y_i|^p\right)^{1/p} \leqq \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p\right)^{1/p}, \quad p > 1.$$

Write $|x_i + y_i|^p = |x_i + y_i||x_i + y_i|^{p/q} \leqq |x_i||x_i + y_i|^{p/q} + |y_i||x_i + y_i|^{p/q}$, and then apply Hölder's inequality, once with $a_i = |x_i|$ and once with $a_i = |y_i|$. Observe that Minkowski's inequality obviously holds when $p = 1$.

6.9 / QUADRATIC FORMS

Another application of Lagrange multipliers is to the study of central conics and their generalizations. A central conic is simply one that has a center—that is, an ellipse or a hyperbola. The parabola, which has no center, will not be treated in this approach. Our problem is that of recognizing and describing a central conic from its equation. Since this problem for the parabola is comparatively easy, its omission from our consideration here is not serious. When we describe a central conic in terms of a co-ordinate system whose origin coincides with the center, the equation looks like

$$Ax^2 + 2Bxy + Cy^2 = k. \tag{6.9–1}$$

The left-hand side,

$$Q(x, y) = Ax^2 + 2Bxy + Cy^2$$

is called a *quadratic form*. We can define a quadratic form without reference to conics by saying that it is a homogeneous polynomial of second degree in two or more variables. Homogeneity means the same here as in §6.53. A quadratic form in three variables looks like

$$Q(x, y, z) = Ax^2 + By^2 + Cz^2 + 2Dxy + 2Exz + 2Fyz, \tag{6.9–2}$$

and so on for any number of variables. We shall be mainly concerned with two variables, but quadratic forms in three variables will be given some attention.

Beginning with the simplest case, the problem is to plot the graph of the equation (6.9–1). This is easy if B happens to be zero and (6.9–1) reduces to

$$Ax^2 + Cy^2 = k.$$

If A and C are positive we have an ellipse if k is positive, the point $(0, 0)$ if $k = 0$, and no graph at all if $k < 0$. If A and C are of opposite signs, we have a hyperbola if $k \neq 0$ and two straight lines if $k = 0$. The case where $A = 0$ or $C = 0$ is trivial.

The ease with which these special cases can be treated suggests that when confronted by (6.9–1) we try to find some other co-ordinate system in terms of which the coefficient B is reduced to zero. If we hold the origin fixed and rotate the axes through an angle θ, then the co-ordinates in the new system—say the x', y'-system—are related to those in the x, y-system by the equations

$$\begin{aligned} x &= x' \cos\theta - y' \sin\theta, \\ y &= x' \sin\theta + y' \cos\theta. \end{aligned} \tag{6.9–3}$$

Substituting for x and y in (6.9–1) the new equation of the graph can be written (after some simplification) in the form

$$A'(x')^2 + 2B'x'y' + C'(y')^2 = k$$

where A', B' and C' are functions of A, B, C, and θ.

At this point in an introductory analytic geometry course, two very im-

portant results are established. The first is that by proper choice of θ, B' can be made equal to zero and the equation thereby reduced to canonical form,

$$A'(x')^2 + C'(y')^2 = k. \tag{6.9–4}$$

The second is a far-reaching and surprising identity relating the original coefficients and the new ones. Although A', B', and C' won't look at all like A, B, and C, it is nevertheless true that

$$A'C' - (B')^2 = AC - B^2 = \begin{vmatrix} A & B \\ B & C \end{vmatrix} \tag{6.9–5}$$

no matter what angle of rotation is used. This result is called the invariance of the discriminant because it says that $AC^2 - B^2$—the *discriminant* of the quadratic form—remains unchanged under all rotations of the co-ordinate axes. These facts from analytic geometry play an essential role in what is to follow.

We recall from our elementary courses that finding the proper angle of rotation and then going through the steps of actually reducing a quadratic form to canonical form is rather tedious. By using Lagrange multipliers we are now going to develop a way of obtaining the canonical form without going through so much computation.

Suppose that we are given equation (6.9–1) and asked to describe its graph. We can begin by observing that since quadratic forms are continuous, and the unit circle is closed and bounded, it follows from Theorem II, §5.3 that $Q(x, y)$, subject to the constraint $x^2 + y^2 = 1$, has both a maximum and minimum. Let (x_1, y_1) be the co-ordinates in the x, y-system of a point where $Q(x, y)$ takes on one of these extreme values. We could then rotate the co-ordinate axes to where the positive axis of abscissas passed through the point, as in Fig. 47. Then in the rotated co-ordinate system the new co-ordinates of this point would be $x' = 1$, $y' = 0$. The quadratic form $Q(x, y)$ would be transformed into

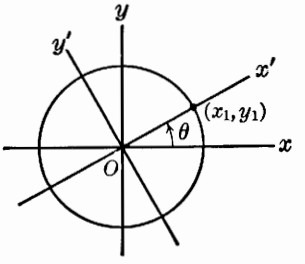


Fig. 47.

$$Q'(x', y') = A'(x')^2 + 2B'x'y' + C'(y')^2.$$

The equation of the unit circle would be the same in the rotated co-ordinate system as in the original system, $(x')^2 + (y')^2 = 1$, and we would know that the function $Q'(x', y')$ would have an extreme value, subject to the constraint $(x')^2 + (y')^2 = 1$ at $x' = 1$, $y' = 0$. In terms of the associated Lagrange function

$$L'(x', y') = A'(x')^2 + 2B'x'y' + C'(y')^2 - \lambda[(x')^2 + (y')^2]$$

this means that the two equations

$$\begin{aligned} \frac{1}{2}\frac{\partial L'}{\partial x'} &= (A' - \lambda)x' + B'y' = 0 \\ \frac{1}{2}\frac{\partial L'}{\partial y'} &= B'x' + (C' - \lambda)y' = 0 \end{aligned} \tag{6.9–6}$$

must be satisfied by $x' = 1$, $y' = 0$. Substituting these two values into the second equation reveals that $B' = 0$ and therefore that we have hit upon a rotation which reduces $Q(x, y)$ to the canonical form

$$Q'(x', y') = A'(x')^2 + C'(y')^2,$$

and the transformed equation of the conic is the canonical form (6.9–4).

It is the invariance of the discriminant which now allows us to find A' and C' easily. Notice that the Lagrange function, $L(x, y) = Ax^2 + 2Bxy + Cy^2 - \lambda(x^2 + y^2)$ associated with $Q(x, y)$ is itself a quadratic form having as its discriminant $(A - \lambda)(C - \lambda) - B^2$. This can be written as the determinant

$$\begin{vmatrix} A - \lambda & B \\ B & C - \lambda \end{vmatrix}.$$

We have just seen that the rotation which transforms the co-ordinates (x_1, y_1) into $(1, 0)$ transforms $L(x, y)$ into

$$L'(x', y') = A'(x')^2 + C'(y')^2 - \lambda[(x')^2 + (y')^2]$$

which is a quadratic form with discriminant $(A' - \lambda)(C' - \lambda)$. By (6.9–5) we have that

$$\begin{vmatrix} A - \lambda & B \\ B & C - \lambda \end{vmatrix} = (A' - \lambda)(C' - \lambda).$$

The discriminant of $L(x, y)$ is clearly a second-degree polynomial in λ whose leading coefficient is 1. Each such polynomial has the unique factorization $(\lambda - r_1)(\lambda - r_2)$ where r_1 and r_2 are the zeros of the polynomial. From this it follows that the coefficients A' and C' in the canonical form of the quadratic form $Q(x, y)$ are simply the two roots of the quadratic equation

$$\begin{vmatrix} A - \lambda & B \\ B & C - \lambda \end{vmatrix} = 0. \tag{6.9–7}$$

These two roots can of course be found as soon as equation (6.9–1) is given, and the equation of the conic reduced immediately to canonical form, without going through again each time our fairly lengthy argument which justifies the process.

Since there are two roots, the reader might wonder which to call A' and which C'. The answer is that it doesn't matter—there are two equally good canonical forms to which every quadratic form in two variables (and every central conic) can be reduced. This ambiguity was presaged when we began by picking a point (x_1, y_1) where one of the extreme values of $Q(x, y)$ on the unit circle was taken on. For our purposes at that time—the elimination of the xy-term—it made no difference whether this point gave a maximum or a minimum value to Q. The only difference it makes now is that if this point, whose co-ordinates in the primed system are $x' = 1$, $y' = 0$, is a point of maximum, then A' is the larger of the two roots. If this point (x_1, y_1) to which we rotate the positive axis of abscissas gives a minimum, then A' is the smaller of the two roots. We see this from the obvious fact that if $\alpha > \beta$, then the maximum value

of $\alpha x^2 + \beta y^2$ on the unit circle is α, and this value is taken on at $(1, 0)$—as well as at $(-1, 0)$ of course. If $\alpha < \beta$ then α is the minimum value, and it is taken on at $(1, 0)$ and $(-1, 0)$. The following example shows how easy it is to find the canonical form of the equation of a central conic by this method.

Example 1. Find the dimensions of the ellipse $73x^2 + 72xy + 52y^2 = 100$.
Here $A = 73$, $B = 36$, $C = 52$.
The equation (6.9–7) becomes

$$\lambda^2 - 125\lambda + 2500 = 0,$$

with roots $\lambda = 25$, $\lambda = 100$. This means, therefore, that the equation of the ellipse can be put in the form

$$25x'^2 + 100y'^2 = 100, \text{ or } \frac{x'^2}{4} + y'^2 = 1,$$

by a rotation of axes. The principal semiaxes of the ellipse are therefore 2 and 1.

The generalization to the case of three variables is now a rather easy matter. The essential statement of the generalization is this: Given the quadratic form $Q(x, y, z)$ as in (6.9–2), it is possible to make a rotation of the co-ordinate axes so that $Q(x, y, z)$ becomes

$$G(x', y', z') = \lambda_1 x'^2 + \lambda_2 y'^2 + \lambda_3 z'^2, \tag{6.9–8}$$

and the λ's are the roots of the equation

$$\begin{vmatrix} A - \lambda & D & E \\ D & B - \lambda & F \\ E & F & C - \lambda \end{vmatrix} = 0. \tag{6.9–9}$$

A proof of this statement is indicated in Exercise 14. If one wishes, matters may be arranged so that $\lambda_1 \geqq \lambda_2 \geqq \lambda_3$.

Example 2. Reduce the quadratic form $xy + xz - yz$ to the form (6.9–8) and so identify the surface $xy + xz - yz = -1$.
The equation (6.9–9) becomes

$$\begin{vmatrix} -\lambda & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\lambda & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\lambda \end{vmatrix} = 0$$

in this case. On expansion and simplification this becomes

$$-\tfrac{1}{4} + \tfrac{3}{4}\lambda - \lambda^3 = 0,$$

the roots of which are found to be $\frac{1}{2}$, $\frac{1}{2}$, -1. Thus, after a suitable rotation, our equation becomes

$$\tfrac{1}{2}x'^2 + \tfrac{1}{2}y'^2 - z'^2 = -1.$$

This defines a hyperboloid of two sheets with circular cross sections perpendicular to the z'-axis (for $|z'| > 1$).

A more symmetrical notation for quadratic forms is in some ways extremely desirable. If we write x_1, x_2, x_3 instead of x, y, z, a quadratic form in x_1, x_2, x_3 will have terms of all possible types x_ix_j, with i and j assuming the values 1, 2, 3. If we write a_{ij} for the coefficient of x_ix_j, the quadratic form will have the appearance

$$\begin{aligned} F(x_1, x_2, x_3) = \quad & a_{11}x_1^2 + a_{12}x_1x_2 + a_{13}x_1x_3 \\ & + a_{21}x_2x_1 + a_{22}x_2^2 + a_{23}x_2x_3 \\ & + a_{31}x_3x_1 + a_{32}x_3x_2 + a_{33}x_3^2. \end{aligned} \tag{6.9–10}$$

Since $x_1x_2 = x_2x_1$, we agree to make $a_{12} = a_{21}$ = half the total coefficient of x_1x_2; similarly for a_{13} and a_{23}. The *discriminant* of the form is, by definition, the determinant

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}. \tag{6.9–11}$$

The determinant appearing in (6.9–9) is seen to be the discriminant of the form $Q(x, y, z) - \lambda(x^2 + y^2 + z^2)$.

EXERCISES

1. Find the dimensions of the ellipse $41x^2 - 24xy + 34y^2 = 25$.

2. Show that $F(x, y) = x^2 - 4xy - 2y^2 = 1$ is the equation of a hyperbola. Find a point on the unit circle at which $F(x, y)$ is a maximum. Then draw the xy-axes, the axes of symmetry of the hyperbola, and the hyperbola itself.

3. Find the maximum and minimum values of $F(x, y) = 9x^2 - 6xy + y^2$ on the circle $x^2 + y^2 = 1$. If λ_1 is the maximum value, show that $F(x, y) = \lambda_1$ is the equation of two lines, each tangent to the circle at a point where the maximum occurs.

4. Let $F(x, y, z) = y^2 + z^2 - \sqrt{2}xy + \sqrt{2}xz + 2yz$. Find the maximum and minimum values of this function on the surface of the unit sphere. What does (6.9–8) become in this case?

5. Reduce each of the following quadratic forms to the standard form (6.9–8);

(a) $y^2 + z^2 - \sqrt{2}xy - \sqrt{2}xz + 2yz$.

(b) $13x^2 + 13y^2 + 10z^2 + 8xy - 4xz - 4yz$.

6. **(a)** Put $F(x, y, z) = xy + yz + zx$ in the form (6.9–8).

(b) What is the maximum value of F on the unit sphere?

(c) At what points does it occur?

7. Determine the signs of $\lambda_1, \lambda_2, \lambda_3$ for each of the following quadratic forms, and so identify the type of each quadric surface. You need not find the actual value of the λ's.

(a) $x^2 + xy + yz = 1$.

(b) $yz + xz - xy = 1$.

(c) $x^2 + 2y^2 + 3z^2 - 2xy - 2yz = 2$.

8. Find maximum and minimum values of $17x^2 - 30xy + 17y^2$ when $5x^2 - 6xy + 5y^2 = 4$.

9. Find the minimum value of $x^2 + y^2 + xy$ subject to the condition $2x^2 + 6xy + 2y^2 = 9$.

10. Suppose that the locus of $Ax^2 + 2Bxy + Cy^2 = 1$ is an ellipse. Consider the problem of locating the maximum and minimum values of $x^2 + y^2$ on the ellipse. Apply Lagrange's method, starting with the expression $x^2 + y^2 - \lambda(Ax^2 + 2Bxy + Cy^2)$. Show that, in the resulting equations for locating the extreme values, λ must be a root of the equation

$$\begin{vmatrix} 1 - \lambda A & -B\lambda \\ -B\lambda & 1 - C\lambda \end{vmatrix} = 0,$$

and that the roots of this equation are the extreme values of $x^2 + y^2$. What is the relation between these roots and the semiaxes of the ellipse?

11. If x, y, λ are solutions of the simultaneous equations

$$\begin{aligned} (A - \lambda)x + By &= 0, \\ Bx + (C - \lambda)y &= 0, \end{aligned} \qquad x^2 + y^2 = 1,$$

show that $Ax^2 + 2Bxy + Cy^2 = \lambda$. Hence show that, in the notation used in the text, λ_1 and λ_2 are respectively the maximum and minimum values of $F(x, y)$ when $x^2 + y^2 = 1$.

12. **(a)** Assume that $F(x, y) = k (k > 0)$ defines a family of ellipses. Let λ_1 and λ_2 be the roots of (6.9–7), with $\lambda_1 \geqq \lambda_2$. Show that the ellipse $F(x, y) = \lambda_1$ is externally tangent to the circle $x^2 + y^2 = 1$ at the ends of the minor axis of the ellipse, and that the ellipse $F(x, y) = \lambda_2$ is internally tangent to the circle at the ends of the major axis of the ellipse. Draw a figure showing these two ellipses, the circle, and the $x'y'$-axes. What can you infer from the figure about maximum and minimum values of $F(x, y)$ on the circle?

(b) Assume that $F(x, y) = k$ defines a family of hyperbolas. If λ_1 and λ_2 are the roots of (6.9–7) with $\lambda_1 \geqq \lambda_2$, explain why $\lambda_1 > 0$ and $\lambda_2 < 0$. Draw a figure showing the $x'y'$-axes, the unit circle, the hyperbolas $F(x, y) = \lambda_1$, $F(x, y) = \lambda_2$, and other members of the family.

13. Prove (6.9–5) by actually substituting (6.9–3) in (6.9–1) and computing A', B', C'. Also prove that $A' + C' = A + C$.

14. **(a)** Suppose that $Q(x, y, z)$ in (6.9–2) becomes a new form $G(x', y', z')$ with coefficients $A', B', \ldots, F'$ when we shift to new axes $x'y'z'$ obtained by a rotation from the xyz-system. Write the equations which correspond to (6.9–6) for the problem of making $G(x', y', z')$ a maximum on the unit sphere. If the new axes are chosen so that this maximum occurs at $x' = 1$, $y' = z' = 0$, show that $D' = E' = 0$, and that $G(x', y', z') = \lambda_1 x'^2 + B'y'^2 + C'z'^2 + 2F'y'z'$, where λ_1 is the maximum value of G on the unit sphere. Now explain how it is possible to choose a new set of axes x'', y'', z'', by a rotation from the $x'y'z'$-system, rotating about the x'-axis, in such a way that the form becomes $\lambda_1 x''^2 + \lambda_2 y''^2 + \lambda_3 z''^2$, where λ_2 and λ_3 are respectively the maximum and minimum values of G subject to the two constraints $x' = 0$, $y'^2 + z'^2 = 1$.

(b) It may be proved algebraically that the discriminant of a quadratic form $Q(x, y, z)$ is equal to the discriminant of the new form $G(x', y', z')$ which is obtained from $Q(x, y, z)$ by a rotation of axes. Use this fact to prove that the numbers λ_1, λ_2, λ_3 described in part **(a)** of this problem are the roots of the cubic equation (6.9–9).

MISCELLANEOUS EXERCISES

1. Choose a and b so that $\int_0^1 (\sqrt{x} - a - bx)^2 \, dx$ is as small as possible; $a + bx$ is then called a "least-square" approximation to $\sqrt{x}$ in the interval $[0, 1]$.

2. If $\phi(u, v) = f(x, y)$, where $u = y^2 - x^2$, $v = y^2 + x^2$, show that

$$\frac{1}{4xy}\frac{\partial^2 f}{\partial x\,\partial y} = \frac{\partial^2 \phi}{\partial v^2} - \frac{\partial^2 \phi}{\partial u^2}.$$

3. (**a**) Show by a diagram the part of the xy-plane in which $f(x, y) = (a - x)(a - y)(x + y - a) \geqq 0$. Assume $a > 0$. (**b**) Find all points of the plane at which $f_1(x, y) = f_2(x, y)$. (**c**) Which of these points yield relative maxima or minima of $f(x, y)$, and which do not? (**d**) Does f have any absolute extrema?

4. Let $S = 32\sqrt{3}\left(\frac{1}{x} + \frac{1}{y}\right) + xy(\sec\theta - \frac{1}{2}\tan\theta)$. Find the minimum value of S in the region $x > 0$, $y > 0$, $0 \leqq \theta < \pi/2$ of $xy\theta$-space. How do you know that the minimum does not occur when $\theta = 0$?

5. Find the point of occurrence of the maximum value of x^2yz^3 on the part of the plane $x + y + z = 24$ that lies in the first octant.

6. Find the maximum and minimum values of $x^2 + y^2 + z^2$ subject to the two conditions $(x^2/25) + (y^2/25) + (z^2/9) = 1$, $x + y + 2z = 0$.

7. Find the ratios y/x and x/z to make $x^2 + 2xy + x\sqrt{x^2 + z^2}$ a minimum when $3x^2y + x^2z$ has a prescribed positive value (x, y, z all > 0).

8. Find the maximum and minimum values of z subject to the conditions $x^2 + y^2 + z^2 = 25$, $(x - 2\sqrt{6})^2 + (y - 2\sqrt{3})^2 + (z - 6)^2 = 13$.

9. Suppose a, b, c, p given all positive, and $p < 1$. Find the maximum of $ax^p + by^p + cz^p$ subject to the conditions $x + y + z = 1$, x, y, $z > 0$.

10. Consider $S = 2(xy + yz + zx)$, subject to the conditions $ax + by + cz = 2A$ and x, y, z are all positive, where $A > 0$ and $0 < a < b + c$, $0 < b < c + a$, $0 < c < a + b$. Show that the maximum possible value of S is $8A^2[2(ab + bc + ca) - (a^2 + b^2 + c^2)]^{-1}$.

11. Find the absolute maximum and minimum values of

$$f(x, y) = 3x^2 - 2(y + 1)x + 3y - 1$$

in the square $0 \leqq x \leqq 1$, $0 \leqq y \leqq 1$.

12. (**a**) Find the minimum distance from the point (3, 4, 15) to the cone $4z^2 = x^2 + y^2$.
(**b**) Find the minimum distance from the point (9, 12, −5) to the cone $4z^2 = x^2 + y^2$.
(**c**) Are there any relative extrema in addition to the absolute minimum in part (**a**)? in part (**b**)?

7 / GENERAL THEOREMS OF PARTIAL DIFFERENTIATION

7 / PRELIMINARY REMARKS

This chapter is not primarily concerned with the technique of partial differentiation, but with statements and proofs of some of the important theorems about differentials and partial differentiation. We have separated the material of the chapter from that of Chapter 6 for a number of reasons. In studying the subject of partial differentiation the student needs first of all to get acquainted with the new ideas which the subject presents to him. He needs to assimilate these ideas through the medium of examples and problems. He will want a reasoned development of the subject, but he will be more interested in mastery of technique and appreciation of some applications than in the details of the longer proofs, particularly as regards the proofs of theorems which he is quite willing to take for granted in the early stages. The theorems of §§7.1, 7.2 and 7.3 have been referred to in Chapter 6. The student needs to know these theorems, but he can very well go through Chapter 6 without studying their proofs.

Sections 7.4 and 7.5 are on a somewhat different footing. Every student who uses advanced calculus is likely to have need of Taylor's formula for a function of several variables. One meets references to the formula frequently in the literature of applied mathematics and in various branches of higher analysis. The law of the mean is merely a special case of Taylor's formula. We have put this material here rather than in Chapter 6 because its applications are not so immediate.

The final section of the chapter deals with tests for maximum or minimum values of a function in terms of the second partial derivatives. These tests are of great importance—both practical and theoretical.

Within the limits of time of an ordinary year course in advanced calculus the instructor may wish to make only a selection from Chapter 7 with such a degree of emphasis on the proofs as he or she sees fit. The various sections are practically independent of each other, except that §§7.4 and 7.5 go together.

The definition of differentiability given in §6.4 can be reformulated slightly in equivalent ways that we shall find useful. We shall give the reformulations for a function of n variables. We use the notation

$$\|h\| = (h_1^2 + \cdots + h_n^2)^{1/2} \tag{7-1}$$

introduced in connection with (6.4–17).

A function f of n variables that is defined in a neighborhood of the point $(x_1, \ldots, x_n)$ is differentiable at $(x_1, \ldots, x_n)$ if and only if (1) all of the first partial derivatives of f exist at $(x_1, \ldots, x_n)$ and (2) we can write the formula

$$f(x_1 + h_1, \ldots, x_n + h_n) - f(x_1, \ldots, x_n) = \sum_{i=1}^{n} f_i(x_1, \ldots, x_n)\, h_i + \epsilon \|h\|, \qquad (7\text{–}2)$$

where ϵ is a variable quantity depending on $h_1, \ldots, h_n$, with value 0 when $h_1 = \cdots = h_n = 0$, and such that $\epsilon \to 0$ when $\|h\| \to 0$.

This formulation is obviously equivalent to that stated in connection with (6.4–17), because the A_i's in (6.4–17) must of necessity be given by $A_i = f_i(x_1, \ldots, x_n)$ [see (6.4–18)].

Another equivalent formulation is obtained if we replace $\|h\|$ in (7–2) by $\|h\|_*$, where

$$\|h\|_* = |h_1| + \cdots + |h_n|. \qquad (7\text{–}3)$$

The reason for the equivalence is that

$$\|h\| \le \|h\|_* \le \sqrt{n}\,\|h\|, \qquad (7\text{–}4)$$

or, in more explicit form.

$$(h_1^2 + \cdots + h_n^2)^{1/2} \le |h_1| + \cdots + |h_n| \le \sqrt{n}\,(h_1^2 + \cdots + h_n^2)^{1/2} \qquad (7\text{–}5)$$

The first of these inequalities is easily proved, for if we calculate $(|h_1| + \cdots + |h_n|)^2$, we obtain all of the terms $h_1^2, \ldots, h_n^2$ plus additional terms, none of them negative. The second inequality in (7–5) is a consequence of Cauchy's inequality, given in Exercise 29, §6.8. (In that inequality put $a_i = |h_i|$ and $b_i = 1$.)

Because of the foregoing we can see that it does not matter whether we write $\epsilon\|h\|$ or $\epsilon_*\|h\|_*$ in (7.2) (with $\epsilon_* \to 0$ as $\|h\|_* \to 0$). For we see by (7.4) that $\|h\| \to 0$ is equivalent to $\|h\|_* \to 0$, and if $\epsilon\|h\| = \epsilon_*\|h\|_*$, we see that $\epsilon \to 0$ is equivalent to $\epsilon_* \to 0$ because $\epsilon \le \epsilon_*\sqrt{n}$ and $\epsilon_* \le \epsilon$ (with $\epsilon = \epsilon_* = 0$ if all the h_i's are 0).

7.1 / SUFFICIENT CONDITIONS FOR DIFFERENTIABILITY

The concept of differentiability for a function of several variables was defined in §6.4. To be differentiable at a given point a function must have first partial derivatives at that point. But this alone is not enough. We may have a function $f(x, y)$ such that $f_1(0, 0)$ and $f_2(0, 0)$ exist, and yet such that f is not differentiable at $(0, 0)$; for an illustration see Exercise 7, §6.4. The following theorem deals with sufficient conditions for differentiability:

THEOREM I. *Suppose the function $f(x, y)$ is defined in some neighborhood of the point (a, b). Suppose one of the partial derivatives, say $\dfrac{\partial f}{\partial x}$, exists at each point of the neighborhood and is continuous at (a, b), while the other partial derivative is defined at least at the point (a, b). Then f is differentiable at (a, b).*

Proof. We shall use one of the formulations of differentiability from the preceding section. We shall show that we can write

$$f(a+h, b+k) - f(a, b) = f_1(a, b)h + f_2(a, b)k + \epsilon(|h| + |k|), \qquad (7.1\text{–}1)$$

where $\epsilon \to 0$ as $(h, k) \to (0, 0)$. We work with small values of h and k, and we express the left side of (7.1–1) as the sum of two differences:

$$f(a+h, b+k) - f(a, b) = f(a+h, b+k) - f(a, b+k) + f(a, b+k) - f(a, b). \qquad (7.1\text{–}2)$$

Next we apply the law of the mean (Theorem IV, §1.2) to $f(x, b+k)$ as a function of x alone. The result is that there is a point between $x = a$ and $x = a + h$, which we may denote by $x = a + \theta h$, where $0 < \theta < 1$, such that

$$f(a+h, b+k) - f(a, b+k) = f_1(a+\theta h, b+k)h.$$

Because f_1 is assumed to be continuous at (a, b), we can write

$$f_1(a+\theta h, b+k) = f_1(a, b) + \epsilon_1,$$

where $\epsilon_1 \to 0$ as $h \to 0$ and $k \to 0$. Thus we have

$$f(a+h, b+k) - f(a, b+k) = f_1(a, b)h + \epsilon_1 h,$$

and we can put the expression on the right of the equality sign here in place of the first difference on the right in (7.1–2). For the other difference on the right in (7.1–2) we use the definition of $f_2(a, b)$ as the limit of a quotient to write (when $k \neq 0$)

$$\frac{f(a, b+k) - f(a, b)}{k} = f_2(a, b) + \epsilon_2,$$

where $\epsilon_2 \to 0$ as $k \to 0$. We define ϵ_2 to be 0 if $k = 0$. This permits us to replace the second difference on the right in (7.1–2) by $f_2(a, b)k + \epsilon_2 k$. The result is that we have

$$f(a+h, b+k) - f(a, b) = f_1(a, b)h + f_2(a, b)k + \epsilon_1 h + \epsilon_2 k.$$

To bring this equation into the form of (7.1–1) we define

$$\epsilon = \frac{\epsilon_1 h + \epsilon_2 k}{|h| + |k|} \quad \text{if} \quad |h| + |k| \neq 0$$

and $\epsilon = 0$ if $|h| + |k| = 0$. Then, if $|h| + |k| \neq 0$,

$$|\epsilon| \leq |\epsilon_1| \frac{|h|}{|h| + |k|} + |\epsilon_2| \frac{|k|}{|h| + |k|} \leq |\epsilon_1| + |\epsilon_2|,$$

because, clearly, $|h| \leq |h| + |k|$ and $|k| \leq |h| + |k|$. But now it is evident that $\epsilon \to 0$ when $|h| + |k| \to 0$, because $\epsilon_1 \to 0$ and $\epsilon_2 \to 0$. The proof is now complete.

It will be observed that the conditions of the theorem are not symmetrical as regards x and y. We might equally well have assumed the mere existence of $f_1(a, b)$, while assuming the continuity of $f_2(x, y)$ at (a, b). In general, for a

function of more than two variables, we assume the mere existence of one of the first partial derivatives, and the continuity of the other first partial derivatives. We may then conclude that the function is differentiable. The proof is similar to that of Theorem I. For most practical purposes the following statement is sufficient:

THEOREM II. *A function of several variables is differentiable at a point if the function and all its first partial derivatives are defined in some neighborhood of the point and if these derivatives are continuous at the point.*

EXERCISES

1. Let $f(x, y) = (x^4 + y^4)/(x^2 + y^2)$ if $x^2 + y^2 \neq 0$, and define $f(0, 0) = 0$. Show that f has first partial derivatives at all points, satisfying the inequalities $|f_1(x, y)| \leqq 6|x|$, $|f_2(x, y)| \leqq 6|y|$. Is f differentiable at $(0, 0)$?

2. Let $f(x, y) = (x^3 - y^3)/(x^2 + y^2)$ if $x^2 + y^2 \neq 0$, and define $f(0, 0) = 0$. Show that f has first partial derivatives at all points, but that these derivatives are discontinuous at $(0, 0)$. The function is *not* differentiable at $(0, 0)$. To prove this, show first that if it *were* differentiable, one could write

$$\frac{x^3 - y^3}{x^2 + y^2} = x - y + \epsilon(|x| + |y|),$$

where $\epsilon \to 0$ as $(x, y) \to (0, 0)$. Then show that this is impossible. SUGGESTION: Consider the situation when $y = -x$.

3. Let $f(x, y) = (x^2 + y^2) \sin \dfrac{1}{\sqrt{x^2 + y^2}}$ if $x^2 + y^2 \neq 0$, and define $g(0, 0) = 0$. Show that f has first partial derivatives at all points, but that these derivatives are discontinuous at $(0, 0)$. Show that $|f_1(x, y)| \leqq 2|x| + 1$. Prove that f is differentiable at $(0, 0)$.

This example shows that the hypotheses in Theorems I and II are sufficient, but not necessary, conditions for differentiability.

4. Prove that the function

$$f(x, y) = \frac{xy}{(x^2 + y^2)^2} \quad \text{if} \quad (x, y) \neq (0, 0)$$

$$f(0, 0) = 0$$

satisfies Laplace's equation, $\dfrac{\partial^2 f}{\partial x^2} + \dfrac{\partial^2 f}{\partial y^2} = 0$, everywhere, but that f is not even continuous (let alone differentiable) at the origin.

7.2 / CHANGING THE ORDER OF DIFFERENTIATION

We mentioned at the outset of Chapter 6 that we ordinarily find the relation

$$\frac{\partial^2 f}{\partial y\, \partial x} = \frac{\partial^2 f}{\partial x\, \partial y} \tag{7.2–1}$$

to be valid for the functions $f(x, y)$ which we meet in everyday use of calculus. The relation (7.2–1) may be false in particular cases, however, and so it is well to know something of the conditions sufficient to guarantee its validity.

THEOREM III. *Let the function $f(x, y)$ be defined in some neighborhood of the point (a, b). Let the partial derivatives f_1, f_2, f_{12} and f_{21} also be defined in this neighborhood, and suppose that f_{12} and f_{21} are continuous at (a, b). Then $f_{12}(a, b) = f_{21}(a, b)$. In other words, (7.2–1) holds at the point (a, b).*

Proof. We shall work entirely inside a square having its center at (a, b), and lying inside the neighborhood referred to in the theorem. Let h be a number different from zero such that the point $(a + h, b + h)$ is inside the square just referred to. Consider the expression

$$D = f(a + h, b + h) - f(a + h, b) - f(a, b + h) + f(a, b).$$

If we introduce the function

$$\phi(x) = f(x, b + h) - f(x, b),$$

we can express D in the form

$$D = \phi(a + h) - \phi(a). \tag{7.2–2}$$

Now ϕ has the derivative

$$\phi'(x) = f_1(x, b + h) - f_1(x, b).$$

Hence ϕ is continuous, and we may apply the law of the mean to (7.2–2), with the result

$$D = h\phi'(a + \theta_1 h) = h[f_1(a + \theta_1 h, b + h) - f_1(a + \theta_1 h, b)], \tag{7.2–3}$$

where $0 < \theta_1 < 1$. Next, let

$$g(y) = f_1(a + \theta_1 h, y).$$

The function g has the derivative

$$g'(y) = f_{12}(a + \theta_1 h, y).$$

Now we can write (7.2–3) in the form

$$D = h[g(b + h) - g(b)]$$

and apply the law of the mean. The result is

$$D = h^2 g'(b + \theta_2 h) = h^2 f_{12}(a + \theta_1 h, b + \theta_2 h),$$

where $0 < \theta_2 < 1$.

We might instead have started by expressing D in the form

$$D = \psi(b + h) - \psi(b),$$

where

$$\psi(y) = f(a + h, y) - f(a, y).$$

This procedure would have led to an expression

$$D = h^2 f_{21}(a + \theta_4 h, b + \theta_3 h),$$

with $0 < \theta_3 < 1$, $0 < \theta_4 < 1$. On comparing the two expressions for D we see that

$$f_{12}(a + \theta_1 h, b + \theta_2 h) = f_{21}(a + \theta_4 h, b + \theta_3 h). \tag{7.2–4}$$

If we now make $h \to 0$, the points at which the derivatives in (7.2–4) are evaluated both approach (a, b). Hence, by the assumed continuity of f_{12} and f_{21}, we conclude that $f_{12}(a, b) = f_{21}(a, b)$. This completes the proof.

The conditions of Theorem III are not the only known sufficient conditions for the truth of (7.2–1). The theorem provides a useful working criterion, however. Another criterion is furnished by the following theorem:

THEOREM IV. *Let $f(x, y)$ and its first partial derivatives f_1, f_2 be defined in a neighborhood of the point (a, b), and suppose that f_1 and f_2 are differentiable at that point. Then $f_{12}(a, b) = f_{21}(a, b)$.*

This theorem requires more of the function f in some ways, and less in others, than Theorem III. We omit the proof, which begins very much like that of Theorem III.

We may use Theorem III or Theorem IV to prove that a mixed partial derivative of order higher than the second is independent of the order of performing the differentiations, provided we make appropriate assumptions about the continuity or differentiability. Thus, for example, suppose that $f(x, y)$ and all its partial derivatives of orders one, two, and three are continuous. (This is more than we actually need.) Then

$$\frac{\partial^3 f}{\partial y^2\, \partial x} = \frac{\partial^3 f}{\partial y\, \partial x\, \partial y} = \frac{\partial^3 f}{\partial x\, \partial y^2}.$$

For $$f_{122} = (f_{12})_2 = (f_{21})_2 = f_{212},$$

and $$f_{212} = (f_2)_{12} = (f_2)_{21} = f_{221}.$$

By similar arguments we can deal with functions of more than two independent variables.

EXERCISES

1. Let $f(x, y) = xy\left(\frac{x^2 - y^2}{x^2 + y^2}\right)$ if $x^2 + y^2 \neq 0$, and define $f(0, 0) = 0$. Show that $f_1(0, y) = -y$ and $f_2(x, 0) = x$ for all x and y. Then show that $f_{12}(0, 0) = -1$ and $f_{21}(0, 0) = 1$.

2. Define $f(x, y) = x^2 \tan^{-1}(y/x) - y^2 \tan^{-1}(x/y)$ if neither x nor y is zero, and $f(x, y) = 0$ if either $x = 0$ or $y = 0$ (or both). Show as in Exercise 1 that $f_{12}(0, 0) = -1$, $f_{21}(0, 0) = 1$.

3. Theorem III can be improved as follows:

Assume that f is defined in some neighborhood of the point (a, b), and that the partial derivatives f_1, f_2, f_{12} are also defined in this neighborhood. Suppose finally that f_{12} is continuous at (a, b). Then it is true that f_{21} is defined at (a, b) and that $f_{12}(a, b) = f_{21}(a, b)$.

(This theorem is due to H. A. Schwarz.)

Prove the theorem with the assistance of the following suggestions: Let h, k be small

numbers different from zero. Let

$$\Delta = f(a+h, b+k) - f(a+h, b) - f(a, b+k) + f(a, b).$$

Show that there are numbers θ_1, θ_2 between 0 and 1, depending on h and k, such that

$$\Delta = hkf_{12}(a+\theta_1 h, b+\theta_2 k).$$

If $\epsilon > 0$, choose $\delta > 0$ so that $\left|\frac{\Delta}{hk} - f_{12}(a, b)\right| < \epsilon$ if $0 < |h| < \delta$ and $0 < |k| < \delta$. Why is this possible? Find the limit of $\frac{\Delta}{hk}$ with h fixed, as $k \to 0$, and so conclude that

$$\left|\frac{f_2(a+h, b) - f_2(a, b)}{h} - f_{12}(a, b)\right| \leqq \epsilon$$

if $0 < |h| < \delta$. Now complete the proof of Schwarz's theorem.

4. To prove Theorem IV, start as in the proof of Theorem III, and obtain (7.2–3). Then, from the fact that f_1 is differentiable at (a, b), one can write

$$f_1(a+\theta_1 h, b+h) = f_1(a, b) + f_{11}(a, b)\theta_1 h + f_{12}(a, b)h + \epsilon_1|h|,$$

where $\epsilon_1 \to 0$ as $h \to 0$. Explain why this is so.

Then go on to explain how to obtain the expression

$$D = h^2 f_{12}(a, b) + \epsilon|h|h,$$

where $\epsilon \to 0$ as $h \to 0$. Explain the derivation of the similar expression (where $\epsilon' \to 0$ as $h \to 0$)

$$D = h^2 f_{21}(a, b) + \epsilon'|h|h,$$

using the fact that f_2 is differentiable at (a, b). Now complete the proof of Theorem IV.

7.3 / DIFFERENTIALS OF COMPOSITE FUNCTIONS

In §6.4 we made the statement that a differentiable function of differentiable functions is differentiable. We now formulate this proposition in precise terms, and prove it. There may be any number of independent variables in each of the functions involved. For simplicity we deal with the case of two variables throughout.

THEOREM V. *Let $F(x, y)$ be defined in some neighborhood of the point (a, b), and let it be differentiable there. Let $f(s, t)$ and $g(s, t)$ be defined in some neighborhood of (s_0, t_0), and differentiable at that point. Assume further that*

$$f(s_0, t_0) = a, \; g(s_0, t_0) = b,$$

and consider the composite function

$$G(s, t) = F(f(s, t), g(s, t)).$$

Then G is differentiable at (s_0, t_0). Its differential may be written

$$dG = \frac{\partial F}{\partial x}\,dx + \frac{\partial F}{\partial y}\,dy, \tag{7.3–1}$$

where

$$dx = \frac{\partial f}{\partial s}\,ds + \frac{\partial f}{\partial t}\,dt, \tag{7.3–2}$$

$$dy = \frac{\partial g}{\partial s}\,ds + \frac{\partial g}{\partial t}\,dt.$$

It is to be understood that the partial derivatives of F are evaluated at (a, b), those of f and g at (s_0, t_0), and that ds, dt are independent variables.

Proof. Let us write $u = G(s, t)$, $x = f(s, t)$, $y = g(s, t)$. For arbitrary Δs, Δt write $\Delta u = G(s_0 + \Delta s, t_0 + \Delta t) - G(s_0, t_0)$, with corresponding meanings for Δx, Δy. Note that $\Delta u = F(a + \Delta x, b + \Delta y) - F(a, b)$. Accordingly, using one of the formulations of the differentiability condition discussed in §7, we may write

$$\begin{aligned} \Delta u &= F_1\,\Delta x + F_2\,\Delta y + \epsilon(|\Delta x| + |\Delta y|), \\ \Delta x &= f_1\,\Delta s + f_2\,\Delta t + \delta(|\Delta s| + |\Delta t|), \\ \Delta y &= g_1\,\Delta s + g_2\,\Delta t + \eta(|\Delta s| + |\Delta t|), \end{aligned} \tag{7.3–3}$$

where $\epsilon \to 0$ as Δx and $\Delta y \to 0$, while $\delta \to 0$ and $\eta \to 0$ as Δs and $\Delta t \to 0$. Here, for convenience and brevity, we are using F_1 to stand for $F_1(a, b)$, f_1 for $f_1(x_0, t_0)$, and so on. From (7.3–3) we see that

$$\Delta u = F_1(f_1\,\Delta s + f_2\,\Delta t) + F_2(g_1\,\Delta s + g_2\,\Delta t) + (F_1\delta + F_2\eta)(|\Delta s| + |\Delta t|) + \epsilon(|\Delta x| + |\Delta y|).$$

This may be rewritten in the form

$$\begin{aligned} \Delta u = {} & (F_1 f_1 + F_2 g_1)\,\Delta s + (F_1 f_2 + F_2 g_2)\,\Delta t \\ & + \left\{ F_1\delta + F_2\eta + \epsilon\,\frac{|\Delta x| + |\Delta y|}{|\Delta s| + |\Delta t|} \right\} (|\Delta s| + |\Delta t|). \end{aligned}$$

If we can show that

$$\lim_{(\Delta s,\,\Delta t)\to(0,\,0)} \left\{ F_1\delta + F_2\eta + \epsilon\,\frac{|\Delta x| + |\Delta y|}{|\Delta s| + |\Delta y|} \right\} = 0, \tag{7.3–4}$$

we shall have proved that G is differentiable, with the differential

$$dG = (F_1 f_1 + F_2 g_1)\,ds + (F_1 f_2 + F_2 g_2)\,dt.$$

This last form is equivalent to the combination of (7.3–1) and (7.3–2). Hence it remains only to prove (7.3–4).

Now F_1 and F_2 denote constants in (7.3–4), so that $F_1\delta + F_2\eta \to 0$. The functions f and g are continuous at (s_0, t_0) by Theorem II, §6.4. Hence Δx and $\Delta y \to 0$, and therefore $\epsilon \to 0$ as Δs and $\Delta t \to 0$. Let M be a constant larger than the greatest of the numbers $|f_1|$, $|f_2|$, $|g_1|$, $|g_2|$. Then by (7.3–3) we see that

$$|\Delta x| \leqq (M + \delta)(|\Delta s| + |\Delta t|),\ |\Delta y| \leqq (M + \eta)(|\Delta s| + |\Delta t|).$$

Hence

$$\epsilon\,\frac{|\Delta x| + |\Delta y|}{|\Delta s| + |\Delta t|} \leqq \epsilon(2M + \delta + \eta).$$

This quantity approaches zero as Δs and Δt approach zero. The assertion (7.3–4) is there established.

7.4 / THE LAW OF THE MEAN

The student is already familiar with the law of the mean for functions of a single independent variable, in the form

$$f(a+h)-f(a)=hf'(a+\theta h),\ 0<\theta<1 \tag{7.4–1}$$

(see (1.2–4) and Theorem IV, §1.2). It is useful to have a generalization of this result for functions of several independent variables. In seeking an appropriate form for such a generalization, we look upon (7.4–1) as furnishing a convenient formula for the difference between the values of the function f at two points of the x-axis, namely $x=a$ and $x=a+h$. This leads us, in the case of a function of two variables, to search for a means of expressing the difference

$$F(a+h,\ b+k)-F(a,\ b),$$

where the line segment joining the points $(a,\ b)$, $(a+h,\ b+k)$ lies in the region of definition of the function F.

THEOREM VI. *Let F be defined in a region R of the xy-plane. Let L be the line segment with ends $(a,\ b)$, $(a+h,\ b+k)$. We suppose that L lies in R and that all points of L except possibly the ends are interior points of R. Finally, we assume that F is continuous at each point of L and differentiable at each such point with the possible exception of the ends. Then, for a certain value of θ, such that $0<\theta<1$, we have*

$$F(a+h,\ b+k)-F(a,\ b)=hF_1(a+\theta h,\ b+\theta k)+kF_2(a+\theta h,\ b+\theta k). \tag{7.4–2}$$

Proof. By introducing a parametric representation of the line segment L,

$$x=a+th,\ y=b+tk,\ 0\leqq t\leqq 1,$$

we are able to regard the value of F along L as a function of the parameter t. Let us write

$$f(t)=F(a+th,\ b+tk). \tag{7.4–3}$$

The derivative of this composite function is

$$f'(t)=hF_1(a+th,\ b+tk)+kF_2(a+th,\ b+tk). \tag{7.4–4}$$

Applying the ordinary law of the mean in the form (7.4–1), we have

$$f(1)-f(0)=f'(\theta),\ 0<\theta<1. \tag{7.4–5}$$

On setting $t=0$, 1 successively in (7.4–3), and $t=\theta$ in (7.4–4), we see that (7.4–5) becomes the formula (7.4–2). The proof is therefore complete.

Observe that the point $(a + \theta h, b + \theta k)$ is a point of the segment L somewhere between its ends (see Fig. 48).

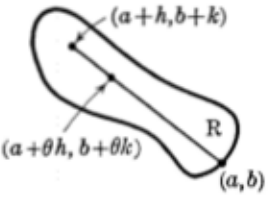


Fig. 48.

When a set has the property that for each pair of points belonging to it the straight line segment connecting them consists entirely of points which also belong to the set, then the set is said to be *convex*.

If the domain R of the function F in Theorem VI is both open and convex, then clearly (7.4–2) holds for every pair of points (a, b) and $(a + h, b + k)$ in R. It follows immediately that if both first partial derivatives of F vanish throughout R, then F has the same value at $(a + h, b + k)$ as at (a, b). By holding a and b fixed while letting h and k vary, $(a + h, b + k)$ can be made to represent each point in R. Therefore we see that F must take on the same value at each point of its domain—in other words, F must be constant. This result is an analogue, for functions of two variables, of Theorem V, §1.2.

Actually, we can get a more general analogue if we replace the condition of convexity by a weaker condition called connectedness. For our purpose here it will suffice to define connectedness for sets that are open.

Definition. *An open set S in the plane (or in space of three dimensions) is called connected in case each pair of points in S can be joined by a path consisting of a finite number of straight line segments joined end to end consecutively, the whole path lying entirely in S and not crossing itself anywhere. Such a path may be called a polygonal arc.*

Later, in §17.7, we shall give a general definition of connectedness, applicable to any set, open or not. When that definition is applied to open sets, it turns out to be equivalent to the definition which we are using here.

As an example of a nonconnected set, let S consist of all points of the plane for which $x^2 > 1$, that is, all points except those for which $-1 \leqq x \leqq 1$. Plainly S consists of two separated parts (see Fig. 49). Two points, one in each part, cannot be joined by a broken-line path lying entirely in S. This particular set S comes naturally to our attention if we study the function

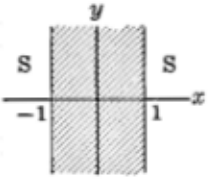


Fig. 49.

$$f(x, y) = y + \sqrt{x^2 - 1}.$$

Now we come to the theorem.

THEOREM VII. *Let $F(x, y)$ be a function which is defined and differentiable throughout a connected open set S, and suppose that the first partial derivatives of F vanish at each point of S. Then $F(x, y)$ is constant in S.*

Proof. Suppose A and B are any two points of S. Let them be joined by a path consisting of segments $AP_1, P_1P_2, \ldots, P_{n-1}P_n, P_nB$, all lying in S (see Fig. 50). By the comment just after the proof of Theorem VI we see that F has the same value at A as at P_1, the same value at P_1 as at P_2, and so on, so that F has the same value at B as at A. This means that F is constant in S.

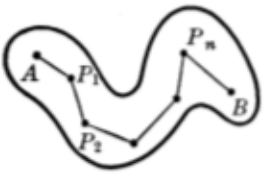


Fig. 50.

Theorems VI and VII admit of immediate extension to functions of more than two independent variables. The extension of the law of the mean for three independent variables is

$$F(a+h, b+k, c+l) - F(a, b, c) = hF_1(\bar{x}, \bar{y}, \bar{z}) + kF_2(\bar{x}, \bar{y}, \bar{z}) + lF_3(\bar{x}, \bar{y}, \bar{z}), \tag{7.4–6}$$

where $\bar{x} = a + \theta h, \bar{y} = b + \theta k, \bar{z} = c + \theta l.$

Formula (7.4–2) can also be written in the form

$$F(x, y) - F(a, b) = F_1(X, Y)(x - a) + F_2(X, Y)(y - b), \tag{7.4–7}$$

where (X, Y) is a certain point on the line joining (a, b) and (x, y).

In §2.7 we had occasion to refer to the fact that a set may be empty. For logical reasons we need to be aware that, even though a set is empty, it may have certain properties "by default." For example, we cite the fact that the empty set is open. The logic of the situation is that a set S is called open if for every point P in S there is a neighborhood of P contained in S. If S is empty there is no point P in S and hence the requirement about a neighborhood of P has no force as a restriction on S. We say that the requirement "is satisfied vacuously," which means that it is satisfied by default. Hence S is open. Similarly, an empty set is connected.

EXERCISES

1. Let $F(x, y) = xy^2 - x^2y$. Find the appropriate value of θ in (7.4–2) if **(a)** $a = b = 0, h = 1, k = 2$; **(b)** $a = b = 0, h = 3, k = 2$; **(c)** $a = b = 1, h = 3, k = 2$.

2. Let $F(x, y)$ be the quadratic function $Ax^2 + 2Bxy + Cy^2$. Show that (7.4–7) holds with $X = \frac{1}{2}(x + a)$, $Y = \frac{1}{2}(y + b)$. What does this mean about the value of θ in (7.4–2)?

3. Taking $F(x, y) = \sin x \cos y$, prove that for some θ between 0 and 1 it is true that $\frac{3}{4} = \frac{\pi}{3}\cos\frac{\pi\theta}{3}\cos\frac{\pi\theta}{6} - \frac{\pi}{6}\sin\frac{\pi\theta}{3}\sin\frac{\pi\theta}{6}$.

4. **(a)** Write out formula (7.4–2) for $F(x, y) = \log(xe^{y^2})$, with $a = 1, b = 0, h = e - 1, k = 1$. **(b)** Write out (7.4–7) for this same function, with a, b, x, y arbitrary, except that $a > 0, x > 0$.

5. If $x \neq a$ in (7.4–7), show that $Y = b + \frac{y - b}{x - a}(X - a)$. Hence show that, under suitable conditions, $F(1, 0) - F(0, 1) = F_1(X, 1 - X) - F_2(X, 1 - X)$.

6. Let $F(x, y) = (1 - 2xy + x^2)^{-1/2}$. As a result of considering $F(1, 0) - F(0, 1)$, show

that there is a number θ such that $0 < \theta < 1$ and

$$1 - \sqrt{2} = \sqrt{2}(1 - 3\theta)(1 - 2\theta + 3\theta^2)^{-3/2}.$$

7. Let $F(x, y, z) = xyz$. Find the appropriate value of θ in (7.4–6) if **(a)** $a = b = c = 0$, $h = k = l = 1$; **(b)** $a = b = 0$, $c = 1$, $h = k = 1$, $l = -1$; **(c)** $a = c = 0$, $b = 1$, $h = l = 1$, $k = 0$.

8. Show that the open disc $\{(x, y): x^2 + y^2 < 1\}$ is convex.

HINT: Begin by showing that the line segment determined by any two points (a, b) and $(a + h, b + k)$ in the disc is the set of all points of the form $\{a + th, b + tk\}$ where $0 \le t \le 1$.

(Actually, we still have a convex set if we join to this open disc some or all of its boundary.)

9. Explain why the empty set is connected, and why every set consisting of just one point is connected.

7.5 / TAYLOR'S FORMULA AND SERIES

Just as we extended the ordinary law of the mean to functions of several variables, so we may extend the version of Taylor's formula given in §4.3. The method is the same as that employed in the proof of Theorem VI, §7.4. We write

$$f(t) = F(a + th, b + tk) \tag{7.5–1}$$

and apply Taylor's formula to $f(t)$, using the two values $t = 0, t = 1$. From (4.3–7) with $a = 0, h = 1$ we have

$$f(1) = f(0) + f'(0) + \cdots + \frac{f^{(n)}(0)}{n!} + \frac{f^{(n+1)}(\theta)}{(n+1)!}, \quad 0 < \theta < 1. \tag{7.5–2}$$

The assumptions are that F and its partial derivatives of orders 1 to n inclusive are differentiable at all points along the line joining (a, b) and $(a + h, b + k)$. The main problem now is that of calculating the higher derivatives of f from (7.5–1). The first derivative is given by (7.4–4) in the previous section. Working from that formula, we see that

$$f''(t) = h[hF_{11} + kF_{12}] + k[hF_{21} + kF_{22}],$$

where all the partial derivatives on the right are evaluated at $(a + th, b + tk)$. Since $F_{12} = F_{21}$ (Theorem IV, §7.2), we have

$$f''(t) = h^2F_{11} + 2hkF_{12} + k^2F_{22}.$$

This is sometimes written in the form

$$f''(t) = \left[\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^2 F(x, y)\right]_{x=a+th,\, y=b+tk},$$

it being understood that

$$\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^2 F(x, y) = h^2\frac{\partial^2 F}{\partial x^2} + 2hk\frac{\partial^2 F}{\partial x\,\partial y} + k^2\frac{\partial^2 F}{\partial y^2}.$$

The analogy with the pattern of the binomial expansion is now evident. We have

$$f'''(t) = h^3F_{111} + 3h^2kF_{112} + 3hk^2F_{122} + k^3F_{222}$$
$$= \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^3 F(x, y),$$

with x and y set equal to $a + th$ and $b + tk$, respectively, in the partial derivatives. The general formula is

$$f^{(n)}(t) = \left[\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^n F(x, y)\right]_{x=a+th,\, y=b+tk}. \tag{7.5–3}$$

To get Taylor's formula for $F(x, y)$ we use (7.5–1) and (7.5–3) to substitute in (7.5–2). For $n = 1$ the result is

$$F(a + h, b + k) = F(a, b) + hF_1(a, b) + kF_2(a, b) + R_2, \tag{7.5–4}$$

where

$$R_2 = \frac{1}{2!}\left[\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^2 F(x, y)\right]_{x=a+\theta h,\, y=b+\theta k},$$

and $0 < \theta < 1$. For $n = 2$ the result is

$$F(a + h, b + k) = F(a, b) + hF_1(a, b) + kF_2(a, b)$$
$$+ \frac{1}{2!}[h^2F_{11}(a, b) + 2hkF_{12}(a, b) + k^2F_{22}(a, b)] + R_3, \tag{7.5–5}$$

where

$$R_3 = \frac{1}{3!}\left[\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^3 F(x, y)\right]_{x=a+\theta h,\, y=b+\theta k},$$

and $0 < \theta < 1$. The extension to higher values of n is obvious. Observe that from $f^{(n)}(0)$ we get a homogeneous polynomial of degree n in h and k, the coefficient of $h^{n-p}k^p$ being $\dfrac{n!}{p!(n-p)!}\dfrac{\partial^n F}{\partial x^{n-p}\,\partial y^p}$, with the partial derivative evaluated at $x = a$, $y = b$.

Under certain conditions on the function F, the point (a, b), and the size of h and k, it may happen that $F(a + h, b + k)$ can be represented as the infinite series

$$F(a + h, b + k) = F(a, b) + \sum_{n=1}^{\infty} \frac{1}{n!}\left[\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^n F(x, y)\right]_{x=a,\, y=b}. \tag{7.5–6}$$

This is the form of Taylor's series for a function of two variables. If we use only a specified number of terms of this series we get an approximate expression for $F(a + h, b + k)$. Frequently it is more convenient to write (x, y) in place of $(a + h, b + k)$. The typical term of the series then becomes a homogeneous polynomial in $(x - a)$ and $(y - b)$.

Example. Write Taylor's formula (7.5–4) with $n = 1$, and carry the series (7.5–6) through the term in $n = 2$, if $F(x, y) = 1/(xy)$, $a = 1$, $b = -1$.

From $F(x, y) = x^{-1}y^{-1}$ we readily find

$$\frac{\partial F}{\partial x} = -x^{-2}y^{-1}, \frac{\partial F}{\partial y} = -x^{-1}y^{-2},$$

$$\frac{\partial^2 F}{\partial x^2} = 2x^{-3}y^{-1}, \frac{\partial^2 F}{\partial x\,\partial y} = x^{-2}y^{-2}, \frac{\partial^2 F}{\partial y^2} = 2x^{-1}y^{-3}.$$

It is now easily seen that (7.5–4) becomes

$$\frac{1}{(1+h)(-1+k)} = -1 + h - k + R_2,$$

$$R_2 = \frac{h^2}{(1+\theta h)^3(-1+\theta k)} + \frac{hk}{(1+\theta h)^2(-1+\theta k)^2} + \frac{k^2}{(1+\theta h)(-1+\theta k)^3}.$$

The series (7.5–6) begins

$$\frac{1}{(1+h)(-1+k)} = -1 + (h-k) + (-h^2 + hk - k^2) + \cdots.$$

Detailed verification should be supplied by the student as he reads this example.

If we write $x = 1 + h$, $y = -1 + k$, the last formula becomes

$$\frac{1}{xy} = -1 + [(x-1) - (y+1)] + [-(x-1)^2 + (x-1)(y+1) - (y+1)^2] + \cdots.$$

We shall not investigate the precise limitations on $(x-1)$ and $(y+1)$ which are necessary in this expansion.

There are situations in which we need the Taylor series with remainder for functions of more than two variables. Suppose for example that we wish to expand a function $F(x_1, x_2, \ldots, x_n)$ about the point $a_1, a_2, \ldots, a_n$. One way of writing the expansion to terms of degree k, followed by a remainder, is

$$F(a_1 + h_1, a_2 + h_2, \ldots, a_n + h_n) = F(a_1, a_2, \ldots, a_n) + \sum_{r=1}^{k} \frac{1}{r!}\left(h_1\frac{\partial}{\partial x_1} + h_2\frac{\partial}{\partial x_2} + \cdots + h_n\frac{\partial}{\partial x_n}\right)^r F + \frac{1}{(k+1)!}\left(h_1\frac{\partial}{\partial x_1} + h_2\frac{\partial}{\partial x_2} + \cdots + h_n\frac{\partial}{\partial x_n}\right)^{k+1} F,$$

where all the derivatives up through those of order k are evaluated at the point $(a_1, a_2, \ldots, a_n)$, and those of order $k+1$ are evaluated at $a_1 + \theta h, a_2 + \theta h_2, \ldots, a_n + \theta h_n$ where θ is some number properly between 0 and 1. The line through $(a_1, a_2, \ldots, a_n)$ and $(a_1 + h_1, a_2 + h_2, \ldots, a_n + h_n)$ has parametric equations $x_1 = a_1 + th_1$, $x_2 = a_2 + th_2, \ldots, x_n = a_n + th_n$. Therefore the point at which the $(k+1)$st order partial derivatives in the remainder term are evaluated is between the two points where $t = 0$ and $t = 1$.

Another way of saying the same thing results from letting $x_1 = a_1 + h_1$, $x_2 = a_2 + h_2, \ldots, x_n = a_n + h_n$. Replacing the h's by their equivalents in terms of

the x's gives

$$F(x_1, x_2, \ldots, x_n) = F(a_1, a_2, \ldots, a_n) + \sum_{r=1}^{k} \frac{1}{r!} \left[(x_1 - a_1) \frac{\partial}{\partial x_1} + \cdots + (x_n - a_n) \frac{\partial}{\partial x_n} \right]^r F$$
$$+ \frac{1}{(k+1)!} \left[(x_1 - a_1) \frac{\partial}{\partial x_1} + \cdots + (x_n - a_n) \frac{\partial}{\partial x_n} \right]_{P^*}^{k+1} F,$$

where all derivatives or order less than or equal to k are evaluated at $(a_1, a_2, \ldots, a_n)$ and the derivatives in the remainder term, all of which are of order $(k+1)$, are evaluated at some point P^* on the line segment connecting $(a_1, \ldots, a_n)$ and $(x_1, \ldots, x_n)$.

In the theory of maxima and minima of functions of several variables, we are interested in the case where $k = 1$. The above formula then reduces to

$$F(x_1, x_2, \ldots, x_n) = F(a_1, a_2, \ldots, a_n) + (x_1 - a_1)F_1(a_1, \ldots, a_n) + \cdots + \tag{7.5–7}$$
$$+ (x_n - a_n)F_n(a_1, a_2, \ldots, a_n) +$$
$$+ \frac{1}{2!} \left[(x_1 - a_1) \frac{\partial}{\partial x_1} + \cdots + (x_n - a_n) \frac{\partial}{\partial x_n} \right]_{P^*}^{2} F$$
$$= F(a_1, a_2, \ldots, a_n) + \sum_{i=1}^{n} (x_i - a_i)F_i(a_1, a_2, \ldots, a_n) +$$
$$+ \frac{1}{2!} \sum_{i,j=1}^{n} (x_i - a_i)(x_j - a_j)F_{ij}(P^*)$$

where F_{ij} is evaluated at some point of the form $[a_1 + \theta(x_1 - a_1), a_2 + \theta(x_2 - a_2), \ldots, a_n + \theta(x_n - a_n)]$—in other words, at some point on the line segment connecting $(a_1, a_2, \ldots, a_n)$ and $(x_1, x_2, \ldots, x_n)$. The remainder term in this special case can be recognized as a quadratic form in the n variables $x_1 - a_1, x_2 - a_2, \ldots, x_n - a_n$. See §6.9, and recall that by Theorem III $F_{ij} = F_{ji}$ under very general hypotheses on F.

EXERCISES

1. Write Taylor's formula (7.5–5) for $F(x, y) = \sin x \sin y$, using $a = 0$, $b = 0$, and $n = 2$.

2. Write Taylor's formula for $F(x, y) = \cos x \cos y$, using $a = 0$, $b = 0$, and $n = 2$.

3. Write Taylor's formula with $a = 3$, $b = 3$, $n = 3$ for $F(x, y) = x^3 + y^3 - 9xy + 27$.

4. Let $F(x, y) = \log(x + e^y)$. Expand according to Taylor's series in powers of $x - 1$ and y, going far enough to include all terms of degree 2 in these quantities.

5. Follow the instructions of Exercise 4 for $F(x, y) = \sin(e^y + x^2 - 2)$.

6. Write Taylor's series for $e^x \cos y$ in powers of x and y, going far enough to include all terms of third degree.

7. Write Taylor's series for $e^{-y^2 + 2xy}$ in powers of x and y, going far enough to include all terms of fourth degree.

8. Write Taylor's series for $x^2y + xy^2 + 1$ in powers of $x - 1$ and $y - 1$.

9. Write Taylor's series for $xy^3 - y^2 + y + 2$ in powers of $x - 1$ and $y - 2$.

10. (**a**) Find a linear function of x and y which is a good approximation for $F(x, y) = \tan^{-1}\left(\frac{x-y}{1+xy}\right)$ when x and y are small.

(**b**) Write the constant and linear terms in Taylor's series of $F(x, y)$ in powers of $x - 3$ and $y - \frac{1}{2}$.

11. Write out in full the expression

$$\frac{1}{4!}\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^4 F(x, y).$$

What does the expression become (**a**) if $F(x, y) = x^4 - x^2y^2 + y^4$; (**b**) if $F(x, y) = \sin xy$ and if one sets $x = y = (\pi/2)^{1/2}$ after doing the differentiation?

12. (**a**) Carry on the work of the illustrative example in the text, showing that

$$\frac{\partial^n F}{\partial x^{n-p}\,\partial y^p} = (-1)^n(n-p)!p!\frac{1}{x^{n-p+1}y^{p+1}},$$

and that the polynomial of degree n in h and k in the Taylor's series is

$$(-1)^{n-1}[h^n - h^{n-1}k + h^{n-2}k^2 - \cdots + (-1)^n k^n].$$

(**b**) Assuming that $|h| < 1$ and $|k| < 1$, write

$$\begin{aligned}\frac{1}{(1+h)(-1+k)} &= -(1+h)^{-1}(1-k)^{-1} \\ &= -(1-h+h^2-h^3+\cdots)(1+k+k^2+\cdots).\end{aligned}$$

Then multiply the two series together term by term, and collect together the terms of like degree. Compare with the result found in (**a**).

7.6 / SUFFICIENT CONDITIONS FOR A RELATIVE EXTREME

In §6.3 we discussed relative maxima and minima for a function $f(x, y)$. In Theorem I of that section we reached the important conclusion that if f attains a relative extreme value at an interior point of its region of definition, then necessarily the partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ vanish at the point (provided these derivatives exist, of course). The conditions

$$\frac{\partial f}{\partial x} = 0, \frac{\partial f}{\partial y} = 0 \tag{7.6–1}$$

at a point do not in themselves guarantee a relative extreme, however. In this section we wish to develop criteria which, taken together with conditions (7.6–1), will guarantee a relative extreme, and enable us to distinguish a relative maximum from a relative minimum.

It will be helpful if we begin by reviewing briefly the analogous considerations for a function of one variable. Suppose we have a function $y = f(x)$ defined on some interval having $x = a$ as an interior point. We suppose f to be differentiable on the interval, and we assume that the second derivative exists at $x = a$.

THEOREM VIII. *Under the conditions on f as just stated, suppose that $f'(a) = 0$. Then*

(i) *If $f''(a) > 0$, f has a relative minimum at $x = a$;*

(ii) *If $f''(a) < 0$, f has a relative maximum at $x = a$;*

(iii) *If $f''(a) = 0$ no conclusion may be drawn; f may have a relative maximum or minimum, or it may have neither.*

Proof. Consider the value of f at any point $x = a + h$ near $x = a$. We take $h \neq 0$, but it may be either positive or negative. By the law of the mean,

$$f(a + h) - f(a) = hf'(a + \theta h), \quad 0 < \theta < 1. \tag{7.6–2}$$

Now, by the definition of $f''(a)$, we have

$$\lim_{\Delta x \to 0} \frac{f'(a + \Delta x) - f'(a)}{\Delta x} = f''(a).$$

Since $f'(a) = 0$ we can write this in the form

$$\frac{f'(a + \Delta x)}{\Delta x} = f''(a) + \epsilon,$$

where ϵ is a variable quantity such that $\epsilon \to 0$ as $\Delta x \to 0$. Consequently, if we choose $\Delta x = \theta h$, we have

$$f'(a + \theta h) = (f''(a) + \epsilon)\theta h,$$

and (7.6–2) becomes

$$f(a + h) - f(a) = (f''(a) + \epsilon)\theta h^2. \tag{7.6–3}$$

If we now suppose that $f''(a) \neq 0$, we see that, as soon as h is small enough to insure that $|\epsilon| < |f''(a)|$, the sign of the left member of (7.6–3) is the same as the sign of $f''(a)$. Consequently, if $f''(a) > 0$, we conclude that $f(a + h) > f(a)$ for all sufficiently small values of h different from zero. This means that f has a relative minimum at $x = a$. We have thus proved part (i) of the theorem; the same kind of argument is used for part (ii). If $f''(a) = 0$ no conclusion is reached, however, since we do not know the sign of ϵ in (7.6–3). The three examples $y = x^4$, $y = -x^4$, $y = x^3$, with $a = 0$, show that any of the possibilities mentioned in part (iii) may arise.

Let us now turn to the case of two independent variables. We have the following corresponding theorem:

THEOREM IX. *Suppose that $F(x, y)$ is defined and differentiable throughout a region R of which (a, b) is an interior point, and suppose that the first partial derivatives of F vanish at that point. Suppose further that the partial derivatives F_1 and F_2 are differentiable at (a, b). Let us write*

$$A = F_{11}(a, b), \; B = F_{12}(a, b), \; C = F_{22}(a, b).$$

Then

(i) *If* $B^2 - AC < 0$ *and* $A > 0$, F *has a relative minimum at* (a, b);
(ii) *If* $B^2 - AC < 0$ *and* $A < 0$, F *has a relative maximum at* (a, b);
(iii) *If* $B^2 - AC > 0$, F *has neither a maximum nor a minimum at the point;*
(iv) *If* $B^2 - AC = 0$, *no conclusion may be drawn and any of the behaviors of* F *described in parts* (i)–(iii) *may occur.*

Proof. Note that part (iii) of Theorem IX has no counterpart in Theorem VIII. The method of proof of Theorem IX is similar to that of Theorem VIII. We consider the point $(a + h, b + k)$, where h and k are both small, but not both zero. By the law of the mean we have (with $0 < \theta < 1$)

$$F(a + h, b + k) - F(a, b) = hF_1(a + \theta h, b + \theta k) + kF_2(a + \theta h, b + \theta k).$$

Since we assumed F_1 and F_2 differentiable, we may write (see the discussion of differentiability near the end of §7)

$$F_1(a + \theta h, b + \theta k) - F_1(a, b) = \theta h F_{11}(a, b) + \theta k F_{12}(a, b) + \epsilon_1(|\theta h| + |\theta k|),$$

where $\epsilon_1 \to 0$ as h and $k \to 0$. A similar expression may be written for F_2, with some quantity ϵ_2 in place of ϵ_1. Since $F_1(a, b) = F_2(a, b) = 0$, by hypothesis, we have

$$F(a + h, b + k) - F(a, b) = \theta[Ah^2 + 2Bhk + Ck^2) + (|h| + |k|)(\epsilon_1 h + \epsilon_2 k)]. \quad (7.6\text{–}4)$$

This formula is the counterpart of (7.6–3). To get the information we desire from it, however, it is more convenient to express h and k in terms of polar co-ordinates with origin at (a, b). Let us write

$$h = r\cos\phi, k = r\sin\phi.$$

When these expressions are substituted in (7.6–4), a factor r^2 can be taken out, and we get

$$F(a + h, b + k) - F(a, b) = \theta r^2[G(\phi) + \delta], \quad (7.6\text{–}5)$$

where for abbreviation we have set

$$G(\phi) = A\cos^2\phi + 2B\sin\phi\cos\phi + C\sin^2\phi,$$
$$\delta = (|\cos\phi| + |\sin\phi|)(\epsilon_1\cos\phi + \epsilon_2\sin\phi).$$

Here $0 < \theta < 1$, and $\delta \to 0$ as $r \to 0$. Since $G(\phi)$ is independent of r, it is clear that, if $G(\phi) \neq 0$, the sign of the left member of (7.6–5) will be the same as the sign of $G(\phi)$ when r is sufficiently small. Moreover, $G(\phi)$ is continuous, whence it follows that, if $G(\phi)$ is never zero, it always has the same sign, and, when r is sufficiently small, the sign of $G(\phi) + \delta$ will be the same as the sign of $G(\phi)$. Therefore, if $G(\phi)$ is always positive, F has a minimum at (a, b), while if $G(\phi)$ is always negative, F has a maximum at (a, b). On the other hand, if $G(\phi)$ is sometimes positive and sometimes negative, F has neither a maximum nor a minimum at (a, b). We shall show that the cases (i)–(iii) in the statement of Theorem IX lead to exactly these three types of behavior for $G(\phi)$.

Observe that

$$Ah^2 + 2Bhk + Ck^2 = r^2G(\phi).$$

This shows that the sign of $G(\phi)$ is the same as the sign of the quadratic function

$$f(h, k) = Ah^2 + 2Bhk + Ck^2.$$

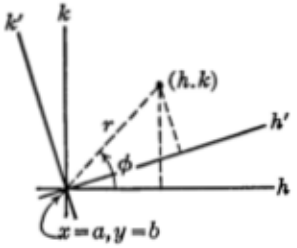


Fig. 51.

Now let us regard h, k as rectangular co-ordinates in a system with origin at the point $x = a$, $y = b$. Let h', k' denote rectangular co-ordinates in a system obtained from the hk-system by a rotation about the origin of the system (see Fig. 51). As we saw in §6.9, it is possible to choose the rotation in such a way that $f(h, k)$ becomes

$$\lambda_1 h'^2 + \lambda_2 k'^2 = r^2 G(\phi), \tag{7.6–6}$$

where λ_1 and λ_2 are roots of the equation

$$\begin{vmatrix} A-\lambda & B \\ B & C-\lambda \end{vmatrix} = \lambda^2 - (A+C)\lambda + AC - B^2 = 0. \tag{7.6–7}$$

Here $h^2 + k^2 = h'^2 + k'^2 = r^2$. Observe that the product of the roots of (7.6–7) is

$$\lambda_1\lambda_2 = AC - B^2, \tag{7.6–8}$$

and that the sum is

$$\lambda_1 + \lambda_2 = A + C. \tag{7.6–9}$$

Everything now depends on the sign of the expression (7.6–6). It is clear that if λ_1 and λ_2 are both positive, $G(\phi)$ is positive, and we have the case of a minimum at (a, b), whereas if λ_1 and λ_2 are both negative, we have the case of a maximum at (a, b). Let us now consider cases (i) and (ii) of the theorem. The hypothesis $B^2 - AC < 0$ implies that A and C are of the same sign, and also, by (7.6–8), that λ_1 and λ_2 are of the same sign. Consequently, from (7.6–9) we see that $B^2 - AC < 0$ and $A > 0$ imply that λ_1 and λ_2 are positive, whereas $B^2 - AC < 0$ and $A < 0$ imply that λ_1 and λ_2 are negative. The conclusions in cases (i) and (ii) are therefore established by the foregoing arguments.

In case (iii), $B^2 - AC > 0$ implies that λ_1 and λ_2 are of opposite signs. Now we can choose ϕ so as to make $h' = 0$, $k' \neq 0$, $r^2G(\phi) = \lambda_2 k'^2$, and we can also choose ϕ so as to make $h' \neq 0$, $k' = 0$, $r^2G(\phi) = \lambda_1 h'^2$. Thus $G(\phi)$ can change sign, and so we have neither a maximum nor a minimum at (a, b).

Finally, if $B^2 - AC = 0$, at least one of the roots λ_1 and λ_2 is zero, by (7.6–8). No conclusion about a maximum or minimum can be drawn in this case. The reasons for this are clear from (7.6–6) and (7.6–5). As examples we cite the three functions

$$x^2, -y^2, x^4 - y^2.$$

Each of these functions comes under case (iv) of Theorem IX, with $a = b = 0$. The first has a minimum at $(0, 0)$, the second a maximum, while the third has neither.

A point at which all the first partial derivatives of a function are equal to zero is called a *critical point* of the function. If (a, b) is a critical point of $F(x, y)$, and if $B^2 - AC \neq 0$ at (a, b) (in the notation of Theorem IX), we call the critical point *nondegenerate.* The point is called a *saddle point* if $B^2 - AC > 0$.

It should be noted that, if we assume that F has partial derivatives of all orders, and that it can be represented by its Taylor's series (see §7.5), then, when (a, b) is a critical point, the Taylor's series starts as follows:

$$F(a + h, b + k) = F(a, b) + \frac{1}{2!}[Ah^2 + 2Bhk + Ck^2] + \cdots.$$

Observe also that $AC - B^2$ is the discriminant of the quadratic form

$$Ah^2 + 2Bhk + Ck^2.$$

This gives us a clue to the proper method of defining nondegenerate critical points for functions of more than two variables. We illustrate briefly for the case of three variables. Suppose the origin is a critical point of the function $F(x_1, x_2, x_3)$, so that $F_1(0, 0, 0) = F_2(0, 0, 0) = F_3(0, 0, 0) = 0$. Let $a_{11} = F_{11}(0, 0, 0)$, $a_{12} = F_{12}(0, 0, 0)$, $a_{13} = F_{13}(0, 0, 0)$, and so on.
Then Taylor's series is

$$F(x_1, x_2, x_3) = F(0, 0, 0) + \frac{1}{2!}\sum_{i,j=1}^{3} a_{ij}x_i x_j + \cdots.$$

The critical point is called nondegenerate in case the determinant

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

is not zero. For nondegenerate critical points there is a generalization of Theorem IX. We consider the roots λ_1, λ_2, λ_3 of the equation

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & a_{13} \\ a_{21} & a_{22} - \lambda & a_{23} \\ a_{31} & a_{32} & a_{33} - \lambda \end{vmatrix} = 0.$$

If these roots are all of the same sign, F has a *minimum* at the critical point if the roots are all *positive,* and a *maximum* there if they are all *negative.* But, if some roots are positive and some negative, there is neither a maximum nor a minimum at the critical point. These facts follow quickly from the discussion of quadratic forms in §6.9. It is not actually necessary to assume the convergence of Taylor's series. All that is needed is that the first derivatives of F be differentiable at the critical point.

We shall now indicate how the foregoing critical point theory can be developed for functions of more than two variables. We suppose that we have a function $F(x_1, \ldots, x_n)$ of n independent variables, defined and differentiable at all points in a region including $(a_1, \ldots, a_n)$ as an interior point. We shall assume that $(a_1, \ldots, a_n)$ is a critical point of F that is, that all the first partial derivatives of F vanish at this point. And finally we assume that each of the first partial derivatives of F is differentiable at least at this one point $(a_1, \ldots, a_n)$. By an easy extension of Theorem IV to n variables, these hypotheses assure us that the second partial derivatives exist at the critical point, and that the symmetrical relations of the type $\dfrac{\partial^2 F}{\partial x_i\, \partial x_j} = \dfrac{\partial^2 F}{\partial x_j\, \partial x_i}$ are all valid. As usual, we use the notation $\dfrac{\partial^2 F}{\partial x_j\, \partial x_i} = F_{ij}$. From now on we restrict our attention to the points contained within some ball centered at $(a_1, \ldots, a_n)$ where the radius of the ball is small enough so that all points within the ball belong also to the region in which our assumptions about F hold.

To compare the value of F at $(x_1, \ldots, x_n)$ with the value at $(a_1, \ldots, a_n)$ we use the law of the mean. This was proved for functions of two variables in Theorem VI, and later in §7.4 it was pointed out that the natural extension to functions of n variables is also valid. For convenience let $h_i = x_i - a_i$. Then there is some θ, $(0 < \theta < 1)$, such that

$$F(x_1, \ldots, x_n) - F(a_1, \ldots, a_n) = \sum_{i=1}^{n} h_i F_i(a_1 + \theta h_1, \ldots, a_n + \theta h_n) \qquad (7.6\text{–}10)$$

Next, we use the differentiability of all the F_i's at $(a_1, \ldots, a_n)$. It enables us to write

$$F_i(a_1 + \Delta x_1, \ldots, a_n + \Delta x_n) - F_i(a_1, \ldots, a_n)$$
$$= \sum_{j=1}^{n} F_{ij}(a_1, \ldots, a_n)\, \Delta x_j + \epsilon_i \sqrt{(\Delta x_1)^2 + \cdots + (\Delta x_n)^2} \qquad (7.6\text{–}11)$$

where ϵ_i depends on $\Delta x_1, \ldots, \Delta x_n$ in such a way that $\epsilon_i \to 0$ when all the Δx_i's $\to 0$. See (7.2). If we now put $\Delta x_i = \theta h_i$ in (7.6–11) and combine with (7.6–10), bearing in mind that the first order partial derivatives all vanish at the critical point, we see that

$$F(x_1, \ldots, x_n) - F(a_1, \ldots, a_n) = \theta \left[\sum_{i,j=1}^{n} F_{ij}(a_1, \ldots, a_n)\, h_i h_j \right]$$
$$+ \theta (h_1^2 + \cdots + h_n^2)^{1/2} \sum_{i=1}^{n} \epsilon_i h_i. \qquad (7.6\text{–}12)$$

The value of θ is unknown but we do know that it is properly between 0 and 1 and hence positive. Consequently, the sign of the difference $F(x_1, \ldots, x_n) - F(a_1, \ldots, a_n)$ is the same as the sign of

$$\sum_{i,j=1}^{n} F_{ij}(a_1, \ldots, a_n)\, h_i h_j + (h_1^2 + \cdots + h_n^2)^{1/2} \sum_{i=1}^{n} \epsilon_i h_i. \qquad (7.6\text{–}13)$$

The first summation is a quadratic form in the variables $h_1, \ldots, h_n$. The absolute value of the second part is small in comparison with $(h_1^2 + \cdots + h_n^2)$ when the latter is small, because the ϵ_i's approach zero as the h_i's $\to 0$. In fact, by Cauchy's inequality, (see §6.8, Exercise 8),

$$\left|\sum_{i=1}^{n} \epsilon_i h_i\right| \leq \left(\sum_{i=1}^{n} \epsilon_i^2\right)^{1/2} \left(\sum_{i=1}^{n} h_i^2\right)^{1/2},$$

$$\left|(h_1^2 + \cdots + h_n^2)^{1/2} \sum_{i=1}^{n} \epsilon_i h_i\right| \leq \left(\sum_{i=1}^{n} \epsilon_i^2\right)^{1/2} \left(\sum_{i=1}^{n} h_i^2\right), \tag{7.6–14}$$

justifying the assertion made in the preceding sentence.

This brings us to the essential aspect of our theory, and we focus attention on the quadratic form

$$Q(h_1, \ldots, h_n) = \sum_{i,j=1}^{n} F_{ij}(a_1, \ldots, a_n) h_i h_j.$$

If this quadratic form remains always of the same sign, no matter how the h_i's are chosen, so long as they are not all zero, then as we shall see in a moment, the sign of the quadratic form and the sign of $F(x_1, \ldots, x_n) - F(a_1, \ldots, a_n)$ will be the same when $(x_1, \ldots, x_n)$ is sufficiently close to $(a_1, \ldots, a_n)$, that is, when $(h_1^2 + \cdots + h_n^2)^{1/2}$ is sufficiently small. Then we have a relative minimum at $(a_1, \ldots, a_n)$ if the difference is positive and a relative maximum at $(a_1, \ldots, a_n)$ if the difference is negative.

The proof of what has been said depends on a straightforward generalization from $F(x, y, z)$ to $F(x_1, \ldots, x_n)$ in Exercise 14 of §6.9. This shows that by a suitable rotation of axes we can transform the quadratic form to one having only squares,

$$\sum_{i,j=1}^{n} F_{ij}(a_1, \ldots, a_n) h_i h_j = \sum_{i=1}^{n} \lambda_i k_i^2, \tag{7.6–15}$$

where the coefficients are roots of the equation

$$\begin{vmatrix} F_{11} - \lambda & F_{12} & \ldots & F_{1n} \\ F_{21} & F_{22} - \lambda & \ldots & F_{2n} \\ \vdots & & & \\ F_{n1} & F_{n2} & \ldots & F_{nn} - \lambda \end{vmatrix} = 0 \tag{7.6–16}$$

and F_{ij}, of course, is an abbreviation for $F_{ij}(a_1 \ldots a_n)$. It is not hard to see that this determinant is a polynomial of degree n in the variable λ, and therefore we can think of the equation in the form

$$c_n \lambda^n + c_{n-1} \lambda^{n-1} + \cdots + c_1 \lambda + c_0 = 0, \tag{7.6–16$'$}$$

in which it is not hard to see that $c_n = (-1)^n$, while the constant term, c_0, is given

by

$$\begin{vmatrix} F_{11} & F_{12} & \dots & F_{1n} \\ F_{21} & F_{22} & \dots & F_{2n} \\ \vdots & & & \\ F_{n1} & F_{n2} & \dots & F_{nn} \end{vmatrix}.$$

If this determinant is not zero, then we say that the critical point is *non-degenerate.* The significant thing about nondegenerate critical points is immediately apparent from (7.6–16)′, that *none of the* λ_i*'s can be zero.* It is also clear that if the critical point is *degenerate,* that is, if the above determinant is zero at the critical point, then at least one of the λ_i's is zero.

We should perhaps remind ourselves at this point that it follows from the generalization of Exercise 14 of §6.9 that, whether the critical point is degenerate or nondegenerate, all the λ_i's have to be real. This is seen from the fact that the λ_i's arise there as a sequence of extreme values of a real function under a sequence of increasingly limiting constraints. To those who have studied matrix theory it is also obvious from the fact that the λ_i's are characteristic values of a real symmetric matrix.

It is obvious from (7.6–15) that the quadratic form cannot change sign if all the λ_i's are positive or if they are all negative. However, if some are positive and others are negative then the quadratic form does change sign. Quadratic forms which are positive (negative) except when all the variables are zero are said to be *positive (negative) definite.* If a form is never negative (positive) but may vanish when not all the variables are zero, it is called *positive (negative) semidefinite.* Quadratic forms which take both positive and negative values are said to be *indefinite.*

The ordered n-tuple $(k_1, \dots, k_n)$ in (7.6–15) represents the same vector after the rotation of axes as was represented by $(h_1, \dots, h_n)$ in the original coordinate system. The individual components of a vector change under a rotation of axes but length does not. Therefore $\sum_{i=1}^{n} h_i^2 = \sum_{i=1}^{n} k_i^2$. Using this fact we can infer from (7.6–14) and (7.6–13) that the sign of $F(x_1, \dots, x_n) - F(a_1, \dots, a_n)$ is the same as the sign of $\sum_{i=1}^{n} \lambda_i k_i^2 + R$, where $|R| \leq \left(\sum_{i=1}^{n} \epsilon_i^2\right)^{1/2} \sum_{i=1}^{n} k_i^2$. Consider first the case where all the λ_i's are positive and let the smallest one be m. Then from the above we see that

$$\sum_{i=1}^{n} \lambda_i k_i^2 + R \geq \left[m - \left(\sum_{i=1}^{n} \epsilon_i^2\right)^{1/2} \right] \sum_{i=1}^{k} k_i^2,$$

and the ϵ_i's will be so small that this is positive if we stay close enough to $(a_1, \dots, a_n)$. Similarly, turning now to the case where all the λ's are negative and the largest is $-m$, we can write

$$\sum_{i} k_i^2 + R \leq \left[-m + \left(\sum_{i=1}^{n} \epsilon_i^2\right)^{1/2} \right] \sum_{i=1}^{n} k_i^2.$$

The right-hand side will be negative if we constrain the ϵ_i's to be small enough by staying sufficiently close to $(a_1, \ldots, a_n)$.

So in summary we can say that if the λ's are all positive, $F(x_1, \ldots, x_n) > F(a_1, \ldots, a_n)$ for all $(x_1, \ldots, x_n)$ sufficiently close to $(a_1, \ldots, a_n)$ and thus the critical point is one where F has a local minimum. Similarly, if all the λ_i's are negative, then $F(x_1, \ldots, x_n) < F(a_1, \ldots, a_n)$ throughout some neighborhood of the critical point and hence F has a local maximum. Actually we have proved a little bit more than this. By definition, F has a local minimum at $(a_1, \ldots, a_n)$ in case there is some neighborhood of this point at every other point of which the value of the function is merely greater than *or equal to* the value at $(a_1, \ldots, a_n)$. We have shown that at a critical point where all the λ_i's are positive, there is a neighborhood at all other points of which F is strictly greater than $F(a_1, \ldots, a_n)$. Such a point is not just a local minimum but a *strict local minimum.* In a similar way we distinguish between a local maximum and a *strict local maximum.* Now, what we have actually proved is that at a critical point where all the λ_i's are positive the function has a strict local minimum and at a critical point where all the λ_i's are negative, the function has a strict local maximum.

These criteria which we have just obtained are not very practical because they give no way of telling whether the λ_i's are all positive or all negative and it would usually not be feasible to compute all these numbers by solving (7.6–16)′. Fortunately there is a method, sometimes known as Gundelfinger's rule, which can, without finding all the λ_i's, tell us whether they are all of the same sign. To use it we begin by computing $d_1, d_2, \ldots, d_n$ defined as follows:

$$d_1 = F_{11}; \; d_2 = \begin{vmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{vmatrix}; \ldots; \; d_n = \begin{vmatrix} F_{11} & F_{12} & \ldots & F_{1n} \\ F_{21} & F_{22} & \ldots & F_{2n} \\ \ldots\ldots & & & \\ F_{n1} & F_{n2} & \ldots & F_{nn} \end{vmatrix}$$

Gundelfinger's rule tells us that:

(i) A necessary and sufficient condition that all the λ_i's be positive is that all the d_i's be positive.

(ii) A necessary and sufficient condition that all the λ_i's be negative is that the d_i's alternate in sign with d_1 being negative.

Both (i) and (ii) apply only to critical points where all the λ_i's are different from zero, that is, to nondegenerate critical points, but they do enable us to classify nondegenerate critical points by looking at the easily obtainable d_i's instead of the hard-to-find λ_i's. Our previous arguments allow us to assert immediately that:

(i) A necessary and sufficient condition that a nondegenerate critical point give a strict local minimum is that all the d_i's be greater than zero.

(ii) A necessary and sufficient condition that a nondegenerate critical point give a strict local maximum is that the d_i's alternately less than zero and greater than zero with d_1 being less than zero.

(iii) A nondegenerate critical point which is neither a strict local minimum nor a strict local maximum is a saddle point, where the function has a strict local maximum from some directions and a strict local minimum from others.

There are extensions of Gundelfinger's rule which enable us to classify degenerate critical points, where the maxima and minima are nonstrict, but we shall not include these details.

EXERCISES

1. Find all the critical points of each of the following functions. Test each critical point by Theorem IX, and state your conclusion.

(a) $y^2+3x^4-4x^3-12x^2+24$.

(b) $x^2-12y^2+4y^3+3y^4$.

(c) $x^4+y^4-2x^2+4xy-2y^2$.

(d) $x^2y^2-5x^2-8xy-5y^2$.

(e) $xy(12-3x-4y)$.

(f) $x^3y^2(a-x-y)$, $a>0$.

(g) $(1-x)(1-y)(x+y-1)$.

(h) $x^2y(24-x-y)^3$.

(i) $\dfrac{1}{xy}-\dfrac{4}{x^2y}-\dfrac{8}{xy^2}$.

(j) $\frac{5}{4}(x^2+y^2)-18x-24y+5\sqrt{x^2+y^2}+250$.

(k) $5(x^2+y^2)-24x-32y-60\sqrt{x^2+y^2}+1000$.

(l) $12x\sin y-2x^2\sin y+x^2\sin y\cos y$.

2. If a and b are positive, show that $(a/x)+(b/y)+xy$ has a minimum at its only critical point. What is the situation if a and b are both negative? if they have opposite signs?

3. How many critical points has the function $(ax^2+by^2)e^{-x^2-y^2}$ if $b>a>0$? Discuss the nature of each of them.

4. Find the shortest distance from the point $(1,-1,1)$ to the surface $z=xy$. Set up the squared distance as a function of x, y, find the critical points of the function, and test them by Theorem IX.

5. Discuss the problem of finding the shortest distance from the point $(0,0,a)$ to the surface $z=xy$, where $a>0$. Proceed as directed in Exercise 4. Separate the case $0<a\leqq 1$ and $1<a$.

6. If z is defined as a function of x, y by the equation $2x^2+2y^2+z^2+8xz-z+8=0$, find the points (x, y, z) at which z has a relative extreme, and test by second derivatives for a maximum or minimum in each case.

7. Proceed as directed in Exercise 6, starting from the equation

$$x^2+2y^2+3z^2-2xy-2yz=2.$$

8. Suppose that F is defined in the neighborhood of (a,b), and that $F_1(a,b)=0$, $F_{11}(a,b)<0$. Why is it impossible for F to have a relative minimum at (a,b)?

9. Locate the critical points of the function $xyz(x+y+z-1)$. Show that there are six lines all of whose points are degenerate critical points, and one nondegenerate critical

point. Is this a maximum or a minimum point? Can you answer this last question without second derivative tests?

10. Show that every critical point of the function $\dfrac{x^3+y^3+z^3}{xyz}$ is degenerate.

11. Locate the critical points of the function

$$F(x, y, z) = (ax^2 + by^2 + cz^2)e^{-x^2-y^2-z^2},$$

where $a > b > c > 0$. Show that there are two points of maximum value of F, one point of minimum value, and four critical points at which there is neither a maximum nor a minimum.

12. Study the function $F(x, y) = (y^2 - x^2)(y^2 - 2x^2)$. Show that there are four lines which divide the plane into eight regions, in each of which F has a constant sign. Discuss the critical point of the function. Are there any relative extrema?

13. Discuss the sign of the function $F(x, y) = (2x^2 - y)(x^2 - y)$ at various points of the plane, by appropriate consideration of the regions into which the plane is divided by the two parabolas $y = x^2$, $y = 2x^2$. Discuss the critical points of the function. Show that, along every straight line through the origin, the values of F reach a minimum at $(0, 0)$ but that F has neither a maximum nor a minimum at $(0, 0)$.

MISCELLANEOUS EXERCISES

1. Generalize Theorem I of §6.3 to functions of n variables.

2. Suppose $f(x, y)$ is differentiable at (a, b), with $A = f_1(a, b)$, $B = f_2(a, b)$. Let $F(r, \theta) = f(a + r\cos\theta, b + r\sin\theta)$. Then $F_1(0, \theta)$ exists and is equal to $A\cos\theta + B\sin\theta$, for every θ. Prove this directly from the definition of differentiability of f and the fact that $F_1(0, \theta) = \lim_{r\to 0}(1/r)\{F(r, \theta) - F(0, \theta)\}$.

3. If $F(x, y) = (1 - x)(1 - y)(x + y - 1)$, $a = b = \frac{2}{3}$, write Taylor's series for $F(a + h, b + k)$. What do you conclude about the sign of the difference $F(a + h, b + k) - F(a, b)$ when h and k are small?

4. Define $f(x, y) = (x^3 - y^3)/(x^2 + y^2)$ if $x^2 + y^2 \neq 0$, and $f(0, 0) = 0$. If we introduce cylindrical co-ordinates (r, θ, z) in the usual way, the surface $z = f(x, y)$ is represented by $z = r(\cos^3\theta - \sin^3\theta)$. Observe that the surface consists of a bundle of half-lines; the half-line corresponding to a fixed value of θ starts at the origin and passes through the cylinder $x^2 + y^2 = 1$ at a point for which $z = \cos^3\theta - \sin^3\theta$.

By plotting the curve $z = \cos^3\theta - \sin^3\theta$ with θ and z treated as plane rectangular co-ordinates, and then rolling up the plane to form a cylinder, one can visualize the surface. Do this. Does the surface have a tangent plane at the origin?

5. Suppose that f and ϕ are functions of a single variable, and that each function has continuous first and second derivatives. We shall suppose that $\phi(a) = c \neq 0$ and that $\phi'(a) \neq 0$. Let $F(x, y, z) = f(x) + f(y) + f(z)$, $G(x, y, z) = \phi(x)\phi(y)\phi(z)$. Consider the extremal problem for $F(x, y, z)$ subject to the constraint $G(x, y, z) = c^2$. Show that a possible solution of the problem occurs when $x = y = z = a$, and that the extreme will be a relative minimum if

$$f'(a)\left\{\frac{\phi''(a)}{\phi'(a)} - \frac{\phi'(a)}{\phi(a)}\right\} < f''(a),$$

and a relative maximum if the inequality is reversed. As instances consider: first, $f(x) = x^2$, $\phi(x) = x$, $a > 0$; and second, $f(x) = e^{-x^2}$, $\phi(x) = x$.

8 / IMPLICIT-FUNCTION THEOREMS

8 / THE NATURE OF THE PROBLEM OF IMPLICIT FUNCTIONS

We have already acquired some familiarity with implicit functions, in §§6.1 and 6.6. Thus far what we have learned about implicit functions has been concerned almost entirely with techniques for differentiating such functions in concrete special cases (§6.1), or with general formulas for the derivatives of implicit functions in terms of functional notation (§6.6). In all this earlier work we have taken for granted the existence and differentiability of the implicit functions. In this chapter we shall inquire into these matters which have been taken for granted.

Consider then an equation in three variables,

$$F(x, y, z) = 0. \tag{8–1}$$

In certain situations we say that an equation of this form has a solution

$$z = f(x, y). \tag{8–2}$$

Our present purpose is to examine the following questions: *What does it mean to say that* (8–2) *is a solution of* (8–1)? *Under what conditions is a solution possible*? *What can be said about the differentiability of the function f in terms of what may be known about the function F*? The answers to these and related questions will occupy us in this chapter.

Questions of the same sort can be asked about other implicit-function situations. We may have $y = f(x)$ as a solution of $F(x, y) = 0$, or $z = f(u, v, w, x, y)$ as a solution of $F(u, v, w, x, y, z) = 0$. Or, there may be several functions which arise as solutions of a system of several equations. Under suitable conditions a set of r equations in $n + r$ variables may determine r of the variables as functions of the remaining n variables.

The nature of the problem is most easily understood, and the explanation of the theory is the simplest, in the case of an implicit function arising as a solution of a single equation in a small number of variables. We shall discuss the case of two variables first, and then the case of three variables. The discussion in this section is intended to provide motivation for, and understanding of, the formal statements of theorems in later sections.

Let F be a function of x and y, defined in a certain region of the xy-plane. Consider the equation

$$F(x, y) = 0. \tag{8–3}$$

This equation expresses a condition which a point (x, y) may or may not satisfy. If there are some points which satisfy the condition, the set of all such points may be called the locus defined by (8–3). We know that, in many cases, the locus is some kind of curve. For instance, if $F(x, y) = 4x^2 + 9y^2 - 36$, the locus is an ellipse. Now let $f(x)$ be a function defined for a certain set of values of x. We say that $y = f(x)$ is a solution of (8–3) if all the points $(x, f(x))$ are part of the locus defined by (8–3), that is, if $F(x, f(x)) = 0$ for all the values of x which are involved in the definition of the function f. We assume without any further explicit mention that all functions which enter the discussion are single-valued.

In our work with implicit functions we do not attempt to get a solution of (8–3) in the form $y = f(x)$ such that it gives us *all* the locus defined by (8–3), for this may be impossible. Thus in the case of the ellipse $4x^2 + 9y^2 - 36 = 0$, part of the locus is given by the graph of $y = \frac{1}{3}(36 - 4x^2)^{1/2}$, and another part of the locus by the graph of $y = -\frac{1}{3}(36 - 4x^2)^{1/2}$. What we *do* attempt is to start with a particular point (x_0, y_0) of the locus defined by (8–3), and then to obtain a function $f(x)$, defined in some interval $x_0 - a < x < x_0 + a$, such that $y = f(x)$ is a solution of (8–3), and such that, *in a sufficiently restricted neighborhood of* (x_0, y_0), *all* the points for which $F(x, y) = 0$ are given by $y = f(x)$. This *localization* of the problem to a neighborhood of a particular point (x_0, y_0) is characteristic of all the treatment of implicit-function problems in this chapter.

Fig. 52 shows how, in the case of the equation $4x^2 + 9y^2 - 36 = 0$, localization of attention to a suitable neighborhood of one particular point of the locus leads to the determination of a solution $y = f(x)$ whose graph comprises all that part of the locus which is in the neighborhood. In Fig. 52 two such localizations are shown, the neighborhood being rectangular in each case. Observe that, if the center of one of the rectangular neighborhoods is a point (x_0, y_0) on the ellipse, the permissible size of the rectangle is governed by the consideration that every line parallel to the y-axis and passing through the interior of the rectangle shall intersect the ellipse exactly once inside the rectangle. It is always possible to choose the rectangle so as to satisfy this condition provided that the point (x_0, y_0) is not one of the points $(\pm 3, 0)$. These are the points where $\partial F/\partial y = 0$, with $F(x, y) = 4x^2 + 9y^2 - 36$. In our study of functions $y = f(x)$ defined implicitly as solutions of $F(x, y) = 0$ we shall always localize the problem within a neighborhood of a point at which $\partial F/\partial y \neq 0$. If we wanted to solve for x terms of y instead of for y in terms of x, we would impose the requirement $\partial F/\partial x \neq 0$.

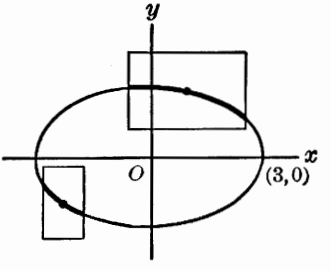


Fig. 52.

Let us turn now to a consideration of equations in three variables. Suppose that $F(x, y, z)$ is defined in a certain region in three-dimensional space. We consider the locus defined by the equation

$$F(x, y, z) = 0, \tag{8–4}$$

and suppose that (x_0, y_0, z_0) is a point of this locus. We then confine our attention to points near (x_0, y_0, z_0), and ask the following question: Is it possible to find a rectangular box defined by certain inequalities

$$|x - x_0| < a, |y - y_0| < b, |z - z_0| < c$$

such that every line parallel to the z-axis and passing through the interior of the box intersects the locus defined by (8–4) exactly once inside the box? If so, then to each pair (x, y) for which $|x - x_0| < a$ and $|y - y_0| < b$ there corresponds a unique z such that $|z - z_0| < c$ and $F(x, y, z) = 0$. This defines z as a single-valued function of (x, y), say $z = f(x, y)$, and gives a solution of (8–4), that is, $F(x, y, f(x, y)) = 0$.

Under certain conditions it is possible to choose a box of the sort just described. We shall explain the plausibility of this statement from a geometrical point of view. Suppose that the locus defined by (8–4) is a surface, such as an ellipsoid, a hyperboloid, or a cone. Suppose that the surface is smooth, and that the tangent plane at the point (x_0, y_0, z_0) is not parallel to the z-axis. Then we expect the part of the surface near (x_0, y_0, z_0) to be almost like the nearby part of the tangent plane; it should then be represented by an equation $z = f(x, y)$, since each line parallel to the z-axis may be expected to intersect the surface exactly once in the vicinity of (x_0, y_0, z_0), provided that $x - x_0$ and $y - y_0$ are sufficiently small (see Fig. 53).

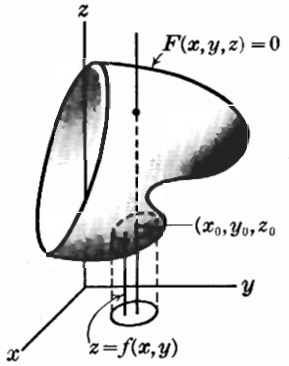


Fig. 53.

The condition on F in order that the plane tangent to the surface at (x_0, y_0, z_0) shall not be parallel to the z-axis is $\frac{\partial F}{\partial z} \neq 0$ at the point, for the ratios $\frac{\partial F}{\partial x} : \frac{\partial F}{\partial y} : \frac{\partial F}{\partial z}$ define the direction of the normal to the surface; therefore $\frac{\partial F}{\partial z} = 0$ means that the normal is perpendicular to the z-axis.

Our discussions of the implicit-function problem for the two cases $F(x, y) = 0$ and $F(x, y, z) = 0$, with geometrical evidence for the solutions $y = f(x)$ and $z = f(x, y)$ in a localized form of the problem, suggest the kind of answers we may expect to some of the questions raised at the beginning of the present section. We do not yet have any real *proofs*, however. The proofs must come out of the analytical situation, for we do not really know the facts about geometry of the curves and surfaces except by an examination of the functions and equations.

8.1 / THE FUNDAMENTAL THEOREM

The following theorem is concerned with a precise statement bearing on the questions raised at the outset in §8.

THEOREM I. *Let $F(x, y, z)$ be a function defined in an open set S containing the point (x_0, y_0, z_0). Suppose that F has continuous first partial derivatives in S. Furthermore assume that*

$$F(x_0, y_0, z_0) = 0, F_3(x_0, y_0, z_0) \neq 0.$$

Under these conditions there exists a box-like region defined by certain inequalities

$$|x - x_0| < a, |y - y_0| < b, |z - z_0| < c,$$

lying in the region S and such that the following assertions are true:

Let R be the rectangular region $|x - x_0| < a, |y - y_0| < b$ in the xy-plane. Then

1. *For any (x, y) in R there is a unique z such that*

$$|z - z_0| < c \quad \text{and} \quad F(x, y, z) = 0.$$

Let us express this dependence of z on (x, y) by writing

$$z = f(x, y) \tag{8.1-1}$$

2. *The function f is continuous in R.*
3. *The function f has continuous first partial derivatives given by*

$$f_1(x, y) = -\frac{F_1(x, y, z)}{F_3(x, y, z)}, f_2(x, y) = -\frac{F_2(x, y, z)}{F_3(x, y, z)},$$

where z is given by (8.1–1).

Proof. The first part of the proof is concerned with determining suitable values for the positive constants a, b, c which are mentioned in the theorem. Let A be a rectangular parallelpiped (box) with center at (x_0, y_0, z_0) such that the whole of A is entirely in the region S, and such that, moreover, $F_3(x, y, z)$ has everywhere in A the same sign which it has at (x_0, y_0, z_0). This choice of A is possible since S is open and F_3 is continuous (see Theorem III, §5.3).

For definiteness let us assume $F_3 > 0$ in A. Consider the top and bottom faces of the box A. If we denote the height of the box by $2c$, these faces will lie in the planes $z = z_0 \pm c$. Since $F_3 > 0$, the value of F increases as we go upward along any line parallel to the z-axis. Since $F(x_0, y_0, z_0) = 0$, it follows that

$$F(x_0, y_0, z_0 + c) > 0 \quad \text{and} \quad F(x_0, y_0, z_0 - c) < 0.$$

Because of the continuity of F we see that F will be positive in a small rectangle with center at $(x_0, y_0, z_0 + c)$ in the plane $z = z_0 + c$, and negative in a small rectangle with center at $(x_0, y_0, z_0 - c)$ in the plane $z = z_0 - c$. Let us choose positive numbers a, b, so that these rectangles are determined by the inequalities

$$|x - x_0| < a, |y - y_0| < b.$$

We also take care to choose a and b so that the box B defined by the

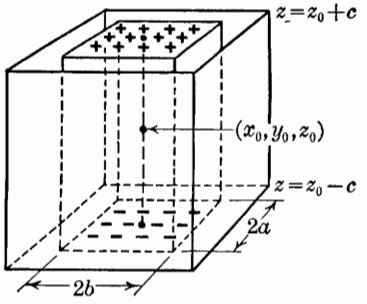


Fig. 54. Boxes A and B.

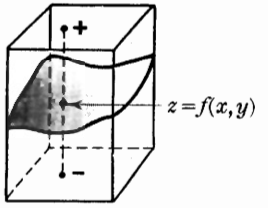


Fig. 55. Box B.

inequalities

$$|x - x_0| < a, |y - y_0| < b, |z - z_0| < c$$

is no larger than the box A (see Fig. 54).

Now consider the value of F along the segment in which a line parallel to the z-axis intersects the box B. As we go up along this segment the value of F increases. At the lower end of the segment, $F < 0$, while at the upper end $F > 0$. Hence there is *just one* point on the segment at which $F = 0$; for a given pair (x, y), the z-co-ordinate of this point is denoted by $z = f(x, y)$ (see Fig. 55). Assertion 1 of the theorem is now proved. Observe that thus far we have made no use of the partial derivatives F_1, F_2.

Having obtained the function $f(x, y)$ by the foregoing argument, let us now prove that it is continuous. To prove continuity at (x_0, y_0) we must show that, given $\epsilon > 0$, there exists $\delta > 0$ such that

$$|f(x, y) - f(x_0, y_0)| < \epsilon \quad \text{when } |x - x_0| < \delta \quad \text{and} \quad |y - y_0| < \delta. \tag{8.1-2}$$

We may assume that $\epsilon \leqq c$. Now $f(x_0, y_0) = z_0$; also,

$$F(x_0, y_0, z_0 + \epsilon) > 0, F(x_0, y_0, z_0 - \epsilon) < 0.$$

Hence, by the very argument used in proving assertion 1 of the theorem, we see that if we choose $\delta > 0$ so that

$$F(x, y, z_0 + \epsilon) > 0 \text{ and } F(x, y, z_0 - \epsilon) < 0$$

when

$$|x - x_0| < \delta \quad \text{and} \quad |y - y_0| < \delta, \tag{8.1-3}$$

then to each (x, y) satisfying (8.1–3) corresponds a unique z such that $|z - z_0| < \epsilon$ and $F(x, y, z) = 0$. This z must be equal to $f(x, y)$, however, by 1. Thus (8.1–2) holds, and the continuity at (x_0, y_0) is proved.

To prove continuity of f at any other point (x_1, y_1) of R, let $z_1 = f(x_1, y_1)$. Observe that F satisfies at (x_1, y_1, z_1) the same hypotheses as are stated in the theorem relative to the point (x_0, y_0, z_0). Hence, by what has already been proved

(as applied to this new situation), all the points (x, y, z) in the vicinity of (x_1, y_1, z_1) such that $F(x, y, z) = 0$ are furnished by a single-valued function (let us call it $g(x, y)$) which is continuous at (x_1, y_1). However, since all these points are in the box B, the uniqueness clause of conclusion 1 of the theorem assures us that $f(x, y) = g(x, y)$ when (x, y) is near (x_1, y_1). Hence f is continuous at (x_1, y_1).

It remains only to prove assertion 3 of the theorem. We shall deal with $\frac{\partial f}{\partial x}$; the treatment of $\frac{\partial f}{\partial y}$ is different only in the letters used. We employ the law of the mean (§7.4). Let (x, y) be a point of R, and let $z = f(x, y)$. We wish to show that

$$\lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x} = -\frac{F_1(x, y, z)}{F_3(x, y, z)}. \tag{8.1–4}$$

We assume that Δx is so small that $(x + \Delta x, y)$ is also in R, and write

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

Now, considering F as a function of x and z only, we have by the law of the mean,

$$F(x + \Delta x, y, z + \Delta z) - F(x, y, z) = \Delta x F_1(X, y, Z) + \Delta z F_3(X, y, Z), \tag{8.1–5}$$

where

$$X = x + \theta\, \Delta x, Z = z + \theta\, \Delta z, 0 < \theta < 1.$$

The left member of (8.1–5) is zero, by the definition of the function f. Hence

$$\frac{\Delta z}{\Delta x} = -\frac{F_1(X, y, Z)}{F_3(X, y, Z)}. \tag{8.1–6}$$

Now $\Delta z \to 0$ when $\Delta x \to 0$, by the continuity of f; therefore $X \to x$ and $Z \to z$. The truth of (8.1–4) is now seen to follow from (8.1–6), because F_1 and F_3 are continuous. The formula

$$f_1(x, y) = -\frac{F_1(x, y, f(x, y))}{F_3(x, y, f(x, y))}$$

has now been established. From it we see that f_1 is continuous (see Theorem IV, §5.3). This completes the proof.

8.2 / GENERALIZATION OF THE FUNDAMENTAL THEOREM

In the theorem of §8.1 we dealt with the existence of a function $z = f(x, y)$ defined implicitly by an equation $F(x, y, z) = 0$. We chose to state the theorem for the case of three variables, but nothing essential in the theorem or its proof is really dependent upon the particular number three. The analytical details and the geometrical language presented in §8.2 may all be modified easily to meet the situation of a different number of variables. We shall state the theorem formally in the general case ($n + 1$ variables). The proof will be omitted. The theorem tells

us what we can be certain of, under appropriate conditions, in speaking of a function

$$z = f(x_1, \ldots, x_n)$$

defined implicitly by an equation of the form

$$F(x_1, x_2, \ldots, x_n, z) = 0.$$

We use geometrical language in speaking of the regions of definition of the above functions.

THEOREM II. *Let $F(x_1, \ldots, x_n, z)$ be defined in an $(n+1)$-dimensional neighborhood of the point $(a_1, \ldots, a_n, c)$. Suppose that F has continuous partial derivatives in this neighborhood, and furthermore, assume that*

$$F(a_1, \ldots, a_n, c) = 0, \; F_{n+1}(a_1, \ldots, a_n, c) \neq 0.$$

Under these conditions there exists a box-like region defined by certain inequalities

$$|x_1 - a_1| < A_1, \ldots, |x_n - a_n| < A_n, |z - c| < C,$$

lying in the above neighborhood, and such that the following assertions are true:

Let R be the n-dimensional region

$$|x_1 - a_1| < A_1, \ldots, |x_n - a_n| < A_n$$

in the space of the variables $x_1, \ldots, x_n$. Then

1. *For any $(x_1, \ldots, x_n)$ in R there is a unique z such that*

$$|z - c| < C \quad \text{and} \quad F(x_1, \ldots, x_n, z) = 0.$$

Let us express this dependence of z on $(x_1, \ldots, x_n)$ by writing

$$z = f(x_1, \ldots, x_n).$$

2. *The function f is continuous in R.*
3. *The function f has continuous first partial derivatives given by*

$$\frac{\partial}{\partial x_i} f(x_1, \ldots, x_n) = -\frac{F_i(x_1, \ldots, x_n, z)}{F_{n+1}(x_1, \ldots, x_n, z)}, \; i = 1, \ldots, n,$$

where $z = f(x_1, \ldots, x_n)$.

EXERCISES

1. It is true that the part of the locus defined by $x + y + z - \sin xyz = 0$ near the point $(0, 0, 0)$ can be represented in the form $z = f(x, y)$?

2. How can you be sure that the equation $e^z(x^2 + y^2 + z^2) - \sqrt{1 + z^2} + y = 0$ has a solution $z = f(x, y)$ which is continuous at $x = 1$, $y = 0$, with $f(1, 0) = 0$? Using the tangent plane as an approximation to the surface, calculate $f(1 + h, k)$ approximately when h and k are small.

3. Can the equation $(x^2+y^2+z^2)^{1/2}-\cos z=0$ be solved uniquely for y in terms of x and z in the neighborhood of the point $(0,1,0)$? Can it be solved uniquely for z in terms of x and y in such a neighborhood?

4. Does there exist a function $f(x,y)$ continuous at $(1,-1)$, with $f(1,-1)=0$, and such that $x^3+y^3+[f(x,y)]^3=3xyf(x,y)$ at all points of a neighborhood of $(1,-1)$?

5. Show that $z^3+(x^2+y^2)z+1=0$ has a unique solution $z=f(x,y)$ defined for all x, y, and that f has continuous first partial derivatives everywhere.

6. Does the function $x(y-1)\sqrt{z}+x^2z^3+\sin x$ satisfy the hypotheses of Theorem I in a neighborhood of the point $(0,0,0)$?

7. For the purposes of this exercise let us make the following definition: A point (x_0, y_0, z_0) of the locus defined by $F(x,y,z)=0$ is called a *regular point* of the locus if the part of the locus is a sufficiently small neighborhood of the point can be represented in at least one of the three forms $z=f(x,y)$, $x=g(y,z)$, $y=h(z,x)$, where f is defined and continuous for all values of (x,y) sufficiently close to (x_0,y_0), $f(x_0,y_0)=z_0$, and corresponding requirements are placed on the functions g, h.

(a) Now suppose that F has continuous first partial derivatives in a neighborhood of (x_0,y_0,z_0), that $F(x_0,y_0,z_0)=0$, and that $\left(\frac{\partial F}{\partial x}\right)^2+\left(\frac{\partial F}{\partial y}\right)^2+\left(\frac{\partial F}{\partial z}\right)^2>0$ at the point. Prove that (x_0,y_0,z_0) is a regular point of the locus.

(b) Is the origin a regular point of the cone $x^2+y^2-z^2=0$?

(c) What is the locus defined by the equation $(x^2+y^2+z^2)^2-a^2(x^2+y^2+z^2)=0$, $a\neq 0$? Are there any non-regular points of the locus?

(d) If a, b, c are positive, show that all points except $(0,0,0)$ of the locus defined by $(x^2+y^2+z^2)^2-a^2x^2-b^2y^2-c^2z^2=0$ are regular.

8. In particular cases we may be able to arrive at information about $z=f(x,y)$ as a solution of $F(x,y,z)=0$ by methods quite different from those used to prove Theorem I. As an example, consider the equation

$$ye^z+xz-x^2-y^2=0.$$

Suppose that x and y are fixed, with $y\neq 0$. Consider w and z as variables, and draw the curve $w=e^z$ and the line

$$w=-\frac{x}{y}z+\frac{x^2+y^2}{y}$$

on the same wz-co-ordinate axes. If the line intersects the curve, the z-co-ordinate of the point of intersection depends on the parameters x, y, say $z=f(x,y)$, and this gives a solution of $F(x,y,z)=0$, where $F(x,y,z)=ye^z+xz-x^2-y^2$.

(a) Plot the graphs of the curve and the line when $y\neq 0$ and $x/y>0$, and show that there is a unique intersection.

(b) Show that if $x/y\leqq 0$ there may be no intersection, one intersection (tangency), or two intersections. Show that, if x, y, z are such that the line is tangent to the curve, then $\frac{\partial F}{\partial z}=0$.

9. **(a)** Give an example of an equation $F(x,y)=0$, where F is continuous but the equation is not satisfied by any points (x,y).

(b) Give another example in which $F(x,y)=0$ is satisfied by a certain point (x_0,y_0), but not by any other points near (x_0,y_0).

(c) Give an example in which the locus defined by $F(x, y, z) = 0$ is not a surface but a curve or a straight line.

10. Formulate an exactly worded theorem corresponding to Theorem I, but for the case of $y = f(x)$ defined as a solution of $F(x, y) = 0$, localized near a point (x_0, y_0). Give a detailed proof similar to the proof of Theorem I, and supply appropriate diagrams analogous to Figs. 54 and 55.

11. Is the locus of $y^2 + x^2 e^y = 0$ a curve?

12. What is the locus defined by

$$(e^{\sin x} - 1)^2 + (\sin y - 1)^2 = 0?$$

13. Suppose $F(x, y)$ has continuous first partial derivatives in a neighborhood of (a, b), that $F(a, b) = 0$, and that $\left(\frac{\partial F}{\partial x}\right)^2 + \left(\frac{\partial F}{\partial y}\right)^2 > 0$ at (a, b). Explain how you know that the part of the locus $F(x, y) = 0$ near (a, b) is a smooth curve without self-intersections.

14. If $F(x, y) = x^2 + y^2 - x^3$, find the solution $y = f(x)$ of $F(x, y) = 0$: **(a)** near the point $(5, 10)$; **(b)** near the point $(10, -30)$. **(c)** Near what points of the locus $F(x, y) = 0$ is there no solution $y = f(x)$ of the type considered in the discussion of Fig. 52?

15. If $F(x, y) = (y - x^2)^2 - x^5$, find the solution $y = f(x)$ of $F(x, y) = 0$: **(a)** near the point $(1, 0)$; **(b)** near the point $(1, 2)$. **(c)** What is the situation near the point $(0, 0)$?

16. What condition on a and b is sufficient to guarantee that the equation

$$yx^2 - y^3 + x^3 - a^2 b - a^3 + b^3 = 0$$

has a solution $y = f(x)$ where f is continuous at $x = a$ and $f(a) = b$?

17. Can the part of the curve $y^n(x - y) - (x + y) = 0$ near the origin be represented in the form $y = f(x)$? Assume that $n \geqq 1$. Draw a straight line segment which represents the curve approximately near the origin.

18. If $x_0 \neq 0$ and $x_0 \neq 1$, show that, if (x, y) is sufficiently near $(x_0, 0)$, the equation $\sin x^2 y - xy = 0$ is equivalent to $y = 0$.

19. **(a)** Find the unique solution $y = f(x)$ of $\tan y - xy = 0$ in the neighborhood of $(x_0, 0)$, if $x_0 \neq 1$.

(b) Show that, if (x_0, y_0) satisfies $\tan y - xy = 0$, and if $y_0 \neq 0$ and $\cos y_0 \neq 0$, the equation $\tan y - xy = 0$ defines uniquely a solution $y = f(x)$ such that f is continuous at x_0 and $y_0 = f(x_0)$.

20. Can the part of the curve $xe^y - y + 1 = 0$ near the point $(e^{-2}, 2)$ be represented in the form $y = f(x)$? Represent it in the form $x = g(y)$ and sketch the curve.

21. Make the following assumptions: (1) $F(x, y)$ is defined when $|x - x_0| < h$ and $|y - y_0| < k$; (2) $F(x_0, y_0) = 0$; (3) $F(x, y)$ is a continuous function of x for each y, and a continuous function of y for each x; (4) for each fixed x, $F(x, y)$ increases as y increases. Deduce that there exists a number c, $0 < c \leqq h$, and a continuous function $f(x)$, defined when $|x - x_0| < c$, such that $|f(x) - y_0| < k$, and such that the graph of $y = f(x)$ comprises all points inside the rectangle $|x - x_0| < c$, $|y - y_0| < k$ at which $F(x, y) = 0$.

8.3 / SIMULTANEOUS EQUATIONS

In this section we shall discuss the problem of implicit functions as it arises in connection with simultaneous equations. We have already considered examples of the technique of finding the partial derivatives of functions defined implicitly

by simultaneous equations (see §§6.1, 6.6). Our present concern is with *existence theorems* analogous to the fundamental theorems I, II given in §§8.1, 8.2.

Suppose we are given two equations in five variables. Experience with algebraic and trigonometric problems of this kind leads us to expect that, in certain cases at least, we may solve for two of the variables in terms of the remaining three. Thus, for instance, from equations

$$F(x, y, z, u, v) = 0,\ G(x, y, z, u, v) = 0, \tag{8.3–1}$$

we may be able to solve for u, v in terms of x, y, z:

$$u = f(x, y, z),\ v = g(x, y, z). \tag{8.3–2}$$

Or, again, if we have three equations in five variables, we may perhaps be able to express three of the variables in terms of the remaining two. The functional-notation expression of this situation would be that the equations

$$F(x, y, u, v, w) = 0,\ G(x, y, u, v, w) = 0,\ H(x, y, u, v, w) = 0$$

give rise to functions

$$u = f(x, y),\ v = g(x, y,),\ w = h(x, y).$$

The general situation suggested by these cases is that in which we are given r equations in $n + r$ variables (n and r positive integers). In certain cases, at least, such a set of equations will define r of the variables as functions of the remaining n variables. In seeking to understand the implicit-function problem for r simultaneous equations, it is easiest to begin with the case $r = 2$. The total number of variables makes very little difference. For definiteness we consider the case of two equations in five variables, say in the form (8.3–1).

The simplest simultaneous-equation systems are the linear systems. Let us, as a preliminary special case, suppose that equations (8.3–1) are linear in u and v, but not necessarily linear in x, y and z. Then they will have the form

$$\begin{aligned} A_1u + B_1v + C_1 &= 0 \\ A_2u + B_2v + C_2 &= 0, \end{aligned} \tag{8.3–3}$$

where the coefficients $A_1, B_1, \ldots, C_2$ all depend on x, y, z. This system of equations can be solved for u, v in terms of x, y, z provided the determinant

$$D = \begin{vmatrix} A_1 & B_1 \\ A_2 & B_2 \end{vmatrix}$$

does not vanish. Now, regarding (8.3–3) as a particular instance of (8.3–1), we see that

$$F(x, y, z, u, v) = A_1u + B_1v + C_1,$$

and thus

$$\frac{\partial F}{\partial u} = A_1,\ \frac{\partial F}{\partial v} = B_1.$$

Similar equations hold for G, and thus the determinant D is the same as the

Jacobian determinant

$$\frac{\partial(F, G)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial F}{\partial u} & \frac{\partial F}{\partial v} \\ \frac{\partial G}{\partial u} & \frac{\partial G}{\partial v} \end{vmatrix} \tag{8.3–4}$$

It will be seen by referring back to §6.6 that this same Jacobian arises in the denominators of the expressions for the partial derivatives of u and v as functions of x, y, z, assuming that such functions are defined by equations (8.3–1).

The foregoing considerations indicate that if we expect to solve equations (8.3–1) for u, v, we should make the assumption that the Jacobian (8.3–4) is different from zero. It must be kept in mind that in the general (nonlinear) case we are concerned not with the actual solution for u, v in the elementary sense of expressing u, v in terms of x, y, z by more or less simple formulas, but with the solution in the theoretical sense of knowing certainly that there exist functions (8.3–2) satisfying equations (8.3–1). A qualified guarantee of the existence of solutions in this theoretical sense is furnished by the following theorem:

THEOREM III. *Let S be a neighborhood of the point $P_0: (x_0, y_0, z_0, u_0, v_0)$ in the 5-dimensional space of the co-ordinates x, y, z, u, v. Suppose that the functions F, G occurring in the system* (8.3–1) *are continuous and have continuous first partial derivatives in S. Also assume that both functions vanish at the point P_0 but that the Jacobian* (8.3–4) *does not vanish at the point. Under these conditions there exists a box-like region lying in S, defined by certain inequalities*

$$|x - x_0| < a, |y - y_0| < b, |z - z_0| < c, \tag{8.3–5}$$

$$|u - u_0| < \alpha, |v - v_0| < \beta, \tag{8.3–6}$$

such that the following assertions are true:

Let R be the region defined, in the 3-dimensional space of the co-ordinates x, y, z, by the inequalities (8.3–5). *Then*

1. *To any (x, y, z) in R there corresponds a unique pair of values u, v such that the inequalities* (8.3–6) *are satisfied and the functions, F, G vanish (i.e., equations* (8.3–1) *are satisfied). This correspondence defines u and v as functions of x, y, z, say*

$$u = f(x, y, z), v = g(x, y, z).$$

2. *The functions f, g are continuous in R.*
3. *The functions f, g have continuous partial derivatives given by*

$$\frac{\partial f}{\partial x} = -\frac{1}{J}\frac{\partial(F, G)}{\partial(x, v)}, \frac{\partial g}{\partial x} = -\frac{1}{J}\frac{\partial(F, G)}{\partial(u, x)} \tag{8.3–7}$$

and similar formulas with x replaced by y, where

$$J = \frac{\partial(F, G)}{\partial(u, v)}. \tag{8.3–8}$$

Proof. The proof rests very heavily on use of Theorem II, §8.2. To begin with, we observe that the two partial derivatives $\frac{\partial F}{\partial v}, \frac{\partial G}{\partial v}$ cannot *both* vanish at the point P_0, for if they did, the Jacobian J would vanish there, contrary to the hypothesis. For definiteness assume that $\frac{\partial F}{\partial v}$ does not vanish at P_0. We are now able to apply Theorem II to the equation $F(x, y, z, u, v) = 0$, taking $n = 4$,

$$x_1 = x, x_2 = y, x_3 = z, x_4 = u, z = v.$$

As a result we obtain a function

$$v = \phi(x, y, z, u)$$

defined for (x, y, z, u) in a neighborhood of (x_0, y_0, z_0, u_0), and furnishing a solution of the equation $F(x, y, z, u, v) = 0$ for v. Next we substitute in G, writing

$$H(x, y, z, u) = G(x, y, z, u, \phi(x, y, z, u)).$$

The equation $G(x, y, z, u, v) = 0$ is thus replaced by the equation $H(x, y, z, u) = 0$. We seek to solve this equation for u. As a condition for being able to solve, we need to know that $\frac{\partial H}{\partial u} \neq 0$. Now, by the rule for composite functions,

$$\frac{\partial H}{\partial u} = \frac{\partial G}{\partial u} + \frac{\partial G}{\partial v}\frac{\partial v}{\partial u}.$$

But we know from Theorem II that

$$\frac{\partial v}{\partial u} = \frac{\partial \phi}{\partial u} = -\frac{\dfrac{\partial F}{\partial u}}{\dfrac{\partial F}{\partial v}}.$$

Hence

$$\frac{\partial H}{\partial u} = \frac{\partial G}{\partial u} - \frac{\partial G}{\partial v}\frac{\dfrac{\partial F}{\partial u}}{\dfrac{\partial F}{\partial v}} = \frac{\dfrac{\partial G}{\partial u}\dfrac{\partial F}{\partial v} - \dfrac{\partial G}{\partial v}\dfrac{\partial F}{\partial u}}{\dfrac{\partial F}{\partial v}},$$

or

$$\frac{\partial H}{\partial u} = -\frac{J}{\dfrac{\partial F}{\partial v}},$$

In a small neighborhood of the point P_0 neither J nor $\frac{\partial F}{\partial v}$ will vanish. We can

then apply Theorem II to the equation $H(x, y, z, u) = 0$ to obtain a solution $u = f(x, y, z)$. Finally, substituting in ϕ we obtain v as a function of x, y, z:

$$v = g(x, y, z) = \phi(x, y, z, f(x, y, z)).$$

We shall omit the exact details of the limitation of the magnitudes of the differences $x - x_0$, $y - y_0, \ldots$ in order to validate all the foregoing arguments. In applying Theorem II we are assured that the functions ϕ, f have continuous first partial derivatives. The function g, as a composite function, will then have continuous first partial derivatives also. The formulas (8.3–7) for the partial derivatives of f and g have already been obtained (see (6.6–11)). We can appeal to this earlier derivation now that we have proved the existence of f and g and the fact that they do possess continuous partial derivatives.

We shall not take space to state formally the analogue of Theorem III for systems of more than two equations. The nonvanishing of the appropriate Jacobian is the key condition. The proof of the general theorem for r equations may be made by mathematical induction on r. The proof for $r = 1$ is that of Theorem II; hence all that is necessary is to make the step from r to $r + 1$. This is not difficult, and may be patterned after the proof of Theorem III, which is the step from $r = 1$ to $r = 2$. Suggestions for this work are contained in Exercises 10, 11.

EXERCISES

1. Do there exist functions $f(x, y)$, $g(x, y)$, continuous in a neighborhood of $(0, 1)$, such that $f(0, 1) = 1$, $g(0, 1) = -1$, and such that

$$[f(x, y)]^3 + xg(x, y) - y = 0,$$
$$[g(x, y)]^3 + yf(x, y) - x = 0?$$

Explain your answer.

2. Suppose that the three equations

$$u^2 + v^2 + w^2 - x^2 = 0,\ u^2 + v^2 - y^2 = 0,\ u^2 + w^2 - z^2 = 0$$

are satisfied by a particular set of values $(x_0, y_0, z_0, u_0, v_0, w_0)$ of the variables. **(a)** What condition on this set of values is sufficient to insure that all "nearby" sets (x, y, z, u, v, w) satisfying the three equations are given by equations $u = f(x, y, z)$, $v = g(x, y, z)$, $w = h(x, y, z)$, where f, g, h are single-valued and continuous, with values u_0, v_0, w_0 respectively at (x_0, y_0, z_0)? **(b)** Solve the given equations explicitly for u^2, v^2, w^2, and from the solutions so found explain what happens if the sufficient condition in part **(a)** is not satisfied.

3. Suppose that $(x_0, y_0, z_0, u_0, v_0, w_0)$ satisfy the equations

$$u^2 + v^2 + w^2 = 1, \frac{u^2}{x^2} + \frac{v^2}{y^2} + \frac{w^2}{z^2} = 1.$$

What are sufficient conditions which guarantee that all "nearby" sets satisfying these equations can be represented in the form

$$u = f(x, y, z, w),\ v = g(x, y, z, w)?$$

4. Let (x_0, y_0, z_0, u_0) satisfy the equations

$$f(x)+f(y)+f(z)=F(u)$$
$$g(x)+g(y)+g(z)=G(u)$$
$$h(x)+h(y)+h(z)=H(u),$$

where all the functions involved have continuous derivatives. **(a)** State a sufficient condition for being able to solve for x, y, z in terms of u in the neighborhood of the given point. **(b)** What does the condition amount to in case $f(x)=x$, $g(x)=x^2$, $h(x)=x^3$?

5. The locus defined by the equations $x^2+y^2+z^2-r^2=0, x+y+z-c=0$ may be interpreted as the circle of intersection of a plane and a sphere. If (x_0, y_0, z_0) is a point of this locus, under what sufficient condition on x_0, y_0, z_0 may the part of the locus near (x_0, y_0, z_0) be represented in the form $y=f(x)$, $z=g(x)$?

6. Let (x_0, y_0, z_0) be a point of the locus defined by $z^2+xy-a=0$, $z^2+x^2-y^2-b=0$. **(a)** Under what sufficient conditions on x_0, y_0 may the part of the locus near (x_0, y_0, z_0) be represented in the form $x=f(z)$, $y=g(z)$? **(b)** What are sufficient conditions on x_0, y_0, z_0 for representing this part of the locus in the form $x=f(y)$, $z=g(y)$?

7. State carefully, in full detail, the analogue of Theorem III for the implicit-function problem arising from the equations

$$F(x, y, u, v, w)=0,\ G(x, y, u, v, w)=0,\ H(x, y, u, v, w)=0,$$

where it is desired to find solutions

$$u=f(x, y),\ v=g(x, y),\ w=h(x, y)$$

in the neighborhood of $(x_0, y_0, u_0, v_0, w_0)$. For the proof assume that $\dfrac{\partial(F, G)}{\partial(v, w)}\neq 0$ and suppose that the equation $F=0$, $G=0$ are solved for v and w in terms of x, y, u. Let $K(x, y, u)=H(x, y, u, v, w)$ when v and w are expressed in terms of x, y, u. Show that

$$\frac{\partial K}{\partial u}=\frac{\partial(F, G, H)}{\partial(u, v, w)}\Big/\frac{\partial(F, G)}{\partial(v, w)}.$$

With these suggestions, write out a complete proof.

8. Suppose that $F(x, y, u, v)=x^2+y^2-2ux+1$,
$G(x, y, u, v)=x^2+y^2+2vy-1$.

(a) Interpret u and v as parameters, and plot the curves $F=0$, $G=0$ in the xy-plane, assuming $u^2\geqq 1$. **(b)** Now suppose x_0, y_0, u_0, v_0 satisfy the equations $F=0$, $G=0$, and that $u_0^2>1$. Explain geometrically why it is reasonable to expect that, if u and v differ but slightly from u_0 and v_0, respectively, the equations $F=0$, $G=0$ will determine a *unique* point (x, y), if this point is required to be sufficiently near (x_0, y_0). **(c)** Show that a set of values x_0, y_0, u_0, v_0 cannot satisfy the three equations $F=0$, $G=0$, $\dfrac{\partial(F, G)}{\partial(x, y)}=0$ unless $u_0^2=1$. Use this result and an appropriate version of the implicit-function theorem for simultaneous equations to give an analytical explanation of the situation described in part **(b)**.

9. Suppose that $(x_0, y_0, z_0, u_0, v_0, w_0)$ satisfy the equations

$$\frac{x^2}{a^2+u}+\frac{y^2}{b^2+u}+\frac{z^2}{c^2+u}=1,$$
$$\frac{x^2}{a^2+v}+\frac{y^2}{b^2+v}+\frac{z^2}{c^2+v}=1,$$
$$\frac{x^2}{a^2+w}+\frac{y^2}{b^2+w}+\frac{z^2}{c^2+w}=1,$$

where $0 < c < b < a$ and $-c^2 < u_0$, $-b^2 < v_0 < -c^2$, $-a^2 < w_0 < -b^2$. Show that the equations have a unique solution $x = f(u, v, w)$, $y = g(u, v, w)$, $z = h(u, v, w)$, where f, g, h are continuous at (u_0, v_0, w_0) and take on the values x_0, y_0, z_0 respectively there, provided that $x_0 y_0 z_0 \neq 0$. Prove conclusively that the appropriate Jacobian is not zero.

10. Assuming that $\dfrac{\partial(F_1, \ldots, F_{r-1})}{\partial(u_1, \ldots, u_{r-1})} \neq 0$,

let $u_i = \phi_i(x_i, \ldots, x_n, u_r)$ $(i = 1, \ldots, r-1)$ be solutions of the first $r-1$ equations in the system

$$\begin{aligned} &F_1(x_1, \ldots, x_n, u_1, \ldots, u_r) = 0 \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ &F_r(x_1, \ldots, x_n, u_1, \ldots, u_r) = 0 \end{aligned}$$

for $u_1, \ldots, u_{r-1}$ in terms of $x_1, \ldots, x_n, u_r$. Let

$$G(x_1, \ldots, x_n, u_r) = F_r(x_1, \ldots, x_n, \phi_1, \ldots, \phi_{r-1}, u_r),$$

where for convenience we have written merely ϕ_i instead of $\phi_i(x_1, \ldots, x_n, u_r)$. Show that

$$\frac{\partial G}{\partial u_r} \frac{\partial(F_1, \ldots, F_{r-1})}{\partial(u_1, \ldots, u_{r-1})} = \frac{\partial(F_1, \ldots, F_r)}{\partial(u_1, \ldots, u_r)}$$

is an identity in $x_1, \ldots, x_n, u_r$.

SUGGESTION: Differentiate the identities

$$\begin{aligned} &F_1(x_1, \ldots, x_n, \phi_1, \ldots, \phi_{r-1}, u_r) = 0 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &F_{r-1}(x_1, \ldots, x_n, \phi_1, \ldots, \phi_{r-1}, u_r) = 0 \\ &F_r(x_1, \ldots, x_n, \phi_1, \ldots, \phi_{r-1}, u_r) - G(x_1, \ldots, x_n, u_r) = 0 \end{aligned}$$

with respect to u_r and regard the resulting equations as a set of r linear equations in $\dfrac{\partial \phi_1}{\partial u_r}, \ldots, \dfrac{\partial \phi_{r-1}}{\partial u_r}, \dfrac{\partial G}{\partial u_r}$ as unknowns. Solving this system for $\dfrac{\partial G}{\partial u_r}$ by determinants gives the required result.

11. State the theorem analogous to Theorem III for functions $u_i = \phi_i(x_1, \ldots, x_n)$ defined by equations $F_i(x_1, \ldots, x_n, u_1, \ldots, u_r) = 0$, where $\dfrac{\partial(F_1, \ldots, F_r)}{\partial(u_1, \ldots, u_r)} \neq 0$. Use the results of Exercise 10 to prove the induction step from $r-1$ to r, and so prove the theorem.

9 / THE INVERSE FUNCTION THEOREM WITH APPLICATIONS

9 / INTRODUCTION

In Chapter 8 we discussed implicit function theory, which deals with problems in which we have a system of m simultaneous (not necessarily linear) equations in n unknowns, where $n > m$. The question is, "When can the system be solved to express some m of the variables as functions of the other $n - m$?" In §8.1 we gave detailed consideration to the case in which $m = 1$ and $n = 3$. In §8.2 we extend this to the case of one equation in $n + 1$ variables, and in Theorem III of §8.3 we went as far as showing how the same methods could provide an answer for a system of two equations, each having five variables. From this point, one could infer what the general theorem must be for systems of m equations in n variables.

In this chapter we switch our attention slightly to what is called inverse function theory. The problem here arises when we have a system of n simultaneous equations which transform one ordered n-tuple of numbers into another, that is, a system of the following kind:

$$\begin{aligned} f^{(1)}(x_1, x_2, \ldots, x_n) &= y_1 \\ f^{(2)}(x_1, x_2, \ldots, x_n) &= y_2 \\ &\cdots\cdots\cdots\cdots \\ f^{(n)}(x_1, x_2, \ldots, x_n) &= y_n \end{aligned} \tag{9–1}$$

When we substitute an ordered n-tuple of numbers into the functions on the left, the n values which we get form another ordered n-tuple of numbers on the right. The system (9–1) transforms the n-tuple $(x_1, x_2, \ldots, x_n)$ into $(y_1, y_2, \ldots y_n)$, and we shall find it convenient to speak of system (9–1) as a *transformation*. The problem which we wish to consider is, "Given transformation (9–1), when it is possible, in theory at least, to solve for the x's in terms of the y's?"

Notice that we already have the answer in the special case where the given functions on the left are linear. The transformation then reduces to

$$\begin{aligned} f^{(1)}(x_1, \ldots, x_n) &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = y_1 \\ f^{(2)}(x_1, \ldots, x_n) &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = y_2 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ f^{(n)}(x_1, \ldots, x_n) &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = y_n \end{aligned} \tag{9–2}$$

Cramer's rule tells us the following: If the coefficient determinant

$$\Delta = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \tag{9–3}$$

is different from zero, then for all values of the given y's, the x-values are uniquely determined. Furthermore, the x's turn out to be linear combinations of the y's, just as in (9–2) the y's are linear combinations of the x's. This means that the solution has the form

$$\begin{aligned} &b_{11}y_1 + b_{12}y_2 + \cdots + b_{1n}y_n = x_1 \\ &b_{21}y_1 + b_{22}y_2 + \cdots + b_{2n}y_n = x_2 \\ &\cdots \\ &b_{n1}y_1 + b_{n2}y_2 + \cdots + b_{nn}y_n = x_n \end{aligned} \tag{9–4}$$

The transformation represented by (9.4) is called the *inverse* of that represented by (9–2). If one is interested in knowing just what values the b's must have in (9–4), they can be found by taking the solution of (9–2) given by Cramer's rule, that is,

$$x_1 = \frac{\Delta_1}{\Delta}, \qquad x_2 = \frac{\Delta_2}{\Delta}, \ldots, \qquad x_n = \frac{\Delta_m}{\Delta},$$

where Δ_m is the determinant obtained from Δ by replacing the mth column of Δ by the column of y's on the right in (9–2). Expanding each of the Δ_m by minors of the y's in the mth column gives us (9–4).

So Cramer's rule is the best known example of an inverse function theorem, and it is the most satisfactory, in the sense that it gives a complete answer to the inverse function question, in those cases to which it applies. But unfortunately it applies only to those cases of (9–1) in which all the functions are linear, that is, in those cases where (9–1) reduces to (9–2). Before leaving this very special case, it is important to notice a certain important feature which it exhibits. In (9–2) $f_j^{(i)} = \dfrac{\partial f^{(i)}}{\partial x_j} = a_{ij}$. Therefore, the coefficient determinant, Δ, can also be written in the following form,

$$\begin{vmatrix} f_1^{(1)} & f_2^{(1)} & \cdots & f_n^{(1)} \\ f_1^{(2)} & f_2^{(2)} & \cdots & f_n^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ f_1^{(n)} & f_2^{(n)} & \cdots & f_n^{(n)} \end{vmatrix}$$

which is seen to be $\dfrac{\partial(f^{(1)}, f^{(2)}, \ldots, f^{(n)})}{\partial(x_1, x_2, \ldots, x_n)}$, that is, the Jacobian of the f's with respect to the x's. This same Jacobian determinant is going to turn up again in the nonlinear case.

Mathematicians frequently try to solve problems in very general language so that their solutions will apply to as large a class of problems as possible. We have been discussing the linear version of the inverse function theorem in language sufficiently general to include any finite number of variables. We shall now illustrate what we have been saying in an extremely simple case, with $n = 2$.

Example 1. Does the transformation

$$\begin{aligned} 2u + 3v &= x \\ u + 2v &= y \end{aligned} \tag{9–5}$$

have an inverse? If so, find it.

The coefficient determinant, Δ, is

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

The fact that an inverse transformation does exist follows immediately from the fact that $\Delta \neq 0$. We easily find the inverse to be

$$\begin{aligned} 2x - 3y &= u \\ -x + 2y &= v. \end{aligned} \tag{9–6}$$

In this example, we tried to distinguish between two problems—showing that an inverse transformation exists and then actually finding it. The reason is that when we came to the nonlinear case, we shall frequently find ourselves in the position of being able to prove that an inverse transformation exists, but not having any idea how to find it.

Before leaving Example 1, it is worthwhile to remind ourselves what it means to say that in equations (9–6) we have solved the given equations (9–5) for u and v in terms of x and y. It means that if we substitute in equations (9–5) for u and v their values as given by (9–6) the result is an identity. If we make these substitutions we get

$$\begin{aligned} 2(2x - 3y) + 3(-x + 2y) &= x, \\ \text{and} \\ (2x - 3y) + 2(-x + 2y) &= y. \end{aligned} \tag{9–7}$$

These equations immediately reduce to

$$\begin{aligned} x &= x \\ \text{and} \\ y &= y, \end{aligned} \tag{9–8}$$

which is the identity transformation in the xy-plane. In (9–7) we have performed the composition of the transformations (9–5) and (9–6) by first performing (9–6) on a point (x, y) and then performing (9–5) on the result. Equations (9–8) show that we ended up right where we started—with the point (x, y), and therefore (9–6) followed by (9–5) is really just the identity transformation. The student will now find it easy to show that if he or she starts with a point (u, v) and transforms

it by (9–5), and then transforms that result by (9–6) he or she will end up with (u, v) again. So whichever one of the transformations (9–5) and (9–6) we start with, if we immediately follow it by applying the other to the output of the first, the result of all this is to leave everything exactly as it was—which is why we call it the identity transformation. From this we see that (9–5) is the inverse of (9–6) for exactly the same reason that (9–6) is the inverse of (9–5).

It is frequently convenient to introduce a symbol to denote a transformation. If we denote by T the transformation defined by equations (9–5), then we can write (9–5) simply as $T(u, v) = (x, y)$. The standard notation for the inverse of a transformation T is T^{-1}, so (9–6) becomes $T^{-1}(x, y) = (u, v)$. What we have said at greater length about the relation of a transformation to its inverse (for the linear case) is

$$T^{-1}[T(u, v)] = (u, v) \quad \text{for all} \quad (u, v)$$

and

$$T[T^{-1}(x, y)] = (x, y) \quad \text{for all} \quad (x, y).$$

Since $T^{-1}[T(a, b)] = T[T^{-1}(a, b)] = (a, b)$ for all ordered pairs (a, b), we say that

$$T^{-1}T = TT^{-1} = I$$

where I is the identity on the set of all ordered pairs of real numbers.

Much of what we have said in the preceding paragraph, growing out of our study of the linear transformation (9–5), is also true of nonlinear transformations, but more attention to the domains of the transformations is required. This is illustrated in one dimension by considering the familiar nonlinear function $f(x) = 1 + x^2$. This transforms any real x into $u = 1 + x^2$. If we solve for x we obtain two solutions: $x = g(u) = \sqrt{u - 1}$ and $x = h(u) = -\sqrt{u - 1}$. (We are considering real numbers only, so we must have $u \geq 1$.) Observe that $f(g(u)) = 1 + (u - 1) = u$ and $f(h(u)) = u$ whenever $u \geq 1$. It is not the case, however, that both g and h are inverses of f. In fact, neither is an inverse of f. We see that $f(x)$ is always in the domain of g as well as in the domain of h. But

$$g(f(x)) = \sqrt{1 + x^2 - 1} = \sqrt{x^2} = |x|$$

so that $g(f(x)) = x$ if $x \geq 0$, but $g(f(x)) = -x$ if $x < 0$. Thus it is *not* true that $g(f(x)) = x$ for all x in the domain of f. Accordingly, g is *not* an inverse of f, even though $f(g(u)) = u$ for all u in the domain of g. The trouble is that f has too big a domain to have an inverse. There is a standard way of getting around this kind of difficulty. We restrict f to a smaller domain, doing it in such a way that in the restricted domain f never takes on the same value at two different points. If we define $f_1(x) = f(x)$ with restriction $x \geq 0$, then $g(f_1(x)) = x$ for all x in the domain of f_1 and $f_1(g(u)) = u$ for all u in the domain of g. Thus f_1 and g are inverse to each other. Similarly, if we define $f_2(x) = f(x)$ when $x \leq 0$, then it may be verified that f_2 and h are inverse to each other.

Observe that the range R of f_1 is the domain of g and the range of g is the domain D of f_1. There are two distinct identity functions here. We shall denote them

by I_R and I_D, respectively. I_R is defined by

$$I_R(u) = f_1(g(u)) = u \quad \text{for all} \quad u \text{ in } R$$

and I_D is defined by

$$I_D(x) = g(f_1(x)) = x \qquad \text{for all } x \text{ in } D.$$

These equations may be written more briefly as $I_R = f_1 \circ g$ and $I_D = g \circ f_1$. The small circle between the functional symbols denotes the process of composition of two functions.

Another example of the principle of restriction of the domain of a function in order to obtain a function that has an inverse is furnished by the definition of the inverse sine function. If we start with $f(x) = \sin x$ and by restriction define $f_1(x) = \sin x$ with domain composed of x's for which $|x| \leqq \frac{\pi}{2}$, then $y = f_1(x)$ is equivalent to $x = \sin^{-1} y$, with $|y| \leqq 1$. Here the inverse function is $g(y) = \sin^{-1} y$.

9.1 / THE INVERSE FUNCTION THEOREM IN TWO DIMENSIONS

We are now ready for the main subject of this chapter, which is the inverse function theorem for nonlinear transformations in spaces of dimension two and higher. Beginning in two dimensions we can just as well simplify the notation of (9.1) to

$$f(u, v) = x, \qquad g(u, v) = y. \tag{9.1–1}$$

We want to know under what conditions we can solve for u and v as functions of x and y. When this can be done, we shall have

$$u = F(x, y), \qquad v = G(x, y). \tag{9.1–2}$$

where F and G are functions such that

$$f[F(x, y), G(x, y)] = x, \qquad g[F(x, y), G(x, y)] = y, \tag{9.1–3}$$

at least under certain restrictions on the domains of the functions involved. And under these restrictions we shall say that the transformation defined by (9.1–2) is the inverse of that defined in (9.1–1).

We shall deal with our problem by treating it as a special case of the implicit function problem that is solved by Theorem III, §8.3. For this purpose we regard equations (9.1–1) as a pair of equations of the form

$$\phi(x, y, u, v) = 0, \qquad \psi(x, y, u, v) = 0,$$

which we look upon as analogues of equations (8.3–1). Here, however, there is no z. This difference is not important, for the considerations of §8.3 apply equally well if the functions in (8.3–1) do not depend on z. We observe that if $\phi(x, y, u, v) = f(u, v) - x$ and $\psi(x, y, u, v) = g(u, v) - y$, we have

$$\frac{\partial(\phi, \psi)}{\partial(u, v)} = \frac{\partial(f, g)}{\partial(u, v)} = J \tag{9.1–4}$$

as the Jacobian to be considered in our application of Theorem III, §8.3 to the present situation. Because of the difference in notation between our present situation and that of §8.3, the analogues of the functions f, g of §8.3 will be the functions F, G of (9.1–2). By applying the theorem of §8.3 to our present situation we obtain the following theorem.

THEOREM I. *Suppose that the functions $f(u, v)$, $g(u, v)$ are continuous, with continuous first partial derivatives, in some neighborhood of (u_0, v_0), and suppose that the Jacobian J in (9.1–4) is not zero at (u_0, v_0). Let $x_0 = f(u_0, v_0)$, $y_0 = g(u_0, v_0)$.*

Then, for suitably restricted rectangular neighborhoods R in the xy-plane, with center (x_0, y_0), and S in the uv-plane, with center (u_0, v_0) (and with the sides of the rectangles parallel to a co-ordinate axis in each case), it can be arranged so that $J \neq 0$ when (u, v) is in S and that for each (x, y) in R there is a unique (u, v) in S for which the equations (9.1–1) are satisfied. In this way there are defined two functions $F(x, y)$ and $G(x, y)$ with domain R such that equations (9.1–3) are satisfied for every (x, y) in R. Moreover, if (x, y) is in R the equations (9.1–2) provide the only (u, v) in S such that equations (9.1–1) are satisfied. In addition, F and G are continuous and have continuous first partial derivatives in R, given by the formulas

$$F_1 = \frac{g_2}{J}, \qquad G_1 = -\frac{g_1}{J}, \qquad F_2 = -\frac{f_2}{J}, \qquad G_2 = \frac{f_1}{J} \tag{9.1–5}$$

where F_1 stands for $F_1(x, y)$, g_2 stands for $g_2[F(x, y), G(x, y)]$ (and likewise for the other partial derivatives), and J is evaluated at (u, v) with u and v expressed in terms of x and y.

Finally, the Jacobian $j = \dfrac{\partial(F, G)}{\partial(x, y)}$ is related to J by the formula

$$j = \frac{1}{J}. \tag{9.1–6}$$

We comment briefly on (9.1–5) and (9.1–6). Equations (9.1–5) are obtained from (8.3–7) by working out what the formulas in (8.3–7) become when applied to the notations of our present situation. The very important equation (9.1–6) results at once when we substitute the values of F_1, G_1, F_2, G_2 from (9.1–5) into the determinant which is j.

To get from Theorem I the precise situation of a transformation T and its inverse T^{-1} we must restrict the domain of the transformation defined by the equations (9.1–1). The theorem tells us there is a one-to-one correspondence between a *part* of S and *all* of R. The part of S in question (we shall call it E) is the set of those points (u, v) in S given by (9.1–2) as the point (x, y) varies over R when F and G are the functions defined in Theorem I. If T denotes the transformation from E to R defined by (9.1–1) when the domain of the transformation is E, the inverse T^{-1} is the transformation from R to E defined by (9.1–2) with R as its domain.

We have been talking about *transformations* and their inverses, but we have referred to Theorem I as an inverse *function* theorem. The occurrence of the two words "function" and "transformation" makes this an appropriate place to speak about an enormous broadening of the function concept that enables us to regard a transformation as a particular kind of function. Heretofore we have been speaking about functions as having for domain a set of points on the real line or in space of two or more dimensions, and as range a set of numbers. These are numerically valued functions. But the general notion of a function, say f, is that of a correspondence involving two sets of things, one set called the *domain* of f and the other the *range* of f, such that to each member P of the domain corresponds a well-determined member of the range, denoted by $f(P)$ and called the *value* of f at P. The members of the domain do not have to be numbers or points. There is no logical need to restrict the type of objects that compose the domain of a function. The same comments apply to the range. A transformation such as that defined by (9.1–1) is a function whose domain is a set in the uv-plane and whose range is a set in the xy-plane. If we consider the equations (9–1) at the beginning of Chapter 9, we can regard the n functions $f^{(1)}$, $f^{(2)}, \ldots, f^{(n)}$ as defining a single function whose domain is a certain set of ordered n-tuples $(x_1, \ldots, x_n)$ and whose range is another set of ordered n-tuples $y_1, \ldots, y_n$. Ordered n-tuples can be regarded as points in a space of n-dimensions. They can also be regarded as vectors in a vector space (see Chapters 10 and 11). Therefore a transformation such as that defined by (9–1) is a function that is defined on a set of points (or vectors) and has points (or vectors) as values. We can also have a transformation with domain in a space of n-dimensions and range in a space of m-dimensions, where m can be different from n. In this chapter we consider mainly the case in which $n = m = 2$.

Let us look further at Theorem I. There is nothing in its statement that tells us explicitly anything about the range E of the inverse transformation T^{-1} (as defined in a preceding paragraph). The domain R of T^{-1} is the interior of a rectangle with center (x_0, y_0); therefore R is an open set, all of its points being interior points. (Open sets and interior points are defined in §5.1.) However, we can infer that E is an open set, also, so that (u_0, v_0) and all the other points of E are interior points of E. We can see this as follows by applying what we have learned from Theorem I, for Theorem I can be applied to the transformation T^{-1} by reversing the roles of the uv-plane and the xy-plane. We know from (9.1–6) that the Jacobian j of (9.1–2) is not zero when (x, y) is in R. Hence, if we start with any point (x_1, y_1) in R and the point (u_1, v_1) of E into which T^{-1} transforms (x_1, y_1), the application of Theorem I tells us that there is a rectangular neighborhood S_1 of (u_1, v_1) and a rectangular neighborhood R_1 of (x_1, y_1) (contained in R) such that to each point (u, v) in S_1 there is a unique point (x, y) in R_1 such that T^{-1} transforms (x, y) into (u, v). But, from what we already know, (u, v) must be in E. Therefore S_1 is in E, and so (x_1, y_1) is an interior point of E. This shows that E is an open set.

The set E is also connected in the sense of §7.4. We shall not prove this here. It is a consequence of the connectedness of R and the continuity of T^{-1}. See Miscellaneous Exercise 3 at the end of this chapter.

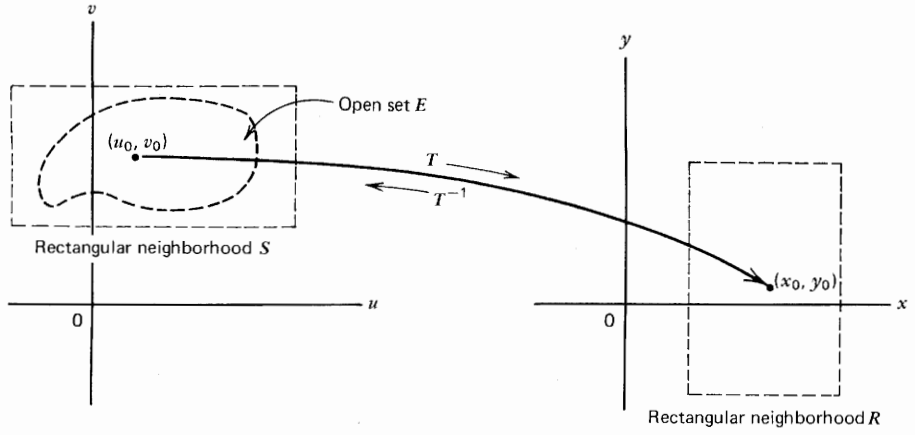


Fig. 56.

We illustrate the discussion of T, T^{-1}, and E in Fig. 56.

In the following Example 2 we have a situation in which it is possible, because of the comparative simplicity of the transformations, to show precisely the appearance of the set E into which T^{-1} transforms R.

Example 2. A transformation which we may call T is given by the equations

$$x = \tfrac{1}{2}(u^2 + v^2), \qquad y = \tfrac{1}{2}(u^2 - v^2). \tag{9.1–7}$$

If we try to solve for u and v we are led immediately to

$$u^2 = x + y \quad \text{and} \quad v^2 = x - y,$$

which shows that as long as we consider T as being defined for all values of u and v, the equations (9.1–7) simply do not determine u and v uniquely for given values of x and y. For each pair (x, y) we have to consider the following set of possibilities for (u, v): $(\sqrt{x+y}, \sqrt{x-y})$, $(\sqrt{x+y}, -\sqrt{x-y})$, $(-\sqrt{x+y}, \sqrt{x-y})$, and $(-\sqrt{x+y}, -\sqrt{x-y})$. Actually, none of these possibilities exists unless the two radicands are both nonnegative. This means that unless both $x + y \geq 0$ and $x - y \geq 0$, there is no pair (u, v) which T transforms into (x, y). It is also obvious from the first equation of (9.1–7) that $x \geq 0$. The transformation T transforms all points in the (u, v)-plane into points in the (x, y)-plane satisfying the three conditions $y \geq -x$, $y \leq x$ and $x \geq 0$. The points satisfying these conditions are those belonging to the wedge-shaped region W in the right half plane with vertex at the origin, upper edge lying along $y = x$ and lower edge along $y = -x$. (See Fig. 57*a*.) If (x, y) is an interior point of W, then all four possibilities for (u, v) are realized—one in each quadrant of the (u, v)-plane. Each point on the boundary of W is the image of two points in the (u, v) plane, except for the vertex which is the image of only $u = 0$, $v = 0$.

Suppose now that we restrict the domain of T to those points (u, v) lying interior to the first quadrant. Denote this restricted domain by Q_1 and let T_1

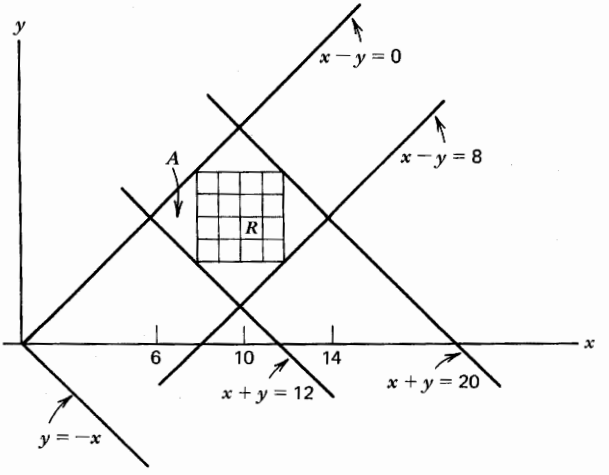


Fig. 57a.

denote the restriction of T to Q_1; T_1 transforms Q_1 into the interior of the wedge W in a one-to-one manner, so T_1 does have an inverse, namely the transformation T_1^{-1} defined by $u = \sqrt{x+y}$ and $v = \sqrt{x-y}$, where (x, y) is any interior point of W. Thus $T_1^{-1}[T_1(u, v)] = (u, v)$ for all (u, v) in Q_1 and $T_1[T_1^{-1}(x, y)] = (x, y)$ for all (x, y) interior to W.

We can make other restrictions T_2, T_3, T_4 of T by limiting (u, v) to the interior of the second, third, and fourth quadrants, respectively. Each of these

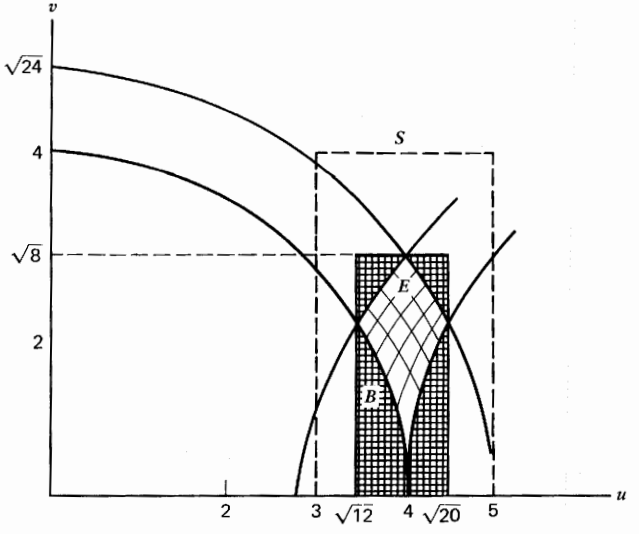


Fig. 57b.

restrictions has an inverse. For example, the inverse T_2^{-1} is defined by $u = -\sqrt{x+y}$, $v = \sqrt{x-y}$, where (x, y) is in the interior of W.

Let us examine further the nature of the restricted transformation T_1 and its inverse T_1^{-1}. The Jacobian in this case is found to be $J = -2uv$, so $J = 0$ when $u = 0$ or $v = 0$. Hence $J \neq 0$ in the interior of each of the quadrants of the uv-plane. For the study of T_1 we consider just the first quadrant. Let us choose the point $u_0 = 4$, $v_0 = 2$. Then $x_0 = 10$, $y_0 = 6$, so that $T_1(4, 2) = (10, 6)$. Theorem I speaks about rectangles R in the xy-plane and S in the uv-plane, centered at (x_0, y_0) and (u_0, v_0), respectively. The theorem tells us nothing about the sizes of R and S, but with our explicit formulas for T_1 and T_1^{-1} let us see what we can find out about possible choices of R and S. Since R cannot extend outside of W, and since its center is at $(10, 6)$, the greatest possible expansion of R would be obtained by placing its upper left hand corner on the line $y = x$ that forms part of the boundary of W. If this R is made narrow and tall, that corner will approach the point $(10, 10)$; if R is made wide, with small height, that corner will approach the point $(6, 6)$. In any case, the corners of R will lie on the lines $x - y = 0$, $x + y = 20$, $x - y = 8$, and $x + y = 12$. See Fig. 57*a*, in which R is represented in the special case when it is a square. It is inscribed in another square A whose sides lie on the foregoing four lines. We consider A as the open set forming the interior of the square.

We shall investigate what happens to R and A when they are mapped into the uv-plane by the transformation T_1^{-1}. We shall see that A is mapped into a rectangle B and that R is mapped into a certain set within B. If R is any rectangle with center $(10, 6)$ inscribed in A (not merely the special case when R is square), a suitable choice for S can be found, as we shall describe (see Fig. 57*b*).

Consider the points in A on any line $x + y = a$, where $12 < a < 20$. The value of u in $(u, v) = T_1^{-1}(x, y)$ on any such line segment is always equal to $\sqrt{a}$, but $v = \sqrt{x-y}$. Now the lines $x - y = b$ running through A are those for which $0 < b < 8$ (see Fig. 57*a*), and so $0 < v < \sqrt{8}$. As (x, y) takes on all positions in A, (u, v) takes on all positions inside the rectangle B whose sides are the lines $u = \sqrt{12}$, $u = \sqrt{20}$, $v = 0$, $v = \sqrt{8}$, and the correspondence is one to one. Thus T_1^{-1} maps A onto all of the open set which is the interior of the rectangle B.

Because R is a part of A it follows that the image $E = T_1^{-1}(R)$ is a subset of B. Hence we can take for S the interior of any rectangle centered at $(4, 2)$, with each side parallel to either the u-axis or v-axis, provided that S lies in the first quadrant and contains B. The rectangle with sides $u = 3$, $u = 5$, $v = 0$, $v = 4$ would do.

We can determine the set E as follows, by considering the families of circles

$$u^2 + v^2 = 2x$$

and hyperbolas

$$u^2 - v^2 = 2y.$$

To be specific, as x varies from 6 to 14, each of the lines on which x is constant,

with $6 < x < 14$, intersects A in a segment that is carried (by T_1^{-1}) into an arc of a circle in the first quadrant of the uv-plane. The arc is that part of the circle in the interior of the rectangle B. When $6 < x < 10$, the circle cuts the two adjacent sides $v = 0$ and $u = \sqrt{12}$ of B; when $10 < x < 14$, the circle cuts the sides $u = \sqrt{20}$ and $v = \sqrt{8}$. See Figures 57*a* and 57*b*. Similarly, as y varies from 2 to 10, each of the lines on which y is constant, with $2 < y < 10$, intersects A in a segment that is transformed by T_1^{-1} into an arc of hyperbola cutting through the interior of B. When $2 < y < 6$, the hyperbolic arc intersects the sides $u = \sqrt{12}$ and $v = \sqrt{8}$ of B, and when $6 < y < 10$, the arc intersects the sides $v = 0$ and $u = \sqrt{20}$. Now, to find the part of B that corresponds to R (we denote it by E), we must locate in B just those portions of the circular arcs that correspond to the segments of the lines $x =$ constant that cut through the interior of R. It is clear that these portions of the circular arcs will be cut off between the two hyperbolic arcs that correspond to the lines $y =$ constant which form the top and bottom of R. In this way we see that the set E is the interior of a curvilinear quadrilateral lying in B and bounded by two circular arcs and two hyperbolic arcs. Figure 57*b* shows the case of the E corresponding to R when R is a square. Incidentally, one sees that the part of B not in E or on its boundary consists of four separate pieces, each of which is the image of a triangular piece of A exterior to R.

EXERCISE

In the equations $x = f(u, v)$, $y = g(u, v)$ regard u, v as first-class variables and x, y as second-class variables, the relation between the classes being that expressed by (9.1–1).

From this point of view the chain rule gives

$$1 = \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x},$$

where $\frac{\partial u}{\partial x}$ means $\frac{\partial F}{\partial x}$ and $\frac{\partial v}{\partial x}$ means $\frac{\partial G}{\partial x}$. Carry on with the chain rule in this manner to obtain three other equations, and then solve them, thus obtaining equations (9.1–5).

9.2 / MAPPINGS

Suppose that

$$x = f(u, v), \quad y = g(u, v) \tag{9.2–1}$$

is some transformation and

$$u = F(x, y), \quad v = G(x, y) \tag{9.2–2}$$

is its inverse. In the previous section we have interpreted (x, y) as rectangular co-ordinates in one plane, and (u, v) as rectangular co-ordinates in another plane. We sometimes find it convenient to say that the point (u, v) is *mapped* into the point (x, y). Equations (9.2–1) are said to define a *mapping*, or a point transformation. If the functions f, g are defined in a certain region of the uv-plane, we say that this region is mapped into the xy-plane. The configuration in the second

plane is called the *image* of the configuration in the first plane, and that in the first plane is sometimes called the *pre-image* of the one in the second.

The concept of a mapping does not require that the transformation have an inverse. Nor is it necessary, for the mere concept of a mapping, to deal with continuous or differentiable transformations. However, we shall study continuously differentiable transformations for which the Jacobian is in general nonzero. The inverse transformation (9.2–2) then defines a mapping of a portion of the xy-plane into the uv-plane.

It is of interest, in studying mappings, to consider what becomes of a configuration of points in one plane, such as a curve or a region, when it is mapped into the other plane. When the mapping is continuous, with continuous inverse, the mapping process may be conceived intuitively as a deformation, with stretching and shrinking of varying amounts, but no tearing or puncturing. If the Jacobian of the mapping is zero at a certain point, however, there may fail to be an inverse mapping. This can occur, for example, because several points in one plane are mapped into the same point in the other plane.

Again, we emphasize that even if there are points where the Jacobian vanishes, this does not necessarily mean that no inverse exists in any neighborhood of these points. The following extremely simple example helps to guard against this unwarranted assumption. If $x = u^3$ and $y = v$, then the Jacobian of the transformation is $3u^2$, which vanishes at all points where $u = 0$, that is, at all points of the v-axis. Nevertheless the transformation maps the entire uv-plane onto the entire xy-plane in a one-to-one manner and therefore has an inverse—$u = \sqrt[3]{x}$, $v = y$ which is valid everywhere—we do not even have to restrict the domain of the original transformation.

The following example guards against an error of the opposite sort. It shows that even if the Jacobian is different from zero at every point, we may still have to restrict the domain of the transformation considerably in order to get an inverse.

Example 1. Consider the mapping

$$u = e^x \cos y, \qquad v = e^x \sin y, \tag{9.2–3}$$

of the xy-plane into the uv-plane. We shall investigate the nature of the mapping by finding out what happens to lines $x =$ constant or $y =$ constant.

We observe that

$$u^2 + v^2 = e^{2x}, \qquad v \cos y = u \sin y. \tag{9.2–4}$$

Thus the image of any point on the line $x = x_0$ is a point on the circle in the uv-plane, with center at the origin and radius e^{x_0}. This circle is defined parametrically by equations (9.2–3) with $x = x_0$ and y as parameter. The point (u, v) goes counterclockwise once around the circle each time that y increases by 2π. The image of a point on the line $y = y_0$ lies on a straight line through the origin in the uv-plane, with slope $\tan y_0$. The image of the entire line $y = y_0$ is just one of the two rays composing the line $v \cos y = u \sin y$, however. For instance,

if $y_0 = \pi/3$, the point (u, v) corresponding to (x, y_0) is given by

$$u = \tfrac{1}{2}e^x, \qquad v = \frac{\sqrt{3}}{2}e^x,$$

so that u and v are always positive. The mapping of the line $y = \pi/3$ onto the ray is indicated in Fig. 58, the part of the ray corresponding to $x < 0$ is shown by a dotted line.

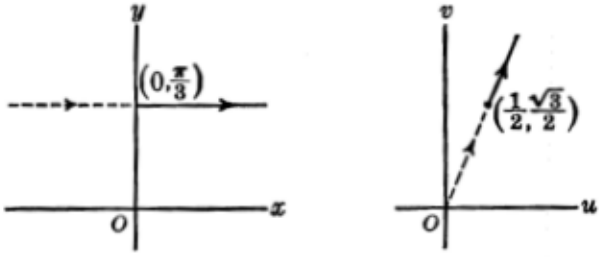


Fig. 58.

If we now consider the strip in the xy-plane between the lines $y = 0$, $y = 2\pi$, we see that the image of the strip is the entire uv-plane with the exception of the origin. Line segments $x = x_0$ crossing the strip map into circles centered at the origin, of radius less than one if $x_0 < 0$, and of radius greater than one if $x_0 > 0$. Lines $y = y_0$ map into rays, the angle between the ray and the positive u-axis being y_0. The origin in the uv-plane is not obtained as the image of any point in the xy-plane, but $(u, v) \to (0, 0)$ as $x \to -\infty$. The nature of the mapping is suggested by Fig. 59*a* and Fig. 59*b* in which certain corresponding areas are indicated by similar shading.

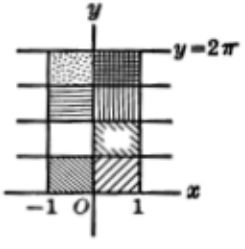


Fig. 59a.

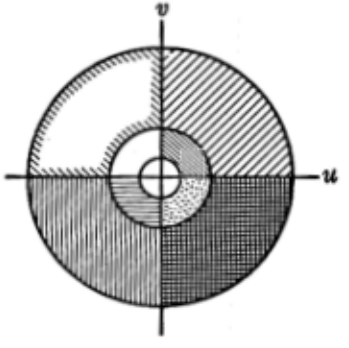


Fig. 59b.

The equations (9.2–3) admit the period 2π for the variable y. Hence any other strip of width 2π parallel to the x-axis is mapped onto the uv-plane in a manner similar to that just described.

The Jacobian of the mapping is

$$\frac{\partial(u, v)}{\partial(x, y)} = \begin{vmatrix} e^x \cos y - e^x \sin y \\ e^x \sin y \quad e^x \cos y \end{vmatrix} = e^{2x}.$$

It does not vanish for any value of x. The nonvanishing of the Jacobian means that if we fix our attention on any point (x_0, y_0) and its image (u_0, v_0), the mapping sets up a one-to-one correspondence between the points of a sufficiently small neighborhood of (x_0, y_0) and those of some neighborhood of (u_0, v_0). This is described by saying that the mapping is *locally* one-to-one, or one-to-one *in the small.* The xy-plane as a whole is not mapped in a one-to-one manner on the uv-plane, however, for many points in the xy-plane have the same image in the uv-plane (two such points are at least 2π units apart, however). Hence we say that the mapping is *not* one-to-one *in the large.*

Example 2. Consider the mapping

$$u = x^2 - y^2, \qquad v = 2xy. \tag{9.2–5}$$

In this example let us investigate the configurations in the xy-plane which map into lines u = constant or v = constant in the uv-plane. This is equivalent to regarding u and v as curvilinear co-ordinates and studying the u-curves and v-curves. (For a general discussion of curvilinear co-ordinates see §9.5. A u-curve is a curve in the xy-plane on which u is a constant and v is a parameter.) We see readily that the u-curves are rectangular hyperbolas, with foci on the x-axis if $u > 0$, and on the y-axis if $u < 0$. The v-curves are rectangular hyperbolas having the x and y axes as asymptotes. For $u = 0$ we have the pair of lines $y = \pm x$, and for $v = 0$ the pair of lines $x = 0$, $y = 0$.

The image of the hyperbola $u_0 = x^2 - y^2$ is the line $u = u_0$. Note, however, that each branch of the hyperbola maps onto the entire line. Likewise, each branch of the hperbola $v_0 = 2xy$ maps onto the entire line $v = v_0$ (see Fig. 60*a*, Fig. 60*b*).

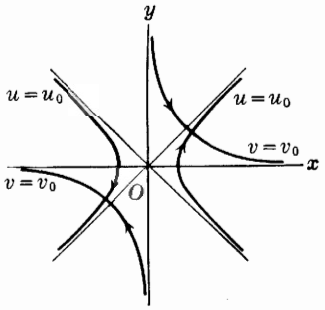


Fig. 60a.

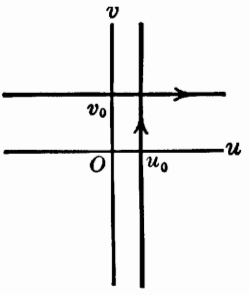


Fig. 60b.

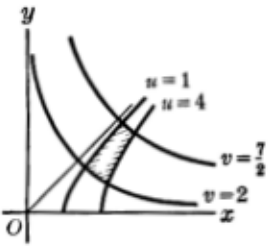


Fig. 61a.

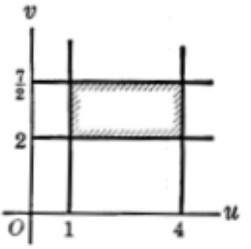


Fig. 61b.

We now see that certain cells with curved sides in the xy-plane are mapped into rectangular cells in the uv-plane (see Fig. 61*a*, 61*b*).

The Jacobian of the mapping (9.2–5) is

$$\frac{\partial(u, v)}{\partial(x, y)} = 4(x^2 + y^2).$$

Hence the mapping is locally one-to-one except in the neighborhood of the origin in the xy-plane.

The notion of a point transformation, or mapping, is not limited to two-dimensional problems. We may, for example, speak of mappings from the xyz-space into uvw-space.

Observe that in the case of mappings from (x, y) to (u, v) which are locally one-to-one, we may interpret u and v as curvilinear co-ordinates in the xy-plane, or we may interpret x and y as curvilinear co-ordinates in the uv-plane. Similar remarks apply when there are more variables in each set.

The sign of the Jacobian has a significance which deserves mention. Consider a mapping with Jacobian not zero at (x_0, y_0), so that the mapping is, at least locally, one-to-one. If C is a small closed curve enclosing the point (x_0, y_0) in the xy-plane, the image of C in the uv-plane will be a small closed curve C' enclosing the point (u_0, v_0) which corresponds to (x_0, y_0). If the point (x, y) goes around C in the counterclockwise sense, the image point (u, v) will go around C'. But will (u, v) go counterclockwise also? The answer depends on the sign of the Jacobian of the mapping. If $J > 0$, (u, v) will go around C' in the same sense that (x, y) goes around C; but if $J < 0$, (u, v) will go in the sense opposite to that of (x, y). We shall not prove this fact just now. See Exercise 7 and §15.32.

EXERCISES

1. Consider the mapping $x = au$, $y = bv$, where $a > 0$, $b > 0$. Find out what region in the uv-plane corresponds to the region in the xy-plane bounded by the ellipse $(x^2/a^2) + (y^2/b^2) = 1$.

2. Let R be the region in the xy-plane bounded by the lines $x - y = 0$, $x + y = 0$, $x - 2y = 2$. Find the region R' in the uv-plane onto which R is mapped by the equations $u = 2x - y$, $v = x - 2y$.

3. Find the image in the uv-plane of the triangular region bounded by $x = 0$, $y = 0$, $x + y = 1$, if $u = x + y$, $v = x - y$.

4. Find the image in the xy-plane of the rectangle in the uv-plane bounded by $u = 1$, $u = 4$, $v = 9$, $v = 16$, if the mapping is $x = \left(\frac{u+v}{2}\right)^{1/2}$, $y = \left(\frac{v-u}{2}\right)^{1/2}$.

5. Consider the mapping $u = \frac{1}{2}(x - y)$, $v = \sqrt{xy}$. Draw a number of the u-curves and v-curves in the xy-plane. What regions in the xy-plane map into the rectangle in the uv-plane bounded by $u = -\frac{3}{2}$, $u = \frac{3}{2}$, $v = 2$, $v = 4$?

6. Study the mapping $x = u - uv$, $y = uv$. Find the inverse transformation. Identify the u-curves and v-curves in the xy-plane, and the x-curves and y-curves in the uv-plane.

7. Let R be the region in the xy-plane bounded by the lines $y - x = 2$, $y - x = 6$, $y + x = 4$, $y + x = 8$. Let R' be the region in the uv-plane onto which R is mapped by the transformation $u = \frac{1}{2}(x + y)$, $v = \frac{1}{2}(x - y)$. Examine the direction in which a point goes around the boundary of R' when its correspondent goes around the boundary of R counterclockwise. Compare with the sign of the Jacobian and note the remark at the end of §9.2.

8. A mapping from the xy-plane into the uv-plane is defined by

$$u = \frac{x}{1 - x - y}, \quad v = \frac{y}{1 - x - y}.$$

(a) Find $\frac{\partial(u, v)}{\partial(x, y)}$. **(b)** Find the inverse transformation and $\frac{\partial(x, y)}{\partial(u, v)}$. **(c)** What are the u-curves and v-curves in the xy-plane? **(d)** What are the x-curves and y-curves in the uv-plane? **(e)** Find the region R in the xy-plane corresponding to the square in the uv-plane bounded by the lines $u = -\frac{1}{2}$, $u = -1$, $v = -1$, $v = -\frac{3}{2}$.

9. A mapping is defined by

$$u = \frac{x^2}{1 - x^2 - y^2}, \qquad v = \frac{y^2}{1 - x^2 - y^2}.$$

(a) What are the u-curves in the xy-plane if $u > 0$? if $-1 < u < 0$? if $u < -1$? **(b)** Answer similar questions for the v-curves. **(c)** Find where $\frac{\partial(u, v)}{\partial(x, y)} = 0$. **(d)** Draw the curves $u = \frac{1}{2}$, $u = 3$, $v = \frac{1}{3}$, $v = 1$ and mark four regions in the xy-plane which correspond to the rectangle bounded by the lines $u = \frac{1}{2}$, etc., in the uv-plane. **(e)** Follow directions similar to those in **(d)** for $u = -\frac{4}{3}$, $u = -2$, $v = -\frac{1}{2}$, $v = -\frac{1}{3}$.

9.3 / SUCCESSIVE MAPPINGS

Suppose we have a transformation mapping (x, y) into (u, v), and a further transformation mapping (u, v) into (ξ, η). The effect of performing these two transformations in succession is to give a mapping from the xy-plane into the $\xi\eta$-plane.

Example 1. Suppose

$$u = x + y, \qquad v = x - y,$$

and

$$\xi = uv, \qquad \eta = u + v.$$

Then

$$\xi = x^2 - y^2, \qquad \eta = 2x.$$

The single mapping which is produced by carrying out two successive transformations is called the *resultant*, or *product*, of the two transformations.

In the above example, we have

$$\frac{\partial(u, v)}{\partial(x, y)} = -2, \qquad \frac{\partial(\xi, \eta)}{\partial(u, v)} = v - u,$$

$$\frac{\partial(\xi, \eta)}{\partial(x, y)} = 4y,$$

as the student should verify for himself. Observe that

$$v - u = -2y.$$

Hence

$$\frac{\partial(\xi, \eta)}{\partial(u, v)} \frac{\partial(u, v)}{\partial(x, y)} = (v - u)(-2) = 4y = \frac{\partial(\xi, \eta)}{\partial(x, y)}.$$

This illustrates a general truth which we now state formally.

THEOREM II. *Let T_1 denote a transformation from the xy-plane into the uv-plane, and let T_2 denote a transformation from the uv-plane into the $\xi\eta$-plane. Let the resultant transformation from the xy-plane into the $\xi\eta$-plane be denoted by T_3. Then the Jacobian of T_3 is the product of the Jacobians of T_1 and T_2, that is,*

$$\frac{\partial(\xi, \eta)}{\partial(x, y)} = \frac{\partial(\xi, \eta)}{\partial(u, v)} \frac{\partial(u, v)}{\partial(x, y)}. \tag{9.3–1}$$

It is assumed that the transformations T_1 and T_2 are continuously differentiable. It is further assumed that the transformation T_2 is defined for points (u, v) obtained by application of the transformation T_1 to points (x, y) in some region R of the xy-plane.

Proof of the Theorem. We regard ξ, η as functions of the first-class variables u, v, which are in turn functions of the second-class variables x, y. Thus

$$\frac{\partial \xi}{\partial x} = \frac{\partial \xi}{\partial u} \frac{\partial u}{\partial x} + \frac{\partial \xi}{\partial v} \frac{\partial v}{\partial x},$$

$$\frac{\partial \xi}{\partial y} = \frac{\partial \xi}{\partial u} \frac{\partial u}{\partial y} + \frac{\partial \xi}{\partial v} \frac{\partial v}{\partial y},$$

with similar equations for $\frac{\partial \eta}{\partial x}$, $\frac{\partial \eta}{\partial y}$. It is then a matter of straightforward

multiplication to verify that

$$\frac{\partial \xi}{\partial x}\frac{\partial \eta}{\partial y}-\frac{\partial \xi}{\partial y}\frac{\partial \eta}{\partial x}=\left(\frac{\partial \xi}{\partial u}\frac{\partial \eta}{\partial v}-\frac{\partial \xi}{\partial v}\frac{\partial \eta}{\partial u}\right)\left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial y}-\frac{\partial u}{\partial y}\frac{\partial v}{\partial x}\right).$$

This is exactly the relation (9.3–1). The result is recognized with less effort if one is familiar with the rule for writing the product of two determinants as another determinant. This is of especial importance in dealing with the generalization of the theorem for transformations in three or more variables.

EXERCISES

1. Use Theorem II to prove (9.1–6).

2. Verify the correctness of Theorem II when the transformations T_1 and T_2 are defined by $u = e^x \cos y$, $v = e^x \sin y$ and $\xi = u^2 + v^2$, $\eta = v/u$, respectively. Begin by finding the equations of the transformation T_3.

3. Let a_{ij} be the term in the ith row and jth column of a determinant of nth order whose value is A. Let b_{ij} and B refer in similar manner to a second determinant. Let C be the value of the determinant with elements c_{ij}, where $c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$. It is a theorem from algebra that $C = AB$. With this fact in mind, consider the generalization of Theorem II for transformations on n variables. Let T_1 be a transformation in which $u_1, \ldots, u_n$ are differentiable functions of $x_1, \ldots, x_n$, and let T_2 be a transformation in which $\xi_1, \ldots, \xi_n$ are differentiable functions of $u_1, \ldots, u_n$. Let T_3 be the resultant transformation, in which $\xi_1, \ldots, \xi_n$ are considered as functions of $x_1, \ldots, x_n$. Use the chain rule and the theorem T_2.

4. Consider the transformation (9.2–1) and its inverse (9.2–2), in the neighborhood of a point where $\frac{\partial F}{\partial x} \neq 0$. Then $u = F(x, y)$ has a solution $x = \phi(u, y)$. Let $G(\phi(u, y), y) = \Phi(u, y)$. Now consider an rs-plane. Show that the mapping (9.2–2) is the resultant of the mapping $r = F(x, y)$, $s = y$ from the xy-plane to the rs-plane, and the mapping $u = r$, $v = \Phi(r, s)$ from the rs-plane to the uv-plane. The Jacobians of these last two transformations are $\frac{\partial F}{\partial x}$ and $\frac{\partial \Phi}{\partial s}$ respectively. Verify in the case of each of these mappings the correctness of the statement made at the end of §9.2 about the significance of the sign of the Jacobian. This may be done by a direct inspection, since the mappings are of such simple types. Then use Theorem II to prove the correctness of the statement in question for the mapping from the xy-plane to the uv-plane. A similar procedure may be used when $\frac{\partial F}{\partial y} \neq 0$.

5. **(a)** Suppose $n > p$. Let $u_1, \ldots, u_n$ be differentiable functions of $x_1, \ldots, x_n$, with $u_{p+1}, \ldots, u_n$ the simple functions defined by $u_i = x_i$, $i = p + 1, \ldots, n$. Show that

$$\frac{\partial(u_1, \ldots, u_n)}{\partial(x_1, \ldots, x_n)} = \frac{\partial(u_1, \ldots, u_p)}{\partial(x_1, \ldots, x_p)}.$$

(b) If $u_1 = x_1, \ldots, u_p = x_p$, and the rest of the u's are differentiable functions of $x_1, \ldots, x_n$, show that

$$\frac{\partial(u_1, \ldots, u_n)}{\partial(x_1, \ldots, x_n)} = \frac{\partial(u_{p+1}, \ldots, u_n)}{\partial(x_{p+1}, \ldots, x_n)}.$$

6. Suppose $F_1, \ldots, F_n$ are continuously differentiable functions of $x_1, \ldots, x_n$, and that $\dfrac{\partial(F_1, \ldots, F_p)}{\partial(x_1, \ldots, x_p)} \neq 0$, where p is a fixed integer, $1 \leqq p < n$. Suppose the equations $u_1 = F_1(x_1, \ldots, x_n), \ldots, u_p = F_p(x_1, \ldots, x_n)$ are solved for $x_1, \ldots, x_p$ in terms of $u_1, \ldots, u_p$ and $x_{p+1}, \ldots, x_n$, and that these values of $x_1, \ldots, x_p$ are then substituted into $F_{p+1}, \ldots, F_n$, giving rise to functions $\psi_i(u_1, \ldots, u_p, x_{p+1}, \ldots, x_n)$, $i = p+1, \ldots, n$. Show that

$$\frac{\partial(F_1, \ldots, F_n)}{\partial(x_1, \ldots, x_n)} = \frac{\partial(F_1, \ldots, F_p)}{\partial(x_1, \ldots, x_p)} \frac{\partial(\psi_{p+1}, \ldots, \psi_n)}{\partial(x_{p+1}, \ldots, x_n)},$$

where it is assumed that $u_1, \ldots, u_p$ are replaced by

$$F_1(x_1, \ldots, x_n), \ldots F_p(x_1, \ldots, x_n)$$

after calculating the derivatives in the last Jacobian.

SUGGESTIONS OF METHOD: Note first of all that $\psi_i(F_1, \ldots, F_p, x_{p+1}, \ldots, x_n) = F_i(x_1, \ldots, x_n)$, $i = p+1, \ldots, n$. Let T_1 and T_2 be defined as follows:

T_1	T_2
$\xi_1 = F_1(x_1, \ldots, x_n)$	$u_1 = \xi_1$
. . . . .	. . . . .
$\xi_p = F_p(x_1, \ldots, x_n)$	$u_p = \xi_p$
$\xi_{p+1} = x_{p+1}$	$u_{p+1} = \psi_{p+1}(\xi_1, \ldots, \xi_n)$
. . . . .	. . . . .
$\xi_n = x_n$	$u_n = \psi_n(\xi_1, \ldots, \xi_n)$.

Now apply Theorem II (for the case of n variables) and Exercise 5.

9.4 / TRANSFORMATIONS OF CO-ORDINATES

The formulas connecting the rectangular co-ordinates (x, y) of a point and the rectangular co-ordinates (x', y') of the same point in a rotated system (see Fig. 62) are

$$\begin{aligned} x &= x' \cos\phi - y' \sin\phi, \\ y &= x' \sin\phi + y' \cos\phi. \end{aligned} \tag{9.4–1}$$

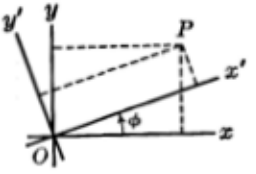


Fig. 62.

The transformation (9.4–1) has the inverse

$$\begin{aligned} x' &= x \cos\phi + y \sin\phi, \\ y' &= -x \sin\phi + y \cos\phi, \end{aligned} \tag{9.4–2}$$

as the student should verify for himself. The student should also verify that

$$\frac{\partial(x, y)}{\partial(x', y')} = \frac{\partial(x', y')}{\partial(x, y)} = 1. \tag{9.4–3}$$

The familiar formulas connecting rectangular and polar co-ordinates in the plane are

$$x = r \cos\theta, \qquad y = r \sin\theta. \tag{9.4–4}$$

The Jacobian of this transformation is

$$\frac{\partial(x, y)}{\partial(r, \theta)} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r. \tag{9.4–5}$$

For convenience we shall assume that $r \geqq 0$ in our dealings with polar co-ordinates. If we attempt to solve for r, θ in terms of x, y, we find

$$r = (x^2 + y^2)^{1/2}, \qquad \sin\theta = \frac{y}{(x^2 + y^2)^{1/2}}, \qquad \cos\theta = \frac{x}{(x^2 + y^2)^{1/2}}. \tag{9.4–6}$$

To get this far we assume that $r \neq 0$, that is, that the Jacobian is not zero. Note that a unique formula for θ cannot be found, because θ may be changed by any multiple of 2π without affecting (9.4–6). This non-uniqueness does not violate the statement of Theorem I, §9.1, however. (Why not?)

It is worth while to consider briefly the question of an explicit formula for θ in terms of x and y. From (9.4–6) one is tempted to write

$$\theta = \tan^{-1}\left(\frac{y}{x}\right).$$

There are many things wrong with this formula, however, if one is seeking the transformation inverse to (9.4–4). If we follow the usual custom of restricting the inverse tangent to its principal value, we limit ourselves to the range $-\pi/2 < \theta < \pi/2$. Any other convention about principal values leads to a similar difficulty. On the other hand, if we ignore principal values, and regard the inverse tangent as a multiple-valued function, we get into trouble, because the values we then get for θ will not all satisfy (9.4–6). Thus, if $x = -1$, $y = 1$, the values of θ such that $\tan\theta = -1$ are either in the second or fourth quadrant. The second-quadrant values satisfy (9.4–6), while the fourth-quadrant values do not. Still another difficulty is that $\tan^{-1}(y/x)$ is discontinuous at points on the y-axis, whereas the transformation (9.4–4) has a continuous inverse in the neighborhood of such a point (if the point is not the origin).

The inverse tangent (or the inverse cotangent) can be used if care is taken, of course. Thus, if $r_0 = \sqrt{2}$, $\theta_0 = \frac{3}{4}\pi$, $x_0 = -1$, $y_0 = 1$, the transformation inverse to (9.4–4) can be written

$$r = (x^2 + y^2)^{1/2}, \qquad \theta = \tan^{-1}\left(\frac{y}{x}\right) + \pi$$

as long as $x < 0$; here the principal value of the inverse tangent is to be used.

Let us now examine the nature of a rectangular or polar co-ordinate system from a point of view which will be fruitful in our systematic study of transformations. We consider the xy-plane with a fixed set of axes. A point (x_0, y_0) in the plane is located by the statement that is at the intersection of the two lines $x = x_0$, $y = y_0$. A co-ordinate system is basically a method of locating points by setting up a correspondence between points and sets of numbers; the numbers corresponding to a point are called co-ordinates of the point. For an arbitrary rectangular co-ordinate system we need two one-parameter families of parallel

lines, the lines of one family intersecting the lines of the other family at right angles. Let the equations of the two families be

$$\begin{aligned} a_1x + b_1y &= u \\ a_2x + b_2y &= v, \end{aligned} \tag{9.4–7}$$

where the a's and b's are fixed, and u and v are parameters of the families. For the required perpendicularity we must have $b_1 : a_1 = -a_2 : b_2$, or

$$a_1a_2 + b_1b_2 = 0. \tag{9.4–8}$$

If we now pick out values u_0, v_0, the two lines $u = u_0$, $v = v_0$ intersect at a unique point (x_0, y_0) (see Fig. 63). This point is determined by (u_0, v_0), and we may take these latter numbers as co-ordinates of the point in a new co-ordinate system. The xy-system and the uv-system are then related by the transformation (9.4–7). The uv-system will not in general be obtained from the xy-system by a rigid motion, for the units of distance along the u-axis and the v-axis will in general be different from each other and from the common unit of distance along the x and y axes.

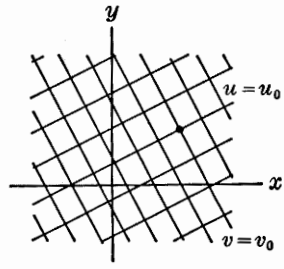


Fig. 63.

If we drop the requirement (9.4–8), but insist that the two families in (9.4–7) not be identical, we get, in general, an *oblique* co-ordinate system.

The system of polar co-ordinates is also based on locating a point at an intersection, but in this case we have a family of concentric circles and a family of rays through the common center of the circles (see Fig. 64). The equation of the circles is

$$x^2 + y^2 = r^2, \tag{9.4–9}$$

with parameter r. The family of rays is given by taking in each case the suitable half of the line

$$y \cos \theta = x \sin \theta, \tag{9.4–10}$$

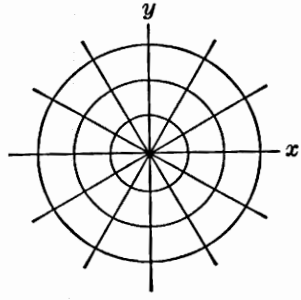


Fig. 64.

with θ as parameter. A point is located by specifying that it is at the intersection of a circle $r = r_0$ and a ray $\theta = \theta_0$. It will be observed that in general a point is at the intersection of just one circle and just one ray. The exceptional point is the origin. This is precisely the point at which the Jacobian $\dfrac{\partial(x, y)}{\partial(r, \theta)} = 0$. It is called a *singular point* of the polar co-ordinate system.

In §9.5 we shall study more general co-ordinate systems arising from the study of two one-parameter families of curves of a fairly arbitrary character. Co-ordinate systems of such a general type are called *curvilinear* co-ordinate systems.

9.5 / CURVILINEAR CO-ORDINATES

A system of curvilinear co-ordinates can be defined if we have two one-parameter families of curves subject to suitable conditions. The conditions may not be satisfied over the whole plane, but only in certain regions. Among the conditions which are normally required is the condition that each point in a region under consideration should lie on one and only one curve of each of the two families, and that these curves should not be tangent at the point. Before discussing the general theory it will be helpful to consider an example.

Example 1. Suppose that the two families of curves are

$$y^2 = -u^2(x - u^2) \tag{9.5–1}$$

$$y^2 = v^2(x + v^2), \tag{9.5–2}$$

where the parameters are u, v.

The curves of both families are parabolas with the x-axis as axis of symmetry. The parabolas (9.5–1) open to the left; the parabolas of the other family open to the right. The vertex of (9.5–1) is at $(u^2, 0)$, and its focus is a distance $\frac{1}{4}u^2$ from the vertex. The y-intercepts are $(0, \pm u^2)$. A number of the curves of both families are shown in Fig. 65.

The parameters u, v may be used as co-ordinates; that is, a point can be located by giving the values of u and v corresponding to the two parabolas which pass through the point. Actually, a u-parabola and a v-parabola intersect twice, so that a pair of values (u, v) does not determine a unique point. Conversely, a given point (x, y) does not determine u and v uniquely, because the equations of the parabolas are not affected by a change in the sign of u or v.

Let us attempt to solve equations (9.5–1) and (9.5–2) for x and y. Subtracting (9.5–2) from (9.5–1), we find

$$0 = -x(u^2 + v^2) + u^4 - v^4,$$

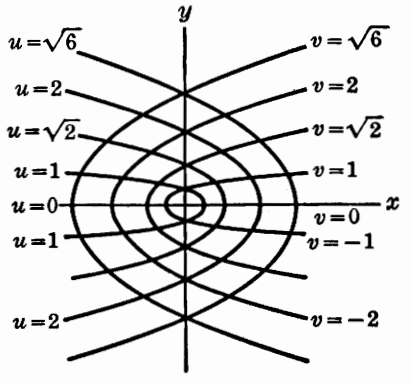


Fig. 65.

or

$$x = \frac{u^4 - v^4}{u^2 + v^2} = u^2 - v^2,$$

provided u and v are not both zero. If this result is substituted in (9.5–2), we get

$$y^2 = u^2 v^2, \quad \text{or} \quad y = \pm uv.$$

There are thus two possible differentiable transformations,

$$x = u^2 - v^2, \qquad y = uv, \tag{9.5–3}$$

and

$$x = u^2 - v^2, \qquad y = -uv, \tag{9.5–4}$$

which may be used to determine a point (x, y) by giving values of u, v. Either transformation allows us to use u and v as curvilinear co-ordinates. Suppose, for example, that we use (9.5–3). There is some arbitrariness in the choice of signs for u and v, but the product uv must have the same sign as y. We may, for instance, use $u \geqq 0$ all the time; then we must use $v > 0$ when $y > 0$ and $v < 0$ when $y < 0$. Figure 65 has been labeled in accord with this choice. Other choices are possible, however. From (9.5–3) we find

$$\frac{\partial(x, y)}{\partial(u, v)} = 2(u^2 + v^2).$$

This is zero only when $u = v = 0$, which by (9.5–3) is equivalent to $x = y = 0$. The origin is called a *singular point* of the uv-curvilinear co-ordinate system. It is not possible to use u and v as co-ordinates, throughout a region having the origin as an interior point, in such a way as to have a one-to-one correspondence between (x, y) and (u, v) with the transformation from (u, v) to (x, y) and the inverse transformation from (x, y) to (u, v) both continuous throughout the region.

Now let us consider the general theory of curvilinear co-ordinates in the plane. Suppose we have given a continuously differentiable transformation

$$x = f(u, v), \qquad y = g(u, v). \tag{9.5–5}$$

such that the Jacobian

$$J = \frac{\partial(f, g)}{\partial(u, v)}$$

is not equal to zero for a certain pair of values u_0, v_0. Denote the corresponding values of x and y by x_0, y_0. The following discussion will relate to pairs (x, y) sufficiently near (x_0, y_0) and pairs (u, v) sufficiently near (u_0, v_0) so that we can use the conclusions described in Theorem I, §9.1. In particular, there is a continuously differentiable inverse transformation

$$u = F(x, y), \qquad v = G(x, y). \tag{9.5–6}$$

Let us denote the Jacobian of the inverse transformation by j. By (9.1–6) of Theorem I we know that $j = \dfrac{1}{J}$.

The one-to-one correspondence between (x, y) and (u, v), as established by the transformation and its inverse, makes it possible for us to use (u, v) as co-ordinates for the point (x, y), since the point determines its co-ordinates by (9.5–6), and the co-ordinates determine the point by (9.5–5). It is desirable, however, to have a geometric interpretation of the uv-co-ordinate system. A geometric interpretation can be given in terms of two families of curves which form a mesh of quadrilaterals in somewhat the same way that the lines $x = \text{constant}$, $y = \text{constant}$, form a rectangular mesh. If we regard u as constant and let v vary, equations (9.5–5) are parametric equations of a curve, v being the parameter. We call such a curve a u-curve. Similarly, a v-curve is defined by (9.5–5) with v fixed and u as a parameter. Alternatively, these curves may be thought of as defined by (9.5–6). With u constant, a u-curve is defined by $F(x, y) = u$. If we confine our attention to a small neighborhood of a point where $J \neq 0$, there will be exactly one u-curve and one v-curve through each point of the neighborhood, and these two curves will not be tangent at the point. We shall now show that the nontangency is a consequence of the fact that $J \neq 0$. The slope of a u-curve may be found from (9.5–5), with u constant and v the parameter. The slope is

$$\frac{dy}{dx} = \frac{\partial g}{\partial v} \Big/ \frac{\partial f}{\partial v}.$$

Likewise, the slope of a v-curve is

$$\frac{dy}{dx} = \frac{\partial g}{\partial u} \Big/ \frac{\partial f}{\partial u}.$$

These slopes are not equal, for their equality would imply that

$$\frac{\partial g}{\partial v}\frac{\partial f}{\partial u} = \frac{\partial g}{\partial u}\frac{\partial f}{\partial v}, \quad \text{or} \quad J = 0,$$

contrary to our assumption.

We can define curvilinear co-ordinates in three dimensions also. The basis for a set of such co-ordinates u, v, w is a transformation

$$x = f(u, v, w), \qquad y = g(u, v, w), \qquad z = h(u, v, w), \tag{9.5–8}$$

with nonvanishing Jacobian, and the inverse transformation

$$u = F(x, y, z), \qquad v = G(x, y, z), \qquad w = H(x, y, z). \tag{9.5–9}$$

The equations (9.5–9) define three one-parameter families of surfaces, called u-surfaces, v-surfaces, and w-surfaces respectively. A u-surface and a v-surface intersect in a curve, which we shall call a uv-curve. Similarly, there are vw-curves and wu-curves. The fact that the Jacobian of the transformation (9.5–8) is not zero implies that the three curves, one of each type, intersecting at a point, are pair-wise nontangent there.

The most familiar examples of curvilinear co-ordinates in three dimensions are cylindrical co-ordinates and spherical co-ordinates. A great variety of

three-dimensional systems can be generated by starting with a two-dimensional system and rotating the plane about a line in the plane. The two families of curves in the plane generate surfaces in space. Half-planes through the axis of rotation form a third set of surfaces. Spherical co-ordinates are derived from plane polar co-ordinates in this way.

Example 2. Consider the transformation

$$x = uv\cos\theta, \qquad y = uv\sin\theta, \qquad z = u^2 - v^2, \tag{9.5–10}$$

with Jacobian

$$\frac{\partial(x, y, z)}{\partial(u, v, \theta)} = 2uv(u^2 + v^2). \tag{9.5–11}$$

Let us assume $u \geqq 0$, $v \geqq 0$, and write $r = (x^2 + y^2)^{1/2} = uv$. Then we see from (9.5–10) that

$$x = r\cos\theta, \qquad y = r\sin\theta.$$

The equations

$$r = uv, \qquad z = u^2 - v^2 \tag{9.5–12}$$

can be regarded as defining a set of plane curvilinear co-ordinates (u, v) in the rz-plane. With rotation about the z-axis, using r, θ as polar co-ordinates in the xy-plane, we obtain equations (9.5–10), which we can use to establish u, v, θ as curvilinear co-ordinates in space. The u-curves and v-curves in the rz-plane are parabolas (compare with equations (9.5–3) and (9.5–1), (9.5–2)), a few of which are shown in Fig. 66. In the three-dimensional system, the u-surfaces and v-surfaces are paraboloids of revolution about the z-axis, and the θ-surfaces are half-planes with the z-axis as edge (see Fig. 67). The transformation has a continuously differentiable inverse in the neighborhood of any point not on the z-axis, for such points are obtained only when neither u nor v is zero, and the Jacobian (9.5–11) is then not equal to zero. All points of the z-axis are singular points of the $uv\theta$-coordinate system.

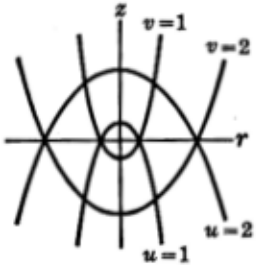


Fig. 66.

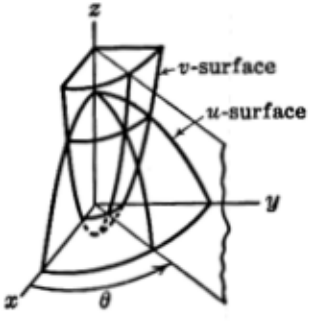


Fig. 67.

EXERCISES

1. (**a**) Sketch the u-curves and v-curves in the xy-plane if $x = u\cosh v$, $y = u\sinh v$. (**b**) Does the uv-system of curvilinear co-ordinates have a singular point? (**c**) To what part of the xy-plane does the system not apply?

2. Discuss the co-ordinates (u, v) related to rectangular co-ordinates (x, y) by the equations $2x - y + u = 0$, $x - 2y + 2v = 0$. Draw and describe the u-curves and the v-curves. Find the transformation and the inverse transformation corresponding to (9.5–5) and (9.5–6) respectively. Verify Theorem III in this case.

3. Show that u, v can be used as curvilinear co-ordinates when $0 < x < \pi/2$ and $y > 0$ if $u = y/\tan x$, $v = y/\sin x$. Solve for x and y in terms of u and v. Draw the curves $u = 1$, $v = 2$ in the xy-plane. Compute $\dfrac{\partial(u, v)}{\partial(x, y)}$ and $\dfrac{\partial(x, y)}{\partial(u, v)}$ and verify that they are reciprocals.

4. Find the inverse of the transformations (9.5–3), assuming u and v positive.

5. Suppose

$$x = \frac{2u}{u^2 + v^2}, \qquad y = \frac{-2v}{u^2 + v^2}.$$

(**a**) Show that

$$x^2 + y^2 = \frac{4}{u^2 + v^2},$$

and hence that

$$u = \frac{2x}{x^2 + y^2}, \qquad v = \frac{-2y}{x^2 + y^2}.$$

(**b**) Show that the u-curves are circles through the origin with centers on the x-axis, and that the v-curves are circles tangent to the x-axis at the origin. Draw several curves of each family. Locate the points with the following curvilinear co-ordinates: $u = 1$, $v = -2$; $u = 0$, $v = 1$; $v = 0$, $u = -2$.

6. Let $x = \cos u\cosh v$, $v = \sin u\sinh v$, where $0 \leqq u \leqq \pi$ and v is arbitrary. (**a**) Describe the u-curves and the v-curves by name, and show how variation in the sizes of u and v affects the curves. Draw a number of curves of each family. (**b**) What are the singular points of the uv-curvilinear co-ordinate system? (**c**) Describe the points for which $u = 0$; for which $u = \pi/2$; for which $v = 0$.

7. Suppose that two families of curves in the xy-plane are defined by

$$\phi(x, y, u) = 0, \qquad \psi(x, y, v) = 0,$$

where ϕ and ψ are continuously differentiable functions such that

$$\frac{\partial(\phi, \psi)}{\partial(x, y)} \neq 0, \qquad \frac{\partial\phi}{\partial u}\frac{\partial\psi}{\partial v} \neq 0$$

for values of the variables under consideration. Show that these families of curves can be used to establish u and v as curvilinear co-ordinates. If the notations J, j are used as in (9.1–6), show that

$$\frac{\partial(\phi, \psi)}{\partial(x, y)} = j\,\frac{\partial\phi}{\partial u}\frac{\partial\psi}{\partial v}.$$

8. (**a**) If $x = r\sin\phi\cos\theta$, $y = r\sin\phi\sin\theta$, $z = r\cos\phi$, find $\dfrac{\partial(x, y, z)}{\partial(r, \phi, \theta)}$.

(**b**) What is the geometric nature of an r-surface? a ϕ-surface? a θ surface?

(c) Solve for r, ϕ, θ in terms of x, y, z, assuming $r>0$, $0<\phi<\pi/2$, $0<\theta<\pi/2$ for convenience. Compute $\dfrac{\partial(r,\phi,\theta)}{\partial(x,y,z)}$ and verify that it is the reciprocal of the Jacobian found in (a).

9. Discuss the three-dimensional system of curvilinear co-ordinates (u, v, θ), where $x = r\cos\theta$, $y = r\sin\theta$,

$$z = \frac{\sin v}{\cosh u + \cos v}, \qquad r = \frac{\sinh u}{\cosh u + \cos v},$$

and $r^2 = x^2 + y^2$. Begin by discussing and sketching the u-curves and v-curves in an rz-plane. The relevant equations to be obtained are

$$r^2 + z^2 - 2r\,\mathrm{ctnh}\,u + 1 = 0,$$
$$r^2 + z^2 + 2z\,\mathrm{ctn}\,v - 1 = 0.$$

Then rotate around the z-axis. Find the singular points of the co-ordinate system. The co-ordinates (u, v, θ) are known as *toroidal*, or *ring*, co-ordinates. Describe the u-surfaces and v-surfaces by name.

10. What is the analogue of (9.1–6) for a transformation $x = f(u)$ where there is just one variable in each set?

9.6 / IDENTICAL VANISHING OF THE JACOBIAN. FUNCTIONAL DEPENDENCE

In this section we inquire into the state of affairs when the Jacobian determinant of a transformation is equal to zero throughout a region. First we consider briefly the significance of this phenomenon in the linear case. After that we move to nonlinear mappings from the plane to the plane, with reasoning which could be used to extend the theorems to higher dimensions.

The two linear functions $ax + by$ and $cx + dy$ are said to be linearly dependent in case there are two numbers k_1 and k_2, not both of which are zero, such that

$$k_1(ax + by) + k_2(cx + dy) \equiv 0.$$

Those who have had some linear algebra will recall that a necessary and sufficient condition for this is that

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = 0$$

This determinant is the Jacobian of the linear transformation

$$u = ax + by,$$
$$v = cx + dy.$$

So the vanishing of the Jacobian of the transformation is equivalent to the existence of two numbers k_1 and k_2 such that

$$k_1u + k_2v = 0 \quad \text{for all} \quad (x, y). \tag{9.6–1}$$

This means that the entire xy-plane is mapped into the straight line (9.6–1) through the origin in the u, v-plane.

This should lead us to wonder if there is some analogous theorem in the case of nonlinear transformations. Let us begin with the transformation

$$u = F(x, y), \quad v = G(x, y) \tag{9.6–2}$$

and see what conclusion we can deduce from the hypothesis that $\dfrac{\partial(F, G)}{\partial(x, y)}$ vanishes identically throughout some region R of the x, y-plane. Since our equations here are nonlinear we would not expect to find that (9.6–2) maps R into a straight line but rather, perhaps, into a curve. This is roughly what the following theorem says—at least locally.

THEOREM III. *Suppose that F and G are continuously differentiable functions of x and y in some open set R and that*

$$\frac{\partial(F, G)}{\partial(x, y)} = 0 \qquad \textit{for all } (x, y) \textit{ in } R. \tag{9.6–3}$$

Suppose further that (x_0, y_0) is a point of R at which either $\dfrac{\partial F}{\partial x} \neq 0$ or $\dfrac{\partial F}{\partial y} \neq 0$. Let $u_0 = F(x_0, y_0)$. Then there exists an interval of the u-axis centered at u_0 and a function $\phi(u)$ defined thereon, such that

$$G(x, y) = \phi[F(x, y)] \tag{9.6–4}$$

throughout a neighborhood of (x_0, y_0).

Before taking up the proof, we comment on the theorem. First, like all our theorems on nonlinear functions, this is a "local" theorem—it gives a result holding in neighborhoods of certain points rather than throughout the entire domain of the transformation. Second, we could just as well have taken $\dfrac{\partial G}{\partial x}$ or $\dfrac{\partial G}{\partial y}$ to be different from zero instead of $\dfrac{\partial F}{\partial x}$ or $\dfrac{\partial F}{\partial y}$. This would have led to the conclusion

$$F(x, y) = \psi[G(x, y)] \tag{9.6–5}$$

instead of (9.6–4). Geometrically speaking, (9.6–4) means that some neighborhood of (x_0, y_0) is mapped into a curve given by $v = \phi(u)$, whereas (9.6–5) gives a curve of the form $u = \psi(v)$. So unless all four of the first partial derivatives, F_1, F_2, G_1, and G_2 vanish at (x_0, y_0), some entire neighborhood of this point is mapped into a curve in the u, v-plane. We could investigate conditions under which these curve segments in the u, v-plane join up to form one big curve, but this would lead to difficulties which we prefer to avoid.

Proof of Theorem III. Consider the equation $u = F(x, y)$. Suppose $\dfrac{\partial F}{\partial x} \neq 0$ at (x_0, y_0); then the implicit-function theorem guarantees the existence of a solution $x = f(u, y)$ giving all triples (x, y, u) near (x_0, y_0, u_0) for which $u = F(x, y)$. Moreover,

$$\frac{\partial f}{\partial y} = -\frac{\dfrac{\partial F}{\partial y}}{\dfrac{\partial F}{\partial x}}. \tag{9.6–6}$$

All this applies when the differences $x - x_0$, $y - y_0$, $u - u_0$ are sufficiently small. Consider the function $G(x, y)$ as a function of u and y, with $x = f(u, y)$. The partial derivative with respect to y is

$$\frac{\partial G}{\partial x}\frac{\partial f}{\partial y} + \frac{\partial G}{\partial y} = \frac{\dfrac{\partial (F, G)}{\partial (x, y)}}{\dfrac{\partial F}{\partial x}} = 0$$

because of (9.6–6) and (9.6–3). Thus $G(f(u, y), y)$ is actually independent of y. Let us write $\phi(u) = G(f(u, y), y)$. Since $x = f(u, y)$ is equivalent to $u = F(x, y)$ for the values of the variables here in question, $\phi(u) = G(f(u, y), y)$ is equivalent to (9.6–4). If we assume $\dfrac{\partial F}{\partial y} \neq 0$ instead of $\dfrac{\partial F}{\partial x} \neq 0$, similar reasoning again leads to (9.6–4).

***Example* 2.** The argument is illustrated by $F(x, y) = x^2y^2$, $G(x, y) = -xy$, $x_0 = y_0 = 1$, $f(u, y) = \sqrt{u}/y$, $\phi(u) = -\sqrt{u}$. Observe that, with these same functions F, G, but with $x_0 = -1$, $y_0 = 1$, we obtain $f(u, y) = -\sqrt{u}/y$, $\phi(u) = \sqrt{u}$.

Theorem III can be generalized to n functions of n variables. For $n = 3$ the hypotheses that

$$\frac{\partial (F, G, H)}{\partial (x, y, z)} = 0 \tag{9.6–7}$$

in a neighborhood of (x_0, y_0, z_0), and that at least one of the Jacobians

$$\frac{\partial (F, G)}{\partial (x, y)}, \quad \frac{\partial (F, G)}{\partial (y, z)}, \quad \frac{\partial (F, G)}{\partial (z, x)}$$

is not zero at this point, leads to a conclusion of the form

$$H(x, y, z) = \phi(F(x, y, z), G(x, y, z)), \tag{9.6–8}$$

where $\phi(u, v)$ is defined near $u_0 = F(x_0, y_0, z_0)$, $v_0 = G(x_0, y_0, z_0)$.

Theorem III says that under certain conditions the identical vanishing of the Jacobian implies that neighborhoods (two-dimensional subsets) in the xy-plane are mapped into curves (one-dimensional subsets) in the uv-plane. Our next theorem will be somewhat like a converse of this. We shall assume that for all

(x, y) in R,

$$\Phi[F(x, y), G(x, y)] = 0. \tag{9.6–9}$$

Notice that this includes as special cases both $G(x, y) = \phi[F(x, y)]$ and $F(x, y) = \psi[G(x, y)]$. When such a function as Φ exists, F and G are said to be *functionally* dependent in R. The term is not completely defined until we specify what other properties Φ must have. We shall require that it be continuous.

THEOREM IV. *Suppose that*

$$u = F(x, y), \qquad v = G(x, y) \tag{9.6–10}$$

is a continuously differentiable mapping of a region R of the xy-plane into the uv-plane. Let R′ be a region in the uv-plane which contains the image of R. Let $\Phi(u, v)$ be a function which is defined and continuous in R′, such that there is no neighborhood of a point in R′ throughout which $\Phi(u, v) = 0$. Finally, suppose that

$$\Phi(F(x, y), G(x, y)) = 0 \tag{9.6–11}$$

whenever (x, y) is a point of R. Then the Jacobian $\dfrac{(F, G)}{(x, y)}$ vanishes at all points of R.

Proof. Suppose the equation (9.6–3) fails to hold at some point (x_0, y_0) in R. Then the mapping (9.6–10) is locally one-to-one, and maps a neighborhood of (x_0, y_0) onto a neighborhood of the corresponding point (u_0, v_0). In view of this, (9.6–11) means that $\Phi(u, v) = 0$ throughout a neighborhood of (u_0, v_0). Since this violates the hypothesis, (9.6–3) must hold at every point of R.

Example 1. Theorem IV is illustrated by

$$F(x, y) = \sqrt{y}\sin x, \qquad G(x, y) = y\cos^2 x - y, \qquad \Phi(u, v) = u^2 + v,$$

with R the half-plane $y > 0$, R' the entire uv-plane.

When a pair of functions $F(x, y)$, $G(x, y)$ are such that the conditions stated in the hypotheses of Theorem IV are satisfied, F and G are said to be *functionally dependent* in the region R. The functional dependence is expressed by (9.6–11).

EXERCISES

1. Are the functions $x + y$, $x^2 + y^2$ functionally dependent in the neighborhood of any point?

2. Under what condition will the linear functions $ax + by$, $cx + dy$ be functionally dependent?

3. (a) Let $F(x, y, z) = x + y - z$, $G(x, y, z) = x - y + z$, $H(x, y, z) = x^2 + y^2 + z^2 - 2yz$. Verify that (9.6–7) holds in this case. Solve the equations $u = F(x, y, z)$, $v = G(x, y, z)$ for

x and y in terms of u, v, z, and substitute the solutions for x and y in $H(x, y, z)$. Observe that the result is independent of z. From this work find the function $\phi(u, v)$ such that (9.6–8) holds.

(b) Show that this line of reasoning is applicable in the general case, provided

$$\frac{\partial(F, G)}{\partial(x, y)} \neq 0.$$

4. Show in each case that the functions are functionally dependent, and find the way in which the third function depends on the first two. Use the method of Exercise 3.

(a) $u = x + y + z$, $v = xy + yz + zx$, $w = x^2 + y^2 + z^2$.

(b) $u = x/(y - z)$, $v = y/(z - x)$, $w = z/(x - y)$.

5. Without using the inverse function theorem, prove the variation on Theorem IV obtained by replacing the third sentence of that theorem by the following: Let $\Phi(u, v)$ be a differentiable function in R' for which $\Phi_1^2 + \Phi_2^2 > 0$ at each point of R'. HINT: All you need is the chain rule and the fact that a homogeneous linear system has a nontrivial solution if and only if the coefficient determinant is zero.

MISCELLANEOUS EXERCISES

1. (a) Find the inverse of the transformation

$$u = \frac{x}{r^2}, v = \frac{y}{r^2}, w = \frac{z}{r^2}, \text{ where } r^2 = x^2 + y^2 + z^2.$$

(b) What are the u-surfaces?

(c) Calculate $\dfrac{\partial(u, v, w)}{\partial(x, y, z)}$.

2. If $t = r\cos\phi$, $x = r\sin\phi\cos\psi$, $y = r\sin\phi\sin\psi\cos\theta$, $z = r\sin\phi\sin\psi\sin\theta$, show that $t^2 + x^2 + y^2 + z^2 = r^2$ and find $\dfrac{\partial(t, x, y, z)}{\partial(r, \phi, \psi, \theta)}$.

This indicates how spherical co-ordinates may be introduced in four-dimensional space.

3. The notations in this exercise are those of the discussion preceding Example 2 in §9.1. Show that the set E, which is the image of the open rectangle R under the inverse transformation, is connected. HINT: Let A denote that subset of E consisting of (u_0, v_0) and those points of E which can be reached from (u_0, v_0) by a polygonal path made up of finitely many line segments and lying altogether in E. A is not empty. Let B be the part of E which is not in A. We want to show that B is empty. Show that A is open and B must be open. Then $T(A)$ and $T(B)$ are two disjoint open sets whose union is R. By §5.1, Example 5, they cannot both be nonempty. Therefore $T(B)$ is empty, so B is empty and E is connected.

10 / VECTORS AND VECTOR FIELDS

10 / PURPOSE OF THE CHAPTER

This chapter has two purposes. One is to prepare the way for Chapter 11, in which we study linear transformations from one Euclidean vector space to another. For this purpose we need to explain the basic notions about vectors in Euclidean space and the algebra of Euclidean vector spaces. The second purpose is to concentrate on the notions of gradient, divergence, and curl in relation to the notions of scalar fields and vector fields. Much of this second purpose of the chapter is related to the use of vector-valued functions in the study of curves and surfaces in Chapter 14 and the study of the divergence theorem and Stokes' theorem in Chapter 15. Nearly all of the important applications of these theorems arise in connection with the reduction of "applied problems" (such as problems in electrostatics, electricity and magnetism, elasticity, gravitational theory, diffusion, and hydrodynamics) to mathematical problems formulated in terms of partial differential equations.

Much of the usefulness of vectors comes from the fact that we are able to avoid cumbersome and perhaps irrelevant references to a point by the specification of its co-ordinates in some co-ordinate system. Instead, we identify the point by a vector. The convenience in being able to use one symbol instead of several for a mathematical object is significant. But there is also another important consideration: that of achieving the clarity of thought and understanding that comes from expressing things in a way that makes no reference to a co-ordinate system.

10.1 / VECTORS IN EUCLIDEAN SPACE

In §2.6 we talked about the "axis of reals" as a straight line representing the real number system, each point being thought of as a real number. We call this line the one-dimensional Euclidean space $\mathbb{R}$, or $\mathbb{R}^1$. The *origin* of this space is the point representing number zero.

The two-dimensional Euclidean space $\mathbb{R}^2$ is the plane of plane analytic geometry, with points (x, y) identified by use of a Cartesian co-ordinate system consisting of an x-axis and a y-axis intersecting at right angles. Sometimes we denote points by (x_1, x_2) instead of (x, y). The origin in $\mathbb{R}^2$ is the point $(0, 0)$.

The Euclidean space $\mathbb{R}^3$ is the three-dimensional space of points (x, y, z), where the three co-ordinate axes are mutually perpendicular. The origin in $\mathbb{R}^3$ is the point $(0, 0, 0)$.

There is, of course, a point of view about Euclidean geometry (of three dimensions, let us say, to be specific), in which the geometric theory is constructed without an origin and a co-ordinate system, the fundamental notions and theorems being developed from assumptions about points, lines, planes, and the use of distance and the concepts of parallelism and perpendicularity. In this aspect of Euclidean geometry there is no particular point to be given special recognition as "the origin" and there are no particular lines to be given special status as co-ordinate axes. We shall use this "co-ordinate-free" point of view from time to time and we shall take advantage of our common familiarity with what we may call the "physical reality" of three-dimensional space, the space in which we live and about which we have some useful intuitive perceptions based on experience. But we shall build our systematic treatment of vectors in Euclidean space on the foundation provided by $\mathbb{R}^3$ as a set of things called either points or vectors. The same thing can be done for $\mathbb{R}^2$ as the basic model for Euclidean plane geometry (when treated analytically). And then it is readily seen how to make the generalization to $\mathbb{R}^n$, where n can be any positive integer.

Let us denote elements of $\mathbb{R}^3$ by $\mathbf{A} = (A_1, A_2, A_3)$, $\mathbf{B} = (B_1, B_2, B_3)$, and so on. We shall call them *vectors*, using boldface type for the symbols. We call A_1, A_2, A_3 the components of $\mathbf{A}$. They are simply the co-ordinates of $\mathbf{A}$ if we think of it as a point (which we are free to do, of course). The reason for using the word vector is that we are going to introduce definitions of algebraic operations on the elements of $\mathbb{R}^3$ which make it into what is called a *vector space* in the technical terminology of linear algebra. It is a common usage to represent $\mathbf{A}$ visually as an arrow from $(0, 0, 0)$ to (A_1, A_2, A_3). If P is the tip of the arrow (see Fig. 68), then P *is* $\mathbf{A}$ when we think of $\mathbf{A}$ as a point.

There is an algebra of vectors in $\mathbb{R}^3$ which rests on addition and subtraction of vectors and multiplication of vectors by real numbers (which are often called *scalars*). The vector $(0, 0, 0)$ is called the zero vector, and denoted by $\mathbf{0}$.

The vector sum of $\mathbf{A}$ and $\mathbf{B}$ is defined by

$$\mathbf{A} + \mathbf{B} = (A_1, A_2, A_3) + (B_1, B_2, B_3) = (A_1 + B_1, A_2 + B_2, A_3 + B_3). \quad (10.1\text{–}1)$$

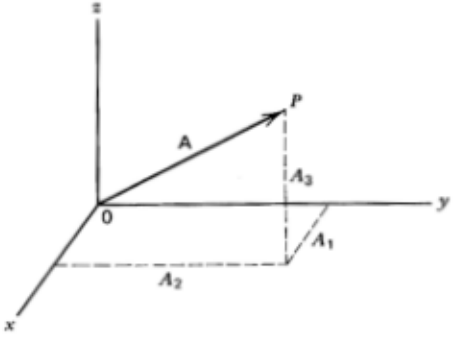


Fig. 68.

The vector difference is defined by

$$\mathbf{A} - \mathbf{B} = (A_1 - B_1, A_2 - B_2, A_3 - B_3). \tag{10.1–2}$$

Multiplying $\mathbf{A}$ by a scalar c is defined by

$$c\mathbf{A} = c(A_1, A_2, A_3) = (cA_1, cA_2, cA_3). \tag{10.1–3}$$

It is easily verified that the following algebraic rules are valid:

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \qquad \text{(commutative law)} \tag{10.1–4}$$

$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) \qquad \text{(associative law)} \tag{10.1–5}$$

$$\mathbf{A} + \mathbf{0} = \mathbf{A} \tag{10.1–6}$$

$$\left.\begin{aligned} c(\mathbf{A} + \mathbf{B}) &= c\mathbf{A} + c\mathbf{B} \qquad (10.1\text{–}7)\\ (a + b)\mathbf{A} &= a\mathbf{A} + b\mathbf{A} \qquad (10.1\text{–}8)\end{aligned}\right\} \text{(distributive laws)}$$

$$a(b\mathbf{A}) = (ab)\mathbf{A} \tag{10.1–9}$$

$$1\mathbf{A} = \mathbf{A} \tag{10.1–10}$$

$$0\mathbf{A} = \mathbf{0}, \tag{10.1–11}$$

Note that the zero on the left in (10.1–11) is a number, while the zero on the right is a vector. Observe also that $\mathbf{A} - \mathbf{B}$ is the same as $\mathbf{A} + (-1)\mathbf{B}$, that $\mathbf{A} - \mathbf{A} = \mathbf{0}$, and that the equation $\mathbf{X} + \mathbf{B} = \mathbf{A}$ can be solved uniquely for $\mathbf{X}$, the solution being $\mathbf{X} = \mathbf{A} - \mathbf{B}$. We regularly write $-\mathbf{A}$ for $(-1)\mathbf{A}$. Also, we sometimes write $\mathbf{A}/c$ for $\left(\frac{1}{c}\right)\mathbf{A}$.

These algebraic rules are so fully consistent with the ordinary rules in the algebra of real numbers that we find it easy to become proficient in vector algebra. Note, however, that we have not, in the foregoing, defined how to form the product of two vectors as another vector. Instead, we define what is called the *scalar product* of two vectors. The usual notation for it employs a dot between the factors, and for that reason $\mathbf{A} \cdot \mathbf{B}$ is called the *dot product* of $\mathbf{A}$ and $\mathbf{B}$. It is also sometimes called the *inner product. The dot product is a scalar.* The definition is made in terms of the components:

$$\mathbf{A} \cdot \mathbf{B} = A_1B_1 + A_2B_2 + A_3B_3. \tag{10.1–12}$$

It is easily verified that the following rules are valid:

$$\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A} \qquad \text{(commutative law)} \tag{10.1–13}$$

$$\mathbf{A} \cdot (\mathbf{B} + \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \cdot \mathbf{C} \qquad \text{(distributive law)} \tag{10.1–14}$$

$$(c\mathbf{A}) \cdot \mathbf{B} = c(\mathbf{A} \cdot \mathbf{B}) \tag{10.1–15}$$

It follows from the choice $c = 0$ in (10.1–15) that $\mathbf{0} \cdot \mathbf{B} = 0$. However it can happen that $\mathbf{A} \cdot \mathbf{B} = 0$ even though neither $\mathbf{A}$ nor $\mathbf{B}$ is the zero vector. For example, $(1, 1, 1) \cdot (1, -2, 1) = 0$.

We see from (10.1–12) that

$$(\mathbf{A} \cdot \mathbf{A})^{1/2} = (A_1^2 + A_2^2 + A_3^2)^{1/2}$$

is the length of the vector $\mathbf{A}$, that is, the distance from the origin to the tip of $\mathbf{A}$ (or to $\mathbf{A}$, thought of as a point). We shall denote the length of $\mathbf{A}$ by $\|\mathbf{A}\|$ and call it the *norm* of $\mathbf{A}$. Thus

$$\|A\| = (\mathbf{A} \cdot \mathbf{A})^{1/2}. \tag{10.1–16}$$

The only vector with zero length is $\mathbf{0}$. We say that each nonzero vector determines (or *has*) a direction in $\mathbb{R}^3$. If $\mathbf{A} \neq \mathbf{0}$ and $\mathbf{B} \neq \mathbf{0}$, we say that $\mathbf{A}$ and $\mathbf{B}$ have the *same direction* if one is a positive multiple of the other $\mathbf{B} = c\mathbf{A}$, where $c > 0$. When $\mathbf{B}$ is a negative multiple of $\mathbf{A}$ we say that $\mathbf{A}$ and $\mathbf{B}$ have opposite directions.

If we subject every vector in $\mathbb{R}^3$ to a transformation by adding to each vector the same nonzero vector $\mathbf{C}$, so that every $\mathbf{A}$ is transformed into $\mathbf{A} + \mathbf{C}$, we call this a *translation* of the space. Every point is carried into a new point which is a distance $\|\mathbf{C}\|$ in the direction of $\mathbf{C}$ from the original point.

There is a convenient way to visualize addition and subtraction of vectors. Let $\mathbf{A}$ and $\mathbf{B}$ be two nonzero vectors, and picture them as arrows emanating from the origin. If $\mathbf{B}$ is not a multiple of $\mathbf{A}$ the two vectors determine a unique plane. The sum $\mathbf{C} = \mathbf{A} + \mathbf{B}$ is then the vector emanating from the origin which forms the diagonal of the parallelogram of which $\mathbf{A}$ and $\mathbf{B}$ are adjacent sides. Another way of describing the situation intuitively is as follows: Displace $\mathbf{B}$ by a translation which brings its initial point (the origin) into coincidence with the terminal point (tip) of $\mathbf{A}$. The sum $\mathbf{A} + \mathbf{B}$ is then the vector from the initial point of $\mathbf{A}$ to the terminal point of the displaced vector $\mathbf{B}$ (see Fig. 69). This mode of representing

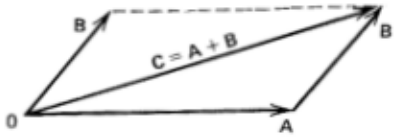


Fig. 69.

vector addition gives rise to the name "the parallelogram law of addition." If $\mathbf{B}$ has the same or opposite direction as $\mathbf{A}$, the addition of $\mathbf{B}$ to $\mathbf{A}$ can be visualized by the same process of displacement of $\mathbf{B}$: $\mathbf{A} + \mathbf{B}$ extends from the initial point of $\mathbf{A}$ to the tip of the displaced vector $\mathbf{B}$.

Vector subtraction can also be displayed visually by use of the parallelogram law. See Fig. 70 and recall that $\mathbf{A} - \mathbf{B}$ is that vector which, when added to $\mathbf{B}$, gives $\mathbf{A}$. Multiplication by scalars can likewise be displayed visually. See Fig. 71.

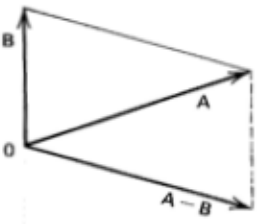


Fig. 70.

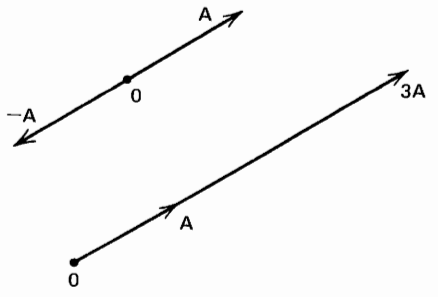


Fig. 71.

The dot product $\mathbf{A} \cdot \mathbf{B}$ can also be given a geometric interpretation that can be displayed visually. To explain this we use the law of cosines and properties of the dot product. We know from (10.1–16) that

$$\|\mathbf{A} - \mathbf{B}\|^2 = (\mathbf{A} - \mathbf{B}) \cdot (\mathbf{A} - \mathbf{B}).$$

But, by the algebraic rules,

$$\begin{aligned}(\mathbf{A} - \mathbf{B}) \cdot (\mathbf{A} - \mathbf{B}) &= (\mathbf{A} - \mathbf{B}) \cdot \mathbf{A} - (\mathbf{A} - \mathbf{B}) \cdot \mathbf{B} = \mathbf{A} \cdot \mathbf{A} - \mathbf{B} \cdot \mathbf{A} - \mathbf{A} \cdot \mathbf{B} + \mathbf{B} \cdot \mathbf{B} \\ &= \|\mathbf{A}\|^2 - 2\mathbf{A} \cdot \mathbf{B} + \|\mathbf{B}\|^2.\end{aligned}$$

Therefore

$$\|\mathbf{A} - \mathbf{B}\|^2 = \|\mathbf{A}\|^2 + \|\mathbf{B}\|^2 - 2\mathbf{A} \cdot \mathbf{B}. \tag{10.1–17}$$

Now consider Fig. 72, in which θ is the angle between the vectors $\mathbf{A}$ and $\mathbf{B}$. (Here we assume that neither $\mathbf{A}$ nor $\mathbf{B}$ is $\mathbf{0}$.) The triangle with vertices at $\mathbf{0}$ and the tips of $\mathbf{A}$ and $\mathbf{B}$ has the side opposite θ as a line segment whose length is the same as that of $\mathbf{A} - \mathbf{B}$. By the law of cosines

$$\|\mathbf{A} - \mathbf{B}\|^2 = \|\mathbf{A}\|^2 + \|\mathbf{B}\|^2 - 2\|\mathbf{A}\| \, \|\mathbf{B}\| \cos \theta \tag{10.1–18}$$

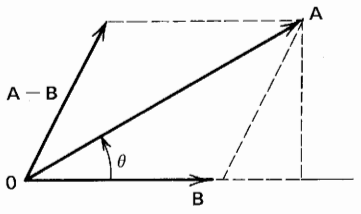


Fig. 72.

On comparing (10.1–18) with (10.1–17) we see that

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos \theta. \tag{10.1–19}$$

This can be expressed as follows: The dot product $\mathbf{A} \cdot \mathbf{B}$ is equal to the product of the length of $\mathbf{B}$ times the signed projection of $\mathbf{A}$ on the directed line of the vector $\mathbf{B}$. From (10.1–19) it is clear that, when neither $\mathbf{A}$ nor $\mathbf{B}$ is $\mathbf{0}$, $\mathbf{A} \cdot \mathbf{B} = 0$ if and only if $\cos \theta = 0$, that is, if and only if $\mathbf{A}$ is perpendicular to $\mathbf{B}$.

Sometimes nonzero vectors in Euclidean space of three dimensions are defined purely geometrically as directed line segments. Thus, if P and Q are distinct points, the ordered pair of points P, Q determines a directed line segment from P to Q and is called a vector, denoted by $\overrightarrow{PQ}$. If R,S is another

ordered pair of points such that $\overrightarrow{PQ}$ and $\overrightarrow{RS}$ have the same length and direction, $\overrightarrow{PQ}$ and $\overrightarrow{RS}$ are said to be equal (or, sometimes, equivalent) vectors. This means that $\overrightarrow{RS}$ can be brought into coincidence with $\overrightarrow{PQ}$ by a translation of the whole space whereby R is moved along a straight line to P, and S is likewise moved, in the same direction, to Q. In physics a vector is often called a *free vector* if it is considered to be the same after any translation of the space. In our presentation of vectors in R^3, however, we think of all vectors as having their initial point at the origin. With this mode of thinking about vectors we call R^3 *a Euclidean vector space.* As mentioned previously, we find it convenient to refer to a vector **A** as either a vector or a point, so that it is needless to have a separate notation for the point that is the tip of **A**.

10.11 / ORTHOGONAL UNIT VECTORS IN R^3

Let a fixed rectangular co-ordinate system be chosen with origin O. It is conventional, particularly in dealing with physical applications, to work with right-handed co-ordinate systems, and we shall ordinarily adhere to this convention. Now let **i**, **j**, **k** be vectors, each of unit length, in the directions of the positive x, y, and z axes, respectively (see Fig. 73). It is clear that if **A** is any vector, we can express it in the form

$$\mathbf{A} = A_1\mathbf{i} + A_2\mathbf{j} + A_3\mathbf{k}, \tag{10.11–1}$$

where A_1, A_2, A_3 are the components of **A** (see Fig. 74). We call **i**, **j**, **k** the

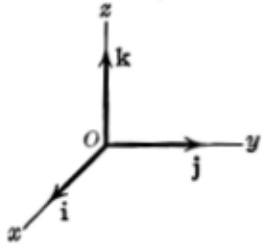


Fig. 73.

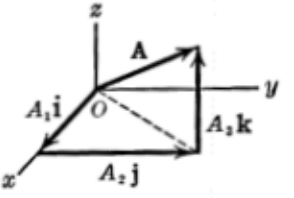


Fig. 74.

fundamental orthonormal triad associated with this particular co-ordinate system. The word "orthonormal" is a combination of "orthogonal" and "normal." The vectors **i**, **j**, **k** form an *orthogonal set*; that is, they are mutually perpendicular. A vector is said to be normalized, or normal, if it is of unit length. The orthonormal character of the triad **i**, **j**, **k** is expressed by the relations

$$\begin{aligned} \mathbf{i}\cdot\mathbf{i} = \mathbf{j}\cdot\mathbf{j} = \mathbf{k}\cdot\mathbf{k} &= 1, \\ \mathbf{i}\cdot\mathbf{j} = \mathbf{j}\cdot\mathbf{k} = \mathbf{k}\cdot\mathbf{i} &= 0. \end{aligned} \tag{10.11–2}$$

EXERCISES

1. Find $\mathbf{A}+\mathbf{B}$, $\mathbf{A}-\mathbf{B}$, and $\mathbf{A}\cdot\mathbf{B}$ in each case. Are $\mathbf{A}$ and $\mathbf{B}$ collinear, perpendicular, or neither?

(a) $\mathbf{A}=(1,1,-1)$, $\mathbf{B}=(3,-2,-1)$.

(b) $\mathbf{A}=(1,4,3)$, $\mathbf{B}=(4,2,-4)$.

(c) $\mathbf{A}=(2,-1,1)$, $\mathbf{B}=(3,-4,-4)$.

(d) $\mathbf{A}=(6,4,-2)$, $\mathbf{B}=(-9,-6,3)$.

2. Find $\|\mathbf{A}\|$ in each case.

(a) $\mathbf{A}=2\mathbf{i}+3\mathbf{j}+6\mathbf{k}$. **(b)** $\mathbf{A}=4\mathbf{i}+2\mathbf{j}-4\mathbf{k}$. **(c)** $\mathbf{A}=2\mathbf{i}+\mathbf{j}-2\mathbf{k}$. **(d)** $\mathbf{A}=4\mathbf{i}+3\mathbf{j}$.

3. If the angle between $\mathbf{A}$ and $\mathbf{B}$ is θ, where $\mathbf{A}=(-2,-2,1)$ and $\mathbf{B}=(1,-2,2)$, find $\cos\theta$.

4. Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ be four vectors in the vector space with origin O, and suppose that $\mathbf{B}-\mathbf{A}=\mathbf{C}-\mathbf{D}$. Show that the tips of these four vectors are consecutive vertices of a parallelogram.

5. Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ be noncoplanar vectors in the vector space with origin O. What is the vector $\mathbf{A}+\mathbf{B}+\mathbf{C}$ in relation to the parallepiped of which $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are edges?

6. Let $\mathbf{A}=(\cos\theta,\sin\theta,0)$, $\mathbf{B}=(\cos\phi,\sin\phi,0)$, where $0\leqq\phi\leqq\theta\leqq 2\pi$. Draw a figure showing the positions of $\mathbf{A}$ and $\mathbf{B}$ in the xy-plane. Use (10.1–12) and (10.1–19) to obtain the trigonometric identity for $\cos(\theta-\phi)$.

7. If $\mathbf{A}$ and $\mathbf{B}$ are of unit length, and θ is the angle between them, express $\|\mathbf{A}-\mathbf{B}\|$ as a function of θ.

8. Let $\mathbf{A}$ be a vector of unit length, and let $\mathbf{B}$ be any vector. Let $\mathbf{C}=(\mathbf{B}\cdot\mathbf{A})\mathbf{A}$, $\mathbf{D}=\mathbf{B}-\mathbf{C}$. Prove that $\mathbf{D}\cdot\mathbf{A}=0$. Draw a figure showing the relation of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$. Work out the special cases **(a)** $\mathbf{A}=\mathbf{i}$, $\mathbf{B}=5\mathbf{i}+2\mathbf{j}$; **(b)** $\mathbf{A}=\frac{1}{5}(3\mathbf{i}+4\mathbf{j})$, $\mathbf{B}=5\mathbf{j}$.

10.12 / THE VECTOR SPACE $\mathbb{R}^n$

By a natural generalization of our definition of $\mathbb{R}^1$, $\mathbb{R}^2$, and $\mathbb{R}^3$ we define $\mathbb{R}^n$, for any positive integer n, as the set of all ordered n-tuples $(x_1,\ldots,x_n)$ of n real numbers. We call the elements of $\mathbb{R}^n$ either points or vectors. Here we shall usually call them vectors, because we are going to be extending the notions that were introduced in §10.1 to enable us to regard $\mathbb{R}^n$ as a Euclidean vector space of n dimensions. When $n>3$ we are in general deprived of our useful resort to physical intuition and the visualization of vector relationships in "ordinary" space of one, two, or three dimensions. However, we continue to use geometric language, even though we deal with matters analytically.

We use the notations $\mathbf{x}=(x_1,\ldots,x_n)$, $\mathbf{y}=(y_1,\ldots,y_n)$, and so on, with a vector designated by a single letter in boldface type and the components of the vector designated by the same letter, with subscripts, but *not* in boldface type. We can use either capital or lower case letters for vectors.

We define $\mathbf{x}+\mathbf{y}$ and $c\mathbf{x}$ by the natural generalizations of (10.1–1) and (10.1–3), adding vectors by adding corresponding components, and forming $c\mathbf{x}$ by multiplying each component of $\mathbf{x}$ by c. We also extend the definition of the dot product, defining

$$\mathbf{x}\cdot\mathbf{y}=x_1y_1+\cdots+x_ny_n. \tag{10.12–1}$$

The *length* of $\mathbf{x}$, called its *norm*, is denoted by $\|\mathbf{x}\|$ and defined by

$$\|\mathbf{x}\| = (x_1^2 + \cdots + x_n^2)^{1/2}. \tag{10.12–2}$$

From (10.12–1) we see that

$$\|\mathbf{x}\|^2 = \mathbf{x} \cdot \mathbf{x}. \tag{10.12–3}$$

With $(0, 0, \ldots, 0)$ as the vector $\mathbf{0}$, the same algebraic laws hold for $\mathbb{R}^n$ as those given for $\mathbb{R}^3$ in (10.1–4) to (10.1–11) inclusive and (10.1–13) to (10.1–15) inclusive.

The following inequality, known as Cauchy's inequality, is very important:

$$\left|\sum_{i=1}^{n} x_i y_i\right| \leqq \left(\sum_{i=1}^{n} x_i^2\right)^{1/2} \left(\sum_{i=1}^{n} y_i^2\right)^{1/2}. \tag{10.12–4}$$

One way of obtaining Cauchy's inequality was indicated in Exercise 29, §6.8. For another way see Exercise 4 at the end of this section. We can rewrite (10.12–4) as

$$|\boldsymbol{x} \cdot \mathbf{y}| \leqq \|\boldsymbol{x}\| \, \|\mathbf{y}\|. \tag{10.12–5}$$

From this it is easy to show that

$$\|\mathbf{x} + \mathbf{y}\| \leqq \|\mathbf{x}\| + \|\mathbf{y}\|. \tag{10.12–6}$$

We leave the derivation of (10.12–6) to the student; see Exercise 5. This inequality is called the triangle inequality. It corresponds to the geometric assertion that in a triangle the length of one side is never larger than the sum of the lengths of the other two sides.

We define *distance* in $\mathbb{R}^n$ by the natural generalization of the formula in $\mathbb{R}^3$. The distance $d(\mathbf{x}, \mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$ (thought of as points) is defined as

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{n} (x_i - y_i)^2\right]^{1/2} = \|\mathbf{x} - \mathbf{y}\|. \tag{10.12–7}$$

Two vectors $\mathbf{x}$, $\mathbf{y}$ in $\mathbb{R}^n$ are said to be *orthogonal* if $\mathbf{x} \cdot \mathbf{y} = 0$. The naturalness of this definition is seen from the discussion accompanying (10.1–19).

There is a set of n mutually orthogonal vectors of unit length in $\mathbb{R}^n$ entirely analogous to the vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ in $\mathbb{R}^3$ that were introduced in §10.11. We define

$$\begin{aligned} \mathbf{e}_1 &= (1, 0, 0, \ldots, 0) \\ \mathbf{e}_2 &= (0, 1, 0, \ldots, 0) \\ &\vdots \\ \mathbf{e}_n &= (0, 0, \ldots, 0, 1). \end{aligned} \tag{10.12–8}$$

It is clear that $\|\mathbf{e}_i\| = 1$ and $\mathbf{e}_i \cdot \mathbf{e}_j = 0$ if $i \neq j$, so we call $\mathbf{e}_1, \ldots, \mathbf{e}_n$ an *orthonormal* set. We see at once that we can write

$$\mathbf{x} = x_1 \mathbf{e}_1 + \cdots + x_n \mathbf{e}_n, \tag{10.12–9}$$

so that each vector $\mathbf{x}$ is a linear combination of the $\mathbf{e}_i$'s, the coefficients being the components of $\mathbf{x}$. This representation of $\mathbf{x}$ is the analogue of the representation of $\mathbf{A}$ in (10.11–1).

Our reason for saying that $\mathbb{R}^n$ is n-dimensional is closely related to (10.12–9). To explain the concept of dimensionality we must first speak about the notions of linear dependence and linear independence of a set of vectors. A set of k vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is called *linearly dependent* if there is a set of k scalars $c_1, \ldots, c_k$ *not all zero*, such that

$$c_1\mathbf{u}_1 + \cdots + c_k\mathbf{u}_k = \mathbf{0}. \tag{10.12–10}$$

If $c_j \neq 0$, we can solve the foregoing equation for $\mathbf{u}_j$, expressing it as a linear combination of the other $\mathbf{u}_i$'s. On the other hand, if the set $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is *not* linearly dependent, we call it *linearly independent.* In this case, an equation of the form (10.12–10) cannot be valid except when all the c_i's are zero.

The vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ are linearly independent, for

$$c_1\mathbf{e}_1 + \cdots + c_n\mathbf{e}_n = (c_1, \ldots, c_n) = \mathbf{0}$$

if and only if $c_1 = \cdots = c_n = 0$. The vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ form what is called a *basis* for $\mathbb{R}^n$; this means that the vectors are linearly independent and that every vector in $\mathbb{R}^n$ is a linear combination of the $\mathbf{e}_i$'s. There are other sets forming a basis for $\mathbb{R}^n$; we call $\mathbf{e}_1, \ldots, \mathbf{e}_n$ *the standard basis* for $\mathbb{R}^n$. In particular, $\mathbf{i}, \mathbf{j}, \mathbf{k}$ is the standard basis for $\mathbb{R}^3$. It is a special case of a general theorem of linear algebra that every basis for $\mathbb{R}^n$ consists of exactly n vectors. It is for this reason that we say that $\mathbb{R}^n$ is n-dimensional.

It is easy to see that a set of k vectors is linearly dependent if one of them is zero, and that an orthonormal set of k vectors is linearly independent (Exercises 1 and 2). As a general test for linear dependence we have the following theorem.

THEOREM I. *A necessary and sufficient condition that the vectors* $\mathbf{v}_1, \ldots, \mathbf{v}_k$ *be linearly dependent is that the determinant* G *defined by*

$$G = \begin{vmatrix} \mathbf{v}_1 \cdot \mathbf{v}_1 & \mathbf{v}_1 \cdot \mathbf{v}_2 & \cdots & \mathbf{v}_1 \cdot \mathbf{v}_k \\ \mathbf{v}_2 \cdot \mathbf{v}_1 & \mathbf{v}_2 \cdot \mathbf{v}_2 & \cdots & \mathbf{v}_2 \cdot \mathbf{v}_k \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{v}_k \cdot \mathbf{v}_1 & \mathbf{v}_k \cdot \mathbf{v}_2 & \cdots & \mathbf{v}_k \cdot \mathbf{v}_k \end{vmatrix} \tag{10.12–11}$$

be equal to zero. (An equivalent statement is that the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ *are linearly independent if and only if* $G \neq 0$.)

Proof. We have to show that $G = 0$ if and only if there exist scalars $c_1, \ldots, c_k$, not all zero, such that

$$c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \mathbf{0}. \tag{10.12–12}$$

If we suppose that the equation (10.12–12) holds for some set of c_i's, then, by forming dot products of the expression in (10.12–12) successively with

$\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$, we obtain the system of equations

$$\begin{aligned} c_1\mathbf{v}_1 \cdot \mathbf{v}_1 + c_2\mathbf{v}_1 \cdot \mathbf{v}_2 + \cdots + c_k\mathbf{v}_1 \cdot \mathbf{v}_k &= 0 \\ c_1\mathbf{v}_2 \cdot \mathbf{v}_1 + c_2\mathbf{v}_2 \cdot \mathbf{v}_2 + \cdots + c_k\mathbf{v}_2 \cdot \mathbf{v}_k &= 0 \\ \cdots & \\ c_1\mathbf{v}_k \cdot \mathbf{v}_1 + c_2\mathbf{v}_k \cdot \mathbf{v}_2 + \cdots + c_k\mathbf{v}_k \cdot \mathbf{v}_k &= 0. \end{aligned} \tag{10.12–13}$$

On the other hand, if equations (10.12–13) are satisfied by a set of c_i's, then so are the equations that result when we multiply both sides of the first equation by c_1, the second by c_2, and so on. But, by use of the algebraic rules governing dot products, the equations thus obtained can be written

$$\begin{aligned} c_1\mathbf{v}_1 \cdot (c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k) &= 0 \\ c_2\mathbf{v}_2 \cdot (c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k) &= 0 \\ \cdots & \\ c_k\mathbf{v}_k \cdot (c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k) &= 0. \end{aligned}$$

On adding these equations, we conclude that

$$(c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k) \cdot (c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k) = 0,$$

and hence that equation (10.12–12) is valid, for a vector is zero if and only its dot product with itself is zero (see (10.12–2) and (10.12–3)).

We see, therefore, that an equation of the form (10.12–12) holds true if and only if the system of homogeneous linear equations (10.12–13) (with the c_i's regarded as unknowns) is satisfied. But, by the algebraic theory of simultaneous linear equations, the homogeneous system (10.12–13) is satisfied by a set of c_i's that are not all zero if and only if the determinant of the system (which is the determinant G in (10.12–11) is equal to zero. This proves the theorem.

The determinant G is called the Gram determinant, or the Gramian, in honor of J. P. Gram, a mathematician of the nineteenth century. We shall presently see his name again.

The members of a linearly independent set of vectors need not be unit vectors, of course, and they need not be mutually orthogonal, but if we are given a linearly independent set of k vectors $\mathbf{A}_1, \ldots, \mathbf{A}_k$, it is possible by a systematic process to construct an orthonormal set of k vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$, each of which is a linear combination of the $\mathbf{A}_i$'s. (Here we are assuming $k \geqq 2$.) This process is called *the Gram–Schmidt process*, in recognition of the work of J. P. Gram and E. Schmidt, a mathematician of the early twentieth century.

The idea of the process is quite simple; the basic method can be interpreted geometrically by dealing with two linearly independent vectors that are not orthogonal. We denote them by **C** and **D** and picture them in a plane. See Fig. 75.

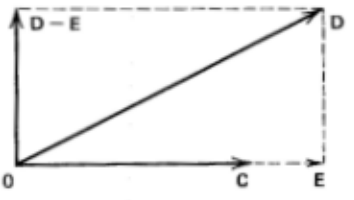


Fig. 75.

Let **E** be defined by

$$\mathbf{E} = \frac{\mathbf{D} \cdot \mathbf{C}}{\|\mathbf{C}\|^2} \mathbf{C}. \tag{10.12–14}$$

We call **E** the vector projection of **D** on **C**. Observe that $\frac{\mathbf{C}}{\|\mathbf{C}\|}$ is a unit vector in the direction of **C** and that **E** is a multiple of this unit vector by the factor

$$\frac{\mathbf{D} \cdot \mathbf{C}}{\|C\|}.$$

According to (10.1–19) this factor is $\|\mathbf{D}\|$ times the cosine of the angle between **C** and **D**; this justifies calling **E** the vector projection of **D** on **C**.

Now consider the vector $\mathbf{D} - \mathbf{E}$ (i.e., **D** minus the vector projection of **D** on **C**). It is orthogonal to **C**. This is clear in Fig. 75, but we can verify it analytically, for, by (10.12–14),

$$(\mathbf{D} - \mathbf{E}) \cdot \mathbf{C} = \mathbf{D} \cdot \mathbf{C} - \frac{\mathbf{D} \cdot \mathbf{C}}{\|\mathbf{C}\|^2} \mathbf{C} \cdot \mathbf{C} = 0,$$

because $\|\mathbf{C}\|^2 = \mathbf{C} \cdot \mathbf{C}$.

The Gram–Schmidt process utilizes repeatedly the process of subtracting from a given vector each of its projections on a certain set of vectors. If $\mathbf{A}_1, \ldots, \mathbf{A}_k$ is the given linearly independent set of vectors, we start by defining the unit vector

$$\mathbf{v}_1 = \frac{\mathbf{A}_1}{\|\mathbf{A}_1\|},$$

and let $\mathbf{B}_2$ be $\mathbf{A}_2$ minus the projection of $\mathbf{A}_1$ on $\mathbf{v}_1$:

$$\mathbf{B}_2 = \mathbf{A}_2 - (\mathbf{A}_2 \cdot \mathbf{v}_1)\mathbf{v}_1;$$

(recall that $\|\mathbf{v}_1\| = 1$). Now $\mathbf{B}_2$ is orthogonal to $\mathbf{v}_1$, just as, above, $\mathbf{D} - \mathbf{E}$ is orthogonal to **C**. Also, $\mathbf{B}_2$ is a linear combination of $\mathbf{A}_1$ and $\mathbf{A}_2$, and therefore is not **0** (because $\mathbf{A}_1$ and $\mathbf{A}_2$ are linearly independent). Hence we can define the unit vector

$$\mathbf{v}_2 = \frac{\mathbf{B}_2}{\|\mathbf{B}_2\|};$$

it is orthogonal to $\mathbf{v}_1$. Next, we define

$$\mathbf{B}_3 = \mathbf{A}_3 - (\mathbf{A}_3 \cdot \mathbf{v}_1)\mathbf{v}_1 - (\mathbf{A}_3 \cdot \mathbf{v}_2)\mathbf{v}_2;$$

here we have subtracted from $\mathbf{A}_3$ its projection on $\mathbf{v}_1$ and its projection on $\mathbf{v}_2$. A calculation shows that $\mathbf{B}_3$ is orthogonal to $\mathbf{v}_1$ and to $\mathbf{v}_2$. Also, it is not zero (why?), and so, if we define

$$\mathbf{v}_3 = \frac{\mathbf{B}_3}{\|\mathbf{B}_3\|},$$

we see that $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ form an orthonormal set. Moreover, $\mathbf{v}_1$ is a multiple of $\mathbf{A}_1$, $\mathbf{v}_2$ is a linear combination of $\mathbf{A}_1$ and $\mathbf{A}_2$, and $\mathbf{v}_3$ is a linear combination of $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$. It is now clear how to continue the process of obtaining more $\mathbf{v}_i$'s as long as there are still $\mathbf{A}_i$'s from which to subtract their projections on $\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}$.

Example. Show that the following four vectors in R^4 are linearly independent. Then apply the Gram–Schmidt process to construct an orthonormal set from the given vectors. The vectors are:

$$\begin{aligned} &\mathbf{A}_1 = (1, 1, 1, 1) \qquad \mathbf{A}_2 = (0, -1, 0, -1), \\ &\mathbf{A}_3 = (0, 0, 1, 1), \qquad \mathbf{A}_4 = (1, -2, -2, 0). \end{aligned} \tag{10.12–15}$$

We illustrate the application of Theorem I. First we calculate

$$\begin{aligned} &\mathbf{A}_1 \cdot \mathbf{A}_1 = 4, \quad \mathbf{A}_1 \cdot \mathbf{A}_2 = -2, \quad \mathbf{A}_1 \cdot \mathbf{A}_3 = 2, \quad \mathbf{A}_1 \cdot \mathbf{A}_4 = -3, \\ &\mathbf{A}_2 \cdot \mathbf{A}_2 = 2, \quad \mathbf{A}_2 \cdot \mathbf{A}_3 = -1 \quad \mathbf{A}_2 \cdot \mathbf{A}_4 = 2, \\ &\mathbf{A}_3 \cdot \mathbf{A}_3 = 2, \quad \mathbf{A}_3 \cdot \mathbf{A}_4 = -2 \\ &\mathbf{A}_4 \cdot \mathbf{A}_4 = 9. \end{aligned}$$

The Gramian is

$$G = \begin{vmatrix} 4 & -2 & 2 & -3 \\ -2 & 2 & -1 & 2 \\ 2 & -1 & 2 & -2 \\ -3 & 2 & -2 & 9 \end{vmatrix}.$$

By standard methods for calculating the value of a determinant we find that $G = 25$. Because $G \neq 0$ we conclude that the four vectors are linear independent. Therefore we can proceed with the Gram–Schmidt process. We list the results by stages, leaving the detailed calculations to be verified by the student.

$$\begin{aligned} &\mathbf{v}_1 = (\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}), \qquad \mathbf{B}_2 = \mathbf{A}_2 + \mathbf{v}_1 = (\tfrac{1}{2}, -\tfrac{1}{2}, \tfrac{1}{2}, -\tfrac{1}{2}), \\ &\mathbf{v}_2 = \mathbf{B}_2, \qquad \mathbf{B}_3 = \mathbf{A}_3 - \mathbf{v}_1 = (-\tfrac{1}{2}, -\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}) \\ &\mathbf{v}_3 = \mathbf{B}_3, \qquad \mathbf{B}_4 = \mathbf{A}_4 + \tfrac{3}{2}\mathbf{v}_1 - \tfrac{1}{2}\mathbf{v}_2 + \tfrac{1}{2}\mathbf{v}_3, \\ &\mathbf{B}_4 = (\tfrac{5}{4}, -\tfrac{5}{4}, -\tfrac{5}{4}, \tfrac{5}{4}), \quad \|\mathbf{B}_4\| = \tfrac{5}{2}, \quad \mathbf{v}_4 = (\tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{1}{2}, \tfrac{1}{2}). \end{aligned}$$

EXERCISES

1. Show directly by the definition that $\mathbf{A}_1, \ldots, \mathbf{A}_k$ is a linearly dependent set of vectors if at least one of the vectors is zero.

2. Show directly by the definition that a set of nonzero and mutually orthogonal vectors is linearly independent.

3. Show that the vectors $\mathbf{A}_1 = 2\mathbf{i}$, $\mathbf{A}_2 = 3\mathbf{i} + 4\mathbf{j}$, $\mathbf{A}_3 = \mathbf{i} + 2\mathbf{j} + 3\mathbf{k}$ are linearly independent, and apply the Gram–Schmidt process to them.

4. (a) Deduce the inequality (10.12–4) in the form (10.12–5) with the aid of the

following suggestions. Consider

$$\sum_{i=1}^{n} (x_i + ty_i)^2$$

and express it as a quadratic polynomial in the real variable t. Since the polynomial can never be negative, it cannot have two distinct real roots (why not?). What does this imply about the expression under the radical sign in the quadratic formula for the roots of the polynomial?

(b) Show from the foregoing considerations that the equality sign will hold in (10.12–5) if and only if $\mathbf{x}$ is a multiple of $\mathbf{y}$.

5. Deduce the inequality $\|\mathbf{x}+\mathbf{y}\| \leqq \|\mathbf{x}\| + \|\mathbf{y}\|$ with the aid of Cauchy's inequality in the form $|\mathbf{x}\cdot\mathbf{y}| \leqq \|\mathbf{x}\|\,\|\mathbf{y}\|$. Under what condition (and only that) is it possible to have $\|\mathbf{x}+\mathbf{y}\| = \|\mathbf{x}\| + \|\mathbf{y}\|$?

6. (a) Show that the vectors $\mathbf{A} = (1, 0, 1)$, $\mathbf{B} = (1, 1, -1)$, and $\mathbf{C} = (-1, 2, 1)$ are mutually orthogonal.

(b) If $\mathbf{V} = (5, -2, 3)$, express $\mathbf{V}$ as a linear combination of $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$. Hint: Try to find α, β, γ so that $\mathbf{V} = \alpha\mathbf{A} + \beta\mathbf{B} + \gamma\mathbf{C}$ and show that α can be found as a multiple of $\mathbf{V}\cdot\mathbf{A}$.

7. Instead of using Theorem I to prove that the vectors in (10.12–15) are linearly independent, one can also use the following general test: A set of vectors $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4$ in $\mathbb{R}^4$ is linearly independent if and only if

$$\begin{vmatrix} \mathbf{A}_1\cdot\mathbf{e}_1 & \mathbf{A}_2\cdot\mathbf{e}_1 & \mathbf{A}_3\cdot\mathbf{e}_1 & \mathbf{A}_4\cdot\mathbf{e}_1 \\ \mathbf{A}_1\cdot\mathbf{e}_2 & \mathbf{A}_2\cdot\mathbf{e}_2 & \mathbf{A}_3\cdot\mathbf{e}_2 & \mathbf{A}_4\cdot\mathbf{e}_2 \\ \mathbf{A}_1\cdot\mathbf{e}_3 & \mathbf{A}_2\cdot\mathbf{e}_3 & \mathbf{A}_3\cdot\mathbf{e}_3 & \mathbf{A}_4\cdot\mathbf{e}_3 \\ \mathbf{A}_1\cdot\mathbf{e}_4 & \mathbf{A}_2\cdot\mathbf{e}_4 & \mathbf{A}_3\cdot\mathbf{e}_4 & \mathbf{A}_4\cdot\mathbf{e}_4 \end{vmatrix} \neq 0$$

where $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4$ are the standard basis vectors in $\mathbb{R}^4$. Prove that this test is valid. Then apply it to the vectors in (10.12–15).

8. Show that the dot product defined in (10.12–1) has the following five properties.

(a) $\mathbf{x}\cdot\mathbf{y} = \mathbf{y}\cdot\mathbf{x}$ for all $\mathbf{x}$ and $\mathbf{y}$.

(b) $\mathbf{x}\cdot(\mathbf{y}+\mathbf{z}) = \mathbf{x}\cdot\mathbf{y} + \mathbf{x}\cdot\mathbf{z}$ for all $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$.

(c) $(c\mathbf{x})\cdot\mathbf{y} = c(\mathbf{x}\cdot\mathbf{y})$ for all $\mathbf{x}$, $\mathbf{y}$ and all scalars c.

(d) $\mathbf{x}\cdot\mathbf{x} \geq 0$ for all $\mathbf{x}$.

(e) $\mathbf{x}\cdot\mathbf{x} = 0$ if and only $\mathbf{x} = \mathbf{0}$.

10.2 / CROSS PRODUCTS IN $\mathbb{R}^3$

In $\mathbb{R}^3$, there is another kind of vector multiplication, in which the product of two vectors is another vector. We shall define this product presently. It is called the *cross product* (or, sometimes, the *vector product*), and is denoted by $\mathbf{A}\times\mathbf{B}$.

Definition. *The cross product of a vector* $\mathbf{A}$ *by the vector* $\mathbf{B}$ *is defined as follows. If either* $\mathbf{A}$ *or* $\mathbf{B}$ *is* $\mathbf{0}$ *we define the product to be* $\mathbf{0}$:

$$\mathbf{0}\times\mathbf{B} = \mathbf{A}\times\mathbf{0} = \mathbf{0}.$$

Otherwise, let θ *be the angle between* $\mathbf{A}$ *and* $\mathbf{B}$. *We define* $\mathbf{A}\times\mathbf{B}$ *to be the vector*

of magnitude

$$\|\mathbf{A} \times \mathbf{B}\| = \|\mathbf{A}\| \, \|\mathbf{B}\| \sin \theta$$

whose line is perpendicular to the plane of **A** *and* **B**, *and whose direction is such that* **A**, **B**, *and* $\mathbf{A} \times \mathbf{B}$ *form a right-handed system (see Fig. 76). If the vectors* **A**, **B** *lie along the same line, they do not determine a plane. In this case* $\sin \theta = 0$, *however and so* $\mathbf{A} \times \mathbf{B} = \mathbf{0}$. *Note that the magnitude of* $\mathbf{A} \times \mathbf{B}$ *is equal to the area of the parallelogram of which* **A**, **B** *are adjacent sides.*

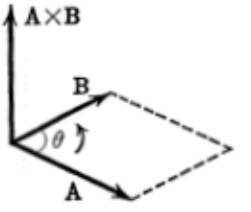


Fig. 76.

The motivation for this definition, and its usefulness, will be better understood by the student after he has seen the occurrence of the cross product in physical applications and in later mathematical developments. It has only very limited analogies with the ordinary product of two numbers. Moreover, the cross product is something peculiar to vectors in three dimensions, having no analogue for vector spaces of dimension other than three.

The principal algebraic rules governing the cross product are

$$\mathbf{A} \times \mathbf{B} = -(\mathbf{B} \times \mathbf{A}), \tag{10.2–1}$$

$$(c\mathbf{A}) \times \mathbf{B} = c(\mathbf{A} \times \mathbf{B}), \tag{10.2–2}$$

$$\mathbf{A} \times (\mathbf{B} + \mathbf{C}) = (\mathbf{A} \times \mathbf{B}) + (\mathbf{A} \times \mathbf{C}). \tag{10.2–3}$$

Multiplication is not commutative, but anticommutative, as we see by (10.2–1). This law is apparent from the definition of the cross product, since **B**, **A** and $-(\mathbf{A} \times \mathbf{B})$ form a right-handed system. The law (10.2–2) is also apparent from the definition of the cross product. The rule

$$\mathbf{A} \times (c\mathbf{B}) = c(\mathbf{A} \times \mathbf{B}) \tag{10.2–4}$$

can be deduced from (10.2–1) and (10.2–2).

Now consider the proof of the distributive law (10.2–3). The law obviously holds if $\mathbf{A} = \mathbf{0}$, so we consider the proof on the assumption that $\mathbf{A} \neq \mathbf{0}$. If **B** is any vector, let **B′** denote the vector projection of **B** on the plane perpendicular to **A** through the origin (see Fig. 77). Clearly $\|\mathbf{B}'\| = \|\mathbf{B}\| \sin \theta$, and therefore $\mathbf{A} \times \mathbf{B} = \mathbf{A} \times \mathbf{B}'$. Now, projecting in this manner, we see that $\mathbf{B}' + \mathbf{C}'$ is the projection of $\mathbf{B} + \mathbf{C}$. Therefore, instead of proving (10.2–3), it is sufficient to prove

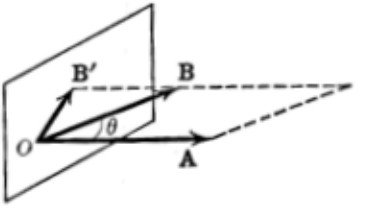


Fig. 77.

$$\mathbf{A} \times (\mathbf{B}' + \mathbf{C}') = \mathbf{A} \times \mathbf{B}' + \mathbf{A} \times \mathbf{C}'. \tag{10.2–5}$$

The advantage here is that the vectors **B′**, **C′** and $\mathbf{B}' + \mathbf{C}'$ either are **0** or are perpendicular to **A**.

Now let us consider a figure (Fig. 78) in which **A** is perpendicular to the plane of the page and is directed out toward the student as he reads the page. We may assume that neither **B′** nor **C′** is **0**, and also that these vectors are not collinear, since in these cases (10.2–5) is certainly true, as we readily see (if **B′** and **C′** are collinear, **C′** is a multiple of **B′**, and (10.2–5) then follows from (10.2–4)). Now, if **V** is any vector in the plane of the page, $\mathbf{A}\times\mathbf{V}$ will also lie in the plane of the page. It will make a 90° angle with **V**, and will have a length $\|\mathbf{A}\|$ times that of **V**. Let us take **V** to be, successively, **B′**, **C′**, $\mathbf{B}'+\mathbf{C}'$. Then we see that the configuration of the vectors $\mathbf{A}\times\mathbf{B}'$, $\mathbf{A}\times\mathbf{C}'$, $\mathbf{A}\times(\mathbf{B}'+\mathbf{C}')$ is similar to the configuration of **B′**, **C′**, $\mathbf{B}'+\mathbf{C}'$, but is $\|\mathbf{A}\|$ times as large and is turned through an angle of 90° in the plane of the page (see Fig. 78). Necessarily, then, $\mathbf{A}\times(\mathbf{B}'+\mathbf{C}')$ will be the diagonal of the parallelogram with adjacent sides $\mathbf{A}\times\mathbf{B}'$, $\mathbf{A}\times\mathbf{C}'$, because $\mathbf{B}'+\mathbf{C}'$ is the diagonal of the parallelogram with adjacent sides **B′**, **C′**. Therefore (10.2–3) is true.

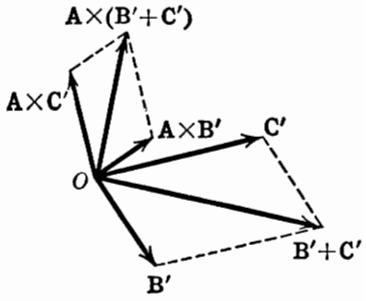


Fig. 78.

A second distributive law

$$(\mathbf{A}+\mathbf{B})\times\mathbf{C}=\mathbf{A}\times\mathbf{C}+\mathbf{B}\times\mathbf{C}$$

follows at once from (10.2–3) and (10.2–1).

Using the distributive law, we may obtain expressions for the components of $\mathbf{A}\times\mathbf{B}$ in terms of the components of the vectors **A** and **B**. We have

$$\mathbf{A}\times\mathbf{B}=(A_1\mathbf{i}+A_2\mathbf{j}+A_3\mathbf{k})\times(B_1\mathbf{i}+B_2\mathbf{j}+B_3\mathbf{k}).$$

We expand this by repeated use of the distributive law, and use (10.2–2) and (10.2–3). Nine terms are obtained in this way. To simplify the result we observe the following multiplication table, which is easily worked out directly from the definitions.

$$\begin{array}{lll}\mathbf{i}\times\mathbf{i}=\mathbf{0} & \mathbf{i}\times\mathbf{j}=\mathbf{k} & \mathbf{i}\times\mathbf{k}=-\mathbf{j}\\ \mathbf{j}\times\mathbf{i}=-\mathbf{k} & \mathbf{j}\times\mathbf{j}=\mathbf{0} & \mathbf{j}\times\mathbf{k}=\mathbf{i}\\ \mathbf{k}\times\mathbf{i}=\mathbf{j} & \mathbf{k}\times\mathbf{j}=-\mathbf{i} & \mathbf{k}\times\mathbf{k}=\mathbf{0}\end{array}$$

As a consequence, we find

$$\mathbf{A}\times\mathbf{B}=(A_2B_3-A_3B_2)\mathbf{i}+(A_3B_1-A_1B_3)\mathbf{j}+(A_1B_2-A_2B_1)\mathbf{k}. \qquad (10.2\text{–}6)$$

It is convenient to remember this formula by noting that if we write $\mathbf{C}=\mathbf{A}\times\mathbf{B}$, then

$$C_1=\begin{vmatrix}A_2 & A_3\\ B_2 & B_3\end{vmatrix},\ C_2=\begin{vmatrix}A_3 & A_1\\ B_3 & B_1\end{vmatrix},\ C_3=\begin{vmatrix}A_1 & A_2\\ B_1 & B_2\end{vmatrix}.$$

These two-row determinants are the cofactors of the elements of the first row of a three-row determinant having A_1, A_2, A_3, and B_1, B_2, B_3 as its second and third

rows, respectively. Accordingly, as a memory device, we sometimes write

$$\mathbf{A}\times\mathbf{B}=\begin{vmatrix}\mathbf{i} & \mathbf{j} & \mathbf{k}\\ A_1 & A_2 & A_3\\ B_1 & B_2 & B_3\end{vmatrix}. \tag{10.2–7}$$

EXERCISES

1. Find the indicated cross products.

(a) $(\mathbf{i}+\mathbf{j}+\mathbf{k})\times(\mathbf{i}+2\mathbf{j}+3\mathbf{k})$;

(b) $(2\mathbf{i}-3\mathbf{j}-\mathbf{k})\times(2\mathbf{i}-5\mathbf{j}+3\mathbf{k})$;

(c) $(\mathbf{i}-2\mathbf{j}-\mathbf{k})\times(\mathbf{i}-3\mathbf{j}+4\mathbf{k})$.

2. Find the area of the parallelogram of which the vectors $\mathbf{A}=\mathbf{i}-\mathbf{j}+2\mathbf{k}$ and $\mathbf{B}=2\mathbf{i}+4\mathbf{j}-\mathbf{k}$ are adjacent sides.

3. **(a)** Let **A**, **B**, **C** be noncoplanar vectors from O with terminal points P, Q, R respectively. Explain why $\frac{1}{2}(\mathbf{B}-\mathbf{A})\times(\mathbf{C}-\mathbf{A})$ is a vector perpendicular to the plane of PQR, and of length equal to the area of the triangle PQR. **(b)** Find the area of the triangle formed by the points $(1, 1, -2)$, $(2, -1, 1)$, $(1, 3, -1)$.

4. Find $\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C})$ and $\mathbf{B}\cdot(\mathbf{A}\times\mathbf{C})$ if

(a) $\mathbf{A}=2\mathbf{i}-3\mathbf{j}+5\mathbf{k}$, $\mathbf{B}=-\mathbf{i}+4\mathbf{j}+2\mathbf{k}$, $\mathbf{C}=2\mathbf{i}+3\mathbf{j}$;

(b) $\mathbf{A}=2\mathbf{i}+3\mathbf{j}+\mathbf{k}$, $\mathbf{B}=\mathbf{i}+2\mathbf{j}+5\mathbf{k}$, $\mathbf{C}=-2\mathbf{i}+4\mathbf{j}+3\mathbf{k}$.

5. **(a)** If $\mathbf{A}=(A_1, A_2, A_3)$, etc., show that

$$\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C})=\begin{vmatrix}A_1 & A_2 & A_3\\ B_1 & B_2 & B_3\\ C_1 & C_2 & C_3\end{vmatrix}.$$

(b) How does it follow from **(a)** that $\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C})=(\mathbf{A}\times\mathbf{B})\cdot\mathbf{C}$?

(c) What is the value of $\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C})$ if any two of the vectors are equal?

(d) Explain why the numerical value of the determinant in **(a)** is equal to the volume of the parallelepiped having the vectors **A**, **B**, **C** as concurrent edges.

(e) If **A**, **B**, **C** are permuted in all possible ways in the product $\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C})$, how many different values can be obtained?

(f) Find the values of $\mathbf{D}\cdot(\mathbf{B}-\mathbf{A})$ and $\mathbf{D}\cdot(\mathbf{C}-\mathbf{A})$, where

$$\mathbf{D}=\mathbf{A}\times\mathbf{B}+\mathbf{B}\times\mathbf{C}+\mathbf{C}\times\mathbf{A}.$$

6. Let the pairs, **A**, **B** and **C**, **D** each determine a plane. Write an equation involving dot and cross products expressing the condition and these two planes be perpendicular.

10.3 / RIGID MOTIONS OF THE AXES

By a *rigid motion* of the axes we mean a shift from a rectangular co-ordinate system xyz with origin O to another rectangular co-ordinate system $x'y'z'$ with origin O', both systems having the same unit of distance, and both systems having the same orientation (i.e., both being right-handed or both left-handed). Such a shift can be accomplished in two stages: by a translation to a new system with origin O' and axes parallel to the original axes, followed by a rotation about O'. The equations for a translation of the co-ordinate system are very simple, and need not concern us right now, since we regard the vector space with origin

O' as not essentially different from the vector space with origin O. Accordingly we confine our attention to a rotation of the co-ordinate axes about a fixed origin O. We deal exclusively with right-handed systems.

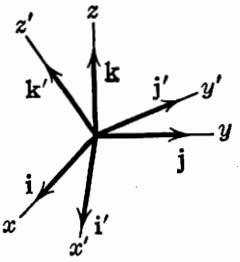


Fig. 79.

Let $\mathbf{i}'$, $\mathbf{j}'$, $\mathbf{k}'$ be the fundamental orthonormal triad associated with the $x'y'z'$ system (see Fig. 79). We recall from analytic geometry the concept of the direction cosines of a directed line. Let the direction cosines of the x'-axis relative to the x-, y-, and z-axes respectively be l_1, m_1, n_1. By the property (10.1–19) of dot products we have

$$\mathbf{i}' \cdot \mathbf{i} = l_1, \qquad \mathbf{i}' \cdot \mathbf{j} = m_1, \qquad \mathbf{i}' \cdot \mathbf{k} = n_1.$$

Likewise

$$\mathbf{j}' \cdot \mathbf{i} = l_2, \qquad \mathbf{j}' \cdot \mathbf{j} = m_2, \qquad \mathbf{j}' \cdot \mathbf{k} = n_2$$

are the direction cosines of the y'-axis relative to the axes of the xyz-system. With similar relations for the direction cosines of the z'-axis relative to the axes of the xyz-system, all the relations can be conveniently and compactly exhibited in a table, as follows.

	$\mathbf{i}$	$\mathbf{j}$	$\mathbf{k}$
$\mathbf{i}'$	l_1	m_1	n_1
$\mathbf{j}'$	l_2	m_2	n_2
$\mathbf{k}'$	l_3	m_3	n_3

(10.3–1)

Observe that the direction cosines of the x-axis relative to the axes of the $x'y'z'$-system are l_1, l_2. l_3 and so forth.

Now consider the vector $\overrightarrow{OP}$, where P is any point, with co-ordinates (x, y, z) and (x', y', z'), respectively, in the two systems. Then

$$\overrightarrow{OP} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k} \tag{10.3–2}$$

and

$$\overrightarrow{OP} = x'i' + y'j' + z'k', \tag{10.3–3}$$

Now observe from (10.3–3) that $x' = \overrightarrow{OP} \cdot \mathbf{i}'$, $y' = \overrightarrow{OP} \cdot \mathbf{j}'$, and so on. If we form the dot product $\overrightarrow{OP} \cdot \mathbf{i}'$ from (10.3–2), we obtain

$$x' = \overrightarrow{OP} \cdot \mathbf{i}' = x\mathbf{i} \cdot \mathbf{i}' + y\mathbf{j} \cdot \mathbf{i}' + z\mathbf{k} \cdot \mathbf{i}',$$

or

$$x' = l_1 x + m_1 y + n_1 z.$$

By similar arguments we obtain the complete set of relations

$$\begin{aligned} x' &= l_1 x + m_1 y + n_1 z \\ y' &= l_2 x + m_2 y + n_2 z \\ z' &= l_3 x + m_3 y + n_3 z. \end{aligned} \tag{10.3–4}$$

These are the equations of transformation for the rotation of the co-ordinate system. The inverse transformation can be found in exactly the same way, starting from (10.3–3). The equations are

$$\begin{aligned} x &= l_1x' + l_2y' + l_3z' \\ y &= m_1x' + m_2y' + m_3z' \\ z &= n_1x' + n_2y' + n_3z'. \end{aligned} \tag{10.3–5}$$

Consider now a vector **A** of the vector space with origin O. This vector will have components A_1, A_2, A_3 in the xyz-system, and A_1', A_2', A_3' in the $x'y'z'$-system. Since the components of **A** are merely the co-ordinates of the terminal point of **A**, we see that the two sets of components are related in exactly the same way that xyz and $x'y'z'$ are related, that is,

$$\begin{aligned} A_1' &= l_1A_1 + m_1A_2 + n_1A_3 \\ A_2' &= l_2A_1 + m_2A_2 + n_2A_3 \\ A_3' &= l_3A_1 + m_3A_2 + n_3A_3, \end{aligned} \tag{10.3–6}$$

with an inverse set of relations corresponding to (10.3–5). The two sets of relations are easily kept in mind by a table similar to (10.3–1):

	A_1	A_2	A_3
A_1'	l_1	m_1	n_1
A_2'	l_2	m_2	n_2
A_3'	l_3	m_3	n_3

It follows from these laws of transformation of components that if we know the components of a vector in one co-ordinate system, we can find its components in any system obtained by a rotation of the axes.

EXERCISES

1. Show that

$$\mathbf{i}' = l_1\mathbf{i} + m_1\mathbf{j} + n_1\mathbf{k}, \quad \mathbf{i} = l_1\mathbf{i}' + l_2\mathbf{j}' + l_3\mathbf{k}',$$

and obtain four other allied relations. Start from the fact that, if **A** is any vector, $\mathbf{A} = (\mathbf{A}\cdot\mathbf{i})\mathbf{i} + (\mathbf{A}\cdot\mathbf{j})\mathbf{j} + (\mathbf{A}\cdot\mathbf{k})\mathbf{k}$.

2. What is the numerical value of $\mathbf{i}\cdot(\mathbf{j}\times\mathbf{k})$? Express this product in terms of $\mathbf{i}'$, $\mathbf{j}'$, $\mathbf{k}'$ by the results of Exercise 1, and deduce that

$$\begin{vmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \\ n_1 & n_2 & n_3 \end{vmatrix} = 1.$$

See Exercise 5a, §10.2.

3. Observe that if one solves (10.3–5) for x' by Cramer's rule, and uses the result of Exercise 2, one finds

$$x' = (m_2n_3 - m_3n_2)x + (n_2l_3 - n_3l_2)y + (l_2m_3 - l_3m_2)z.$$

Comparing with (10.3–4), we surmise that

$$l_1 = \begin{vmatrix} m_2 & m_3 \\ n_2 & n_3 \end{vmatrix}, m_1 = \begin{vmatrix} n_2 & n_3 \\ l_2 & l_3 \end{vmatrix}, n_1 = \begin{vmatrix} l_2 & l_3 \\ m_2 & m_3 \end{vmatrix}.$$

Prove these results directly by considering the cross product $\mathbf{i}' = \mathbf{j}' \times \mathbf{k}'$. What similar results are given as a consequence of the relation $\mathbf{i} = \mathbf{j} \times \mathbf{k}$? Show that all the relations of this type are summarized in the statement: *Any element of the determinant appearing in Exercise 2 is equal to its own cofactor.*

4. Of what dot-product relation is the equation $l_1^2 + m_1^2 + n_1^2 = 1$ the expression? Answer the corresponding question for each of the equations

$$l_1 l_2 + m_1 m_2 + n_1 n_2 = 0, l_1 m_1 + l_2 m_2 + l_3 m_3 = 0.$$

5. Suppose the nine direction cosines in the table (10.3–1) are as follows:

$\frac{2}{7}$	$\frac{3}{7}$	$\frac{6}{7}$
$\frac{3}{7}$	$-\frac{6}{7}$	$\frac{2}{7}$
$\frac{6}{7}$	$\frac{2}{7}$	$-\frac{3}{7}$

(a) Find the components in the $x'y'z'$-system of the vector $14\mathbf{i} - 21\mathbf{j} + 7\mathbf{k}$.

(b) Find the components in the xyz-system of the vector $7\mathbf{i}' + 28\mathbf{j}' - 35\mathbf{k}'$.

6. Suppose the nine direction cosines in the table (10.3–1) are as follows:

$\frac{1}{\sqrt{6}}$	$\frac{-1}{\sqrt{2}}$	$\frac{1}{\sqrt{3}}$
$\frac{1}{\sqrt{6}}$	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{3}}$
$\frac{-2}{\sqrt{6}}$	0	$\frac{1}{\sqrt{3}}$

(a) Find $\sqrt{2}(\mathbf{i} + \mathbf{j}) - \mathbf{k}$ in terms of $\mathbf{i}'$, $\mathbf{j}'$, $\mathbf{k}'$.

(b) Find $\mathbf{i}' - \mathbf{j}' + \sqrt{6}\mathbf{k}'$ in terms of $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$.

10.4 / INVARIANTS

The concept of invariance and of something that is invariant is most simply illustrated (for our purposes) by showing how the expression for the dot product of two vectors in $\mathbb{R}^3$ (as defined in (10.1–12)) is represented in terms of the components of the vectors with respect to an $x'y'z'$-co-ordinate system obtained by a rotation from the xyz-co-ordinate system used in $\mathbb{R}^3$. The relationship

between the components of vectors in the two systems is explained in §10.3. The fact that the components of **A** along the axes of the rotated system are A_1', A_2', A_3' enables us to write

$$\mathbf{A} = A_1'\mathbf{i}' + A_2'\mathbf{j}' + A_3'\mathbf{k}'.$$

There is a similar representation of **B**. Therefore

$$\mathbf{A} \cdot \mathbf{B} = (A_1'\mathbf{i}' + A_2'\mathbf{j}' + A_3'\mathbf{k}') \cdot (B_1'\mathbf{i}' + B_2'\mathbf{j}' + B_3\mathbf{k}').$$

We can calculate this dot product by the algebraic rules governing dot products. Because of the fact that the vectors $\mathbf{i}'$, $\mathbf{j}'$, $\mathbf{k}'$ form an orthonormal system we have relationships such as $\mathbf{i}' \cdot \mathbf{i}' = 1$, $\mathbf{i}' \cdot \mathbf{j}' = 0$, and so on. When we complete the calculations we find that

$$\mathbf{A} \cdot \mathbf{B} = A_1'B_1' + A_2'B_2' + A_3'B_3'. \tag{10.4–1}$$

From the definition of $\mathbf{A} \cdot \mathbf{B}$ in (10.1–12) we now see that

$$A_1B_1 + A_2B_2 + A_3B_3 = A_1'B_1' + A_2'B_2' + A_3'B_3'. \tag{10.4–2}$$

Thus we see that the formula for $\mathbf{A} \cdot \mathbf{B}$ is the same in terms of the primed components as it is in terms of the unprimed components. That is what we mean when we say that the expression on the left in (10.4–2) is *an invariant*, or *is invariant* with respect to a rotation of the co-ordinate axes. We could anticipate this result, of course, because of formula (10.1–19), which expresses $\mathbf{A} \cdot \mathbf{B}$ in geometric terms, using the length of the vectors and the angle between them, which we are already accustomed to think of as being independent of the choice of a particular co-ordinate system.

Another example of invariance is provided by the cross product of two vectors. By this we mean that the cross product $\mathbf{A} \times \mathbf{B}$ can be expressed in the form

$$\mathbf{A} \times \mathbf{B} = (A_2'B_3' - A_3'B_2')\mathbf{i}' + (A_3'B_1' - A_1'B_3')\mathbf{j}' + (A_1'B_2' - A_2'B_1')\mathbf{k}', \tag{10.4–3}$$

which has the same form as (10.2–6) except that everything (components and the set of orthonormal vectors) is referred to the $x'y'z'$-co-ordinate axes. The procedure used to derive (10.2–6) started from a geometrical definition of $\mathbf{A} \times \mathbf{B}$, without the explicit involvement of a co-ordinate system. The derivation of (10.2–6) depended on expressing **A** and **B** as linear combinations of **i**, **j** and **k** and then making use of the algebraic rules (10.2–1), (10.2–2), (10.2–3) and the nine special formulas for the various cross products such as $\mathbf{i} \times \mathbf{i}$, $\mathbf{i} \times \mathbf{j}$, $\mathbf{i} \times \mathbf{k}$, . . ., as exhibited in §10.2. The same method will lead us to the formula (10.4–3) if we express **A** and **B** as linear combinations of $\mathbf{i}'$, $\mathbf{j}'$, and $\mathbf{k}'$, and if we make use of the special formulas for the various cross products such as

$$\mathbf{i}' \times \mathbf{i}' = \mathbf{0}, \qquad \mathbf{i}' \times \mathbf{j}' = \mathbf{k}', \qquad \mathbf{i}' \times \mathbf{k}' = -\mathbf{j}'.$$

and so on. The correctness of these formulas is apparent from the geometric definition of cross products.

We could have proceeded entirely differently, *without* a geometric definition of $\mathbf{A} \times \mathbf{B}$. Let us sketch out this alternate procedure, starting with a definition of $\mathbf{A} \times \mathbf{B}$, entirely analytically, in $\mathbb{R}^3$. This new definition, in terms of components, is:

$$\mathbf{A} \times \mathbf{B} = (A_2B_3 - A_3B_2, A_3B_1 - A_1B_3, A_1B_2 - A_2B_1). \tag{10.4–4}$$

Next, we verify that the algebraic rules

$$\mathbf{A} \times \mathbf{B} = -(\mathbf{B} \times \mathbf{A}), (c\mathbf{A}) \times \mathbf{B} = c(\mathbf{A} \times \mathbf{B}), \tag{10.4–5}$$

and

$$\mathbf{A} \times (\mathbf{B} + \mathbf{C}) = (\mathbf{A} \times \mathbf{B}) + (\mathbf{A} \times \mathbf{C}) \tag{10.4–6}$$

are valid, using only (10.4–4) and the algebraic rules in §10.1. This verification is entirely straightforward; we shall not give the details.

From (10.4–4) and the rules for evaluating a determinant we readily find that

$$(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C} = \begin{vmatrix} A_1 & A_2 & A_3 \\ B_1 & B_2 & B_3 \\ C_1 & C_2 & C_3 \end{vmatrix}. \tag{10.4–7}$$

We verify this by expanding the determinant by minors of elements in the third row. The components of $\mathbf{A} \times \mathbf{B}$, as given in (10.4–4), are the cofactors of C_1, C_2, C_3, respectively, in (10.4–7). Because a determinant is equal to zero if any row in it is the same as, or a multiple of, another row, it is apparent from (10.4–7) that $(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C} = 0$ if two of the three vectors are the same of if one of them is a multiple of another.

The multiplication table for the various cross products among the vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ is easily worked out, using the definition (10.4–4). These results are listed in §10.2 (where they were obtained by use of the former method of definition).

We turn now to consider the effect of the introduction of a new $x'y'z'$-co-ordinate system, related to the basic xyz-system in $\mathbb{R}^3$ by rotating the axes. Our aim is to show that, if $\mathbf{A}$ and $\mathbf{B}$ have components A_1', A_2', A_3' and B_1', B_2', B_3' with respect to the $x'y'z'$-system, then $\mathbf{A} \times \mathbf{B}$ has the components

$$A_2'B_3' - A_3'B_2', A_3'B_1' - A_1'B_3', A_1'B_2' - A_2'B_1'$$

in the new system. If new unit vectors $\mathbf{i}'$, $\mathbf{j}'$, $\mathbf{k}'$ (along the positive x', y', and z' co-ordinate axes, respectively) are introduced, this means that

$$\begin{aligned}(A_2B_3 - A_3B_2)\mathbf{i} + (A_3B_1 - A_1B_3)\mathbf{j} + (A_1B_2 - A_2B_1)\mathbf{k} = \\ (A_2'B_3' - A_3'B_2')\mathbf{i}' + (A_3'B_1' - A_1'B_3')\mathbf{j}' + (A_1'B_2' - A_2'B_1')\mathbf{k},\end{aligned} \tag{10.4–8}$$

so that the cross product $\mathbf{A} \times \mathbf{B}$, even though defined by (10.4–4) in terms of one preferred co-ordinate system, is actually a *vector invariant* with respect to rotations of the co-ordinate axes. To prove this we shall first of all need to know some facts about $\mathbf{i}'$, $\mathbf{j}'$, $\mathbf{k}'$ and their relationships to $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$. We can express $\mathbf{i}'$, $\mathbf{j}'$, $\mathbf{k}'$

as linear combinations of $\mathbf{i}, \mathbf{j}, \mathbf{k}$ as follows:

$$\begin{aligned} \mathbf{i}' &= l_1\mathbf{i} + m_1\mathbf{j} + n_1\mathbf{k} \\ \mathbf{j}' &= l_2\mathbf{i} + m_2\mathbf{j} + n_2\mathbf{k} \\ \mathbf{k}' &= l_3\mathbf{i} + m_3\mathbf{j} + n_3\mathbf{k}. \end{aligned} \tag{10.4–9}$$

Also, because $\mathbf{i}'$, $\mathbf{j}'$, and $\mathbf{k}'$ are linearly independent and $\mathbb{R}^3$ is three dimensional, any vector $\mathbf{A}$ in $\mathbb{R}^3$ is a linear combination of $\mathbf{i}'$, $\mathbf{j}'$ and $\mathbf{k}'$, the coefficient of $\mathbf{i}'$ being $\mathbf{A}\cdot\mathbf{i}'$, and so on. In this way we can see with the aid of (10.4–9) that

$$\begin{aligned} \mathbf{i} &= l_1\mathbf{i}' + l_2\mathbf{j}' + l_3\mathbf{k}' \\ \mathbf{j} &= m_1\mathbf{i}' + m_2\mathbf{j}' + m_3\mathbf{k}' \\ \mathbf{k} &= n_1\mathbf{i}' + n_2\mathbf{j}' + n_3\mathbf{k}'. \end{aligned} \tag{10.4–10}$$

For example, the coefficient of $\mathbf{j}'$ in the expression for $\mathbf{k}$ is $\mathbf{k}\cdot\mathbf{j}' = n_2$, as we can see from the second equation in (10.4–9). There are various relationships between the elements in the determinant

$$D = \begin{vmatrix} l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \\ l_3 & m_3 & n_3 \end{vmatrix}. \tag{10.4–11}$$

For example, $l_1^2 + l_2^2 + l_3^2 = 1$ is the expression of the fact that $\|\mathbf{i}\| = 1$, and $l_1l_2 + m_1m_2 + n_1n_2 = 0$ is the expression of the fact that $\mathbf{i}'\cdot\mathbf{j}' = 0$.

Now let us find the values of $\mathbf{i}'\times\mathbf{i}'$, $\mathbf{i}'\times\mathbf{j}'$, $\mathbf{i}'\times\mathbf{k}'$, and so on. It is clear from the very definition (10.4–4) that $\mathbf{i}'\times\mathbf{i}' = \mathbf{j}'\times\mathbf{j}' = \mathbf{k}'\times\mathbf{k}' = 0$, so we can concentrate on $\mathbf{i}'\times\mathbf{j}'$, $\mathbf{j}'\times\mathbf{k}'$, $\mathbf{k}'\times\mathbf{i}'$ and then obtain $\mathbf{j}'\times\mathbf{i}'$, etc., by using the first result in (10.4–5). We observe that the x' and y' components of $\mathbf{i}'\times\mathbf{j}'$ are $(\mathbf{i}'\times\mathbf{j}')\cdot\mathbf{i}'$ and $(\mathbf{i}'\times\mathbf{j}')\cdot\mathbf{j}'$, both of which are zero because, as we saw from (10.4–7), $(\mathbf{A}\times\mathbf{B})\cdot\mathbf{C} = 0$ if any two of the three vectors are the same. The z' component of $\mathbf{i}'\times\mathbf{j}'$ is $(\mathbf{i}'\times\mathbf{j}')\cdot\mathbf{k}'$, and this is D, as we see from (10.4–7), (10.4–9), and (10.4–11). Thus $\mathbf{i}'\times\mathbf{j}' = D\mathbf{k}'$. We can find $\mathbf{j}'\times\mathbf{k}'$ and $\mathbf{k}'\times\mathbf{i}'$ by the same method. We display the results:

$$\begin{aligned} \mathbf{i}'\times\mathbf{j}' &= D\mathbf{k}' \\ \mathbf{j}'\times\mathbf{k}' &= D\mathbf{i}'. \\ \mathbf{k}'\times\mathbf{i}' &= D\mathbf{j}' \end{aligned} \tag{10.4–12}$$

We don't yet know the value of D, but we will find it. Let us calculate $\mathbf{j}\times\mathbf{k}$, using the second and third formulas in (10.4–10):

$$\mathbf{j}\times\mathbf{k} = (m_1\mathbf{i}' + m_2\mathbf{j}' + m_3\mathbf{k}')\times(n_1\mathbf{i}' + n_2\mathbf{j}' + n_3\mathbf{k}').$$

When this is worked out using the rules (10.4–5) and (10.4–6) and the results (10.4–12), we find

$$\mathbf{j}\times\mathbf{k} = (m_2n_3 - m_3n_2)D\mathbf{i}' + (m_3n_1 - m_1n_3)\,D\mathbf{j}' + (m_1n_2 - m_2n_1)\,D\mathbf{k}'.$$

But $\mathbf{j} \times \mathbf{k} = \mathbf{i} = l_1\mathbf{i}' + l_2\mathbf{j}' + l_3\mathbf{k}'$, and therefore

$$l_1 = D(m_2n_3 - m_3n_2)$$
$$l_2 = D(m_3n_1 - m_1n_3)$$
$$l_3 = D(m_1n_2 - m_2n_1).$$

If we multiply these equations by l_1, l_2, l_3 respectively, and add, we obtain

$$1 = l_1^2 + l_2^2 + l_3^2 = D^2,$$

because, as we see from (10.4–11),

$$l_1(m_2n_3 - m_3n_2) + l_2(m_3n_1 - m_1n_3) + l_3(m_1n_2 - m_2n_1) = D.$$

But then $D = \pm l$ and $\mathbf{i}' \times \mathbf{j}' = \pm\mathbf{k}'$. We must choose the value $D = 1$, rather than $D = -1$ because the $x'y'z'$-system, being obtained from the xyz-system by a rigid rotation, is right handed.

We can now complete the proof of (10.4–8). The left side is $\mathbf{A} \times \mathbf{B}$, which we can write as

$$(A_1'\mathbf{i}' + A_2'\mathbf{j}' + A_3'\mathbf{k}') \times (B_1'\mathbf{i}' + B_2'\mathbf{j}' + B_3'\mathbf{k}').$$

When this cross product is expanded and the results simplified, using (10.4–12) (with $D = 1$) and the other cross-product relationships involving $\mathbf{i}'$, $\mathbf{j}'$, $\mathbf{k}'$, we obtain the expression on the right side of the equation in (10.4–8).

In the rest of this chapter we often talk about a point of Euclidean space as an entity that has an existence quite apart from any co-ordinate system. We can use the co-ordinates (x, y, z) of a point P in some rectangular co-ordinate system, but we make a distinction between the point P and the number triple (x, y, z). This is a different point of view than that in §10.1, where the number triple *was* the point. From this new point of view Euclidean space of three dimensions is not the same as $\mathbb{R}^3$, in which there is a preferred point (the origin) and a preferred rectangular co-ordinate system. When we talk about the Euclidean *vector* space of three dimensions we have introduced the origin as a special point. The points are then the same as vectors, and, if we do not talk about components, we are talking in a co-ordinate-free mode. If we *do* introduce co-ordinates, which it is often convenient to do (perhaps only temporarily), we may nevertheless find it necessary to establish that certain things we are talking about are invariant with respect to specified changes in the co-ordinate system, such as rotation of the axes. It is to familiarize the student with the concept of invariance that we have engaged in this discussion of scalar and vector invariants.

The foregoing remarks are relevant to our discussion of *gradient*, *divergence*, and *curl*, in sections to come.

EXERCISES

1. Show that, if the co-ordinate system is rotated 90° about the z-axis, so that the x'-axis is the same as the y-axis and the y'-axis coincides with the negative x-axis, the

scheme (10.3–7) becomes

	A_1	A_2	A_3
A_1'	0	1	0
A_2'	−1	0	0
A_3'	0	0	1

Then, show by an example that, if **A** and **B** are vectors, A_1B_1 is not a scalar invariant. Likewise show that the triple (A_1B_1, A_2B_2, A_3B_3) does not define a vector invariant, that is, that in general the vector having components (A_1B_1, A_2B_2, A_3B_3) in the xyz-system does not have components $(A_1'B_1', A_2'B_2', A_3'B_3')$ in the $x'y'z'$-system. Consider, e.g., $\mathbf{A} = \mathbf{i}$, $\mathbf{B} = \mathbf{i} + \mathbf{j}$.

2. Is $A_1 + A_2 + A_3$ a scalar invariant if $\mathbf{A} = A_1\mathbf{i} + A_2\mathbf{j} + A_3\mathbf{k}$? Justify your answer.

3. Let the nine direction cosines in the table (10.3–1) be specified as follows:

$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	0
$\frac{1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$
$\frac{1}{\sqrt{6}}$	$\frac{-1}{\sqrt{6}}$	$\frac{-2}{\sqrt{6}}$

Let $\mathbf{A} = \mathbf{i} + \mathbf{j}$, $\mathbf{B} = \mathbf{i} - \mathbf{j} + \mathbf{k}$.

(a) Calculate $\mathbf{A} \times \mathbf{B}$ directly in terms of $\mathbf{i}, \mathbf{j}, \mathbf{k}$.

(b) Calculate **A** and **B** and then $\mathbf{A} \times \mathbf{B}$ in terms of $\mathbf{i}', \mathbf{j}', \mathbf{k}'$.

(c) Reconcile the results in **(a)** and **(b)** by converting $\mathbf{A} \times \mathbf{B}$ from **(b)** to the answer in **(a)** by expressing $\mathbf{i}', \mathbf{j}', \mathbf{k}'$ in terms of $\mathbf{i}, \mathbf{j}, \mathbf{k}$.

10.5 / SCALAR POINT FUNCTIONS

The word "scalar" is used to contrast with the word "vector." It is customary, in any context where vectors and real numbers are both being discussed, to refer to real numbers as scalars. Thus, in this book, a scalar is a real number. The word may also be used as an adjective.

Let us recall the general meaning of the word "function." A function is a correspondence between two classes of objects; these two classes are called respectively the *domain of definition* of the function and the *range of values* of the function, or, more briefly, the *domain* and the *range* of the function. The function itself is the correspondence whereby to each object in the domain is assigned a corresponding object in the range. Let us now consider a case in which the domain is a class of points and the range is a class of real numbers. In such a case we shall call the function a *scalar point function.* If f denotes the

function, P a variable point, and u the value of the function at P, we write $u = f(P)$.

Example 1. At a fixed instant of time consider all the points in the earth's atmosphere, and let u be the temperature (in degrees centigrade) at the point P.

Example 2. On the surface of a sphere of unit radius, let Q be a fixed point. If P is any point on the sphere, let u be the shortest great-circle distance from P to Q.

Example 3. If C is a curve in a plane, and P is any point of C, let u be the radius of curvature of C at P, provided the curve is smooth enough to have a well-defined curvature.

The essential thing to be noted about the definition of a scalar point function is that it is independent of co-ordinate systems. We do not need co-ordinate systems to explain the concept of a scalar point function. And, if a co-ordinate system is used to locate various points, the function is essentially the correspondence between P and u, *not* a correspondence between the co-ordinates of P and u. There is, however, a link between a point function $f(P)$ and certain functions of the co-ordinates of P. This link is the concept of invariance under co-ordinate transformations. Suppose, for instance, that f is a given point function, defined throughout some region of Euclidean three-dimensional space. Choose a system of rectangular co-ordinates x, y, z, and let (x, y, z) be co-ordinates of P. Then the value of f at P will depend on (x, y, z), say $f(P) = F(x, y, z)$, so that the scalar point function f determines F as a real-valued function of the three variables x, y, z. Suppose another rectangular co-ordinate system x', y', z' is obtained from the xyz-system by a rotation of axes. When x, y, z are expressed in terms of x', y', z' by the equations of the form (10.3–5), $F(x, y, z)$ is transformed into some new function $G(x', y', z')$. But, since (x, y, z) and (x', y', z') refer to the same point P, we have $f(P) = G(x', y', z')$. The *scalar point function f is invariant under the transformation of co-ordinates.* We may also express $f(P)$ as a function of cylindrical or spherical co-ordinates.

Example 4. If O is a fixed point of space, consider the vector space with origin O. If P is an arbitrary point, denote the vector $\overrightarrow{OP}$ by $\mathbf{R}$. Let $\mathbf{A}$ be a fixed vector. Then

$$f(P) = \mathbf{A} \cdot \mathbf{R} \quad \text{and} \quad \phi(P) = \|\mathbf{R}\|^2$$

are scalar point functions. If we introduce a fixed rectangular co-ordinate system with O as origin, and let $\mathbf{A} = (A_1, A_2, A_3)$, $\mathbf{R} = (x, y, z)$, then in this co-ordinate system we have

$$f(P) = A_1x + A_2y + A_3z,$$
$$\phi(P) = (x^2 + y^2 + z^2).$$

It is possible to define directly the concepts of continuity and differen-

tiability for scalar point functions, without reference to co-ordinate systems. Alternatively, however, the definitions may be made with reference to some arbitrarily chosen rectangular co-ordinate system. Thus, if $f(P) = F(x, y, z)$ in that system, we can say that f is continuous at P_0 if F is continuous at (x_0, y_0, z_0), with a similar definition for differentiability. Although these definitions are made with reference to a particular rectangular co-ordinate system, they are actually independent of the choice of that system. For example, if $F(x, y, z) = G(x', y', z')$, where the two systems are related by a rotation, and if F is differentiable, then G is differentiable also (by Theorem V, §7.3).

10.51 / VECTOR POINT FUNCTIONS

The concept of a vector point function is similar to the concept of a scalar point function in the matter of being independent of particular choices of co-ordinate systems. The difference is that the function values are vectors instead of scalars. The domain of definition of a vector point function is some set of points P. The range of values of the function is some set of vectors. Let f denote the function, and let $\mathbf{F}$ denote the vector corresponding to P. Then we write $\mathbf{F} = f(P)$ where $f(P)$ depends just on P itself and not on the co-ordinates we happen to be using.

Example 1. Let $\mathbf{R} = \overrightarrow{OP}$, and let $\mathbf{A}$ be a fixed vector. Then each of the expressions

$$\mathbf{R}, \qquad (\mathbf{A} \cdot \mathbf{R})\mathbf{R}, \qquad \mathbf{A} \times \mathbf{R}, \qquad \|\mathbf{R}\|^{-3}\mathbf{R}$$

defines a vector point function.

It is often convenient to introduce a co-ordinate system in order to deal with a vector point function. We shall be concerned mostly with rectangular co-ordinate systems. If we have an xyz-system with origin O, let $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ be the fundamental orthonormal triad associated with the xyz-system. Suppose we have a function $\mathbf{F} = f(P)$; let the components of $\mathbf{F}$ be denoted by $F_1(x, y, z)$, $F_2(x, y, z)$, $F_3(x, y, z)$. Then

$$\mathbf{F} = F_1(x, y, z)\mathbf{i} + F_2(x, y, z)\mathbf{j} + F_3(x, y, z)\mathbf{k}. \tag{10.51–1}$$

In another rectangular system $x'y'z'$, obtained from the xyz-system by rotation, $\mathbf{F}$ will have a different set of components, and with the triad $\mathbf{i}'$, $\mathbf{j}'$, $\mathbf{k}'$ for the new system, $\mathbf{F}$ will be expressed in the form

$$\mathbf{F} = F_1'(x', y', z')\mathbf{i}' + F_2'(x', y', z')\mathbf{j}' + F_3'(x', y', z')\mathbf{k}'. \tag{10.51–2}$$

The components F_1', F_2', F_3' will be related to F_1, F_2, F_3 in the same way that A_1', A_2', A_3' are related to A_1, A_2, A_3 in equations (10.3–6). Also, the two orthonormal triads are related in the manner indicated by table (10.3–1).

The value $\mathbf{F}$ of the vector point function is a vector invariant. Note, however, that an individual component of $\mathbf{F}$, such as $F_1(x, y, z)$, is not a scalar invariant, for in general $F_1(x, y, z) \neq F_1'(x', y', z')$.

If F_1, F_2, F_3 are any three functions of x, y, z, we may regard them as components of a vector point function $\mathbf{F}$ defined by (10.51–1), provided we understand that F_1, F_2, F_3 are the components of $\mathbf{F}$ in one fixed rectangular co-ordinate system, and that the components in any other rectangular system are obtained by transformations of the type (10.3–6).

A vector point function is often called a *vector field*, especially if the domain of definition is a region in two or three dimensions. This terminology springs from a certain mode of representing the vector function. Let the vector $\mathbf{F}$ be drawn in space with its initial point at the point P to which $\mathbf{F}$ corresponds. This allows one to portray the manner in which the magnitude and direction of $\mathbf{F}$ vary with the location of P. The geometrical configuration, whether actually drawn or merely conceived, is called a vector field (see Fig. 80).

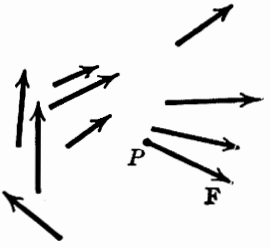


Fig. 80.

Example 2. Consider the gravitational field of the earth. If O is the center of the earth, M is the mass of the earth, and r is the distance from O to an arbitrary point P outside the earth, then a particle of unit mass placed at P is attracted toward O by a force of magnitude

$$k\frac{M}{r^2}, \tag{10.51–3}$$

where k is a constant of proportionality. This force can be represented by a vector of magnitude given by (10.51–3), and of direction opposite to that of the vector $\mathbf{R} = \overrightarrow{OP}$. The magnitude of $\mathbf{R}$ is r. Therefore a vector of unit length in the direction opposite to that of $\mathbf{R}$ is

$$-\frac{1}{r}\mathbf{R}.$$

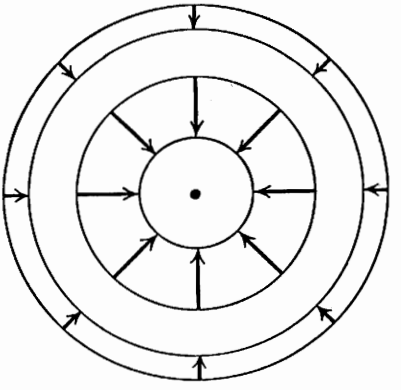


Fig. 81.

Thus the gravitational force $\mathbf{F}$ on the unit mass at P is

$$\mathbf{F} = -k\frac{M}{r^3}\mathbf{R}. \tag{10.51–4}$$

A portrayal of this vector field is suggested by Fig. 81.

EXERCISES

1. The equations

$$x = \frac{1}{\sqrt{2}}(x' - y'),$$

$$y = \tfrac{1}{2}(x' + y') + \frac{1}{\sqrt{2}}z',$$

$$z = -\tfrac{1}{2}(x' + y') + \frac{1}{\sqrt{2}}z'$$

define a rotation of axes.

(a) If $F(x, y, z) = 2x^2 - y^2 - z^2$ is the representation of a scalar point function in the xyz-system, find the representation $G(x', y', z')$ of the function in the $x'y'z'$-system.

(b) What is the table (10.3–7) for this rotation of axes? Find the xyz-representation $F(x, y, z)$ for the scalar point function for which $G(x', y', z') = x' + y' + \sqrt{2}z'$.

(c) Find the components of the vector field $\mathbf{F} = \sqrt{2}z\mathbf{i} + (y + z)\mathbf{j} + \sqrt{2}x\mathbf{k}$ in the $x'y'z'$-system.

(d) Express the vector field $\mathbf{F} = \sqrt{2}y'\mathbf{i}' + z'\mathbf{j}' + x'\mathbf{k}'$ in the xyz-system.

2. The equations

$$x' = \tfrac{1}{7}(2x + 3y + 6z),$$
$$y' = \tfrac{1}{7}(3x - 6y + 2z),$$
$$z' = \tfrac{1}{7}(6x + 2y - 3z)$$

define a rotation of axes.

(a) Express x, y, z in terms of x', y', and z'.

(b) Find the expression of $x'^2 + y'^2 - z'^2$ in the xyz-system.

(c) Express the vector field $-y\mathbf{i} + x\mathbf{j}$ in the $x'y'z'$-system.

(d) Do $x\mathbf{i} + 2y\mathbf{j} + 3z\mathbf{k}$ and $x'\mathbf{i}' + 2y'\mathbf{j}' + 3z'\mathbf{k}'$ represent the same vector field? Justify your answer.

10.6 / THE GRADIENT OF A SCALAR FIELD

A scalar point function f may be thought of by imagining each point P at which f is defined as carrying a label with the value $f(P)$ of the function at that point (see the discussion of the second mode of representing functions in §5.4). When a scalar function is represented in this way, it is often called a *scalar field.*

Let f be a scalar field defined throughout some region R, and suppose that f is differentiable in R. We are going to define the concept of the *gradient* of the field. The gradient of a scalar point function is a vector point function. It is

convenient to use a rectangular co-ordinate system in the process of defining the gradient. However, we must take care to be certain that the definition gives us a result which is independent of the choice of the particular rectangular co-ordinate system.

Let a system of rectangular co-ordinates xyz, with origin O, be selected arbitrarily.

In this co-ordinate system let P have co-ordinates (x, y, z), and let $f(P) = F(x, y, z)$ be the representation of our point function. Form the vector field whose representation in the xyz-system is

$$\frac{\partial F}{\partial x}\mathbf{i} + \frac{\partial F}{\partial y}\mathbf{j} + \frac{\partial F}{\partial z}\mathbf{k}. \tag{10.6–1}$$

This vector field, or vector point function, is what we shall call the gradient of the given scalar function f.

Presently we shall give an interpretation of the gradient which enables us to think of it apart from the co-ordinate system. But first let us show that the definition of the gradient yields the same vector field no matter what rectangular co-ordinate system is chosen. If a second co-ordinate system is selected, the two systems are related in such a way that one can be obtained from the other by either a translation or a rotation, or both. Let us consider the case where both systems have the same origin, and are related by a rotation of axes. This is the case of principal importance for our discussion of the gradient. The case where a translation may be involved is considered in Exercise 13.

Let the co-ordinates of the two systems be xyz and $x'y'z'$, related as in (10.3–4) and (10.3–5), and let the two representations of the scalar field be

$$f(P) = F(x, y, z) = G(x', y', z').$$

We wish to show that

$$\frac{\partial F}{\partial x}\mathbf{i} + \frac{\partial F}{\partial y}\mathbf{j} + \frac{\partial F}{\partial z}\mathbf{k} = \frac{\partial G}{\partial x'}\mathbf{i}' + \frac{\partial G}{\partial y'}\mathbf{j}' + \frac{\partial G}{\partial z'}\mathbf{k}'. \tag{10.6–2}$$

Once this is done, it will be clear that the definition of the gradient of f by expression (10.6–1) is invariant under a rotation of the axes.

To prove (10.6–2) we must show that the triple $\left(\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}, \frac{\partial F}{\partial z}\right)$ is related to the triple $\left(\frac{\partial G}{\partial x'}, \frac{\partial G}{\partial y'}, \frac{\partial G}{\partial z'}\right)$ just as the triple (A_1, A_2, A_3) is related to the triple (A_1', A_2', A_3') in (10.3–6) or (10.3–7). Now, by the chain rule,

$$\frac{\partial G}{\partial x'} = \frac{\partial F}{\partial x}\frac{\partial x}{\partial x'} + \frac{\partial F}{\partial y}\frac{\partial y}{\partial x'} + \frac{\partial F}{\partial z}\frac{\partial z}{\partial x'},$$

with similar equations for $\frac{\partial G}{\partial y'}$ and $\frac{\partial G}{\partial z'}$. From (10.3–5) we see that

$$\frac{\partial x}{\partial x'} = l_1, \frac{\partial y}{\partial x'} = m_1, \frac{\partial z}{\partial x'} = n_1.$$

Therefore

$$\frac{\partial G}{\partial x'} = l_1 \frac{\partial F}{\partial x} + m_1 \frac{\partial F}{\partial y} + n_1 \frac{\partial F}{\partial z}.$$

This is exactly like the first equation in (10.3–6). The other two relations of this type are established in the same way, and thus (10.6–2) is proved.

Frequently the value of a scalar function is denoted by a dependent variable, e.g., $u = f(P)$. In such cases the gradient is denoted by either of the symbols

$$\textbf{grad}\ u \quad \text{or} \quad \textbf{grad}\ f.$$

The symbolism ∇u or ∇f is also employed. Thus, in any rectangular co-ordinate system xyz, with corresponding orthonormal triad $\mathbf{i}, \mathbf{j}, \mathbf{k}$,

$$\textbf{grad}\ u = \nabla u = \frac{\partial u}{\partial x}\mathbf{i} + \frac{\partial u}{\partial y}\mathbf{j} + \frac{\partial u}{\partial z}\mathbf{k}. \tag{10.6–3}$$

Example 1. If

$$u = 2x^2 + 3y^2 + 4z^2,$$

$$\nabla u = 4x\mathbf{i} + 6y\mathbf{j} + 8z\mathbf{k}.$$

Let us now consider some properties of the gradient.

THEOREM I. *At any given point in the region where the differentiable scalar field $u = f(P)$ is defined let a direction be chosen. The rate of change of u (per unit distance) at the given point in the given direction is equal to the component of the gradient of u in that direction. If* **grad** $u \neq \mathbf{0}$ *at the point, the direction of the gradient is that in which the rate of increase of u is greatest.*

Proof. Let a rectangular co-ordinate system be chosen so that the given direction is that of the positive x-axis. The rate of change in the given direction is then $\partial u/\partial x$. But this is precisely the component of ∇u in the x-direction, by (10.6–3). The first assertion in the theorem is now justified. The correctness of the second assertion is an immediate consequence, since, of all components of a non-zero vector, the largest one is that in the direction of the vector itself.

The rate of change of a function at a given point in a given direction is called the *directional derivative.*

The importance of Theorem I is that it enables us to see the significance of the gradient without any reference to co-ordinate systems. A further grasp of the significance of the gradient is obtainable with the aid of the concept of level surfaces (see the latter part of §5.4). Let us assume that the gradient is a continuous vector function, i.e., that $\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial u}{\partial z}$ are continuous functions. Let P_0 be a point at which these three partial derivatives are not all zero; this is the same as requiring that $\nabla u \neq \mathbf{0}$ at P_0. If u_0 is the value of u at P_0, it can be shown by the implicit-function theorem (§8.1) that, at least in some sufficiently small neighborhood of P_0, the points at which $u = u_0$ form a surface having a tangent

plane whose normal varies continuously in direction as the point of tangency varies. This smooth surface is part of the level surface $u = u_0$.

THEOREM II. *At any point where the gradient is not the vector* **0**, *the gradient is perpendicular to the level surface through the point.*

Proof. Let the representation of the scalar field be $u = F(x, y, z)$. The equation of the level surface is $F(x, y, z) = u_0$. But we know (Example 1, §6.6) that the direction of the normal to the surface is given by the ratios $\frac{\partial F}{\partial x} : \frac{\partial F}{\partial y} : \frac{\partial F}{\partial z}$. These partial derivatives are precisely the components of the gradient, and so the theorem is proved.

Example 2. Consider the scalar field $u = r^3$, where r is the distance OP (O fixed, P variable). Clearly u is constant on a surface if and only if r is constant on that surface. Thus the level surfaces are spheres with center O. Evidently u increases when r increases, the rate of increase being

$$\frac{du}{dr} = 3r^2.$$

It follows from Theorem II that ∇u is normal to the sphere at P. From Theorem I we see that the direction of ∇u is outward from the sphere at P, and that its length is $3r^2$. Now a unit vector in the direction of OP is

$$\frac{1}{r}\overrightarrow{OP}.$$

Therefore

$$\nabla u = 3r^2\left(\frac{1}{r}\overrightarrow{OP}\right) = 3r\overrightarrow{OP}$$

in this case. This example illustrates how Theorems I and II sometimes furnish the means of calculating a gradient without resorting to rectangular co-ordinate systems. For a solution of this same problem in terms of rectangular co-ordinates see Exercise 5.

EXERCISES

1. Find ∇f in the case of each of the following functions.

(a) $f(P) = 2x^2 - 3xy + y^2 - 4xz + 6z^2$.

(b) $f(P) = x/(x^2 + y^2) + z/(y^2 + z^2)$.

(c) $f(P) = \log(x^2 + y^2) + z$.

(d) $f(P) = z/(x^2 + y^2 + z^2)^{3/2}$.

(e) $f(P) = e^{-x}(x^2 + y^2 + z^2)$.

2. In what direction from the point $(2, -1, 3)$ is the function $xy + yz + zx + xyz$ increasing most rapidly? Give direction cosines for the answer.

3. Find a vector indicating the direction and magnitude of most rapid *decrease* of the function $(x^2/16)+(y^2/25)-(z^2/9)$ at the point $(8, 25, -9)$. What is this rate of decrease?

4. **(a)** How fast is $4x^2+9y^2+z^2$ changing at the point $(1, 1, 1)$ in a direction tangent to the ellipsoid $4x^2+9y^2+z^2=14$? **(b)** What is the rate of change in the direction of the outward normal to the ellipsoid at the point?

5. Find ∇u by use of (10.6–3) if $u=(x^2+y^2+z^2)^{3/2}$. Compare with Example 2, noting that $r^2=x^2+y^2+z^2$ and $\overrightarrow{OP}=x\mathbf{i}+y\mathbf{j}+z\mathbf{k}$.

6. Find ∇u if $u=x^2$, using Theorems I and II rather than (10.6–3). Check by using the latter formula.

7. Let $\mathbf{R}$ be the vector $\overrightarrow{OP}$ and r the distance OP. Find the gradient of each of the following functions by the method of Example 2. Express the answers in terms of $\mathbf{R}$ and r.

(a) $u=1/r$; **(d)** $u=\log(1/r)$;

(b) $u=r^2$; **(e)** $u=r^n$;

(c) $u=1/\sqrt{r}$; **(f)** $u=e^{-r^2}$.

8. In the notation of Exercise 7 let $u=F(r)$, where F is a differentiable function of r. Show that $\nabla u=(1/r)F'(r)\mathbf{R}$: **(a)** by the method of Example 2; **(b)** by using (10.6–3) with $r^2=x^2+y^2+z^2$, $\mathbf{R}=x\mathbf{i}+y\mathbf{j}+z\mathbf{k}$. **(c)** Find the most general form for $F(r)$ if $\nabla u=-r^{-3}\mathbf{R}$.

9. Let r denote the distance from P to a fixed line in space. **(a)** If $f(P)$ is a function of r alone, what are the level surfaces of the scalar field? **(b)** If $f(P)=F(r)$, where F is differentiable, what can you say about the magnitude and direction of ∇f at P if P is not on the given line? **(c)** Let the fixed line be the z-axis. If P is the point (x, y, z), show that $\nabla f=[F'(r)/r](x\mathbf{i}+y\mathbf{j})$.

10. **(a)** If $u=f(P)$ and $v=g(P)$ are differentiable scalar fields, show that $\nabla(uv)=u\nabla v+v\nabla u$. **(b)** Use this result to find $\nabla(x/r^3)$.

11. Prove Theorem I along the following lines: Let $\mathbf{n}$ be a unit vector in a certain direction at the point $P_0(x_0, y_0, z_0)$. If the direction cosines of the direction are $\cos\alpha$, $\cos\beta$, $\cos\gamma$, then

$$\mathbf{n}=\cos\alpha\,\mathbf{i}+\cos\beta\,\mathbf{j}+\cos\gamma\,\mathbf{k}.$$

If s is distance measured positively along the line through P_0 in the direction of $\mathbf{n}$, with $s=0$ at P_0, the line has parametric equations

$$x=x_0+s\cos\alpha,\ y=y_0+s\cos\beta,\ z=z_0+s\cos\gamma.$$

Show that $\dfrac{du}{ds}=(\nabla u)\cdot\mathbf{n}$, and explain why this formula is equivalent to Theorem I.

12. If ϕ is colatitude and θ is longitude (both in radians) in a system of spherical co-ordinates related to rectangular co-ordinates by the equations $x=r\sin\phi\cos\theta$, $y=r\sin\phi\sin\theta$, $z=r\cos\phi$, what are **(a)** the level surfaces for $u=\phi$? **(b)** the level surfaces for $u=\theta$?

(c) Show that

$$\nabla\phi=\frac{\cos\theta\cos\phi}{r}\mathbf{i}+\frac{\sin\theta\cos\phi}{r}\mathbf{j}-\frac{\sin\phi}{r}\mathbf{k}.$$

This may be done either by using (10.6–3) or by careful use of Theorems I and II

(d) Show that

$$\nabla\theta = -\frac{\sin\theta}{r\sin\phi}\mathbf{i} + \frac{\cos\theta}{r\sin\phi}\mathbf{j} = -\frac{y}{x^2+y^2}\mathbf{i} + \frac{x}{x^2+y^2}\mathbf{j}.$$

The remark in (c) applies here also.

13. Show that (10.6–2) holds if the xyz-system and the $x'y'z'$-system are related by a translation. In this case we have $\mathbf{i} = \mathbf{i}'$, $\mathbf{j} = \mathbf{j}'$, $\mathbf{k} = \mathbf{k}'$, since the vector space with origin O' is considered to be indistinguishable from the vector space with origin O (see §10.1). Then discuss the case where both a translation and a rotation are needed to pass from one system to the other.

10.7 / THE DIVERGENCE OF A VECTOR FIELD

Let

$$\mathbf{F} = F_1(x, y, z)\mathbf{i} + F_2(x, y, z)\mathbf{j} + F_3(x, y, z)\mathbf{k} \tag{10.7–1}$$

be a vector field with differentiable components. We remind the student that F_1, F_2, F_3 are the components of $\mathbf{F}$ in the xyz-co-ordinate system. In another rectangular co-ordinate system there will be another set of components. We also observe that the subscripts 1, 2, 3 do not refer to partial differentiation, as they did in Chapters 6 and 7. The subscripts are simply labels on the separate functions.

The expression

$$\frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z} \tag{10.7–2}$$

is of interest in relation to the vector field, for it has an important physical meaning in certain kinds of vector fields occurring in physical theories. It also occurs in a very important purely mathematical theorem which we shall study later (the divergence theorem, §15.6). The importance of the expression (10.7–2) is closely connected with the fact that it is a *scalar invariant*, in the following sense: the expression (10.7–2) is unchanged in value if we make any rigid motion of the axes (see formula (10.7–4), further on). We shall deal with the case of a rotation of axes, leaving the simpler case of a translation for the student. Let the notation for the vector field in the rotated system be

$$\mathbf{F} = F_1'(x', y', z')\mathbf{i}' + F_2'(x', y', z')\mathbf{j}' + F_3'(x'. y', z')\mathbf{k}'.$$

Then, by (10.3–6) we have

$$F_1' = l_1F_1 + m_1F_2 + n_1F_3, \tag{10.7–3}$$

and other similar formulas. The xyz-system is related to the $x'y'z'$-system by equations (10.3–4) or (10.3–5). Now, by (10.7–3),

$$\frac{\partial F_1'}{\partial x'} = l_1\frac{\partial F_1}{\partial x'} + m_1\frac{\partial F_2}{\partial x'} + n_1\frac{\partial F_3}{\partial x'}.$$

Also, by the chain rule,

$$\frac{\partial F_1}{\partial x'} = \frac{\partial F_1}{\partial x}\frac{\partial x}{\partial x'} + \frac{\partial F_1}{\partial y}\frac{\partial y}{\partial x'} + \frac{\partial F_1}{\partial z}\frac{\partial z}{\partial x'}$$
$$= l_1\frac{\partial F_1}{\partial x} + m_1\frac{\partial F_1}{\partial y} + n_1\frac{\partial F_1}{\partial z},$$

with similar formulas for $\frac{\partial F_2}{\partial x'}$ and $\frac{\partial F_3}{\partial x'}$.

In the same way,

$$\frac{\partial F_2'}{\partial y'} = l_2\frac{\partial F_1}{\partial y'} + m_2\frac{\partial F_2}{\partial y'} + n_2\frac{\partial F_3}{\partial y'},$$
$$\frac{\partial F_1}{\partial y'} = l_2\frac{\partial F_1}{\partial x} + m_2\frac{\partial F_1}{\partial y} + n_2\frac{\partial F_1}{\partial z},$$

and so on. The student should himself write out all the formulas. From $\frac{\partial F_1'}{\partial x'}$ we obtain nine terms:

$$\frac{\partial F_1'}{\partial x'} = l_1^2\frac{\partial F_1}{\partial x} + l_1m_1\frac{\partial F_1}{\partial y} + l_1n_1\frac{\partial F_1}{\partial z}$$
$$+ m_1l_1\frac{\partial F_2}{\partial x} + m_1^2\frac{\partial F_2}{\partial y} + m_1n_1\frac{\partial F_2}{\partial z}$$
$$+ n_1l_1\frac{\partial F_3}{\partial x} + n_1m_1\frac{\partial F_3}{\partial y} + n_1^2\frac{\partial F_3}{\partial z}.$$

The formulas for $\frac{\partial F_2'}{\partial y'}$ and $\frac{\partial F_3'}{\partial z'}$ are obtained by advancing the subscripts to 2 and 3 respectively on l, m, and n. Now

$$l_1^2 + l_2^2 + l_3^2 = 1,$$
$$l_1m_1 + l_2m_2 + l_3m_3 = 0,$$

and so on. Hence it may be seen that

$$\frac{\partial F_1'}{\partial x'} + \frac{\partial F_2'}{\partial y'} + \frac{\partial F_3'}{\partial z'} = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}. \tag{10.7–4}$$

This proves the invariance of (10.7–2).

Definition. *The scalar function* (10.7–2) *is called the* divergence *of the vector field* (10.7–1). *It is denoted by* div **F**:

$$\text{div } \mathbf{F} = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}. \tag{10.7–5}$$

Observe that div **F** is a scalar field associated with the vector field **F**. We shall not try at this point to display the methematical or physical importance of

the concept of divergence. We remark merely that certain kinds of vector fields have the property that their divergence is everywhere zero. Such fields are called *solenoidal.* Among the most important solenoidal vector fields are those fields of force which are produced by the inverse square law of attraction or repulsion, e.g., gravitational or electrostatic fields. Also, if $\mathbf{F}$ denotes the velocity field of an incompressible fluid in a steady state of flow, the incompressibility is expressed by the fact that div $\mathbf{F} = 0$.

Example 1. The electrostatic field produced by a unit positive charge at O is

$$\mathbf{E} = \frac{1}{r^3}\overrightarrow{OP},$$

where $OP = r$. We shall show that the divergence of this field is zero wherever the field is defined (at all points except O). In a rectangular co-ordinate system let P have co-ordinates x, y, z. Then

$$r^2 = x^2 + y^2 + z^2, \quad \overrightarrow{OP} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}.$$

Hence the components of $\mathbf{E}$ are

$$E_1 = \frac{x}{r^3}, \qquad E_2 = \frac{y}{r^3}, \qquad E_3 = \frac{z}{r_3}.$$

Now

$$\frac{\partial E_1}{\partial x} = \frac{r^3 - 3r^2 x\,\dfrac{\partial r}{\partial x}}{r^6}$$

To find $\dfrac{\partial r}{\partial x}$ we have

$$2r\frac{\partial r}{\partial x} = 2x, \qquad \frac{\partial r}{\partial x} = \frac{x}{r}.$$

Hence

$$\frac{\partial E_1}{\partial x} = \frac{r^3 - 3rx^2}{r^6} = \frac{r^2 - 3x^2}{r^5}.$$

By symmetry,

$$\frac{\partial E_2}{\partial y} = \frac{r^2 - 3y^2}{r^5}, \qquad \frac{\partial E_3}{\partial z} = \frac{r^2 - 3z^2}{r^5};$$

therefore

$$\text{div } \mathbf{E} = \frac{3r^2 - 3(x^2 + y^2 + z^2)}{r^5} = 0.$$

There is another common notation for the divergence; it is

$$\text{div } \mathbf{F} = \nabla \cdot \mathbf{F}. \tag{10.7–6}$$

The reason for this notation is one of formal appearance. The symbol ∇ was introduced in the notation for the gradient (see (10.6–3)). The symbol ∇ by itself

is sometimes expressed as

$$\nabla = \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}.$$

In this form ∇ (which is read as "del") is called a *vector differential operator.* We say that the components of the operator ∇ in this particular co-ordinate system are

$$\frac{\partial}{\partial x}, \quad \frac{\partial}{\partial y}, \quad \frac{\partial}{\partial z}.$$

Recalling the formula

$$\mathbf{A} \cdot \mathbf{B} = A_1B_1 + A_2B_2 + A_3B_3$$

for the dot product of two vectors, we see that appearances would lead us to write

$$\nabla \cdot \mathbf{F} = \frac{\partial}{\partial x}F_1 + \frac{\partial}{\partial y}F_2 + \frac{\partial}{\partial z}F_3.$$

The expression on the right here is in fact div $\mathbf{F}$, if the "product" of the symbols $\frac{\partial}{\partial x}F_1$ is understood to mean the derivative $\frac{\partial F_1}{\partial x}$ and so on. We thus have a justification of the notation (10.7–6).

Particular interest attaches to the divergence of the gradient of a scalar field. If the scalar field is $u = f(P)$ we see that

$$\text{div}(\textbf{grad } u) = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}. \tag{10.7–8}$$

In the ∇ notation

$$\text{div}(\textbf{grad } u) = \nabla \cdot \nabla u.$$

It is customary to write $\nabla \cdot \nabla = \nabla^2$, so that

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}. \tag{10.7–9}$$

The left member of (10.7–9) is read as "del-squared of u," or "del-squared u." The equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0$$

is of fundamental importance in many branches of applied mathematics. It is known as *Laplace's equation,* in honor of the researches of the famous French mathematician Pierre Simon de Laplace (1749–1827). Accordingly the expression $\nabla^2 u$ is often called the Laplacian of u.

Since the gradient and the divergence are both invariants with respect to rigid motions of the axes, it follows from (10.7–8) that $\nabla^2 u$ is a scalar invariant

associated with the field u. In other words,

$$\frac{\partial^2 u}{\partial x'^2}+\frac{\partial^2 u}{\partial y'^2}+\frac{\partial^2 u}{\partial z'^2}=\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+\frac{\partial^2 u}{\partial z^2},$$

if the $x'y'z'$-system is obtained from the xyz-system by a rigid motion.

Example 2. The electrostatic potential at P arising from a dipole of unit strength at O, oriented in the direction of the unit vector $\mathbf{n}$, is

$$V=\frac{\mathbf{n}\cdot\overrightarrow{OP}}{r^3}=\frac{\cos\theta}{r^2},$$

where $r=OP$ (see Fig. 82). The electrostatic field itself is $\mathbf{E}=-\nabla V$. Let us calculate $\nabla\cdot\mathbf{E}=-\nabla^2 V$. We choose a coordinate system with origin O so that $\mathbf{n}$ coincides with the direction of the positive z-axis. Then

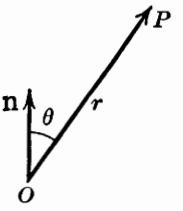


Fig. 82.

$$\mathbf{n}=\mathbf{k},\ \overrightarrow{OP}=x\mathbf{i}+y\mathbf{j}+z\mathbf{k},\ r^2=x^2+y^2+z^2.$$

Hence $\mathbf{n}\cdot\overrightarrow{OP}=z$, and

$$V=\frac{z}{r^3}.$$

Then

$$\frac{\partial V}{\partial x}=-3zr^{-4}\frac{\partial r}{\partial x},\quad 2r\frac{\partial r}{\partial x}=2x,\quad \frac{\partial V}{\partial x}=-\frac{3xz}{r^5}.$$

Likewise, by symmetry,

$$\frac{\partial V}{\partial y}=\frac{-3yz}{r^5}.$$

Next

$$\frac{\partial V}{\partial z}=\frac{1}{r^3}-3zr^{-4}\frac{\partial r}{\partial z}=\frac{1}{r^3}-\frac{3z^2}{r^5}.$$

$$\frac{\partial^2 V}{\partial x^2}=-3z\left(\frac{1}{r^5}-5xr^{-6}\frac{\partial r}{\partial x}\right)=-3z\left(\frac{1}{r^5}-\frac{5x^2}{r^7}\right),$$

and by symmetry we can write down a similar expression for $\frac{\partial^2 V}{\partial y^2}$. Finally,

$$\frac{\partial^2 V}{\partial z^2}=-3r^{-4}\frac{\partial r}{\partial z}-\frac{6z}{r^5}+3z^2 5r^{-6}\frac{\partial r}{\partial z}$$

$$=-\frac{3z}{r^5}-\frac{6z}{r^5}+\frac{15z^3}{r^7}.$$

Collecting, we see that

$$\nabla^2 V=-3z\left(\frac{1}{r^5}-\frac{5x^2}{r^7}\right)-3z\left(\frac{1}{r^5}-\frac{5y^2}{r^7}\right)-3z\left(\frac{3}{r^5}-\frac{5z^2}{r^7}\right)$$

$$=-3z\left(\frac{5}{r^5}-\frac{5(x^2+y^2+z^2)}{r^7}\right)=0.$$

EXERCISES

1. Show that, if u and $\mathbf{F}$ are differentiable scalar and vector fields, respectively, then

$$\nabla \cdot (u\mathbf{F}) = \nabla u \cdot \mathbf{F} + u\nabla \cdot F.$$

2. Show that, granted sufficient differentiability,

$$\nabla \cdot (u\nabla v) = u\nabla^2 v + \nabla u \cdot \nabla v.$$

3. **(a)** Let $\mathbf{F} = z\mathbf{i}$. If this vector field is interpreted to mean that $\mathbf{F}$ at (x, y, z) is the velocity of the particle of a fluid at that point, describe in words the nature of the motion of the fluid. Find $\nabla \cdot \mathbf{F}$. What is happening to the volume of the part of the fluid which at a given instant occupies the cube with sides $x = 0$, $x = 1$, $y = 0$, $y = 1$, $z = 0$, $z = 1$?

(b) As in **(a)**, interpret $\mathbf{F} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ as the velocity at (x, y, z) in a fluid flow. What is the nature of the motion? Find $\nabla \cdot \mathbf{F}$. Let V be the volume of the part of the fluid which at $t = 0$ occupies the sphere of radius r and center O. Find $\frac{dr}{dt}$ and $\frac{dV}{dt}$ in terms of r and show that $\frac{1}{V}\frac{dV}{dt} = \nabla \cdot \mathbf{F}$. This special situation gives some hint of the meaning of $\nabla \cdot \mathbf{F}$ in relation to expansion or contraction of volumes in an arbitrary fluid flow.

4. **(a)** In Exercise 8, §10.6, it was shown that for a scalar field of the form $u = F(r)$, $\nabla u = \frac{1}{r}F'(r)\mathbf{R}$ where $r = OP$ and $\mathbf{R} = \overrightarrow{OP}$. Using this result and the formula in Exercise 1, show that

$$\nabla^2 u = F''(r) + \frac{2}{r}F'(r).$$

(b) Find $F(r)$ if $\nabla^2 u = 0$ when $r > 0$.

5. **(a)** Using the notation r and $\mathbf{R}$ as in Exerxise 4, let $\mathbf{F} = \phi(r)\mathbf{R}$ and show that $\nabla \cdot \mathbf{F} = r\phi'(r) + 3\phi(r)$. **(b)** Find $\phi(r)$ if $\nabla \cdot \mathbf{F} = 0$ when $r > 0$. Show that in this case $\mathbf{F}$ can be interpreted as a force of attraction toward, or repulsion from, O according to the inverse square law.

6. **(a)** If $\mathbf{F} = \mathbf{A} \times \overrightarrow{OP}$, where $\mathbf{A}$ is any constant vector field, show that $\nabla \cdot \mathbf{F} = 0$. SUGGESTION: Choose axes so that $\mathbf{A}$ has the direction of the z-axis. This is not essential, but it simplifies the work.

(b) If a rigid body is rotating around the z-axis with angular speed ω, rotation in the xy-plane being in the counterclockwise sense as viewed by an observer looking down on the plane from the positive z-axis, the velocity of any point P of the body is $\mathbf{V} = \omega\mathbf{k} \times \overrightarrow{OP}$. Verify this. Write out the components of $\mathbf{V}$ and compute $\nabla \cdot \mathbf{V}$.

7. If $u = ax^2 + by^2 + cz^2$, find $\nabla^2 u$. If u is a general homogeneous polynomial of degree two in x, y, z, what is the condition on its coefficients necessary and sufficient to insure $\nabla^2 u = 0$?

10.8 / THE CURL OF A VECTOR FIELD

We have seen that div $\mathbf{F}$ is a scalar invariant which appears *formally* as the dot product $\nabla \cdot \mathbf{F}$ of the vector operator symbol ∇ and the vector $\mathbf{F}$. Let us now ask what we obtain formally from the cross product $\nabla \times \mathbf{F}$. By analogy with (10.2–7)

we write

$$\nabla \times \mathbf{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{vmatrix}.$$

In expanded form

$$\nabla \times \mathbf{F} = \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\right)\mathbf{i} + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\right)\mathbf{j} + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right)\mathbf{k}. \tag{10.8–2}$$

The first question about this expression is: *Is it invariant with respect to rigid motions of the axes*? The student who has thoroughly understood the discussions of the invariance of the gradient and the divergence will appreciate the import of the present question; he will also be ready to carry through the proof of invariance of $\nabla \times \mathbf{F}$ on the basis of the experience he has acquired. In §10.7 we have seen how to deal with the transformation of partial derivatives of the components of $\mathbf{F}$ under a rotation of the axes. We have to prove three relations, of which the relation

$$\frac{\partial F_3'}{\partial y'} - \frac{\partial F_2'}{\partial z'} = l_1\left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\right) + m_1\left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\right) + n_1\left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right) \tag{10.8–3}$$

is typical. To verify (10.8–3) we calculate $\dfrac{\partial F_3'}{\partial y'}$ and $\dfrac{\partial F_2'}{\partial z'}$ by the same methods used to calculate $\dfrac{\partial F_1'}{\partial x'}$ in the discussion that follows (10.7–3). The result thus obtained for the left side of (10.8–3) can be matched up with the right side by making suitable replacements for l_1, m_1, n_1. We can replace l_1 by $m_2n_3 - m_3n_2$ because of the fact that $\mathbf{i}' = \mathbf{j}' \times \mathbf{k}'$, when we look at the components of these latter vectors in the xyz-system. There are corresponding replacements for m_1 and n_1. We leave it to the student to complete the work.

The expression $\nabla \times \mathbf{F}$ defined in (10.8–2) is a vector invariant. Its importance is on the same level as that of the divergence, and for much the same reasons. We call $\nabla \times \mathbf{F}$ the *curl* of the field $\mathbf{F}$, and write

$$\operatorname{curl} \mathbf{F} = \nabla \times \mathbf{F}.$$

The physical or geometrical significance of the curl cannot be explained in a few lines. Certain kinds of vector fields have the property that their curl is everywhere the vector $\mathbf{0}$. Such fields are called *irrotational*. This name comes from hydrodynamics, and is there used to describe the vector field of velocities of particles in certain types of fluid motion. The antithesis of irrotational fluid motion is *vortex* motion, of which a whirlpool affords the simplest example. In the physical applications to fields of force, the irrotational fields are the *conservative* fields, i.e., the fields in which a principle of conservation of energy applies.

Example 1. Consider the vector field $\mathbf{V} = \omega\mathbf{k} \times \overrightarrow{OP}$ representing the velocities of points P in a rigid body rotating about the z-axis (see Exercise 6(b), §10.7). We have

$$\mathbf{V} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 0 & 0 & \omega \\ x & y & z \end{vmatrix} = -\omega y\mathbf{i} + \omega x\mathbf{j}.$$

Here $V_1 = -\omega y$, $V_2 = \omega x$, $V_3 = 0$. Hence

$$\nabla \times \mathbf{V} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ -\omega y & \omega x & 0 \end{vmatrix},$$

$$\nabla \times \mathbf{V} = 2\omega\mathbf{k}.$$

Thus the curl of the velocity field is a vector in the direction of the axis of rotation, having a magnitude equal to twice the angular speed of rotation.

Example 2. Let $\mathbf{F} = \dfrac{c}{r^3}\overrightarrow{OP}$, where c is a constant and r is the distance OP. This may be interpreted as the gravitational field of a mass concentrated at O (compare with Example 2, §10.51). Show that

$$\nabla \times \mathbf{F} = \mathbf{0}.$$

The components of $\mathbf{F}$ are

$$F_1 = \frac{cx}{r^3}, \qquad F_2 = \frac{cy}{r^3}, \qquad F_3 = \frac{cz}{r^3},$$

and $r^2 = x^2 + y^2 + z^2$. Now

$$\frac{\partial F_2}{\partial x} = cy\left(-3r^{-4}\frac{\partial r}{\partial x}\right) = cy(-3r^{-4})\frac{x}{r} = -\frac{3cxy}{r^5}.$$

By a symmetrical calculation

$$\frac{\partial F_1}{\partial y} = -\frac{3cxy}{r^5}.$$

Hence the z-component of $\nabla \times \mathbf{F}$ is

$$\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} = 0.$$

The other components also vanish, by symmetry.

EXERCISES

1. If $\mathbf{F} = \nabla u$, show that $\nabla \times \mathbf{F} = \mathbf{0}$ at all points of the field. Assume that u is twice differentiable.

2. Show that $\nabla \cdot (\nabla \times \mathbf{F}) = 0$ in the case of any twice differentiable vector field.

3. Find $\nabla \times \mathbf{F}$ in each case.

(a) $\mathbf{F} = 2xz\mathbf{i} + 2yz^2\mathbf{j} + (x^2 + 2y^2z - 1)\mathbf{k}$.

(b) $\mathbf{F} = ax\mathbf{i} + by\mathbf{j} + cz\mathbf{k}$.

(c) $\mathbf{F} = y\mathbf{i} + z\mathbf{j} + x\mathbf{k}$.

(d) $\mathbf{F} = \dfrac{x\mathbf{i} + y\mathbf{j}}{x^2 + y^2}$.

(e) $\mathbf{F} = \dfrac{-y\mathbf{i} + x\mathbf{j}}{x^2 + y^2}$.

(f) $\mathbf{F} = -\dfrac{xz}{\sqrt{x^2 + y^2}}\mathbf{i} - \dfrac{yz}{\sqrt{x^2 + y^2}}\mathbf{j} + \sqrt{x^2 + y^2}\,\mathbf{k}$.

4. If u and $\mathbf{F}$ are differentiable scalar and vector fields, respectively, show that $\nabla \times (u\mathbf{F}) = u(\nabla \times \mathbf{F}) + \nabla u \times \mathbf{F}$.

5. If $\mathbf{E}$ and $\mathbf{F}$ are differentiable vector fields, show that

$$\nabla \cdot (\mathbf{E} \times \mathbf{F}) = \mathbf{F} \cdot (\nabla \times \mathbf{E}) - \mathbf{E} \cdot (\nabla \times \mathbf{F}).$$

6. If $\mathbf{F} = \mathbf{A} \times \overrightarrow{OP}$, where $\mathbf{A}$ is a constant vector field, show that $\nabla \times \mathbf{F} = 2\mathbf{A}$.

7. Let $\mathbf{R} = \overrightarrow{OP}$, $r = \|\mathbf{R}\|$. (a) Find $\nabla \times \mathbf{R}$. (b) If $\mathbf{F} = \phi(r)\mathbf{R}$, where ϕ is differentiable, show that $\nabla \times \mathbf{F} = \mathbf{0}$.

8. If $\mathbf{A}$ is a constant vector field, and r, $\mathbf{R}$ have meanings as in Exercise 7, show that $\nabla \times (r^n\mathbf{A} \times \mathbf{R}) = (n + 2)r^n\mathbf{A} - nr^{n-2}(\mathbf{A} \cdot \mathbf{R})\mathbf{R}$.

9. Write out all the details of the proof of (10.8–3).

MISCELLANEOUS EXERCISES

1. Let $\mathbf{A}$ and $\mathbf{B}$ be constant, and let $\mathbf{R} = \overrightarrow{OP}$, $\mathbf{E} = \mathbf{R} - \mathbf{A}$, $\mathbf{F} = \mathbf{R} - \mathbf{B}$. Show that (a) $\nabla \cdot (\mathbf{E} \times \mathbf{F}) = 0$, (b) $\nabla \times (\mathbf{E} \times \mathbf{F}) = 2(\mathbf{B} - \mathbf{A})$, (c) $\nabla(\mathbf{E} \cdot \mathbf{F}) = \mathbf{E} + \mathbf{F}$.

2. Let $\mathbf{R} = \overrightarrow{OP}$, $r = \|\mathbf{R}\|$, and let $\mathbf{A}$ be a constant vector.

(a) Find $\nabla \cdot r^n\mathbf{A})$ and $\nabla \times (r^n\mathbf{A})$.

(b) Find $\nabla \cdot (r^n\mathbf{A} \times \mathbf{R})$.

(c) Show that $\nabla\left(\dfrac{\mathbf{A} \cdot \mathbf{R}}{r^3}\right) = -\nabla \times \left(\dfrac{\mathbf{A} \times \mathbf{R}}{r^3}\right)$.

3. With the notation of Exercise 2, and the additional assumption that $\mathbf{A}$ has unit length, show that

(a) $\nabla \cdot [(\mathbf{A} \cdot \mathbf{R})\mathbf{A}] = 1$, (c) $\nabla \cdot [(\mathbf{A} \times \mathbf{R}) \times \mathbf{A}] = 2$,

(b) $\nabla \times [(\mathbf{A} \cdot \mathbf{R})\mathbf{A}] = 0$, (d) $\nabla \times [(\mathbf{A} \times \mathbf{R}) \times \mathbf{A}] = 0$.

4. Find $\nabla \cdot (\nabla u \times \nabla v)$.

5. If $\mathbf{A}$ is a constant vector, show that $\nabla \cdot (\mathbf{A} \times \mathbf{F}) = -\mathbf{A} \cdot (\nabla \times \mathbf{F})$.

11 / *LINEAR TRANSFORMATIONS*

11 / INTRODUCTION

The development of vector mathematics presented in the preceding chapter took place mostly during the second half of the nineteenth century. One of the leading contributors to the development was the American genius Josiah Willard Gibbs of Yale University. The first part of Chapter 10, which deals with the ways vectors combine with other vectors and with scalars (real numbers), is called vector algebra. During the twentieth century, vector algebra expanded vastly into a subject called linear algebra, which has important applications in much of mathematics, especially advanced calculus.

This chapter will present enough linear algebra to enable us to use the subject to unify and extend the results obtained in the last several chapters. Prerequisite to full understanding of this chapter and the next is some knowledge of simultaneous linear systems involving n equations in n unknowns. In particular, we shall use the following fact: A necessary and sufficient condition that such a system have a unique solution is that the determinant of the coefficient matrix be different from zero. In Chapter 12 we occasionally use some of the most elementary rules for computing determinants.

A big step in the transition from vector algebra to linear algebra is the realization that functions from a vector space to a vector space are in themselves examples of vectors. The first part of this chapter will be devoted to explaining in detail what this means. We can begin with the once popular question, "What is a vector?" The common reply used to be that a vector is a quantity having both magnitude and direction. We shall soon see that neither magnitude nor direction is essential to vectors, and that most things having both (trains, for example) are not vectors. No satisfactory answer was arrived at until it was realized that one should not try to define a vector as an object having certain qualities, but rather as a member of a family of objects governed by certain rules. In particular, the important considerations are how to combine a vector with other vectors and with scalars through certain operations.

The problem is much like that of defining a checker. We might naturally begin by saying that it is a wooden or plastic disk, but such a beginning does not lead to anything satisfactory. If one were asked whether the top of a soft drink bottle is a checker, one's first inclination would probably be to say no. Yet many games of checkers have been played with these objects. The important fact which emerges from all this is that there is no property intrinsic to a thing considered in isolation which can settle the question of whether it is or is not a checker. The

answer depends entirely on whether it is a member of a set of objects which are moved according to certain well-defined rules.

Similarly, a mathematical object is a vector if and only if it is a member of a set whose members combine with each other and with real numbers according to certain well-defined rules—the rules of vector algebra. These rules have already been encountered in our discussion of $\mathbb{R}^3$ in §10.1. We shall collect and present them here as a set of abstract rules. The advantage of this is that it will enable us to apply them to sets of objects which we would be slow to recognize as vectors if we continued to associate vectors with physical or geometrical quantities. The rules actually apply to two sets—the set of real numbers $\mathbb{R}$, whose elements we shall refer to as scalars, and the set $\mathcal{V}$ of vectors. In stating the rules we shall distinguish between the scalars and the vectors by using boldface type to refer to the latter.

On the set of vectors $\mathcal{V}$, there must be a binary operation, denoted by $+$ and called addition, which is commutative and associative, i.e., if $\mathbf{x}$ and $\mathbf{y}$ belong to $\mathcal{V}$, then their sum $\mathbf{x}+\mathbf{y}$ also belongs to $\mathcal{V}$ and

$$\mathbf{x}+\mathbf{y}=\mathbf{y}+\mathbf{x}$$

and

$$\mathbf{x}+(\mathbf{y}+\mathbf{z})=(\mathbf{x}+\mathbf{y})+\mathbf{z}$$

for all $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ in $\mathcal{V}$.

There is a zero vector, denoted by $\mathbf{0}$, with the property that

$$\mathbf{x}+\mathbf{0}=\mathbf{x} \qquad \text{for all } \mathbf{x} \text{ in } \mathcal{V}.$$

The fact that this vector is denoted by the same symbol as the real number zero means that the student must be alert from now on to infer the correct meaning from the context in which the symbol occurs.

For each vector $\mathbf{a}$ in $\mathcal{V}$, there is a unique vector $\mathbf{x}$ such that $\mathbf{a}+\mathbf{x}=0$. This vector $\mathbf{x}$ is denoted by $-\mathbf{a}$, of course.

There is an operation called multiplication of vectors by scalars, which is denoted merely by juxtaposition and which has the following properties.

$$\begin{aligned}
c\mathbf{x}=\mathbf{x}c \text{ is in } \mathcal{V} &\quad \text{for all } \mathbf{x} \text{ in } \mathcal{V} \text{ and all } c \text{ in } \mathbb{R}.\\
c(\mathbf{x}+\mathbf{y})=c\mathbf{x}+c\mathbf{y} &\quad \text{for all } \mathbf{x} \text{ and } \mathbf{y} \text{ in } \mathcal{V} \text{ and all } c \text{ in } \mathbb{R}.\\
(a+b)\mathbf{x}=a\mathbf{x}+b\mathbf{x} &\quad \text{for all } a \text{ and } b \text{ in } \mathbb{R} \text{ and all } \mathbf{x} \text{ in } \mathcal{V}.\\
(ab)\mathbf{x}=a(b\mathbf{x}) &\quad \text{for all } a \text{ and } b \text{ in } \mathbb{R} \text{ and all } \mathbf{x} \text{ in } \mathcal{V}.\\
1\mathbf{x}=\mathbf{x} &\quad \text{for all } \mathbf{x} \text{ in } \mathcal{V}.\\
0\mathbf{x}=\mathbf{0} &\quad \text{for all } \mathbf{x} \text{ in } \mathcal{V}.
\end{aligned}$$

Notice also that neither multiplication nor division of one vector by another is necessarily defined. In many applications of vectors it is convenient to define a dot product (frequently called a scalar product) like that in Chapter 10. And in the study of three-dimensional space the vector product has been seen to be useful, but no concept of multiplying one vector by another is essential in defining vectors, and nowhere will we have occasion to define division for vectors.

We pause here to recall the use of the set membership symbol $\in$, already introduced in §2.7. If $\mathcal{S}$ is any set, the notation $a \in \mathcal{S}$ means "a is a member (or element) of $\mathcal{S}$." It can be variously read as "a belongs to $\mathcal{S}$," or "a is in $\mathcal{S}$." Other slight variations of the verbal rendering of the notation are useful. For instance, we may write one of the foregoing rules about vectors as follows: $(ab)\mathbf{x} = a(b\mathbf{x})$ for each $a, b \in \mathbb{R}$ and each $x \in \mathcal{V}$. In this context $\in$ may be read as "in" or "belonging to."

The entire mathematical structure made up of $\mathcal{V}$, $\mathbb{R}$, the two operations of vector addition and multiplication of a vector by a scalar, and the rules governing them is properly called a *vector space*. To avoid prolixity, however, this term is frequently applied just to the $\mathcal{V}$ when the rest of the structure is understood. The expressions *linear space* and *linear vector space* are frequently used as synonyms for vector space.

Example 1. Let V denote the set $\mathbb{R}^n$ of all ordered n-tuples of real numbers, with addition and scalar multiplication defined as follows: If $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ then

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n),$$

and if c is any scalar, $c\mathbf{x} = (cx_1, cx_2, \ldots, cx_n)$. It can easily be verified that V together with these operations constitutes a vector space. We discussed $\mathbb{R}^n$ in §10.12. Remember that the zero vector $\mathbf{0}$ is $(0, 0, \ldots, 0)$. In Chapter 12 we shall study functions with domain in $\mathbb{R}^n$ and range in $\mathbb{R}^m$, where n and m may or may not be the same.

Example 2. Here, $\mathcal{V}$ is the set of all real-valued continuous functions defined on $[0, 1]$. The reader already knows how to add two functions and how to multiply one by a scalar. He also knows that these operations, performed on continuous functions, always give continuous functions. We merely wish to point out that this long-familiar structure is a vector space, even though we have said nothing about "magnitude and direction" of the vectors.

Example 3. Let $\mathcal{V}$ stand for the collection of all 2×2 matrices. We add matrices simply by adding the elements in corresponding positions, and in order to multiply a matrix by a scalar, we multiply each element of the matrix by that scalar. Even though we have not associated any magnitude or direction with these matrices, they are nonetheless vectors.

An explanation of why $\mathbb{R}^n$ is called n-dimensional was given in §10.12. For the general (abstract) case we call a vector space *finite-dimensional* if there exists a finite set of vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ such that every vector $\mathbf{x}$ can be represented in exactly one way as a linear combination of the $\mathbf{u}_i$'s:

$$\mathbf{x} = c_1\mathbf{u}_1 + \cdots + c_n\mathbf{u}_n.$$

The uniqueness of representation means that the coefficients $c_1, \ldots, c_n$ are uniquely determined by $\mathbf{x}$. This implies that the c_i's are all 0 if $\mathbf{x} = \mathbf{0}$. It also implies that no $\mathbf{u}_i$ is $\mathbf{0}$, for if some $\mathbf{u}_i = \mathbf{0}$, the choice of c_i cannot be uniquely

determined by $\mathbf{x}$. The set of vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ is called a *basis* for the space. In linear algebra it is proved that if $\mathbf{u}_1, \ldots, \mathbf{u}_n$ is a basis and $\mathbf{v}_1, \ldots, \mathbf{v}_m$ is a basis, then $n = m$. The common value of n and m is defined to be the *dimension* of the space. There are vector spaces that are not finite dimensional. Such a space is said to be *of infinite dimension.* The space in Example 2 is infinite dimensional. The space of matrices in Example 3 is of dimension four. One possible basis consists of the four matrices exhibited here:

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

11.1 / LINEAR TRANSFORMATIONS

In Example 2 the vectors are functions—continuous functions from the unit interval [0, 1] to the real numbers. In this section, we introduce another very important kind of function space. The vectors in these spaces are the very simple functions called linear transformations. By a *linear transformation* we mean a function T from a vector space $\mathcal{V}$ to a vector space $\mathcal{S}$ having the following two properties:

1. $T(\mathbf{x}) + T(\mathbf{y}) = T(\mathbf{x} + \mathbf{y})$ for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{V}$, and
2. $T(c\mathbf{x}) = cT(\mathbf{x})$ for every scalar c, and every $\mathbf{x} \in \mathcal{V}$.

Notice that from (1) it follows that $T(\mathbf{0}) + T(\mathbf{0}) = T(\mathbf{0})$, whence $T(\mathbf{0}) = \mathbf{0}$. This is an important fact. The same result can be deduced from (2).

Very often, for convenience, we shall write $T\mathbf{x}$ instead of $T(\mathbf{x})$, omitting the parentheses. This is not done for all functions, but is a common practice with linear transformations.

The statement that f is a function from the set A to the set B is frequently abbreviated to $f: A \to B$. It means that f is defined at every point of A, and that for each a in A, $f(a) \in B$. In other words, the domain of f is all of A, but its range may be either all of B or just some proper subset. Therefore, when we say that T is a linear transformation, we shall imply that $T: \mathcal{V} \to \mathcal{S}$ for some pair of vector spaces $\mathcal{V}$ and $\mathcal{S}$, and that T satisfies Properties (1) and (2). The terms *linear function* and *linear map* are frequently used as synonyms for linear transformation.

In the important special case where $\mathcal{V} = \mathcal{S}$, we shall follow the practice, in this book, of referring to the linear transformation as a *linear operator.* In other words, a linear operator, in our terminology, is a linear transformation which maps a vector space into itself.

If $\mathcal{V}$ and $\mathcal{S}$ are two vector spaces for which there exists a linear transformation $T: \mathcal{V} \to \mathcal{S}$ such that T maps $\mathcal{V}$ in a one-to-one manner onto *all* of $\mathcal{S}$, we say that $\mathcal{V}$ is isomorphic to $\mathcal{S}$. The property of one-to-oneness signifies that if $\mathbf{x} \neq \mathbf{y}$, then $T\mathbf{x} \neq T\mathbf{y}$. Therefore T has an inverse, say L, from $\mathcal{S}$ to $\mathcal{V}$, and since the inverse of a linear transformation is linear (Exercise 3) it follows that if one vector space is isomorphic to a second, then the second is isomorphic to the first, so it is sufficient to say simply that the two spaces are isomorphic.

Two vector spaces can be isomorphic without being identical. For example, the collection of all polynomials of degree not exceeding $n-1$:

$$p(t) = c_0 + c_1 t + c_2 t^2 + \cdots + c_{n-1} t^{n-1}$$

is a vector space if we add polynomials and multiply polynomials by scalars in the usual way. This vector space is isomorphic to $\mathbb{R}^n$ (see Example 1 in the preceding section) by virtue of the mapping which associates $p(t)$ with the vector $(c_0, c_1, \ldots, c_{n-1})$ in $\mathbb{R}^n$. As far as algebraic properties are concerned, two isomorphic vector spaces are indistinguishable.

11.2 / THE VECTOR SPACE $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$

For reasons connected with its role in abstract algebra, the set of all linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$ is denoted by $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$. We wish to prove that this set is a vector space. We define addition the same way we always have for functions; that is, if T_1 and T_2 belong to $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, then their sum T is that function from $\mathbb{R}^n$ to $\mathbb{R}^m$ defined by

$$T(\mathbf{x}) = T_1(\mathbf{x}) + T_2(\mathbf{x}) \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

T is obviously then a well-defined function from $\mathbb{R}^n$ to $\mathbb{R}^m$, but to show that it belongs to $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ we must prove that it is linear. We therefore first verify Property (1).

$$\begin{aligned} T(\mathbf{x}+\mathbf{y}) &= T_1(\mathbf{x}+\mathbf{y}) + T_2(\mathbf{x}+\mathbf{y}) = T_1(\mathbf{x}) + T_1(\mathbf{y}) + T_2(\mathbf{x}) + T_2(\mathbf{y}) \\ &= [T_1(\mathbf{x}) + T_2(\mathbf{x})] + [T_1(\mathbf{y}) + T_2(\mathbf{y})] = T(\mathbf{x}) + T(\mathbf{y}). \end{aligned}$$

Property (2) is checked as follows:

$$T(c\mathbf{x}) = T_1(c\mathbf{x}) + T_2(c\mathbf{x}) = cT_1(\mathbf{x}) + cT_2(\mathbf{x}) = c[T_1(\mathbf{x}) + T_2(\mathbf{x})] = cT(\mathbf{x}).$$

This completes the proof that if T_1 and T_2 are any two elements of $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, their sum must also belong to this set. It is even easier to prove that a scalar multiple of a linear transformation is a linear transformation. The other vector space axioms are obviously satisfied. The zero vector in this space $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ is the linear transformation which maps every vector of $\mathbb{R}^n$ into the zero vector of $\mathbb{R}^m$. The reader should verify that this function is linear.

11.3 / MATRICES AND LINEAR TRANSFORMATIONS

It is assumed that students using this book have some familiarity with the algebra of matrices. No extensive knowedge or expertise is needed, however. We can define the product AB of a matrix A with m rows and n columns (we call it an $m \times n$ matrix) and an $n \times p$ matrix B. The product $C = AB$ is an $m \times p$ matrix. If the elements of A are a_{ij} ($i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$), and if the elements of B are b_{jk} ($j = 1, 2, \ldots, n$; $k = 1, 2, \ldots, p$), then the elements of C are

c_{ik}, given by

$$c_{ik} = \sum_{j=1}^{n} a_{ij}b_{jk},\ i = 1, \ldots, m; \quad k = 1, \ldots, p.$$

We do not define AB unless the number of *columns* of A is the same as the number of *rows* in B. The order of the factors A and B is significant. That is, BA may not be defined, even though AB is; and even if AB and BA are both defined, they need not be equal. It is useful to know that matrix multiplication is associative. That is, $(AB)D = A(BD)$ when all the products shown here are defined.

An $m \times n$ matrix can be used to define a function from $\mathbb{R}^n$ to $\mathbb{R}^m$ as follows. If $\mathbf{x} = (x_1, \ldots, x_n)$ is given, display $\mathbf{x}$ as an $n \times 1$ matrix and multiply A times $\mathbf{x}$ to give an $m \times 1$ matrix $\mathbf{y}$ as shown here:

$$A\mathbf{x} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{pmatrix}$$

$$= \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \mathbf{y}.$$

For each $\mathbf{x}$ in $\mathbb{R}^n$ this defines $\mathbf{y} = A\mathbf{x}$ in $\mathbb{R}^n$. It is easy to show (see Exercise 1) that the function from $\mathbb{R}^n$ to $\mathbb{R}^m$ defined in this way is a linear transformation. Conversely, if T is any element of $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ there is a certain $m \times n$ matrix, A, such that $T\mathbf{x} = A\mathbf{x}$ for every $\mathbf{x}$ in $\mathbb{R}^n$. To prove this we use the standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ in $\mathbb{R}^n$, as defined in (10.12–8). Let us denote the corresponding standard basis in $\mathbb{R}^m$ by $\mathbf{u}_1, \ldots, \mathbf{u}_m$; that is, $u_1 = (1, 0, \ldots, 0), \ldots, \mathbf{u}_m = (0, \ldots, 1)$. With T given, we know that each $T\mathbf{e}_i$ is a linear combination of the $\mathbf{u}_j$'s, thus:

$$\begin{aligned} T\mathbf{e}_1 &= \alpha_{11}\mathbf{u}_1 + \alpha_{12}\mathbf{u}_2 + \cdots + \alpha_{1m}\mathbf{u}_m \\ T\mathbf{e}_2 &= \alpha_{21}\mathbf{u}_1 + \alpha_{22}\mathbf{u}_2 + \cdots + \alpha_{2m}\mathbf{u}_m \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad\quad \vdots \\ T\mathbf{e}_n &= \alpha_{n1}\mathbf{u}_1 + \alpha_{n2}\mathbf{u}_2 + \cdots + \alpha_{nm}\mathbf{u}_m \end{aligned} \tag{11.3–1}$$

for some set of scalar coefficients, α_{ij}. These equations show what T does to the standard basis elements in $\mathbb{R}^n$, and we shall see that this completely determines T. Let $\mathbf{x}$ be any vector in $\mathbb{R}^n$, and suppose $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Then $\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n$, and since T is linear,

$$T\mathbf{x} = T\sum_{i=1}^{n} x_i\mathbf{e}_i = \sum_{i=1}^{n} x_i T(\mathbf{e}_i) = \sum_{i=1}^{n} x_i \sum_{j=1}^{m} \alpha_{ij}\mathbf{u}_j$$

$$= \sum_{j=1}^{m} \left(\sum_{i=1}^{n} \alpha_{ij}x_i\right)\mathbf{u}_j = \sum_{j=1}^{m} y_j\mathbf{u}_j,$$

which means that $\mathbf{y} = (y_1, y_2, \ldots, y_m)$, where

$$\begin{aligned}
\alpha_{11}x_1 + \alpha_{21}x_2 + \cdots + \alpha_{n1}x_n &= y_1 \\
\alpha_{12}x_1 + \alpha_{22}x_2 + \cdots + \alpha_{n2}x_n &= y_2 \\
&\vdots \\
\alpha_{1m}x_1 + \alpha_{2m}x_2 + \cdots + \alpha_{nm}x_n &= y_m.
\end{aligned}$$

This can be written as $A\mathbf{x} = \mathbf{y}$, where the elements a_{ij} of the matrix A are defined by $a_{ij} = \alpha_{ji}$. This can be expressed by saying that the $m \times n$ matrix, A, is the *transpose* of the $n \times m$ coefficient matrix of the system (11.3–1). The transpose of a matrix M is a matrix obtained by interchanging the rows and columns of M. We denote the transpose of M by M^T. In §12.8 we shall use these facts: the transpose of the product of two matrices is the product of the transposes in reverse order; the transpose of the sum is the sum of the transposes; and the transpose of the transpose of a matrix is just the original matrix. The student is asked to verify these properties in Exercise 4.

Observe that the zero element of $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ is represented by the $m \times n$ matrix in which all the entries are zero.

The matrix A which represents T was constructed with the help of what were called the standard bases in the domain and range spaces, $\mathbb{R}^n$ and $\mathbb{R}^m$. There are other bases and if we had used them we would have obtained other matrix representations of T. This suggests that A might more properly be referred to as *the standard representation of* T. But since these are the only bases which we shall use, we shall sometimes speak only of "the" matrix representation of a linear transformation.

It is possible to summarize the results of this section up to this point by the following theorem, which is of basic importance, especially in the next few sections.

THEOREM I. *Each $m \times n$ matrix is the standard representation of a unique linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, and conversely, every element of $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ has a unique standard representation as an $m \times n$ matrix.*

If $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ and $L \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^p)$, then we can define a function $L \circ T$ from $\mathbb{R}^n$ to $\mathbb{R}^p$ as follows:

$$(L \circ T)(\mathbf{x}) = L[T(\mathbf{x})] \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

This is a composite function—the composition of L and T—and is easily proved to be linear (Exercise 2). Therefore, $L \circ T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p)$. It is important to be able to express the matrix which represents $L \circ T$ in terms of the matrices representing L and T. This can be done in the following straightforward way. Let A be the $m \times n$ matrix representing T and let B be the $p \times m$ representation of L. Then,

$$(L \circ T)(\mathbf{x}) = L[T\mathbf{x}] = B(A\mathbf{x}) = B\mathbf{y},$$

where $\mathbf{x}$ is any vector of $\mathbb{R}^n$ and $\mathbf{y}$ is its image under T in $\mathbb{R}^m$.

$$y_i = \sum_{j=1}^{n} a_{ij}x_j \qquad (i = 1, 2, \ldots, m)$$

$$(By)_k = \left(\sum_{i=1}^{m} b_{ki}y_i\right) \qquad (k = 1, 2, \ldots, p)$$

$$= \left(\sum_{i=1}^{m} b_{ki} \sum_{j=1}^{n} a_{ij}x_j\right) \qquad (k = 1, 2, \ldots, p)$$

$$= \sum_{j=1}^{n} \left(\sum_{i=1}^{m} b_{ki}a_{ij}\right)x_j \qquad (k = 1, 2, \ldots, p)$$

$$= \left(\sum_{j=1}^{n} c_{kj}x_j\right) \qquad (k = 1, 2, \ldots, p)$$

$$= C\mathbf{x},$$

where $C = BA$ is the $p \times n$ matrix whose element in the position (k, j) is given by

$$c_{kj} = \sum_{i=1}^{m} b_{ki}a_{ij}.$$

As has been pointed out earlier, if L and T are linear transformations from a space $\mathcal{U}$ to a space $\mathcal{V}$, then their sum $L + T$, defined by

$$(L + T)(\mathbf{x}) = L\mathbf{x} + T\mathbf{x}$$

is also a linear transformation from $\mathcal{U}$ to $\mathcal{V}$. Suppose that the matrix A represents L and the matrix B represents T. Then it will be very easy to define $A + B$ so that it will be a matrix representing $L + T$—we simply take each element of $A + B$ to be the sum of the corresponding elements in A and B. Notice that this defines the sum of two matrices if and only if they have the same number of rows and the same number of columns. Similarly, if we define the product of the matrix A by the scalar α to be the matrix obtained by multiplying each element of A by α, then we shall have it turn out that the matrix αA represents the linear transformation αL. By $A\alpha$ we shall mean the same matrix as αA. It is easy to see that the set of all $m \times n$ matrices form a vector space of dimension m times n.

11.4 / SOME SPECIAL CASES

If $m = 1$, then $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m) = \mathcal{L}(\mathbb{R}^n, \mathbb{R})$ is the vector space of all real-valued linear functions on $\mathbb{R}^n$. The elements of this space are called *linear functionals* on $\mathbb{R}^n$. From Theorem I we know that each such linear functional can be represented by a $1 \times n$ matrix. Such a matrix, consisting of 1 row and n columns, is just an ordered n-tuple of real numbers which acts on the vector $\mathbf{x}$ to give $L\mathbf{x}$ as

follows:

$$L(\mathbf{x}) = (a_1, a_2, \ldots, a_n)\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = a_1x_1 + a_2x_2 + \cdots a_nx_n. \tag{11.4–1}$$

Here we have a $(1 \times n)$ matrix multiplying an $(n \times 1)$ matrix to give the (1×1) matrix $\sum_{i=1}^{n} a_ix_i$, which is just a real number. If we think of the $1 \times n$ matrix that represents L as a vector $\mathbf{a}$, we see that the formula (11.4–1) can be written as $L(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x}$, where we use the dot product in $\mathbb{R}^n$ as defined in (10.12–1).

In Chapter 12 we shall find it useful to think of vectors as special cases of matrices, and to evaluate dot products by matrix multiplication. This is easy as soon as we decide whether to consider a vector as a single column matrix or as a single row matrix. (We shall hereafter omit the word "single" in this context.) The rule is easy: when vectors are being used in matrix operations we consider them to be column matrices unless the transpose sign is attached, in which case we regard the vector as a row matrix. For example, if one wishes to think of the vector $\mathbf{a}$ in $\mathbb{R}^n$ as a matrix, it is taken to be the $(n \times 1)$ matrix

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix},$$

but the transpose, $\mathbf{a}^T$, is of course, $(a_1, a_2, \ldots, a_n)$. Now we have the choice of writing $L(\mathbf{x})$ either as the dot product of vectors, i.e., $L(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x}$ or as an ordinary matrix product, i.e., $L(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$. In summary, we can rewrite (11.4–1) in the following more convenient forms

$$L(\mathbf{x}) = \mathbf{a}^T\mathbf{x} = \mathbf{a} \cdot \mathbf{x}. \tag{11.4–2}$$

Therefore, for each linear functional L on $\mathbb{R}^n$, there is a vector $\mathbf{a}$ in $\mathbb{R}^n$ which represents L in the sense that $L\mathbf{x} = \mathbf{a} \cdot \mathbf{x}$. It is trivial to show that this one-to-one correspondence is linear. Therefore $\mathscr{L}(\mathbb{R}^n, \mathbb{R})$ and $\mathbb{R}^n$ are isomorphic, i.e., algebraically the same.

In the special case $\mathscr{L}(\mathbb{R}, \mathbb{R}^m)$, the linear transformations are, by Theorem I, represented by $m \times 1$ matrices, which are at once identified with vectors in $\mathbb{R}^m$. In other words, if $T \in \mathscr{L}(\mathbb{R}, \mathbb{R}^m)$, then there exists some $\mathbf{a} \in \mathbb{R}^m$ such that, for every real number x,

$$Tx = \mathbf{a}x = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} x = \begin{pmatrix} a_1\lambda \\ a_2x \\ \vdots \\ a_mx \end{pmatrix}.$$

This says that every linear function from $\mathbb{R}$ to $\mathbb{R}^m$ is defined by taking scalar multiples of some fixed vector in $\mathbb{R}^m$. This correspondence between $\mathscr{L}(\mathbb{R}, \mathbb{R}^m)$ and $\mathbb{R}^m$ is easily seen to be one-to-one and linear, giving us again an example of an isomorphism between two vector spaces.

Finally, $\mathscr{L}(\mathbb{R}, \mathbb{R})$ and $\mathbb{R}$ are isomorphic, for all the linear functions from $\mathbb{R}$ to $\mathbb{R}$ are of the form $f(x) = ax$, and the correspondence mapping f into a is a linear transformation from $\mathscr{L}(\mathbb{R}, \mathbb{R})$ to $\mathbb{R}$ which is one to one. Our use here of the term "linear function" is different from its use in high school algebra, where functions of the form $F(x) = ax + b$ are called linear. Such functions are usually called *affine*. They are obtained from linear functions simply by adding some constant. In the more general setting of functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ an affine function $\mathbf{f}$ would be defined by

$$\mathbf{f}(\mathbf{x}) = T\mathbf{x} + \mathbf{b}$$

where $T \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ and $\mathbf{b} \in \mathbb{R}^m$.

11.5 / NORMS

We have remarked in §11 that it is not an essential characteristic of a vector to have magnitude and direction. Nevertheless, in many vector spaces it is possible and useful to define a magnitude, or length, for each vector. When this is done, we denote the *length* of $\mathbf{x}$ by $\|\mathbf{x}\|$ and call it the *norm* of $\mathbf{x}$. (See the definition in (10.1–16) for $\mathbb{R}^3$ and in (10.12–2) for $\mathbb{R}^n$.) Just as we listed in §11 the algebraic rules governing the elements in a vector space, so we now list the rules that we impose on a norm in a vector space V. A function from V to $\mathbb{R}$, with value $\|\mathbf{x}\|$ at $\mathbf{x}$, is called a norm provided that the following rules (axioms) are satisfied:

1. $\|\mathbf{x}\| \geqq 0$ for each $\mathbf{x}$.
2. $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
3. $\|c\mathbf{x}\| = |c|\|\mathbf{x}\|$ for each $\mathbf{x}$ and each scalar c.
4. $\|\mathbf{x} + \mathbf{y}\| \leqq \|\mathbf{x}\| + \|\mathbf{y}\|$ for each $\mathbf{x}$ and $\mathbf{y}$.

For a good many purposes it does not matter exactly how $\|\mathbf{x}\|$ is defined, as long as it has the foregoing properties (1)–(4). We have defined the norm of $\mathbf{x}$ in $\mathbb{R}^n$ in a particular way in (10.12–2). That definition is related to the dot product in $\mathbb{R}^n$. But in general a vector space can have a norm without any relation to a dot product. Other definitions of a norm are possible in $\mathbb{R}^n$, as we shall see presently.

The property (4) is called the triangle inequality, by analogy with the same inequality that is valid in $\mathbb{R}^3$ and $\mathbb{R}^n$. (See (10.12–6), where the reason for the name "triangle inequality" is given.)

We can transform the inequality (4) in various ways. For instance, replace $\mathbf{x}$ by $\mathbf{x} + \mathbf{y}$ and $\mathbf{y}$ by $-\mathbf{y}$ in (4). Then, because $(\mathbf{x} + \mathbf{y}) + (-\mathbf{y}) = \mathbf{x}$ and $\|-\mathbf{y}\| = \|\mathbf{y}\|$, we obtain

$$\|\mathbf{x}\| \leqq \|\mathbf{x} + \mathbf{y}\| + \|\mathbf{y}\|,$$

whence

$$\|\mathbf{x}\| - \|\mathbf{y}\| \leqq \|\mathbf{x} + \mathbf{y}\|. \tag{11.5–1}$$

If we exchange $\mathbf{x}$ and $\mathbf{y}$ here and bear in mind that $\mathbf{y} + \mathbf{x} = \mathbf{x} + \mathbf{y}$, we obtain

$$\|\mathbf{y}\| - \|\mathbf{x}\| \leqq \|\mathbf{x} + \mathbf{y}\| \tag{11.5–2}$$

Since $\big|\|\mathbf{x}\| - \|\mathbf{y}\|\big|$ is either $\|\mathbf{x}\| - \|\mathbf{y}\|$ or $\|\mathbf{y}\| - \|\mathbf{x}\|$, depending on the sign of the difference, we conclude that

$$\big|\|\mathbf{x}\| - \|\mathbf{y}\|\big| \leqq \|\mathbf{x} + \mathbf{y}\|. \tag{11.5–3}$$

Because $\|-\mathbf{y}\| = \|\mathbf{y}\|$, we can change $\mathbf{x} + \mathbf{y}$ to $\mathbf{x} - \mathbf{y}$ in each of the last three inequalities.

Actually, the triangle inequality (4) is derivable from (11.5–1). See Exercise 8. On this account we shall also refer to (11.5–1) as a triangle inequality. Likewise for (11.5–2) and (11.5–3).

If $\|\mathbf{x}\| = 1$ we call $\mathbf{x}$ a *unit vector*. If $\mathbf{x}$ is any nonzero vector, a suitable multiple of $\mathbf{x}$ will be a unit vector. The right multiplier is $1/\|\mathbf{x}\|$, as we see by property (3) of the norm. In any vector space with a norm, the *unit sphere* is defined to be the set of all unit vectors; therefore the equation of the unit sphere is $\|\mathbf{x}\| = 1$.

The norm of §10.12:

$$\mathbf{x} = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$$

is called the *Euclidean norm* in $\mathbb{R}^n$. We could also define other norms as follows:

$$\|\mathbf{x}\| = \sum_{i=1}^{n} |x_i|,$$

or

$$\|\mathbf{x}\| = \text{maximum of } |x_1|, |x_2|, \ldots, |x_n|.$$

Showing that these last two definitions satisfy the conditions (1) to (4) is left for the exercises.

11.6 / METRICS

In preparation for §11.7, where we shall discuss point sets in vector spaces, and continuity of functions with domains of definition and ranges of values in vector spaces, we shall now discuss measurements of distance in a vector space provided with a norm.

In dealing with any vector space, it is common practice to use the words *point* and *vector* interchangeably. Thus we may speak either of the vector $\mathbf{x}$ or the point $\mathbf{x}$. In any vector space provided with a norm we call $\|\mathbf{x} - \mathbf{y}\|$ the distance between $\mathbf{x}$ and $\mathbf{y}$. For vectors in the plane the appropriateness of this definition is shown in Fig. 83, which shows diagrammatically how the equal lengths of the

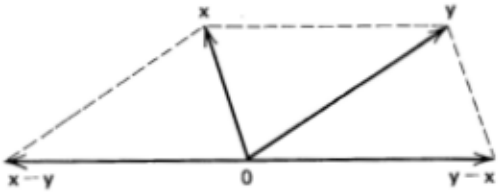


Fig. 83.

vectors $\mathbf{x}-\mathbf{y}$ and $\mathbf{y}-\mathbf{x}$ is the same as the distance between the ends of the vectors $\mathbf{x}$ and $\mathbf{y}$.

Just as we abstracted the notion of the length of a vector by calling the length of $\mathbf{x}$ its norm $\|\mathbf{x}\|$ and listing the four conditions that the norm must satisfy, so we may abstract the idea of distance between pairs of points and list four conditions that we want to be satisfied by a distance function. A distance function is a function that gives us a value $d(\mathbf{x}, \mathbf{y})$ for the distance from $\mathbf{x}$ to $\mathbf{y}$. We require the following four axioms to be satisfied:

D_1: $d(\mathbf{x}, \mathbf{y}) \geqq 0$ for all $\mathbf{x}$ and $\mathbf{y}$.

D_2: $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.

D_3: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$.

D_4: $d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \geqq d(\mathbf{x}, \mathbf{z})$.

Axiom D_4 is called the triangle inequality because it expresses the fact that in a configuration of three points, which we may think of as forming a triangle, the distance along one leg of the triangle is never greater than the sum of the distances along the other two legs.

Our definition $d(x, y) = \|\mathbf{x}-\mathbf{y}\|$ of distance in a vector space with a norm satisfies the four conditions $D_1 - D_4$; D_4 is satisfied as a consequence of the triangle inequality (4) in §11.5.

A distance function satisfying conditions $D_1 - D_4$ is called a *metric*. We have seen how to use a norm to define a metric. The idea of a metric need not be confined to vector spaces, however, for there are no references to addition of vectors or scalar multiples of vectors in the axioms $D_1 - D_4$. In fact, there are metrics which do not come from norms. See Exercises 11 and 12.

11.7 / OPEN SETS AND CONTINUITY

If E is a set and G is a set, then the set consisting of all objects which belong to E or to G, or to both, is called the *union of* E and G, which we denote by $E \cup G$. The set consisting of all members of both E and G is denoted by $E \cap G$ and is called the *intersection* of E and G. If $p(\mathbf{x})$ denotes some proposition involving $\mathbf{x}$, then $\{\mathbf{x}: p(\mathbf{x})\}$ denotes that set consisting of all $\mathbf{x}$ for which the proposition $p(\mathbf{x})$ is true. For example, $\{\mathbf{x}: \mathbf{x} \cdot \mathbf{a} = 3\}$ is the set of all vectors whose inner product with the vector $\mathbf{a}$ is 3. Similarly, $\{\mathbf{x} \in E: p(\mathbf{x})\}$ is that set of vectors belonging to the set E for which the proposition $p(\mathbf{x})$ is true.

For any vector space having a norm, we define the sphere of radius r centered at the vector $\mathbf{a}$, denoted by $S(\mathbf{a}, r)$, as follows:

$$S(\mathbf{a}, r) = \{\mathbf{x} : \|\mathbf{x} - \mathbf{a}\| = r\}.$$

The *open ball*, $B(\mathbf{a}, r)$, of radius r, *centered at* $\mathbf{a}$, is given by

$$B(\mathbf{a}, r) = \{\mathbf{x} : \|\mathbf{x} - \mathbf{a}\| < r\}.$$

The corresponding *closed ball*, $\overline{B(\mathbf{a}, r)}$, is the union of these two sets; that is,

$$\overline{B(\mathbf{a}, r)} = B(\mathbf{a}, r) \cup S(\mathbf{a}, r) = \{\mathbf{x} : \|\mathbf{x} - \mathbf{a}\| \leqq r\}.$$

Notice that in $\mathbb{R}^3$, $S(\mathbf{a}, r)$ is a spherical surface and $B(\mathbf{a}, r)$ consists of the set of points lying inside this surface. $\overline{B(\mathbf{a}, r)}$ is what might be called the solid sphere made up of the surface, together with the set of points lying inside. How would you describe $S(\mathbf{a}, r)$, $B(\mathbf{a}, r)$, and $\overline{B(\mathbf{a}, r)}$ in $\mathbb{R}^2$? in $\mathbb{R}$? Notice that $S(\mathbf{a}, r) \cap B(\mathbf{a}, r)$ is the empty set.

If E is any subset of a normed vector space, we define an *interior point of* E to be a point which belongs to E and which is the center of some open ball contained in E. Notice that this is just a more general statement of our earlier definition in §5.1. An *open set* can still be defined as one consisting entirely of interior points, and a *closed set* is still the complement of an open set.

By a *neighborhood* of a point, we simply mean an open set containing the point. This is a generalization of our earlier usage where the sets which we called neighborhoods were open intervals, open disks, or open rectangles. The really essential property of what we call a neighborhood of a point is that it is a set for which the point is an interior point. The definition which we use is the simplest way to get this property.

A subset E of a normed vector space is said to be *bounded* in case there is some number M such that $\|\mathbf{x}\| < M$ for all $\mathbf{x} \in E$.

Let $\{\mathbf{x}^k\}_{k=1}^{\infty}$ denote a sequence in a vector space $\mathcal{V}$ having a norm $\| \ \|$. The superscript is not an exponent but is simply an index giving the order of the term in the sequence. As one would expect from the theory of sequences of real numbers (see §1.62),

$$\lim_{k \to \infty} \mathbf{x}^k = \mathbf{a}$$

has the meaning that if ϵ is any positive number, then there exists a positive integer N such that $\|\mathbf{x}^k - \mathbf{a}\| < \epsilon$ whenever $N \leqq k$. This is also expressed by saying that the sequence $\{\mathbf{x}^k\}_{k=1}^{\infty}$ converges to $\mathbf{a}$. In terms of our recently introduced notation, it can be expressed by saying that if ϵ is any positive number, then all but at most a finite number of the terms of the sequence lie in the ϵ-ball, $B(\mathbf{a}, \epsilon)$, centered at $\mathbf{a}$.

The definition of continuity still makes sense in all vector spaces having norms. Suppose that $\mathcal{X}$ and $\mathcal{Y}$ are normed vector spaces and $\mathbf{f}$ is a function from some subset D of $\mathcal{X}$ to $\mathcal{Y}$. To say that $\mathbf{f}$ is continuous at the point $\mathbf{a}$ of its domain D means that if ϵ is any positive number, then there exists some positive

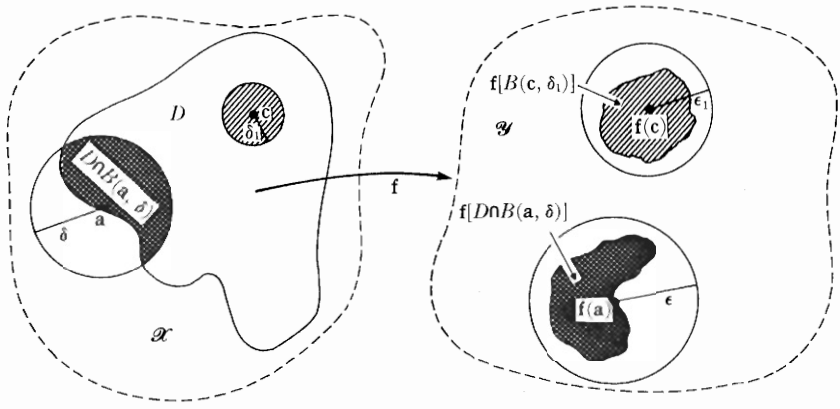


Fig. 84.

number δ such that

$$\|\mathbf{f}(\mathbf{a}) - \mathbf{f}(\mathbf{x})\| < \epsilon \text{ for all } \mathbf{x} \in D \text{ such that } \|\mathbf{x} - \mathbf{a}\| < \delta.$$

In geometrical terms this says that if $B(\mathbf{f}(\mathbf{a}), \epsilon)$ is any open ball, centered at $\mathbf{f}(\mathbf{a})$, in $\mathscr{Y}$, then there is some open ball $B(\mathbf{a}, \delta)$, centered at $\mathbf{a}$, in $\mathscr{X}$ such that $\mathbf{f}$ maps the intersection $D \cap B(\mathbf{a}, \delta)$ into $B(\mathbf{f}(\mathbf{a}), \epsilon)$. If $\mathbf{a}$ happens to be an interior point of D, then we don't have to talk about the intersection of D with $B(\mathbf{a}, \delta)$—we can simply say that for some δ, $\mathbf{f}$ maps $B(\mathbf{a}, \delta)$ into $B(\mathbf{f}(\mathbf{a}), \epsilon)$. Notice that the number δ may depend on both $\mathbf{a}$ and ϵ. See Fig. 84 for a diagram which may help to picture the idea of continuity. The diagram illustrates, with $\mathbf{c}$, δ_1, ϵ_1 in place of $\mathbf{a}$, δ, ϵ, the case in which $\mathbf{c}$ is an interior point of D and the ball $B(\mathbf{c}, \delta_1)$ is in D.

Notice that since continuity is defined with the help of the norm, the norm is automatically a continuous function. In fact, to say that the norm is continuous at $\mathbf{a}$ means only that if ϵ is any positive number, there is some positive number δ such that $|\|\mathbf{x}\| - \|\mathbf{a}\|| < \epsilon$ for all $\mathbf{x}$ such that $\|\mathbf{x} - \mathbf{a}\| < \delta$. By (11.5–2), $|\|\mathbf{x}\| - \|\mathbf{a}\|| \leqq \|\mathbf{x} - \mathbf{a}\|$, which shows that we can take δ to be ϵ.

In §5.3 we stated that the composition of continuous functions is continuous. We shall prove here a form of this statement which is appropriate for our purposes. Here and later we use the notation $g \circ f$ for the composite function whose value at $\mathbf{x}$ is $(g \circ f)(\mathbf{x}) = g(f(\mathbf{x}))$, it being assumed that f is defined at $\mathbf{x}$ and g is defined at $f(\mathbf{x})$. We call $g \circ f$ the composition of g and f (in that order).

THEOREM II. *Suppose that* $\mathbf{f}$ *is a function from some subset of* $\mathscr{X}$ *to* $\mathscr{Y}$ *and that* $\mathbf{g}$ *is a function from some subset of* $\mathscr{Y}$ *to* $\mathscr{Z}$*, where* $\mathscr{X}$*,* $\mathscr{Y}$*, and* $\mathscr{Z}$ *are normed vector spaces. Suppose further that* $\mathbf{f}$ *is defined at* $\mathbf{a}$ *and that* $\mathbf{g}$ *is defined at* $\mathbf{f}(\mathbf{a}) = \mathbf{b}$*. If* $\mathbf{f}$ *is continuous at* $\mathbf{a}$ *and* $\mathbf{g}$ *is continuous at* $\mathbf{b}$*, then the composition* $\mathbf{g} \circ \mathbf{f} = \phi$ *is continuous at* $\mathbf{a}$*.*

Proof. Let $\mathbf{g}(\mathbf{b}) = \mathbf{c}$. Since $\mathbf{g}$ is continuous at $\mathbf{b}$, if $B(\mathbf{c}, \epsilon)$ is any open ball of

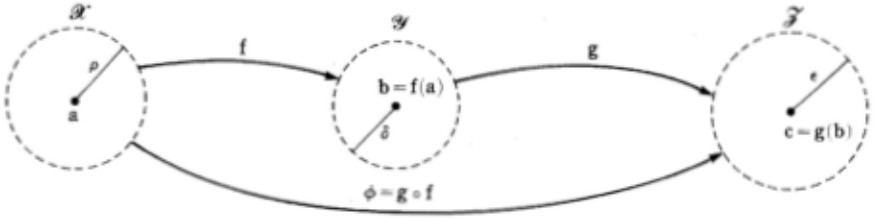


Fig. 85.

radius $\epsilon > 0$, centered at $\mathbf{c}$, there exists some $\delta > 0$ such that $\mathbf{g}(\mathbf{y}) \in B(\mathbf{c}, \epsilon)$ for all $\mathbf{y}$ in $B(\mathbf{b}, \delta)$ at which $\mathbf{g}$ is defined. See Fig. 85.

Since $\mathbf{f}$ is continuous at $\mathbf{a}$, there is some $\rho > 0$ such that $\mathbf{f}(\mathbf{x}) \in B(\mathbf{b}, \delta)$ for all $\mathbf{x} \in B(\mathbf{a}, \rho)$ at which $\mathbf{f}$ is defined. Therefore, for all those $\mathbf{x}$ in $B(\mathbf{a}, \rho)$ at which ϕ is defined, $\phi(\mathbf{x}) \in B(\mathbf{c}, \epsilon)$.

Observe that the points of $B(\mathbf{a}, \rho)$ at which ϕ is defined are those points $\mathbf{x}$ at which $\mathbf{f}$ is defined for which $\mathbf{f}(\mathbf{x})$ belongs to the domain of $\mathbf{g}$.

THEOREM III. *If T is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, then there exists some number M such that*

$$\|T\mathbf{x}\| \leq M\|\mathbf{x}\| \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

Proof. Let the standard matrix representation of T be the $m \times n$ matrix A as in §11.3. Then for each $x \in \mathbb{R}^n$,

$$T\mathbf{x} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

$\mathbf{y} = (y_1, y_2, \ldots, y_m)$ is that vector in $\mathbb{R}^m$ whose ith component y_i is given by

$$y_i = \sum_{j=1}^{n} a_{ij}x_j.$$

By the Cauchy inequality, (see (10.12–4))

$$|y_i| \leq \left(\sum_{j=1}^{n} a_{ij}^2\right)^{1/2}\left(\sum_{1}^{n} x_j^2\right)^{1/2} = \left(\sum_{j=1}^{n} a_{ij}^2\right)^{1/2}\|\mathbf{x}\|.$$

Now, let Q denote the maximum of the m numbers

$$\left(\sum_{j=1}^{n} a_{ij}^2\right)^{1/2} \qquad (i = 1, 2, \ldots, m).$$

Then $|y_i| \leq Q\|\mathbf{x}\|$ for all i and $\|T\mathbf{x}\| = \|\mathbf{y}\| =$

$$\sqrt{y_1^2 + y_2^2 + \cdots + y_m^2} \leq \sqrt{mQ^2\|\mathbf{x}\|^2} = \sqrt{m}\,Q\|\mathbf{x}\|.$$

This proves the theorem and shows that one possible choice for M is $\sqrt{m}Q$.

A transformation T for which there exists a number M such that $\|T\mathbf{x}\| \leqq M\|\mathbf{x}\|$ for all $\mathbf{x}$ is said to be *bounded*, and M is called a bound for the transformation. Later, we shall find it useful to have a bound M for a linear transformation in terms of the dominant element (or elements) of its matrix representation, A. Let

$$K = \max|a_{ij}|,$$

where $i \in (1, 2, \ldots, m)$ and $j \in (1, 2, \ldots, n)$. Then,

$$Q \leqq \sqrt{nK^2} = K\sqrt{n}.$$

Therefore we have that

$$\|T\mathbf{x}\| \leqq \sqrt{mn}K\|\mathbf{x}\| \tag{11.7-1}$$

for all $\mathbf{x} \in \mathbb{R}^n$.

COROLLARY. *If T is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, then T is continuous.*

Proof. By the preceding theorem, T is bounded; that is, there is some number M such that $\|T\mathbf{x}\| \leqq M\|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{R}^n$. Now consider any $\mathbf{a}$ in $\mathbb{R}^n$ and any positive ϵ. Let $\delta = \dfrac{\epsilon}{M}$, so that $\delta M = \epsilon$, and suppose that $\|\mathbf{x} - \mathbf{a}\| < \delta$. Then

$$\|T\mathbf{x} - T\mathbf{a}\| = \|T(\mathbf{x} - \mathbf{a})\| \leqq M\|\mathbf{x} - \mathbf{a}\| < \delta M = \epsilon.$$

Therefore T is continuous at $\mathbf{a}$, by the definition of continuity given earlier in this section. Observe that the choice of δ depends on ϵ but not on $\mathbf{a}$. The fact that δ is independent of $\mathbf{a}$ means that T has a property called *uniform continuity* on all $\mathbb{R}^n$.

11.8 / A NORM ON $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$

It is not at all obvious how to define the magnitude of a linear transformation. An approach which turns out to be quite useful in various connections is based on what the transformation T does to the lengths of vectors. Of course, T will usually have different effects on different vectors—lengthening some, shortening others, and leaving the lengths of others unchanged. To get around this complication, we might try defining the norm of T just in terms of its effect on that vector which it stretches the most; that is, we might consider the ratio $\|T\mathbf{x}\|/\|\mathbf{x}\|$ and take the norm of T to be the maximum of this ratio as x ranges over the space, excluding, of course, the zero vector where the ratio is not defined. This approach can be simplified by noticing that

$$\frac{\|T(\mathbf{x})\|}{\|\mathbf{x}\|} = \left\|T\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)\right\|.$$

Then, because $\mathbf{x}/\|\mathbf{x}\|$ is a unit vector, we see that the set of values taken on by the ratio $\|T\mathbf{x}\|/\|\mathbf{x}\|$ for all nonzero vectors is exactly the same as the set of values taken by the function $\|T(\mathbf{u})\|$ when $\mathbf{u}$ is restricted to lie just on the unit sphere (the set of all vectors having norm 1).

Now, let T be any linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. We want to show that

$$\|T\| = \max_{\|\mathbf{x}\|=1} \|T\mathbf{x}\| \tag{11.8–1}$$

defines a norm on the vector space $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$. But first notice that $\|\ \|$ carries a different meaning in each of its three appearances in (11.8–1). Since $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, $\|\mathbf{x}\|$ involves the length function on $\mathbb{R}^n$. $\|T\mathbf{x}\|$, on the other hand, refers to the length function on $\mathbb{R}^m$, and on the left-hand side, $\|\ \|$ is a function from $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ to the real numbers. It is general practice to use the same symbol, $\|\ \|$, for all norms and leave it to the reader to infer from the context which norm is meant.

Since both T and the norm are continuous functions, their composition, defined by $\|T(\mathbf{x})\|$, is a continuous function from $\mathbb{R}^n$ to $\mathbb{R}$ by Theorem II, §11.7. The unit sphere in $\mathbb{R}^n$ is both closed and bounded (Exercise 17) so by Theorems I and II of §5.3 the above function has a maximum value on this set. This shows that for each T,

$$\max_{\|\mathbf{x}\|=1} \|T(\mathbf{x})\|$$

exists and therefore (11.8–1) does define a function from $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ to the real numbers. It is obviously nonnegative, since the maximum of a nonnegative function is nonnegative. If T_0 is the zero element of $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, then $T_0(\mathbf{x}) = \mathbf{0}$ for all $\mathbf{x}$; hence,

$$\|T_0\| = \max_{\|\mathbf{x}\|=1} \|T_0(\mathbf{x})\| = 0.$$

If T is not the zero element, then there is some vector $\mathbf{v}$ in $\mathbb{R}^n$ such that $T(\mathbf{v}) \neq 0$. Then $\mathbf{u} = \mathbf{v}/\|\mathbf{v}\|$ is a unit vector such that

$$T(\mathbf{u}) = T\left(\frac{\mathbf{v}}{\|\mathbf{v}\|}\right) = \frac{1}{\|\mathbf{v}\|} T(\mathbf{v}) \neq 0,$$

and so $\|T(\mathbf{u})\| > 0$. Therefore

$$\|T\| = \max_{\|\mathbf{x}\|=1} \|T(\mathbf{x})\| \geqq \|T(\mathbf{u})\| > 0.$$

We have now proved that the function $\|\ \|$ defined by (11.8–1) satisfies Property (1) and (2) of §11.5. The fact that it satisfies (3) is easy to prove and will be left as an exercise for the reader. To prove that this function satisfies the fourth property required of every norm is slightly more difficult. What we must

prove is that, for all T_1 and T_2 in $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$,

$$\max_{\|\mathbf{x}\|=1} \|(T_1 + T_2)\mathbf{x}\| \leq \|T_1\| + \|T_2\| \tag{11.8–2}$$

for the left side here is $\|T_1 + T_2\|$. Now, by definition, $(T_1 + T_2)\mathbf{x} = T_1\mathbf{x} + T_2\mathbf{x}$, and so, for any unit vector $\mathbf{x}$,

$$\|(T_1 + T_2)\mathbf{x}\| = \|T_1\mathbf{x} + T_2\mathbf{x}\| \leq \|T_1\mathbf{x}\| + \|T_2\mathbf{x}\| \leq \|T_1\| + \|T_2\|,$$

for certainly $\|T_1\mathbf{x}\| \leq \|T_1\|$ by the definition of $\|T_1\|$, and likewise for T_2. But then we see at once from the foregoing that (11.8–2) is true. We have now proved that the norm defined by (11.8–1) does indeed satisfy the triangle inequality.

An important property of this norm is that, for all $\mathbf{x}$,

$$\|T\mathbf{x}\| \leq \|T\|\,\|\mathbf{x}\|. \tag{11.8–3}$$

The proof of this fact (Exercise 13) is short. The reader should pause long enough now to distinguish clearly among the three meanings which the norm symbol has in (11.8–3).

The norm which we have defined for linear transformations is useful even though we have no way of evaluating it except in certain special cases. It will be sufficient for our purposes to have a practical bound for the norm, and we have already obtained one in the process of proving Theorem III of §11.7. It was found in (11.7–1) that if $T \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ and A is an $m \times n$ matrix representing T, then

$$\|T\| \leq \sqrt{mn}\,K,$$

where $K = \max_{\substack{1\leq i\leq m \\ 1\leq j\leq n}} |a_{ij}|$.

We get an important lower bound for $\|T\|$ by considering what T does to the unit vectors making up the standard basis in $\mathbb{R}^n$. Recall that for each $r \in \{1, 2, \ldots, n\}$, $\mathbf{e}_r$ is that ordered n-tuple consisting of 1 in the rth place and 0's in the other $n - 1$ places. If C_r is the norm of $T\mathbf{e}_r$,

$$\|T\mathbf{e}_r\| = \|A\mathbf{e}_r\| = \|(a_{1r}, a_{2r}, \ldots, a_{mr})\| = \left(\sum_{i=1}^{m} a_{ir}^2\right)^{1/2} = C_r.$$

C_r can be thought of as the length of the vector represented by the rth column in A. Let $C = \max C_r$ for $r \in \{1, 2, \ldots, n\}$. In other words, C is the length of the longest of the n column vectors in A. Since $\|Tu\| = C$ for at least one unit vector, $\|T\| \geq C$. From the definition of C_r we see that C_r is greater than or equal to $\max_i |a_{ir}|$. Hence, $C = \max_r C_r$ is greater than or equal to $\max_r(\max_i |a_{ir}|) = K$, and so $\|T\| \geq C \geq K$. Combining this with our previously obtained upper bound, we have

$$K \leq \|T\| \leq \sqrt{mn}\,K. \tag{11.8–4}$$

The space $\mathscr{L}(\mathbb{R}^n, \mathbb{R})$ is an interesting one from the point of view of computing norms. Recall that the elements of this space are called linear functionals and

that each one of them is represented by a unique vector in $\mathbb{R}^n$. Suppose that T is a linear functional on $\mathbb{R}^n$ and that $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ is its representing vector. The norm $\mathbf{a}$, of course, is $\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$. The norm T is the maximum value of $\|T\mathbf{x}\|$ on the unit sphere, that is, on the set of vectors having unit norm in $\mathbb{R}^n$.

$$\|T\| = \max_{\|\mathbf{x}\|=1} \|T\mathbf{x}\| = \max_{\|\mathbf{x}\|=1} |\mathbf{a} \cdot \mathbf{x}| = \max_{\|\mathbf{x}\|=1} |a_1x_1 + \cdots + a_nx_n|.$$

By the Cauchy inequality,

$$|a_1x_1 + \cdots + a_nx_n| \leqq \left(\sum_1^n a_i^2\right)^{1/2} \|\mathbf{x}\| = \|\mathbf{a}\| \cdot \|\mathbf{x}\|.$$

So for all unit vectors, $\|T\mathbf{x}\| \leq \|\mathbf{a}\|$. Assuming that T is not the zero functional, $\mathbf{a}$ is not the zero vector and $\mathbf{u} = \mathbf{a}/\|\mathbf{a}\|$ is a unit vector. A simple calculation shows that $\|T\mathbf{u}\| = \|\mathbf{a}\|$. Therefore,

$$\max_{\|\mathbf{x}\|=1} \|T\mathbf{x}\| = \|\mathbf{a}\|, \text{ and hence, } \|T\| = \|\mathbf{a}\|.$$

This says that the norm of each nonzero linear functional on $\mathbb{R}^n$ is the same as the $\mathbb{R}^n$ norm of its representing vector. Since this statement about equality of norms is obviously also true for the zero functional, it is true in general.

11.9 / $\mathscr{L}(\mathbb{R}^n)$

In the important special case where m and n are same, $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^n)$ is abbreviated to $\mathscr{L}(\mathbb{R}^n)$, and it denotes the vector space of all linear transformations (in this case operators) from $\mathbb{R}^n$ to $\mathbb{R}^n$. (Recall that in this book we use the term linear operator to denote a linear transformation from a vector space to itself.) From Theorem I we know that the operators in $\mathscr{L}(\mathbb{R}^n)$ are represented by $n \times n$ matrices. An important subset of $\mathscr{L}(\mathbb{R}^n)$ consists of the *invertible* operators. These are one-to-one functions and are also described as *nonsingular.* Saying that T is an invertible operator means that for each $\mathbf{y}$ in $\mathbb{R}^n$, there is a unique $\mathbf{x}$ in $\mathbb{R}^n$ such that $T\mathbf{x} = \mathbf{y}$. The unique $\mathbf{x}$ associated with each $\mathbf{y}$ in this way is denoted by $T^{-1}\mathbf{y}$. This function T^{-1} is called the inverse of T. The linearity of T implies that of T^{-1} (Exercise 3); therefore, if T is an invertible member of $\mathscr{L}(\mathbb{R}^n)$, T^{-1} also belongs to $\mathscr{L}(\mathbb{R}^n)$.

One way of telling whether a linear operator is invertible is to look at the equation $T\mathbf{x} = \mathbf{y}$ in matrix form. If A is a matrix representing T, we have $A\mathbf{x} = \mathbf{y}$, or in scalar form,

$$\begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

We shall assume that the reader has had enough experience with simultaneous linear systems like this to know that: *For each* $\mathbf{y}$ *in* $\mathbb{R}^n$, *a necessary and sufficient condition that there exist a unique solution* $\mathbf{x}$ *is that the determinant of the coefficient matrix be different from zero.* Such matrices are called *nonsingular* or *invertible.* So the invertible linear operators are those which are represented by invertible (or nonsingular) matrices.

Another important characterization of invertible operators is obtained by applying the italicized proposition of the preceding paragraph to the special case where all the y's in the above system are zero. In this case, the linear system is said to be *homogeneous.* It is obvious at a glance that a homogeneous system has the trivial solution $x_1 = x_2 = \cdots = x_n = 0$. If the determinant is different from zero, this trivial solution must be the only one—in other words, a nonsingular operator T maps only the zero vector into the zero vector. But if the determinant is zero, that is, if the operator is singular, then the trivial solution is not unique and this means that T maps some nonzero vector into the zero vector. *Hence, a necessary and sufficient condition that a linear operator be invertible is that it map only the zero vector into the zero vector.* It therefore follows that if a linear operator maps only the zero vector into the zero vector, then it must be invertible, and this is the property which we shall use in §11.10 to prove that certain elements of $\mathscr{L}(\mathbb{R}^n)$ are invertible.

Suppose that T is invertible. We have defined $T^{-1}\mathbf{y}$ to be the unique solution of the equation $T\mathbf{x} = \mathbf{y}$. Since there is a unique solution for all $\mathbf{y}$, we have

$$T(T^{-1}\mathbf{y}) = \mathbf{y} \qquad \text{for all } \mathbf{y}.$$

This means that the composition $T \circ T^{-1} = I$, where I is the identity transformation (i.e., I maps each vector into itself).

As is the case with all invertible functions, T^{-1} is also invertible, and in fact, $(T^{-1})^{-1} = T$. This means that for all $\mathbf{y}$, the equation $T^{-1}\mathbf{x} = \mathbf{y}$ has the unique solution $T\mathbf{y}$ (i.e., $T^{-1}(T\mathbf{y}) = \mathbf{y}$ for all $\mathbf{y}$). Another way to say this is that $T^{-1} \circ T = I$. This and the preceding paragraph together give us that

$$T \circ T^{-1} = T^{-1} \circ T = I.$$

In other words, every invertible linear operator commutes with its inverse and their composition is the identity operator. Conversely, it is easy to show (Exercise 19) that if two linear operators commute and their composition is the identity operator, then each is the inverse of the other. From this and the fact that composition is associative (Exercise 17), it is easy to prove (Exercise 18) that if T and L are invertible then their composition is invertible, and that

$$(T \circ L)^{-1} = L^{-1} \circ T^{-1}.$$

That is, the inverse of the composition of two invertible operators is the composition of their inverses *in the opposite order.*

Suppose that A and A^{-1} are the matrix representations of T and T^{-1}, respectively. Now, matrix multiplication was defined in such a way as to make the product of two matrices represent the composition, in the same order, of the

operators which they represent; it follows, therefore, that

$$AA^{-1} = A^{-1}A = I.$$

Here we have used the same symbol, I, to represent both the identity operator and the so-called identity matrix which represents it. Thus, every nonsingular matrix commutes with its inverse, and the product is the identity matrix. From the fact that

$$I\mathbf{x} = \mathbf{x} \qquad \text{for all } \mathbf{x},$$

it is easily seen that I is the $n \times n$ matrix with 1's along the main diagonal and 0's elsewhere.

If L and T are any two members of $\mathscr{L}(\mathbb{R}^n)$, then $L \circ T \in \mathscr{L}(\mathbb{R}^n)$ and by (11.8–3)

$$\|(L \circ T)(\mathbf{x})\| = \|L(T(\mathbf{x}))\| \leqq \|L\| \cdot \|T(\mathbf{x})\| \qquad \text{for all } \mathbf{x}.$$

Using (11.8–3) again, we get

$$\|(L \circ T)(\mathbf{x})\| \leqq \|L\|\|T\|\|\mathbf{x}\| \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

By the way we defined the norm of a linear transformation, this says that

$$\|L \circ T\| \leqq \|L\|\,\|T\|. \tag{11.9–1}$$

In words, the norm of the composition of two linear operators is less than or equal to the product of their norms. Suppose now that T happens to be invertible. Then $T \circ T^{-1} = I$, and since it is obvious that $\|I\| = 1$, (11.9–1) gives

$$\|T\|\,\|T^{-1}\| \geqq 1, \tag{11.9–2}$$

or, the norm of the inverse of a linear operator is greater than or equal to the reciprocal of the norm of the operator.

Since the inverse T^{-1} of an invertible linear operator T is also linear (Exercise 3), it too has a matrix representation. If A is the matrix which represents T, then the matrix representation of T^{-1} is denoted by A^{-1}. There are ways of constructing A^{-1} when A is given, but we leave most of this to a course in matrix algebra. However, we shall have occasion later to invert a 2×2 matrix. This is easily accomplished by the following three-step process:

(i) Exchange positions of the two main diagonal elements.
(ii) Change the signs of the two off-diagonal elements.
(iii) Now divide each of the four elements by the determinant.

In other words, if

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad \text{then } A^{-1} = \begin{pmatrix} \dfrac{d}{\Delta} & \dfrac{-b}{\Delta} \\ \dfrac{-c}{\Delta} & \dfrac{a}{\Delta} \end{pmatrix},$$

where $\Delta = ad - bc$. Remember that, by Cramer's rule, $\Delta \neq 0$ if and only if A is

invertible. It is easily verified that the displayed matrices are indeed mutually inverse.

11.10 / THE SET OF INVERTIBLE OPERATORS

Let us denote by Ω the set of invertible operators belonging to the vector space $\mathscr{L}(\mathbb{R}^n)$. In Chapter 12 we shall use the important fact that Ω is an open set in $\mathscr{L}(\mathbb{R}^n)$. This means that if $T \in \Omega$ and L is an operator such that $\|T - L\|$ is sufficiently small, then $L \in \Omega$. To prove this we first prove some preliminary results.

LEMMA. *If $T \in \mathscr{L}(\mathbb{R}^n)$ and $\|T\| < 1$, then $I - T$ is invertible, i.e. $(I - T) \in \Omega$, and*

$$\|(I - T)^{-1}\| \leq \frac{1}{1 - \|T\|}.$$

Proof. Consider any $x \neq 0$. We shall show that $(I - T)x \neq 0$, which implies that $I - T$ is invertible. Now $\|(I - T)x\| = \|x - Tx\| \geq \big|\|x\| - \|Tx\|\big|$, by (11.5–3). But $\|Tx\| \leq \|T\|\,\|x\|$. Therefore

$$\|(I - T)x\| \geq \|x\|(1 - \|T\|) > 0 \tag{11.10–1}$$

because $\|T\| < 1$ and $\|x\| > 0$. Therefore $(I - T)^{-1}$ exists.

To estimate the norm of $(I - T)^{-1}$ we can substitute for x in (11.10–1) the vector $(I - T)^{-1}y$, where y is an arbitrary vector in $\mathbb{R}^n$. On the left we get $\|(I - T)(I - T)^{-1}y\| = \|Iy\| = \|y\|$.

Therefore, $\|y\| \geq \|(I - T)^{-1}y\|(1 - \|T\|)$, or

$$\|(I - T)^{-1}y\| \leq \frac{\|y\|}{1 - \|T\|} \qquad \text{for all } y \in \mathbb{R}^n.$$

This gives the inequality for $\|(I - T)^{-1}\|$ stated in the lemma.

COROLLARY. *If $\|I - T\| < 1$, then T is invertible.*

We deduce the corollary as an application of the lemma by putting $I - T$ in place of T in the lemma, observing that $I - (I - T) = T$. Note that the corollary states that the entire open ball centered at I, with radius 1, lies in Ω.

THEOREM IV. *The set of invertible operators is an open set in $\mathscr{L}(\mathbb{R}^n)$. In fact, if $T \in \Omega$ and $\|T - L\| < 1/\|T^{-1}\|$, then $L \in \Omega$. Moreover,*

$$\|L^{-1}\| \leq \frac{\|T^{-1}\|}{1 - \|T^{-1}(T - L)\|}. \tag{11.10–1}$$

Proof. We use the fact that T is invertible to write $L = T - (T - L) = T[I - T^{-1}(T - L)]$. By (11.9–1) and our hypothesis,

$$\|T^{-1}(T - L)\| \leq \|T^{-1}\|\,\|T - L\| < 1.$$

Using the lemma (with $T^{-1}(T-L)$ in place of T) we conclude that $I - T^{-1}(T-L)$ is invertible. Since L is the composition of the two operators T and $I - T^{-1}(T-L)$, both of which are now known to be invertible, we have

$$L^{-1} = [I - T^{-1}(T-L)]^{-1}T^{-1}.$$

By 11.9–1 and an application of the inequality in the lemma, we obtain (11.10–1).

Notice that Theorem IV tells us that each invertible operator T in $\mathscr{L}(\mathbb{R}^n)$ is the center of an open ball whose radius is the reciprocal of the norm of the inverse of T, and each of whose points is itself an invertible operator.

THEOREM V. *The function from Ω to Ω which associates with $T \in \Omega$ its inverse, T^{-1}, is continuous.*

Proof. We have to show that, if $T \in \Omega$ we can make $\|L^{-1} - T^{-1}\|$ as small as we like by requiring $\|L - T\|$ to be suitably small. Theorem IV gives us nearly all we need. To begin with, suppose $\|L - T\| < \dfrac{1}{\|T^{-1}\|}$. Then we know that L^{-1} exists. Observe that

$$T^{-1}(T-L)L^{-1} = (I - T^{-1}L)L^{-1} = L^{-1} - T^{-1},$$

and therefore, by a couple of applications of (11.9–1),

$$\|L^{-1} - T^{-1}\| \leqq \|T^{-1}\|\|T - L\|\|L^{-1}\|.$$

Then, by (11.10–1),

$$\|L^{-1} - T^{-1}\| \leqq \|T^{-1}\|\|T - L\| \frac{\|T^{-1}\|}{1 - \|T^{-1}(T-L)\|}.$$

But $\|T^{-1}(T-L)\| \leqq \|T^{-1}\|\|T-L\|$. Let us bring L still closer to T by requiring that $\|L - T\| < \dfrac{1}{2\|T^{-1}\|}$. Then we see from the foregoing that $1 - \|T^{-1}(T-L)\| > 1/2$ and therefore

$$\|L^{-1} - T^{-1}\| \leqq 2\|T^{-1}\|^2\|T - L\|.$$

From this we immediately draw the desired conclusion.

We conclude by asking: Is Ω a vector space within $\mathscr{L}(\mathbb{R}^n)$? Is the correspondence between T and T^{-1} linear?

EXERCISES

1. Let A be an $m \times n$ matrix. Show that the function $\mathbf{f}: \mathbb{R}^n \to \mathbb{R}^m$ defined in §11.3 by $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$ is a linear transformation.

2. Given that $T \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ and $L \in (\mathbb{R}^m, \mathbb{R}^p)$, prove directly from the definition of linearity that $L \circ T \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^p)$.

3. Prove directly from the definition of a linear transformation that the inverse of an invertible linear operator is a linear operator.

4. (a) Show that the transpose of the sum of two matrices is equal to the sum of their transposes.

(b) If the vector $\mathbf{x}$ in $\mathbb{R}^m$ is thought of as an $(m \times 1)$ matrix, then its transpose $\mathbf{x}^T$ is a $(1 \times m)$ matrix. Show that for any $(n \times m)$ matrix A and any $\mathbf{x} \in \mathbb{R}^m$,

$$(A\mathbf{x})^T = \mathbf{x}^T A^T.$$

Then generalize this to show that

$$(AB)^T = B^T A^T$$

in all cases where A and B are matrices such that their product, AB, is defined.

(c) If A is an $(n \times m)$ matrix, $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$, then

$$(\mathbf{y}^T A)\mathbf{x} = \mathbf{y}^T (A\mathbf{x}),$$

proving that matrix multiplication is associative in this very special case.

(d) Show that for every matrix A,

$$(A^T)^T = A.$$

5. Show that (11.5–1) and property (3) of a norm in §11.5 imply property (4) of that norm by making suitable replacements for $\mathbf{x}$ and $\mathbf{y}$ in (11.5–1).

6. For each $\mathbf{x} \in \mathbb{R}^n$, let

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

Show that for each $p \geq 1$, $\| \ \|_p$ is a norm on $\mathbb{R}^n$. HINT: Use Minkowski's inequality (Exercise 32, §6.8).

7. Let $\| \ \|_1$ and $\| \ \|_m$, with the subscript m standing for "max," be two functions from $\mathbb{R}^n$ to $\mathbb{R}$ defined as follows:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

$$\|\mathbf{x}\|_m = \max_{1 \leq i \leq n} |x_i|.$$

Show that $\| \ \|_1$ and $\| \ \|_m$ are norms on $\mathbb{R}^n$. For the special case where $n = 2$, draw the unit sphere (circle) in these two norms on the same co-ordinate axes with the unit sphere in the Euclidean norm.

8. Prove that if f and g are continuous, real-valued functions on $[a, b]$, then

$$\left| \int_a^b f(x)g(x)\,dx \right| \leq \left(\int_a^b f^2(x)\,dx \right)^{1/2} \left(\int_a^b g^2(x)\,dx \right)^{1/2}.$$

What condition is both necessary and sufficient for equality? HINT: Since

$$\int_a^b [\lambda f(x) + g(x)]^2\,dx \geq 0 \qquad \text{for all } \lambda,$$

the quadratic equation in λ,

$$\lambda^2 \int_a^b f^2(x)\,dx + 2\lambda \int_a^b f(x)g(x)\,dx + \int_a^b g^2(x)\,dx = 0$$

cannot have two distinct real roots, and therefore the discriminant cannot be positive.

This inequality, which is reminiscent of the Cauchy inequality, is known as the Schwarz inequality. Notice that the Cauchy inequality can be proved by the method used here, starting from the fact that $\sum_{i=1}^{n} (\lambda a_i + b_i)^2 = 0$ can be written as a quadratic equation in λ which obviously cannot have distinct real roots. The similarity between these two inequalities has led many authors to lump them together under the same name—the Cauchy-Schwarz inequality.

We have assumed that f, g, f^2, g^2, fg, and $(\lambda f + g)^2$ are all integrable. These things are proved in Chapter 18.

9. Let $\mathscr{C}$ denote the vector space of continuous functions on $[0, 1]$ (see Example 2, §11) and let $\| \ \|_1$ and $\| \ \|_2$ be functions from $\mathscr{C}$ to $\mathbb{R}$ defined as follows:

$$\|f\|_1 = \max_{0 \le x \le 1} |f(x)|,$$

$$\|f\|_2 = \left[\int_0^1 f^2(x)\, dx \right]^{1/2}.$$

Prove that $\| \ \|_1$ and $\| \ \|_2$ are norms on $\mathscr{C}$. HINT: For $\| \ \|_2$ use Exercise 8.

10. Prove that if $\| \ \|$ is any norm on a vector space, $\mathscr{V}$, then the function d defined by

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| \qquad \text{for all } \mathbf{x} \text{ and } \mathbf{y} \text{ in } \mathscr{V}$$

is a metric.

11. Prove that if $\mathscr{V}$ is any vector space and d is the function defined by

$$d(\mathbf{x}, \mathbf{y}) = 0 \qquad \text{if } \mathbf{x} = \mathbf{y}, \qquad \text{and}$$
$$d(\mathbf{x}, \mathbf{y}) = 1 \qquad \text{otherwise,}$$

then d is a metric. Prove that there cannot exist any norm, $\| \ \|$, on $\mathscr{V}$ such that this particular metric is given by $\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathscr{V}$.

12. Let ρ denote the usual distance function in the plane, $\mathbb{R}^2$, and define

$$d(\mathbf{x}, \mathbf{y}) = \frac{\rho(\mathbf{x}, \mathbf{y})}{1 + \rho(\mathbf{x}, \mathbf{y})}.$$

Show that d is also a metric on the plane. Show that it is not possible to define a norm on the plane such that d is expressible in terms of this norm as in Exercise 10.

13. Prove (11.8–3).

14. Let $S(0, 1)$ denote the Euclidean unit sphere in $\mathbb{R}^n$, that is

$$S(0, 1) = \{\mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^{n} x_i^2 = 1\}.$$

Prove that $S(0, 1)$ is closed.

15. Let $\{\mathbf{x}^k\}_{k=1}^{\infty}$ denote a sequence in $\mathbb{R}^n$, that is, $\mathbf{x}^k = (x_1^k, x_2^k, \ldots, x_n^k)$. Prove that a necessary and sufficient condition that

$$\lim_{k \to \infty} \mathbf{x}^k = \mathbf{y} = (y_1, y_2, \ldots, y_n)$$

is that $\lim_{k \to \infty} x_i^k = y_i$ for $i = 1, 2, \ldots, n$.

16. If $\mathbf{x}$ and $\mathbf{y}$ are any two points in $\mathbb{R}^n$, then the equation of the straight line

determined by them can be written

$$\mathbf{z} = \mathbf{x} + t(\mathbf{y} - \mathbf{x}).$$

For $n = 2$ you may have thought of this as the parametric representation of the line, t being the parameter. In general, the right-hand side is a function from $\mathbb{R}$ to $\mathbb{R}^n$. Notice that the function maps the interval $0 \leqq t \leqq 1$ into that linear segment between $\mathbf{x}$ and $\mathbf{y}$ inclusive. This means that we can define *the line segment determined by* $\mathbf{x}$ *and* $\mathbf{y}$ to be the set of points of the form $(1-t)\mathbf{x} + t\mathbf{y}$ for $t \in [0, 1]$. Equivalently, we can define this segment to be the set of points

$$\lambda_1\mathbf{x} + \lambda_2\mathbf{y},$$

where $\lambda_1 \geqq 0$, $\lambda_2 \geqq 0$ and $\lambda_1 + \lambda_2 = 1$. To say that a subset E of a vector space is *convex* means that if $\mathbf{x}$ and $\mathbf{y}$ belong to E, then the line segment determined by $\mathbf{x}$ and $\mathbf{y}$ also lies in E. Prove that every ball—open or closed—is convex.

17. Suppose that $f : A \to B$; $g : B \to C$; and $h : C \to D$. Since composition of functions is a binary operation, we get a function from A to D by forming either $h \circ (g \circ f)$ or $(h \circ g) \circ f$. Show that these two functions are the same. In other words show that although composition of functions is not necessarily commutative, it is always associative.

18. Using Exercise 17 show that if L and T are invertible linear operators, then $L \circ T$ is invertible, and $(L \circ T)^{-1} = T^{-1} \circ L^{-1}$. That is, the inverse of the composition of two invertible operators is the composition of their inverses *in the opposite order.*

19. Show that if T and L are two members of $\mathscr{L}(\mathbb{R}^n)$ such that $T \circ L = L \circ T = I$, then they must be invertible and each is the inverse of the other.

20. Let A and B denote linear operators on $\mathbb{R}^n$. Prove that if B is invertible and A commutes with B, then A commutes with B^{-1}.

21. Show that if $T \in \mathscr{L}(\mathbb{R}^n)$ and $\|T\| < 1$, then

$$(I - T)^{-1} = I + T + T^2 + \cdots + T^k + T^{k+1}(I - T)^{-1}$$

for every positive integer k.

22. Show that the set of invertible operators is not a bounded subset of $\mathscr{L}(\mathbb{R}^n)$.

23. Show that the function $f(T) = T^{-1}$, defined for $T \in \Omega$ (as defined in §11.10), is not uniformly continuous. That is, show that in choosing δ so that $\|T - T_0\| < \delta$ implies $\|T^{-1} - T_0^{-1}\| < \epsilon$, where T_0 and ϵ are preassigned, δ cannot be chosen independently of T_0.

12 / DIFFERENTIAL CALCULUS OF FUNCTIONS FROM $\mathbb{R}^n$ to $\mathbb{R}^m$

12 / INTRODUCTION

The purpose of this chapter is to unify much of our previous work by developing differential calculus in the general setting of functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. This will include the subjects of transformations introduced in Chapter 9, real-valued functions of several variables (i.e., functions from $\mathbb{R}^n$ to $\mathbb{R}$), vector-valued functions of a single variable (i.e., functions from $\mathbb{R}$ to $\mathbb{R}^n$), and even the one-variable functions of elementary calculus as special cases. A function from $\mathbb{R}^n$ to $\mathbb{R}^m$ is an ordered m-tuple of real-valued functions of n real variables, for example,

$$\begin{aligned} y_1 &= f^{(1)}(x_1, x_2, \ldots, x_n) \\ y_2 &= f^{(2)}(x_1, x_2, \ldots, x_n) \\ &\vdots \\ y_m &= f^{(m)}(x_1, x_2, \ldots, x_n). \end{aligned} \tag{12–1}$$

Here we have used superscripts to distinguish the *component functions* (or *co-ordinate functions*, as they are sometimes called), so that we can continue to use subscripts to denote partial derivatives. The vector $(x_1, x_2, \ldots, x_n)$ of $\mathbb{R}^n$ is mapped into the vector $(y_1, y_2, \ldots, y_m)$ of $\mathbb{R}^m$. If we denote the first of these by $\mathbf{x}$ and the second by $\mathbf{y}$, and if we let $\mathbf{f}$ denote the ordered m-tuple of functions $(f^{(1)}, f^{(2)}, \ldots, f^{(m)})$, then we can abbreviate the notation of our function (12–1) to

$$\mathbf{y} = \mathbf{f}(\mathbf{x}).$$

Such functions sometimes go by the name of transformations, as in Chapter 9.

The special case where all of the $f^{(i)}$'s are linear has already been studied, in Chapter 11. In that case, (12–1) reduces to

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \quad & \quad \vdots \quad \quad \vdots \\ y_m &= a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n. \end{aligned} \tag{12–2}$$

This is frequently reduced to the matrix equation

$$\mathbf{y} = A\mathbf{x},$$

where A is the $m \times n$ coefficient matrix. It is important to think of (12–2) as a special case of (12–1), and to see that this special case consists of all the linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$. This is the vector space which we have already studied under the name $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$.

The relevance of linear algebra to differential calculus consists in the fact that it is possible to obtain very good local approximations to quite general functions, such as (12–1), by using linear functions, such as (12–2). We shall see that this enables us to deduce important information about the behavior of a nonlinear function near a point by studying the linear (or affine) functions of best approximation at that point. We begin by extending the idea of a differential to our more general setting.

12.1 / THE DIFFERENTIAL AND THE DERIVATIVE

Our first objective here is to extend in a suitable way for functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ the definition given in §6.4 of differentiability and the differential for functions from $\mathbb{R}^2$ or $\mathbb{R}^n$ to $\mathbb{R}$ [see (6.4–4) and (6.4–17)]. Our second objective is to extend in a suitable way for functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ the relationship between differentials and derivatives that exists in the case of a function from $\mathbb{R}$ to $\mathbb{R}$, as set forth in §1.3. Then we shall show that a differentiable function from $\mathbb{R}^n$ to $\mathbb{R}^m$ is continuous, and we shall state and prove the general chain rule for differentiable functions.

In elementary calculus it has long been customary to introduce the derivative first, and then the differential. For functions from $\mathbb{R}^n$ to $\mathbb{R}^m$, where $n > 1$, it is natural to begin with the differential and come to the derivative afterward. In fact, the concept of the derivative when $n > 1$ is more sophisticated than the concept of the derivative in §1.3; it requires us to think of the derivative as a function from $\mathbb{R}^n$ to $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. But when $n = m = 1$ the more sophisticated point of view is in full harmony with the elementary point of view in §1.3.

In §6.4, and later in §7, we discussed the notion of the differential of a real function of several real variables. The differential of a function f from $\mathbb{R}^n$ to $\mathbb{R}$ is a function of $(x_1, \ldots, x_n)$ and $(dx_1, \ldots, dx_n)$ whose value is

$$\frac{\partial f}{\partial x_1}\, dx_1 + \cdots + \frac{\partial f}{\partial x_n}\, dx_n, \tag{12.1–1}$$

where the partial derivatives are evaluated at $(x_1, \ldots, x_n)$. Thus the differential is a linear function of $(dx_1, \ldots, dx_n)$ when we keep $(x_1, \ldots, x_n)$ fixed. But the definition of the differential requires more of f than merely that it have first partial derivatives with respect to each of the variables $x_1, \ldots, x_n$. If we use the vector notation

$$\mathbf{x} = (x_1, \ldots, x_n), \quad \mathbf{h} = (h_1, \ldots, h_n),$$

the function f from $\mathbb{R}^n$ to $\mathbb{R}$, defined in a neighborhood of $\mathbf{x}$, is said to be differentiable at $\mathbf{x}$ if there exist numbers $A_1, \ldots, A_n$, depending on $\mathbf{f}$ and $\mathbf{x}$, such

that

$$\lim_{\|\mathbf{h}\|\to 0} \frac{|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \sum_{i=1}^{n} A_i h_i|}{\|\mathbf{h}\|} = 0. \tag{12.1–2}$$

It is then necessarily the case that f has first partial derivatives at $\mathbf{x}$ given by

$$\frac{\partial f}{\partial x_i} = A_i. \tag{12.1–3}$$

When f is differentiable, its differential, which is the linear function of $\mathbf{h}$ with value $\Sigma_{i=1}^{n} A_i h_i$, can be written in the form (12.1–1) if we put $h_i = dx_i$. What we have given here is merely a restatement of what was given in connection with (6.4–17).

A student who wishes to review the discussion of differentials earlier in this book will find it appropriate to look at the early part of §1.3, parts of §6.4, and §7. The crucial portions are at (1.3–1), (6.4–4), (6.4–17), and (7–2).

For the general case of a function $\mathbf{f}$ from a neighborhood of some point $\mathbf{x}$ in $\mathbb{R}^n$ to $\mathbb{R}^m$ the definition of differentiability and the differential is modeled very closely after the case in which $m = 1$. The guiding idea is to seek a linear function of $\mathbf{h}$ from $\mathbb{R}^n$ to $\mathbb{R}^m$ to take the place of $\Sigma_{i=1}^{n} A_i h_i$ in (12.1–2), and to use the norm instead of the absolute value used in the numerator of (12.1–2). The essential thing is to regard the linear function of $\mathbf{h}$ as an approximation to the difference $\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})$, and to express the sense in which the approximation is better and better as $\|\mathbf{h}\| \to 0$. Now, a linear function from $\mathbb{R}^n$ to $\mathbb{R}^m$ is simply a member of $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$. So, for a given $\mathbf{x}$, we are seeking a $T \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, *depending on* $\mathbf{x}$, of such a nature that $T\mathbf{h}$ is a good approximation to $\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})$ when $\|h\|$ is small. The requirement we impose is that the difference

$$\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - T\mathbf{h}$$

be small in comparison with $\mathbf{h}$ when $\mathbf{h}$ is small. The precise formulation of the requirement is that

$$\lim_{\|\mathbf{h}\|\to 0} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - T\mathbf{h}\|}{\|\mathbf{h}\|} = 0. \tag{12.1–4}$$

When this condition is satisfied for the given $\mathbf{x}$ we say that $\mathbf{f}$ is differentiable at $\mathbf{x}$, and that the differential of $\mathbf{f}$ is the function of $\mathbf{x}$ and $\mathbf{h}$ whose value is $T\mathbf{h}$. The differential is defined for each $\mathbf{x}$ at which the differentiability requirement is satisfied and for every $\mathbf{h}$ in $\mathbb{R}^n$. The value of the differential is the vector $T\mathbf{h}$ in $\mathbb{R}^m$. We denote the differential as a function by $d\mathbf{f}$ and its value by $d\mathbf{f}(\mathbf{x}, \mathbf{h})$.

It is important to know that, when (12.1–4) is satisfied by a certain T in $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, that T is unique. Such is indeed the case. The proof is left as an exercise (Exercise 1). In a later section we shall see in a different way that T is unique, by finding its standard matrix representation. The elements of the matrix that represents T are first partial derivatives of the component functions of $\mathbf{f}$, as given in (12–1).

As indicated earlier, T may be expected to depend on $\mathbf{x}$. To indicate this dependency we write $T_{\mathbf{x}}$. The use of $\mathbf{x}$ as a subscript is to be distinguished carefully from a notation such as $T\mathbf{x}$ or $T(\mathbf{x})$, which would indicate that T is *acting* on $\mathbf{x}$, transforming it into a vector in $\mathbb{R}^m$.

When $\mathbf{f}$ is differentiable at one or more points there is an associated function from $\mathbb{R}^n$ to $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$. Its domain is the set of those $\mathbf{x}$'s at which $\mathbf{f}$ is differentiable, and the value of the function at $\mathbf{x}$ is $T_{\mathbf{x}}$. We shall call this function the *derivative* of $\mathbf{f}$ and denote it by $\mathbf{f}'$. Thus, by definition,

$$\mathbf{f}'(\mathbf{x}) = T_{\mathbf{x}}. \tag{12.1–5}$$

With this definition we can say that $\mathbf{f}$ has a derivative $\mathbf{f}'(\mathbf{x})$ at $\mathbf{x}$ if and only if $\mathbf{f}'(\mathbf{x})$ is an element of $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ such that

$$\lim_{\|\mathbf{h}\| \to 0} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|} = 0. \tag{12.1–6}$$

We must remember that the notation $\mathbf{f}'(\mathbf{x})\mathbf{h}$ signifies that $\mathbf{f}'(\mathbf{x})$ is a linear transformation that maps $\mathbf{h}$ into the element $\mathbf{f}'(\mathbf{x})\mathbf{h}$ of $\mathbb{R}^m$.

The relation between the differential $d\mathbf{f}$ and the derivative $\mathbf{f}'$ is expressed in the equation

$$d\mathbf{f}(\mathbf{x}, \mathbf{h}) = \mathbf{f}'(\mathbf{x})\mathbf{h}.$$

If we write $d\mathbf{x} = (dx_1, \ldots, dx_n)$ in place of $\mathbf{h} = (h_1, \ldots, h_n)$, we can write the foregoing in the form

$$d\mathbf{f}(\mathbf{x}, d\mathbf{x}) = \mathbf{f}'(\mathbf{x})\, d\mathbf{x}. \tag{12.1–7}$$

Here $d\mathbf{x}$ denotes an arbitrary element of $\mathbb{R}^n$. We are immediately reminded by (12.1–7) of the familiar expression $f'(x)\, dx$ for the differential of a real function of a real variable, in which $f'(x)$ is a real number. Now, when $m = n = 1$, a member of $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ is represented by a 1×1 matrix consisting of a single real number. So, if we think of $\mathbf{f}'(\mathbf{x})\, d\mathbf{x}$ in (12.1–7) as the product of an $m \times n$ matrix times an $n \times 1$ matrix, in the case when $m = n = 1$ we have merely the product of two real numbers. Thus the presentation of the differential in §1.3 is really just a special case of the general presentation we have here.

We come now to the first important theorem about differentiable functions.

THEOREM I. *If* $\mathbf{f}$ *is differentiable at* $\mathbf{a}$, *it is continuous at* $\mathbf{a}$.

We leave the proof as an exercise (Exercise 2).

Another important theorem about differentiable functions is the theorem known as the chain rule. It asserts that the composition of two differentiable functions is differentiable, and it provides the rule for finding the derivative of the composite function. Here is the precise statement.

THEOREM II. (THE CHAIN RULE). *Consider two functions* $\mathbf{f}$ *and* $\mathbf{g}$ *and the composite function* $\mathbf{g} \circ \mathbf{f}$ *in the following situation*: $\mathbf{f}$ *maps an open subset* U *of*

$\mathbb{R}^n$ into $\mathbb{R}^m$, with $\mathbf{x}_0$ in U and $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0)$; $\mathbf{g}$ is defined on an open set V in $\mathbb{R}^m$ that contains $\mathbf{f}(\mathbf{x})$ for all $\mathbf{x}$ in U, and $\mathbf{g}$ maps V into $\mathbb{R}^p$. Assume that $\mathbf{f}$ is differentiable at $\mathbf{x}_0$ and that $\mathbf{g}$ is differentiable at $\mathbf{y}_0$. Then $\mathbf{g} \circ \mathbf{f}$, which maps U into $\mathbb{R}^p$, is differentiable at $\mathbf{x}_0$, and

$$(\mathbf{g} \circ \mathbf{f})'(\mathbf{x}_0) = \mathbf{g}'(\mathbf{y}_0) \circ \mathbf{f}'(\mathbf{x}_0). \tag{12.1–8}$$

Proof. For notational convenience, we shall represent $\mathbf{f}'(\mathbf{x}_0)$ by T and $\mathbf{g}'(\mathbf{y}_0)$ by L. Then $T \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, $L \in \mathscr{L}(\mathbb{R}^m, \mathbb{R}^p)$, and $\mathbf{g}'(\mathbf{y}_0) \circ \mathbf{f}'(\mathbf{x}_0) = LT \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^p)$. What we wish to prove is that $\mathbf{g} \circ \mathbf{f}$ has a derivative at $\mathbf{x}_0$ given by $(\mathbf{g} \circ \mathbf{f})'(\mathbf{x}_0) = LT$; that is, that

$$\lim_{\|\mathbf{h}\| \to 0} \frac{\|\mathbf{g}[\mathbf{f}(\mathbf{x}_0 + \mathbf{h})] - \mathbf{g}[\mathbf{f}(\mathbf{x}_0)] - LT(\mathbf{h})\|}{\|\mathbf{h}\|} = 0. \tag{12.1–9}$$

We can prove this by straightforward calculation, using the differentiability condition for $\mathbf{f}$ in the form (12.1–4) with $\mathbf{x} = \mathbf{x}_0$ and a corresponding condition for $\mathbf{g}$ at $\mathbf{y}_0$. Let us assume $\mathbf{h} \neq \mathbf{0}$ and define

$$\mathbf{F}(\mathbf{h}) = \frac{\mathbf{f}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{f}(\mathbf{x}_0) - T\mathbf{h}}{\|\mathbf{h}\|};$$

we assume that $\|\mathbf{h}\|$ is small enough to keep $\mathbf{x}_0 + \mathbf{h}$ in U. The differentiability of $\mathbf{f}$ at $\mathbf{x}_0$ means that

$$\lim_{\|\mathbf{h}\| \to 0} \|F(\mathbf{h})\| = 0. \tag{12.1–10}$$

If $\mathbf{k} \neq \mathbf{0}$ and if $\|\mathbf{k}\|$ is small enough to keep $\mathbf{y}_0 + \mathbf{k}$ in V we define

$$\mathbf{G}(\mathbf{k}) = \frac{\mathbf{g}(\mathbf{y}_0 + \mathbf{k}) - \mathbf{g}(\mathbf{y}_0) - L\mathbf{k}}{\|\mathbf{k}\|}.$$

The differentiability of $\mathbf{g}$ at $\mathbf{y}_0$ means that

$$\lim_{\|\mathbf{k}\| \to 0} \|\mathbf{G}(\mathbf{k})\| = 0. \tag{12.1–11}$$

Finally, if $\mathbf{h} \neq \mathbf{0}$ we define

$$\mathbf{H}(\mathbf{h}) = \frac{\mathbf{g}[\mathbf{f}(\mathbf{x}_0 + \mathbf{h})] - \mathbf{g}[\mathbf{f}(\mathbf{x}_0)] - LT\mathbf{h}}{\|\mathbf{h}\|}.$$

Then (12.1–9) can be rewritten as

$$\lim_{\|\mathbf{h}\| \to 0} \|\mathbf{H}(\mathbf{h})\| = 0. \tag{12.1–12}$$

Our task is to convert the expression for $\mathbf{H}(\mathbf{h})$ into a form that allows us to get an estimate of the size of $\|\mathbf{H}(\mathbf{h})\|$ and show that (12.1–12) is true. Now, $\mathbf{h}$ is our independent variable; we are going to make $\mathbf{k}$ depend on $\mathbf{h}$, setting

$$\mathbf{k} = \mathbf{f}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{f}(\mathbf{x}_0). \tag{12.1–13}$$

Here we cannot be sure that $\mathbf{k} \neq \mathbf{0}$, so we complete the definition of $\mathbf{G}$, setting $\mathbf{G}(\mathbf{0}) = \mathbf{0}$. Because $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0)$ we can write $\mathbf{f}(\mathbf{x}_0 + \mathbf{h}) = \mathbf{y}_0 + \mathbf{k}$, and so

$$\mathbf{H}(\mathbf{h}) = \frac{\mathbf{g}(\mathbf{y}_0 + \mathbf{k}) - \mathbf{g}(\mathbf{y}_0) - LT\mathbf{h}}{\|\mathbf{h}\|}.$$

From the definition of $\mathbf{G}$ we see that

$$\mathbf{g}(\mathbf{y}_0 + \mathbf{k}) - \mathbf{g}(\mathbf{y}_0) = \|\mathbf{k}\|\mathbf{G}(\mathbf{k}) + L\mathbf{k},$$

and so

$$\mathbf{H}(\mathbf{h}) = \frac{\|\mathbf{k}\|\mathbf{G}(\mathbf{k}) + L\mathbf{k} - LT\mathbf{h}}{\|\mathbf{h}\|}. \tag{12.1–14}$$

Now, from (12.1–13) and the definition of $\mathbf{F}$ we see that

$$\mathbf{k} = \|\mathbf{h}\|\mathbf{F}(\mathbf{h}) + T\mathbf{h}, \tag{12.1–15}$$

and therefore

$$L\mathbf{k} = \|\mathbf{h}\|L[\mathbf{F}(\mathbf{h})] + LT(\mathbf{h}).$$

Substituting this result into (12.1–14) we obtain

$$\mathbf{H}(\mathbf{h}) = \frac{\|\mathbf{k}\|\mathbf{G}(\mathbf{k}) + \|\mathbf{h}\|L[\mathbf{F}(\mathbf{h})]}{\|\mathbf{h}\|}. \tag{12.1–16}$$

From (12.1–15) we see that

$$\|\mathbf{k}\| \leqq \|\mathbf{h}\|\|\mathbf{F}(\mathbf{h})\| + \|T\|\|\mathbf{h}\|, \tag{12.1–17}$$

from which we conclude that $\|\mathbf{k}\| \to 0$ when $\|\mathbf{h}\| \to 0$. Now we put the result of (12.1–17) into the numerator on the right in (12.1–16) and simplify by canceling out $\|\mathbf{h}\|$ from the numerator and denominator. The result is

$$\|\mathbf{H}(\mathbf{h})\| \leqq (\|\mathbf{F}(\mathbf{h})\| + \|T\|)\|\mathbf{G}(\mathbf{k})\| + \|L\|\|\mathbf{F}(\mathbf{h})\|.$$

From this, using (12.1–10) and (12.1–11), we see that (12.1–12) is true. This is what we wished to show, and so Theorem II is proved.

The general form of the chain rule in Theorem II contains as special cases the elementary chain rule of elementary calculus, given as Theorem II in §1.11 and discussed in §1.3, as well as the chain rule of Theorem V in §7.3. Here we see how the use of vector space ideas and co-ordinate-free notation in multivariable calculus enables us to unify some previous results and express the chain rule almost as simply in the general case as in the elementary case of Chapter 1.

12.2 / THE COMPONENT FUNCTIONS AND DIFFERENTIABILITY

When a function $\mathbf{f}$ from $\mathbb{R}^n$ to $\mathbb{R}^m$ is differentiable, we naturally wonder what this implies about the differentiability of the component functions $f^{(1)}, f^{(2)}, \ldots, f^{(m)}$, where $f^{(i)}(\mathbf{x})$ is the ith component of the vector $\mathbf{f}(\mathbf{x})$, so that

$$\mathbf{f}(\mathbf{x}) = (f^{(1)}(\mathbf{x}), \ldots, f^{(m)}(\mathbf{x})). \tag{12.2–1}$$

Also, if **f** is differentiable, how is the derivative of **f** represented by a matrix (in the standard matrix representation of linear transformations discussed in §11.3)? These questions are easily answered in the following theorem.

THEOREM III. *Let* **f** *be a function from an open set* U *in* $\mathbb{R}^n$, *mapping* U *into* $\mathbb{R}^m$. *Let* **f** *have components* $f^{(1)}, \ldots, f^{(m)}$ *from* U *into* $\mathbb{R}$, *as in* (12.2–1). *Then* **f** *is differentiable at a point* **a** *of* U *if and only if each of the component functions is differentiable at* **a**. *When* **f** *is differentiable at* **a**, *the standard matrix representation of* $\mathbf{f}'(\mathbf{a})$, *as a member of* $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, *is the Jacobian matrix*

$$\begin{pmatrix} \dfrac{\partial f^{(1)}}{\partial x_1} & \dfrac{\partial f^{(1)}}{\partial x_2} & \cdots & \dfrac{\partial f^{(1)}}{\partial x_n} \\ \dfrac{\partial f^{(2)}}{\partial x_1} & \dfrac{\partial f^{(2)}}{\partial x_2} & \cdots & \dfrac{\partial f^{(2)}}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial f^{(m)}}{\partial x_1} & \dfrac{\partial f^{(m)}}{\partial x_2} & \cdots & \dfrac{\partial f^{(m)}}{\partial x_n} \end{pmatrix} \tag{12.2–2}$$

in which the partial derivatives are evaluated at **a**.

Proof. Suppose that **f** is differentiable at **a**. Then, by definition,

$$\lim_{\|\mathbf{h}\| \to 0} \frac{\|\mathbf{f}(\mathbf{a}+\mathbf{h}) - \mathbf{f}(\mathbf{a}) - \mathbf{f}'(\mathbf{a})\mathbf{h}\|}{\|\mathbf{h}\|} = 0. \tag{12.2–3}$$

Let the standard matrix representation of $\mathbf{f}'(\mathbf{a})$ have A_{ij} as its element in the ith row and jth column, so that, if $\mathbf{h} = (h_1, \ldots, h_n)$, we have

$$\mathbf{f}'(\mathbf{a})\mathbf{h} = \left(\sum_{j=1}^{n} A_{1j}h_j, \sum_{j=1}^{n} A_{2j}h_j, \ldots, \sum_{j=1}^{n} A_{mj}h_j \right).$$

Now, the ith component of the vector

$$\mathbf{f}(\mathbf{a}+\mathbf{h}) - \mathbf{f}(\mathbf{a}) - \mathbf{f}'(\mathbf{a})\mathbf{h}$$

is

$$f^{(i)}(\mathbf{a}+\mathbf{h}) - f^{(i)}(\mathbf{a}) - \sum_{j=1}^{n} A_{ij}h_j.$$

By the definition of the Euclidean norm in $\mathbb{R}^n$, the absolute value of the ith component of a vector in $\mathbb{R}^n$ is certainly no larger than the norm of that vector. Therefore it is evident, as a consequence of (12.2–3), that

$$\lim_{\|\mathbf{h}\| \to 0} \frac{|f^{(i)}(\mathbf{a}+\mathbf{h}) - f^{(i)}(\mathbf{a}) - \sum_{j=1}^{n} A_{ij}h_j|}{\|\mathbf{h}\|} = 0. \tag{12.2–4}$$

By (12.1–2) and (12.1–3) this means precisely that $f^{(i)}$ is differentiable at **a** and

that

$$A_{ij} = \frac{\partial f^{(i)}}{\partial x_j}, \tag{12.2–5}$$

the derivatives being evaluated at **a**.

Conversely, let us start with the assumption that each component function is differentiable. This implies that the first partial derivatives all exist and that (12.2–4) is satisfied for each i, where the numbers A_{ij} are given by (12.2–5). It follows from this that (12.2–3) is satisfied, because the Euclidean norm of a vector is no larger than the sum of the absolute values of its components. [See the first inequality in (7–5).] This completes the proof.

As we know from examples in Chapter 7, the mere existence of the first partial derivatives of a component function $f^{(i)}$ at the point $\mathbf{x} = \mathbf{a}$ is not sufficient to make $f^{(i)}$ differentiable at **a**. However, as was stated in Theorem II in §7.1, if a function from $\mathbb{R}^n$ to $\mathbb{R}$, defined in a neighborhood of **a**, has first partial derivatives, not just at **a**, but at each point in the neighborhood, and if these derivatives are continuous at **a**, then the function is differentiable at **a**. (A proof of this proposition, for the special case $n = 3$, is asked for in Exercise 3.) As a consequence, we can be assured that the function **f** in Theorem III is differentiable at **a** if all the partial derivatives in the Jacobian matrix (12.2–2) exist throughout a neighborhood of **a** and are continuous at **a**. This is a useful way of testing for differentiability in practice.

There are two particular cases that deserve special mention. One is the case in which $n = 1$; the other is the case in which $m = 1$.

When $n = 1$, the vector **x** becomes a real variable x and the component functions are real functions of a real variable. In this case the Jacobian matrix (12.2–2) has just one column. Its elements are the ordinary (not partial) derivatives with respect to x of the component functions. If the derivatives at $x = a$ of the component functions are $A_1, \ldots, A_m$, the derivative $\mathbf{f}'(a)$ maps the scalar h into the vector $(A_1h, \ldots, A_mh)$ and we can regard $\mathbf{f}'(a)$ as the vector $(A_1, \ldots, A_m)$.

When $m = 1$ we have a scalar function f of the vector **x**. In this case the Jacobian matrix representing $f'(\mathbf{a})$ has one row and n columns, and can be regarded as a vector $(f_1(\mathbf{a}), \ldots, f_n(\mathbf{a}))$, where

$$f_j(\mathbf{a}) = \frac{\partial f}{\partial x_j} \text{ evaluated at } \mathbf{x} = \mathbf{a}.$$

The differential is

$$df(\mathbf{a}, \mathbf{h}) = f'(\mathbf{a})\mathbf{h} = \sum_{j=1}^{n} f_j(\mathbf{a})h_j. \tag{12.2–6}$$

In §10.6 we defined the gradient of a scalar function defined on an open set in $\mathbb{R}^3$. It is natural to extend that definition by defining the gradient of a scalar function f from $\mathbb{R}^n$ to $\mathbb{R}$ as the vector function **grad** f from $\mathbb{R}^n$ to $\mathbb{R}^n$ with

components given by the partial derivatives:

$$\mathbf{grad}\, f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_i}, \ldots, \frac{\partial f}{\partial x_n}\right), \tag{12.2–7}$$

when f is differentiable at $\mathbf{x}$; in (12.2–7) the partial derivatives are evaluated at $\mathbf{x}$. If we view the **grad** $f(\mathbf{x})$ as a one-rowed matrix, we see that it represents $f'(\mathbf{x})$. When $f'(\mathbf{x})$ is applied to a vector $\mathbf{x}$ the resulting scalar can be viewed as a dot product:

$$f'(\mathbf{x})\mathbf{h} = (\mathbf{grad}\, f(\mathbf{x})) \cdot \mathbf{h}. \tag{12.2–8}$$

As a matter of notational convenience we shall denote the value of **grad** $f(\mathbf{x})$ when $\mathbf{x} = \mathbf{a}$ as **grad** $f(\mathbf{a})$.

In the next section we shall discuss an application of the gradient to the problem of finding a point where a scalar function attains a maximum or minimum value.

12.21 / DIRECTIONAL DERIVATIVES AND THE METHOD OF STEEPEST DESCENT

Just as we say that a nonzero vector determines a direction (the direction of the arrow that represents the vector) (in $\mathbb{R}^3$), so we shall say that a nonzero vector in $\mathbb{R}^n$ determines a direction in $\mathbb{R}^n$. Now consider a scalar function f, defined and continuous on some open set in $\mathbb{R}^n$. Because f is continuous, the change in the value of $f(\mathbf{x})$ as we move in any given direction from a particular point will be gradual, and will be small if the change in distance is small. If $\mathbf{u}$ is a unit vector and $\mathbf{a}$ is a point in the domain of f, we define

$$D_{\mathbf{u}} f(\mathbf{a}) = \lim_{t \to 0+} \frac{f(\mathbf{a} + t\mathbf{u}) - f(\mathbf{a})}{t}, \tag{12.21–1}$$

provided the limit exists, as the *directional derivative* of f at $\mathbf{a}$ in the direction of $\mathbf{u}$. In taking the limit in (12.21–1) it is assumed that t is positive and small enough to assure that $\mathbf{a} + t\mathbf{u}$ is always in the domain of f. The directional derivative is to be interpreted as the rate of change per unit of distance, at $\mathbf{a}$, of the value of f, in the direction of $\mathbf{u}$. This is in accord with the standard interpretation of the derivative in elementary calculus, because the distance between $\mathbf{a} + t\mathbf{u}$ and $\mathbf{a}$ is

$$\|(\mathbf{a} + t\mathbf{u}) - \mathbf{a}\| = \|t\mathbf{u}\| = t\|u\| = t.$$

In the case in which f has a first partial derivative with respect to x_j at $\mathbf{a}$, it is readily seen that this partial derivative is equal to the directional derivative at $\mathbf{a}$ in the direction of the standard basis vector $\mathbf{e}_j$, for in that case the only difference between $\mathbf{a} + t\mathbf{e}_j$ and $\mathbf{a}$ is the jth co-ordinates, the jth co-ordinate of $\mathbf{a} + t\mathbf{e}_j$ being $a_j + t$ and that of $\mathbf{a}$ being a_j [where $\mathbf{a} = (a_1, \ldots, a_n)$].

There is an important relationship between directional derivatives and the gradient. For $\mathbb{R}^3$ this was mentioned in §10.6. The situation in $\mathbb{R}^n$ is stated in the next theorem.

THEOREM IV. *Suppose that f, from $\mathbb{R}^n$ to $\mathbb{R}$, is differentiable at* **a**. *Then, for any direction* **u** (*where* **u** *is a unit vector*),

$$D_{\mathbf{u}}f(\mathbf{a}) = [\mathbf{grad}\, f(\mathbf{a})] \cdot \mathbf{u} \tag{12.21–2}$$

Moreover, when $\mathbf{grad}\, f(\mathbf{a})$ *is not zero, its direction is that in which $f(\mathbf{x})$ increases most rapidly at* **a**, *and this greatest rate of increase is equal to* $\|\mathbf{grad}\, f(\mathbf{a})\|$.

Proof. We can obtain (12.21–2) directly from the definition of differentiability of f by applying (12.2–3) to this case with $\mathbf{h} = t\mathbf{u}$ and $t > 0$. Then $\|\mathbf{h}\| \to 0$ means $t \to 0+$. Therefore we obtain

$$\lim_{t\to 0+} \frac{|f(\mathbf{a}+t\mathbf{u}) - f(\mathbf{a}) - f'(\mathbf{a})(t\mathbf{u})|}{t} = 0. \tag{12.21–3}$$

But, by (12.2–8), $f'(\mathbf{a})(t\mathbf{u}) = tf'(\mathbf{a})(\mathbf{u}) = t\ \mathbf{grad}\, f(\mathbf{a}) \cdot \mathbf{u}$, and so we see from (12.21–3) that

$$\lim_{t\to 0+} \left| \frac{f(\mathbf{a}+t\mathbf{u}) - f(\mathbf{a})}{t} - \mathbf{grad}\, f(\mathbf{a}) \cdot \mathbf{u} \right| = 0.$$

From this result and the definition in (12.21–1) we conclude the truth of (12.21–2).

To prove the assertion in the final sentence of Theorem IV we observe that, by (10.2–5),

$$|[\mathbf{grad}\, f(\mathbf{a})] \cdot \mathbf{u}| \leqq \|\mathbf{grad}\, f(\mathbf{a})\| \|u\| = \|\mathbf{grad}\, f(\mathbf{a})\|,$$

because $\|u\| = 1$. Moreover, by Exercise 4b in §10.12, the foregoing inequality becomes an equality if and only if $\mathbf{grad}\, f(\mathbf{a})$ is a multiple of **u**. But, if $\mathbf{grad}\, f(\mathbf{a}) = c\mathbf{u}$, we must have $|c| = \|\mathbf{grad}\, f(\mathbf{a})\|$. Therefore $c = \pm\|\mathbf{grad}\, f(\mathbf{a})\|$. We see in this way that the maximum possible value of $|D_{\mathbf{u}}f(\mathbf{a})|$ is $\|\mathbf{grad}\, f(\mathbf{a})\|$ and that this maximum is attained if and only if

$$\mathbf{u} = \pm \frac{\mathbf{grad}\, f(\mathbf{a})}{\|\mathbf{grad}\, f(\mathbf{a})\|}.$$

With these two choices of sign we get $D_{\mathbf{u}}f(\mathbf{a}) = \pm\|\mathbf{grad}\, f(\mathbf{a})\|$. When the plus sign is chosen **u** has the same direction as $\mathbf{grad}\, f(\mathbf{a})$ and we obtain a positive rate of change which is the greatest possible. The other choice of sign, for the oppositely directed **u**, yields a negative rate of change. In this discussion we have validated the final assertion of the theorem.

The relation of the gradient of f to rates of change of f in various directions, as indicated in Theorem IV, is the basis of a procedure called the *method of steepest descent*, which is of great practical importance in finding extreme values of functions from $\mathbb{R}^n$ to $\mathbb{R}$. The essential ideas can be seen in the case where $n = 2$. Suppose we wish to find a minimum value of the differentiable function $f(x, y)$. The first step is to make as good a guess as we can as to a point where a

minimum value is taken on. Call this initial guess (x_0, y_0) and evaluate **grad** $f(x_0, y_0)$. If (x_0, y_0) is not a critical point of f, **grad** $f(x_0, y_0)$ is not the zero vector, and the value of f will decrease from its value at (x_0, y_0) if we move to a point (x_1, y_1) *not too far from* (x_0, y_0) in the direction of $-$**grad** $f(x_0, y_0)$, this being the direction of most rapid decrease of f at (x_0, y_0). Thus for some positive number α_0, *not too large*,

$$(x_1, y_1) = (x_0, y_0) - \alpha_0 \,\mathbf{grad}\, f(x_1, y_1),$$

where α_0 is some suitably chosen positive number, will be a point where $f(x_1, y_1) < f(x_0, y_0)$. If (x_1, y_1) is not a critical point of f, this process can be repeated, leading us to $(x_2, y_2) = (x_1, y_1) - \alpha_1 \,\mathbf{grad}\, f(x_1, y_1)$, and so on. Continuing in this way we generate a sequence of points $\{(x_n, y_n)\}_{n=0}^{\infty}$ along which the function f decreases as n increases. The points (x_n, y_n) can then be expected to converge to a critical point, if there is one, where f has a minimum value—at least a local minimum. The sequence of values $\{f(x_n, y_n)\}_1^{\infty}$ will then decrease, approaching the desired minimum value as a limit. The sequence of successive approximations is generated by the formula

$$(x_{n+1}, y_{n+1}) = (x_n, y_n) - \alpha_n \,\mathbf{grad}\, f(x_n, y_n), \tag{12.21–4}$$

in which the α_n's must be chosen as advantageously as possible.

In some of the more complicated problems of practical interest, the α_n's are chosen mainly by experimenting on the computer. If the function f is rather simple and twice differentiable, a formula can be derived which sometimes generates a good sequence of α_n's. Suppose that (x_n, y_n) has been found. We need a number α_n such that (12.21–4) will produce a point (x_{n+1}, y_{n+1}) where f has a smaller value than at (x_n, y_n). Notice that

$$f(x_{n+1}, y_{n+1}) = f(x_n - \alpha_n f_x^{(n)}, y_n - \alpha_n f_y^{(n)}),$$

where $f_x^{(n)} = f_x(x_n, y_n)$ and $f_y^{(n)} = f_y(x_n, y_n)$. We now approximate $f(x_{n+1}, y_{n+1})$ by the Taylor series, using formula (7.5–5) with $a = x_n$, $b = y_n$, $h = -\alpha_n f_x^{(n)}$ and $k = -\alpha_n f_y^{(n)}$. Then $x_{n+1} = x_n + h$ and $y_{n+1} = y_n + k$. Taking just the first three terms of the Taylor series we get the following approximation:

$$f(x_{n+1}, y_{n+1}) \simeq f(x_n, y_n) - \alpha_n[(f_x^{(n)})^2 + (f_y^{(n)})^2] + \tfrac{1}{2}\alpha_n^2 Q_n$$

where

$$Q_n = (f_x^{(n)})^2 f_{xx}^{(n)} + 2 f_x^{(n)} f_y^{(n)} f_{xy}^{(n)} + (f_y^{(n)})^2 f_{yy}^{(n)},$$

in which $f_{xx}^{(n)} = \dfrac{\partial^2 f}{\partial x^2}$, $f_{xy}^{(n)} = \dfrac{\partial^2 f}{\partial x \partial y}$ and $f_{yy}^{(n)} = \dfrac{\partial^2 f}{\partial y^2}$ are all evaluated at (x_n, y_n). If we are getting close to the point of minimum value of $f(x, y)$, we may assume that Q_n is positive (see the discussion of the sign of $G(\phi)$ in connection with (7.6–5). Moreover, if α_n is sufficiently small, the term $\frac{1}{2}\alpha_n^2 Q_n$ should be so small that we can be assured that $f(x_{n+1}, y_{n+1}) < f(x_n, y_n)$ if α_n is positive. Under these conditions, a good choice for α_n will be that for which the expression

$$-\alpha_n[(f_x^{(n)})^2 + (f_y^{(n)})^2] + \tfrac{1}{2}\alpha_n^2 Q_n$$

is a minimum. This obviously occurs when

$$\alpha_n = \frac{(f_x^{(n)})^2 + (f_y^{(n)})^2}{Q_n}. \tag{12.21–5}$$

To see that this method actually can be carried out, we go back to Exercise 13 of §6.3. Using a programmable pocket calculator, a program of fewer than 200 steps can be written to generate the sequence (12.21–4) using the α_n's determined by (12.21–5). The program can also have the machine print out $f(x_n, y_n)$ and $\|\mathbf{grad}\, f(x_n, y_n)\|$ at each stage. If we start with the initial guess $(x_0, y_0) = (1, 1)$, the following results are obtained.

n	x_n	y_n	$\|\mathbf{grad}\, f(x_n, y_n)\|$	$f(x_n, y_n)$
0	1	1	10.81665	−25
1	1.6724	1.4483	5.8338	−28.5712
2	1.68995	1.2552	2.64037	−29.5574
3	1.9754	1.10243	1.74862	−29.9168
4	1.95826	1.01917	0.39027	−29.9912
5	1.99880	1.00946	0.16869	−29.9993
6	1.99673	1.00153	0.03104	−29.9999
7	1.99992	1.00069	0.01233	−29.99999602
8	1.99977	1.00011	0.00223	−29.99999972
9	1.99999	1.00005	0.00088	−29.99999998
10	1.99998	1.00001	0.00016	−30
11	1.999999597	1.000003484	0.0000624022	−30
12	1.999998817	1.000000555	0.0000112465	−30
13	1.999999971	1.000000248	0.000004434	−30
14	1.999999916	1.000000039	0.000007991	−30
15	1.999999998	1.000000018	0.000000315	−30
16	1.999999994	1.000000003	0.000000567	−30
17	2.000000000	1.000000001	0.000000223	−30

The minimum value of −30 is found to 10 digits in 10 steps, but to get the critical point with equal accuracy takes 18 steps.

The convergence here is tediously slow. A faster method will be given in §12.3. However, with the method of steepest descent one can get convergence even if the initial guess is not very good, whereas the faster method may not converge at all if one starts from a bad initial guess. This possibility will be illustrated in §12.3.

In conclusion, we make the obvious comment that the same ideas lead easily to a method of steepest ascent which is useful in case one is trying to find a maximum rather than a minimum (Exercise 25). Finally, we should make it clear that formula (12.21–5) is just one of various methods for choosing the α's. There

are several reasons why it can fail to give satisfactory results in certain problems. When this happens, one can frequently make a good guess for α at each stage by looking at the behavior of the points (x_n, y_n) up to that stage.

12.3 / NEWTON'S METHOD

In this section we illustrate the use of the derivative of a function $\mathbf{f}$ from $\mathbb{R}^n$ to $\mathbb{R}^n$ in seeking a solution to the equation $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ by a method of successive approximations known as *Newton's method*, or also as the *Newton-Raphson method*. The work illustrates very nicely how a formula from elementary calculus of functions of one variable can be extended directly to the multivariable case because of the way in which we have defined the derivative of a transformation.

We begin with the elementary case in which $n = 1$. Suppose we have a differentiable real function of the real variable x and we wish to find a solution of the equation $f(x) = 0$. For the type of problem in which we are interested it will generally be known that there is at least one x such that $f(x) = 0$. We may even have some rough notion of the value of such an x; the problem then is to obtain a fairly accurate approximation to its value. For example, if $f(x) = x^3 - 3x^2 + 3$, we can see that $f(2) = -1$ and $f(3) = 3$, so that $f(x) = 0$ for some x such that $2 < x < 3$, and we want to find it, accurate to several places of decimals. Or, if $f(x) = e^x + x - 2$, we see that $f(0) = -1$ and $f(1) = 1.718$, so there is an x such that $0 < x < 1$ and $f(x) = 0$. What is its value more precisely? In many such cases, if we start with a *guessed* value x_0 of the solution, we can use Newton's method to generate a succession of points $x_1, x_2, \ldots, x_n, \ldots$ such that the sequence $\{x_n\}$ will converge to a limit x such that $f(x) = 0$. The idea of the method is to find the line tangent to the curve $y = f(x)$ at the point $(x_0, f(x_0))$, and then find the point x_1 where this tangent intersects the x-axis. Then we can repeat the operation with the line tangent to $y = f(x)$ at $(x_1, f(x_1))$, and so on. See Fig. 86. In each of the particular examples cited above, the portion of the graph near its intersection with the x-axis has a general resemblance to the graph in Fig. 86, and if we start with a guessed value x_0 that is not too far from the abscissa of the intersection, then x_n will converge to the solution. The formula for the successive x_n's is easy to find. The equation of the tangent at $(x_0, f(x_0))$ is

$$y - f(x_0) = f'(x_0)(x - x_0).$$

If we set $y = 0$ and solve for x we find

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

The general formula is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, 2, \ldots \tag{12.3-1}$$

At this point it will be instructive to the student to carry out the successive steps

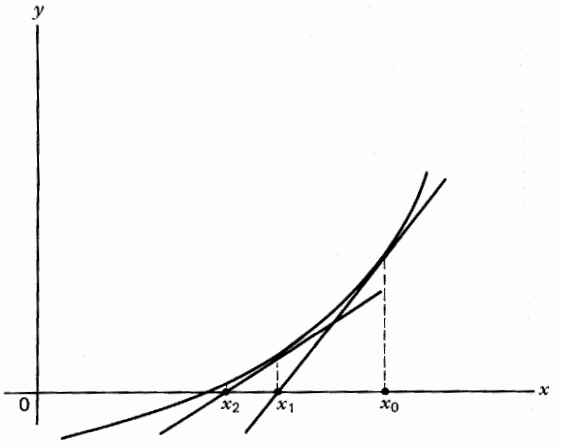


Fig. 86.

of Newton's method on a particular problem. As a simple example, let $f(x) = x^3 - 20$. To find the solution of $x^3 - 20 = 0$ is to find the cube root of 20. Simple calculations with a handheld computer show that the root of $f(x) = 0$ is between 2.7 and 2.8, and closer to 2.7. If we guess $x_0 = 2.7$ and calculate x_1 and x_2 we find $x_2 = 2.71442$, which is a solution accurate to five decimal places. If we make the initial guess $x_0 = 2.8$, we get the same solution as x_3 instead of x_2. For a less simple problem that is still easily handled by a pocket calculator, see Exercise 26.

We shall not deal with theorems about sufficient conditions to insure that the sequence $\{x_n\}$ generated by Newton's method will actually converge to a root of the equation $f(x) = 0$, although there are theorems of this kind. The interested reader will find an excellent brief treatment of such a theorem (for a real function of a real variable) in *Infinitesimal Calculus*, by Jean Dieudonné (Houghton Mifflin, 1971).

When we move to the general case of solving an equation of the form $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, where $\mathbf{f}$ has its domain and range in $\mathbb{R}^n$, the problem is that of finding $\mathbf{x} = (x_1, \ldots, x_n)$ as the solution of a set of simultaneous equations in n unknowns. If the equations are not linear this can be a formidable problem. In general we cannot expect to be able to solve such a problem by algebra, even if the n equations themselves are algebraic.

One way in which such equations arise is when we undertake to find the critical points of a scalar function. Suppose, for example, we want to find the critical points of the function

$$F(x, y) = x^3 + y^3 + 3xy^2 - 15x - 15y$$

in the interior of the first quadrant. The critical points are those points where the

vector

$$\textbf{grad}\, F(x, y) = \left(\frac{\partial F}{\partial x}, \quad \frac{\partial F}{\partial y}\right) = (0, 0). \tag{12.3–2}$$

By direct calculation we find

$$\textbf{grad}\, F(x, y) = 3(x^2 + y^2 - 5,\ y^2 + 2xy - 5),$$

and so the equation (12.3–2) is equivalent to the two simultaneous equations

$$\begin{aligned} x^2 + y^2 - 5 &= 0, \\ 2xy + y^2 - 5 &= 0. \end{aligned} \tag{12.3–3}$$

These equations happen to be easy to solve by algebra. The only critical point in the interior of the first quadrant is (2, 1). Later, in Exercise 27, the student can try out the method of Newton on this problem, and observe how it can lead from a first guess at the solution to a highly accurate approximation to the exact solution.

For the general case of $\mathbf{f}$ from $\mathbb{R}^n$ to $\mathbb{R}^n$, Newton's method proceeds in a manner that is entirely analogous to the special case $n = 1$. A start is made with a guessed approximation x_0 to the solution of $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. Then we use $\mathbf{f}'(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$ as an approximation to $\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)$, as is warranted by the definition of the derivative. Then we set $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ in the approximate formula

$$\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) = \mathbf{f}'(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

and solve for $\mathbf{x}$, denoting the solution by $\mathbf{x}_1$. To achieve the solution, we assume that the linear operator $\mathbf{f}'(\mathbf{x}_0)$ has an inverse, $[\mathbf{f}'(\mathbf{x}_0)]^{-1}$, so that the equation

$$0 - \mathbf{f}(\mathbf{x}_0) = \mathbf{f}'(\mathbf{x}_0)(\mathbf{x}_1 - \mathbf{x}_0) \tag{12.3–4}$$

leads to the formula

$$\mathbf{x}_1 = \mathbf{x}_0 - [\mathbf{f}'(\mathbf{x}_0)]^{-1}\mathbf{f}(\mathbf{x}_0). \tag{12.3–5}$$

We emphasize that in (12.3–4) the operator $\mathbf{f}'(\mathbf{x}_0)$ is acting on the vector $\mathbf{x}_1 - \mathbf{x}_0$ and that in (12.3–5) the operator $[\mathbf{f}'(\mathbf{x}_0)]^{-1}$ is acting on the vector $\mathbf{f}(\mathbf{x}_0)$. We then proceed by a repetition of the process, obtaining a sequence of vectors $\mathbf{x}_1$, $\mathbf{x}_2, \ldots, \mathbf{x}_n, \ldots$, where

$$\mathbf{x}_{n+1} = \mathbf{x}_n - [\mathbf{f}'(\mathbf{x}_n)]^{-1}\mathbf{f}(\mathbf{x}_n). \tag{12.3–6}$$

It is assumed that $\mathbf{f}'(\mathbf{x}_n)$ has an inverse for each $\mathbf{x}_n$. Sufficient conditions for the convergence of this vector form of Newton's method are given in advanced texts on numerical analysis.

We conclude this section with an application of this powerful method to a two-dimensional problem, carried out on a programmable pocket calculator.

Example. We wish to solve the system

$$\begin{aligned} -13 + x - 2y + 5y^2 - y^3 &= 0, \\ -29 + x - 14y + y^2 + y^3 &= 0. \end{aligned}$$

Notice that by the elementary method of eliminating x and factoring the resulting cubic in y, we find immediately the solution $(5, 4)$, and we also see that this is the only solution which involves only real numbers. Thinking of the system as $\mathbf{f}(x, y) = 0$, we get

$$\mathbf{f}'(x, y) = \begin{pmatrix} 1 & -2 + 10y - 3y^2 \\ 1 & -14 + 2y + 3y^2 \end{pmatrix}.$$

The determinant of this matrix is $6y^2 - 8y - 12$, so $\mathbf{f}'$ is a nonsingular linear operator except on the lines $y = \frac{1}{3}(2 \pm \sqrt{22})$, that is, approximately, $y = 2.23$ and $y = -0.897$. The inverse is given by (see §11.9)

$$[\mathbf{f}'(x, y)]^{-1} = \begin{pmatrix} \dfrac{-14 + 2y + 3y^2}{6y^2 - 8y - 12} & \dfrac{2 - 10y + 3y^2}{6y^2 - 8y - 12} \\ \dfrac{-1}{6y^2 - 8y - 12} & \dfrac{1}{6y^2 - 8y - 12} \end{pmatrix},$$

so Newton's method, expressed by (12.3–6), gives the sequence of vectors generated by the following formula

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \end{pmatrix} - [\mathbf{f}'(x_n, y_n)]^{-1} \begin{pmatrix} -13 + x_n - 2y_n + 5y_n^2 - y_n^3 \\ -29 + x_n - 14y_n + y_n^2 + y_n^3 \end{pmatrix}.$$

A program to generate this sequence with a pocket calculator can be written with fewer than 200 steps. Starting with the initial guess $(x_0, y_0) = (10, 8)$ the following results are generated.

n	x_n	y_n
0	10	8
1	−21.805	5.8701
2	− 2.2100	4.6503
3	3.9071	4.1187
4	4.95746	4.00507
5	4.999918991	4.00000987
6	5.000000000	4.000000000

This shows that convergence from the starting point $(10, 8)$ is quite rapid, giving accuracy to 10 significant digits in just six steps. Such success cannot be assured for all starting points however. Notice the very different behavior of the sequence generated when we start with $(x_0, y_0) = (15, -2)$.

12.4 / A FORM OF THE LAW OF THE MEAN FOR VECTOR FUNCTIONS

In this section we obtain (Theorem V) a generalization of the law of the mean (as presented in §1.2 and §7.4), and an application of it in the form of an inequality (Theorem VI); both are applicable to differentiable functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. We

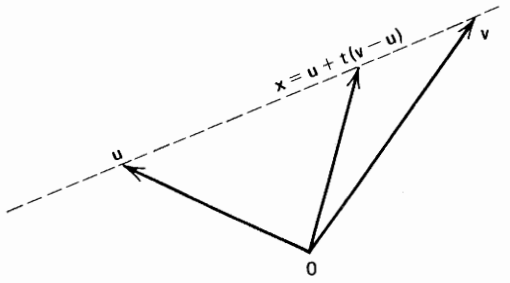


Fig. 87.

need the notion of a line segment in $\mathbb{R}^n$ determined by two points, $\mathbf{u}$ and $\mathbf{v}$. It consists of all points of the form

$$\mathbf{x} = \mathbf{u} + t(\mathbf{v} - \mathbf{u}) \text{ where } 0 \leqq t \leqq 1.$$

(In this connection, see Exercise 16 of Chapter 11.) The above line segment is closed because it contains both end points. We shall sometimes denote it by $[\mathbf{u}, \mathbf{v}]$. The corresponding open segment, which we denote by $(\mathbf{u}, \mathbf{v})$, is obtained by restricting t to the open interval $0 < t < 1$. The situation is portrayed in Fig. 87.

The vector $\mathbf{v} - \mathbf{u}$ extends from $\mathbf{u}$ to $\mathbf{v}$, so, if $0 < t < 1$, $\mathbf{u}$ plus t times $(\mathbf{v} - \mathbf{u})$ terminates at some point on the segment from $\mathbf{u}$ to $\mathbf{v}$. As t increases from 0 to 1, $\mathbf{x}$ moves from $\mathbf{u}$ to $\mathbf{v}$.

THEOREM V. *Let* $\mathbf{f}$ *be a function from an open set* G *in* $\mathbb{R}^n$ *to* $\mathbb{R}^m$. *Let* $\mathbf{u}$ *and* $\mathbf{v}$ *be two points of* G *such that the closed line segment* $[\mathbf{u}, \mathbf{v}]$ *lies in* G. *Suppose that* $\mathbf{f}$ *is continuous at each point of the closed segment and differentiable at each point of the open segment* $(\mathbf{u}, \mathbf{v})$. *Then to each vector* $\mathbf{w}$ *in* $\mathbb{R}^n$ *corresponds a vector* ξ *terminating on the open segment* $(\mathbf{u}, \mathbf{v})$ *such that*

$$[\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})] \cdot \mathbf{w} = [\mathbf{f}'(\xi)(\mathbf{v} - \mathbf{u})] \cdot \mathbf{w}. \tag{12.4–1}$$

Proof. Define a function $\mathbf{F}$ of t from $[0, 1]$ to $\mathbb{R}^m$ by $\mathbf{F}(t) = \mathbf{f}[\mathbf{u} + t(\mathbf{v} - \mathbf{u})]$. It is easily seen that $\mathbf{F}$ is continuous on $[0, 1]$ and differentiable on $(0, 1)$. We can regard $\mathbf{F}'(t)$ as a vector in $\mathbb{R}^m$ (see §12.2). In this particular case, by the chain rule,

$$\mathbf{F}'(t) = \mathbf{f}'[\mathbf{u} + t(\mathbf{v} - \mathbf{u})](\mathbf{v} - \mathbf{u})$$

when $0 < t < 1$, where the value of the derivative of $\mathbf{f}$ is a member of $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ acting on $\mathbf{v} - \mathbf{u}$. Now pick an arbitrary vector $\mathbf{w}$ in $\mathbb{R}^m$ and consider the real-valued function $\phi(t) = \mathbf{w} \cdot \mathbf{F}(t)$. It is easy to see that ϕ is continuous on $[0, 1]$ and differentiable in $(0, 1)$; moreover, it is readily proved (see Exercise 7a) that $\phi'(t) = \mathbf{w} \cdot \mathbf{F}'(t)$. Now obviously,

$$\phi(1) - \phi(0) = [\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})] \cdot \mathbf{w}.$$

Since ϕ is just a differentiable function from $\mathbb{R}$ to $\mathbb{R}$, the ordinary law of the mean tells us that there is some number θ that $0 < \theta < 1$ and

$$\phi(1) - \phi(0) = \phi'(\theta)(1 - 0) = \phi'(\theta).$$

Thus we see that

$$[\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})] \cdot \mathbf{w} = \{\mathbf{f}'[\mathbf{u} + \theta(\mathbf{v} - \mathbf{u})](\mathbf{v} - \mathbf{u})\} \cdot \mathbf{w}.$$

Letting $\boldsymbol{\xi} = \mathbf{u} + \theta(\mathbf{v} - \mathbf{u})$ we obtain (12.4–1), thus concluding the proof.

THEOREM VI. *Under the assumptions of THEOREM V, there is some point* $\boldsymbol{\xi}$ *on the line segment connecting* $\mathbf{u}$ *and* $\mathbf{v}$ *such that*

$$\|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})\| \leqq \|\mathbf{f}'(\boldsymbol{\xi})(\mathbf{v} - \mathbf{u})\|, \tag{12.4–2}$$

where $\| \ \|$ *of course denotes the usual Euclidean norm.*

Proof. If $\mathbf{f}(\mathbf{u}) = \mathbf{f}(\mathbf{v})$, (12.4–2) is obviously true, because $\|\mathbf{0}\| = 0$ and the norm of every vector is nonnegative. If $\mathbf{f}(\mathbf{u}) \neq \mathbf{f}(\mathbf{v})$, we start by taking the absolute value of each side in (12.4–1) and applying Cauchy's inequality (see (10.12–5)) to the right-hand side.

$$|[\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})] \cdot \mathbf{w}| = |[\mathbf{f}'(\boldsymbol{\xi})[\mathbf{v} - \mathbf{u})] \cdot \mathbf{w}| \leqq \|\mathbf{f}'(\boldsymbol{\xi})(\mathbf{v} - \mathbf{u})\| \, \|\mathbf{w}\|. \tag{12.4–3}$$

Then we choose $\mathbf{w}$ as follows:

$$\mathbf{w} = \frac{\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})}{\|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})\|}.$$

As we can see, $\|\mathbf{w}\| = 1$ and

$$|[\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})] \cdot \mathbf{w}| = \frac{\|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})\|^2}{\|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})\|} = \|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})\|.$$

Putting these results in (12.4–3) we obtain (12.4–2).

12.41 / THE HESSIAN AND EXTREME VALUES

In (6.9–10) we had occasion to consider a quadratic form in three variables. In n variables, such a function could be expressed as

$$\begin{aligned} Q(h_1, \ldots, h_n) = {} & a_{11}h_1^2 + a_{12}h_1h_2 + \cdots + a_{1n}h_1h_n \\ & + a_{21}h_2h_1 + a_{22}h_2^2 + \cdots + a_{2n}h_2h_n \\ & \cdots \\ & + a_{n1}h_nh_1 + \cdots + \cdots + a_{nn}h_n^2 \end{aligned} \tag{12.41–1}$$

We can just as well assume that $a_{ij} = a_{ji}$, and this is ordinarily done. Taking advantage of matrix notation and the rules of matrix multiplication, we can more conveniently represent the quadratic form by

$$Q(h_1, \ldots, h_n) = \mathbf{h}^T A \mathbf{h} \tag{12.41–2}$$

where $\mathbf{h}^T$ is regarded as a $(1 \times n)$ matrix, A is the $n \times n$ matrix (a_{ij}), and the vector $\mathbf{h}$ is thought of as an $(n \times 1)$ matrix. The product of three such matrices, in this order, is defined and comes out to be a scalar. The scalar in this case is the quadratic form.

We have encountered quadratic forms earlier in studying Taylor series for functions of several variables (§7.5). In the theory of maxima and minima, these series expansions are usually made about a critical point of the function, where all the first partial derivatives are zero. Suppose we expand $f(x_1, \ldots, x_n)$ about a critical point $\mathbf{a} = (a_1, \ldots, a_n)$. Since the coefficients of all linear terms are zero, the series begins as follows:

$$f(x_1, \ldots, x_n) = f(a_1, \ldots, a_n) + \left[\frac{1}{2}\left(h_1\frac{\partial}{\partial x_1} + \cdots + h_n\frac{\partial}{\partial x_n}\right)^2 f\right]_{\mathbf{x}=\mathbf{a}} + \cdots$$

where $h_k = x_k - a_k$. The second term can be written as $\frac{1}{2}Q(h_1, \ldots, h_n)$, where Q is a quadratic from in the h's, having as its coefficients the second order partial derivatives of f evaluated at $\mathbf{a}$. In fact, it is exactly the quadratic from (12.41–1) if we take a_{ij} to be $\dfrac{\partial^2 x}{\partial x_i \partial x_j}$, evaluated at $\mathbf{a}$. The above series can then be written

$$f(x_1, \ldots, x_n) = f(a_1, \ldots, a_n) + \tfrac{1}{2}\mathbf{h}^T A\mathbf{h} + \cdots.$$

This matrix A is called the *Hessian* of $\mathbf{f}$ at $\mathbf{a}$.

We learned in (§7.6) that $\mathbf{f}$ has local minima at those critical points where $\mathbf{h}^T A\mathbf{h}$ is positive definite, and local maxima at critical points where this quadratic form is negative definite. We then gave, without proof, a rule for deciding if a quadratic form is positive definite, and a rule to tell if it is negative definite. Since the quadratic form Q and the symmetric matrix A which represents it completely determine each other, it is common practice to predicate positive definiteness and negative definiteness of real symmetric matrices as well as of the quadratic forms which they represent. We must bear in mind, of course, that most symmetric matrices and quadratic forms are neither positive definite nor negative definite. This is reflected in the existence of critical points at which there is neither a local minimum nor a local maximum.

We now conclude this section with a reformulation, in terms of the Hessian, of the rules of search for extreme values. Suppose that f is a function from an open subset U of $\mathbb{R}^n$ to $\mathbb{R}$, having continuous partial derivatives of first and second order. To find the local extremes of f in U, we first find the critical points. Then we construct the Hessian at each critical point. At those critical points where the Hessian is positive definite, f takes on values which are local minima; at those where this matrix is negative definite, f has local maxima. The Hessian of f at an arbitrary $\mathbf{x}$ is in fact the value $f''(\mathbf{x})$ of the second derivative of f at $\mathbf{x}$; it is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$. We recall that f is a function from $\mathbb{R}^n$ to $\mathbb{R}$ and that $f'(\mathbf{x})$ is the gradient of f at $\mathbf{x}$, **grad** f being a function from $\mathbb{R}^n$ to $\mathbb{R}^n$. Therefore the derivative of **grad** f is a function from $\mathbb{R}^n$ to $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^n)$; the Jacobian matrix that represents the derivative of **grad** f (and the second derivative of f) is what we have called the Hessian of f. The name comes from a

German mathematician, Otto Hesse (1811–1874). Observe the striking resemblance between our sufficient conditions for a relative minimum of f at $\mathbf{a}$ and the corresponding conditions for the case of a function of one real variable. (See miscellaneous Exercise 2 at the end of Chapter 4.)

12.5 / CONTINUOUSLY DIFFERENTIABLE FUNCTIONS

We now consider further the situation discussed in §12.2 of a function $\mathbf{f}$ from an open set U of $\mathbb{R}^n$ to $\mathbb{R}^m$. We assume that $\mathbf{f}$ is differentiable in U, that is, that $\mathbf{f}'(\mathbf{x})$ exists at each point of U. We have sometimes denoted this derivative by $T_{\mathbf{x}}$ or simply by T to remind us that it is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, that is, $T \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$. When we arrive at the main theorem of this chapter—the inverse function theorem—we shall see the importance of knowing when the derivative $\mathbf{f}'$, from U to $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, is a continuous function. For continuity of $\mathbf{f}'$ at a point $\mathbf{a}$, the requirement is that $\|\mathbf{f}'(\mathbf{x}) - \mathbf{f}'(\mathbf{a})\|$ approach zero as $\|\mathbf{x} - \mathbf{a}\|$ approaches zero. Since $\mathbf{f}'(\mathbf{x}) - \mathbf{f}'(\mathbf{a})$ is a member of $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, we use the norm for that space, which was discussed in §11.8. When the function $\mathbf{f}'$ is continuous, $\mathbf{f}$ is said to be *continuously differentiable.*

It turns out that there is a simple way of testing $\mathbf{f}$ to find out if it is continuously differentiable, by examining the first partial derivatives of the scalar component functions $f^{(1)}, f^{(2)}, \ldots, f^{(m)}$ of the vector-valued function $\mathbf{f}$. It is customary to say that the function $\mathbf{f}$ is of class $C^{(1)}$ in U in case each of the first partial derivatives of each of the component functions is continuous in U. These are the partial derivatives, $m \times n$ in number, in the matrix (12.2–2). The following theorem provides the simple test referred to above for continuous differentiability.

THEOREM VII. *A necessary and sufficient condition that* $\mathbf{f}$ *be continuously differentiable in* U *is that* $\mathbf{f}$ *belong to class* $C^{(1)}$ *in* U.

Proof. Let $\mathbf{x}$ and $\mathbf{a}$ be points of U. Then $\mathbf{f}'(\mathbf{x}) - \mathbf{f}'(\mathbf{a})$ is a linear transformation whose matrix has the difference $f_j^{(i)}(\mathbf{x}) - f_j^{(i)}(\mathbf{a})$ as the element in the ith row and jth column. Now let

$$K(\mathbf{x}) = \max_{\substack{1 \leqq i \leqq m \\ 1 \leqq j \leqq n}} |f_j^{(i)}(\mathbf{x}) - f_j^{(i)}(\mathbf{a})|. \tag{12.5–1}$$

By (11.8–4) we know that

$$K(\mathbf{x}) \leqq \|\mathbf{f}'(\mathbf{x}) - \mathbf{f}'(\mathbf{a})\| \leqq \sqrt{mn}\, K(\mathbf{x}). \tag{12.5–2}$$

The proof of our theorem follows at once from these inequalities; for, saying that the partial derivatives are continuous at $\mathbf{x} = \mathbf{a}$ is equivalent to saying that $K(\mathbf{x}) \to 0$ as $\mathbf{x} \to \mathbf{a}$; we see this from (12.5–1). But from (12.5–2) we see that $K(\mathbf{x}) \to 0$ as $\mathbf{x} \to \mathbf{a}$ if and only if $\|\mathbf{f}'(\mathbf{x}) - \mathbf{f}'(\mathbf{a})\| \to 0$ as $\mathbf{x} \to \mathbf{a}$.

12.6 / THE FUNDAMENTAL INVERSION THEOREM

In this section we come to the central theorem of Chapter 12. It is about a system of n equations by which a vector $(x_1, x_2, \ldots, x_n)$ in $\mathbb{R}^n$ is mapped into the vector $(y_1, y_2, \ldots, y_n)$ also in $\mathbb{R}^n$. We write the system as

$$\begin{aligned} f^{(1)}(x_1, \ldots, x_n) &= y_1 \\ f^{(2)}(x_1, \ldots, x_n) &= y_2 \\ &\cdots \\ f^{(n)}(x_1, \ldots, x_n) &= y_n \end{aligned} \tag{12.6–1}$$

This is the same as the system (12–1) mentioned at the opening of the chapter except that we now assume $m = n$. If the component functions are linear functions of the x_i's (as in (12–2), this system can be handled by linear algebra. But we want to deal with a more general case in which the $f^{(i)}$'s are not necessarily linear. We shall use vector notation and vector methods. We write (12.6–1) as

$$\mathbf{f}(\mathbf{x}) = \mathbf{y}, \tag{12.6–2}$$

and assume that the domain of definition of $\mathbf{f}$ is a neighborhood $\mathcal{N}$ of a certain point $\mathbf{a}$. We assume that $\mathbf{f}$ is continuous at each point of $\mathcal{N}$. Then $\mathbf{a}$ is mapped into $\mathbf{f}(\mathbf{a})$ and $\mathcal{N}$ is mapped into a set (the image of $\mathcal{N}$) that we denote by $\mathbf{f}(\mathcal{N})$ The theorem to which we are coming is about expressing $\mathbf{x}$ as a function of $\mathbf{y}$; or perhaps it would be better to say that the theorem is about a way of knowing that the relation between $\mathbf{x}$ and $\mathbf{y}$ can be regarded as one in which $\mathbf{x}$ is a function of $\mathbf{y}$. Certain assumptions and restrictions must be imposed in order to obtain a useful result. Otherwise the function $\mathbf{f}$ might not establish a one-to-one relationship between $\mathcal{N}$ and $f(\mathcal{N})$, two or more $\mathbf{x}$'s may be mapped into the same $\mathbf{y}$, and in such a case $\mathbf{x}$ is not uniquely determined by $\mathbf{y}$.

In the case of a linear system, (12.6–2) can be written in the form $A\mathbf{x} = \mathbf{y}$, where A denotes an $n \times n$ matrix of constants a_{ij} (see (12–2)). We know that the linear system defines a one-to-one mapping of $\mathbb{R}^n$ onto all of $\mathbb{R}^n$ if the determinant of the matrix A is not zero, and that then $\mathbf{x}$ can be expressed as a function of $\mathbf{y}$ in the form $\mathbf{x} = A^{-1}\mathbf{y}$, where A^{-1} is the inverse matrix.

The first question we ask is this: How can we obtain a condition on the function $\mathbf{f}$ that will be a suitable generalization of the condition that the matrix A have nonzero determinant? We get at the answer by assuming that $\mathbf{f}$ is differentiable. Then, if $\mathbf{x}$ is a point of $\mathcal{N}$ which is close to $\mathbf{a}$, the difference $\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{a})$ is *approximately* equal to $T_0[\mathbf{x} - \mathbf{a}]$, where $T_0 = \mathbf{f}'(\mathbf{a})$ is a member of $\mathcal{L}(\mathbf{R}^n)$. This suggests that we consider whether the linear operator T_0 is invertible. In fact, let us consider what T_0 is in the special case in which $\mathbf{y} = \mathbf{f}(\mathbf{x})$ takes the linear form $\mathbf{y} = A\mathbf{x}$. In that case

$$f^{(i)}(x_1, \ldots, x_n) = a_{i1}x_1 + \cdots + x_{in}x_n,$$

and $\dfrac{\partial f^{(i)}}{\partial x_j} = a_{ij}$, so that the matrix A is exactly the matrix of partial derivatives

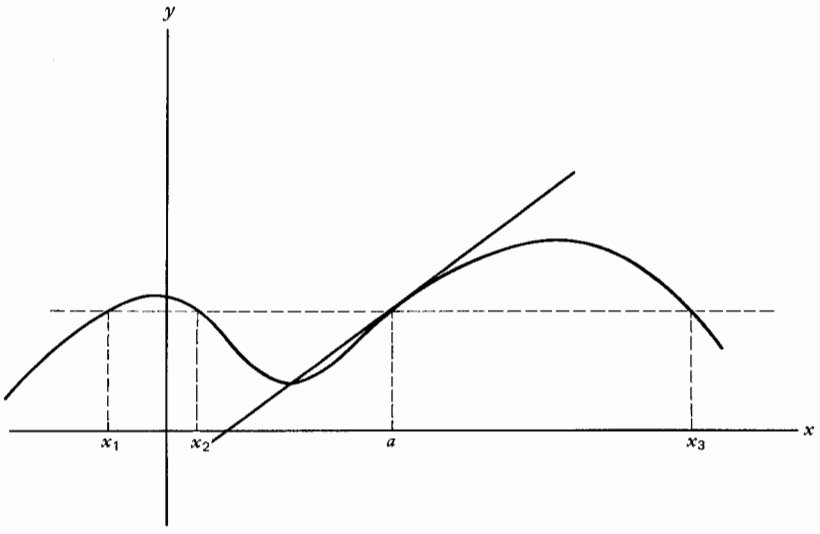


Fig. 88.

evaluated at **a** exhibited in (12.2–2). Thus $A = T_0$ in this particular case. If T_0 is invertible, the equation $\mathbf{y} - \mathbf{f}(\mathbf{a}) = T_0[\mathbf{x} - \mathbf{a}]$ is uniquely solvable for $\mathbf{x} - \mathbf{a}$ in terms of $\mathbf{y} - \mathbf{f}(\mathbf{a})$, and it is reasonable to wonder if this may not imply that $\mathbf{y} = f(\mathbf{x})$ is uniquely solvable for **x** in terms of **y** if we restrict all our considerations to **x**'s near **a** and **y**'s near **f(a)**.

The extremely simple case where $n = 1$ is illuminating. In this case if we assume that the curve $y = f(x)$ is smooth, which means that the derivative $f'(x)$ is continuous, then a small piece of the curve $y = f(x)$ near $x = a$, $y = f(a)$ looks very much like the tangent line $y - f(a) = f'(a)(x - a)$. In this case, if $f'(a) \neq 0$, the linear equation can be solved for x, and it is also true that, for values of x sufficiently close to $x = a$, the relationship between x and y is one-to-one (see Fig. 88).

The figure also shows that the relation between x and y need not be one-to-one for all values of x; in Fig. 88, we see that $f(x_1) = f(a) = f(x_2) = f(x_3)$, and hence there are four values of x for the y value $f(a)$.

We are now ready to state the main theorem. We use a method of proof that depends on **f** being differentiable throughout an open set and having a derivative which is continuous at one point (at least) in that set where **f'** is an invertible operator. To obtain the desired conclusion, we have to restrict the values of **x** to some neighborhood of that point, and this restricted neighborhood will (in general) be only a part of the open set with which we start.

THEOREM VIII. (*Fundamental Inversion Theorem*). *Let* **f** *be a differentiable function from an open set* $\mathcal{N}$ *in* $\mathbb{R}^n$ *to* $\mathbb{R}^n$, *and suppose that there is a point* **a** *in* $\mathcal{N}$ *at which the derivative,* **f'**, *is continuous. Suppose further that the linear*

operator $\mathbf{f}'(\mathbf{a})$ *is invertible. Then there exists in* $\mathcal{N}$ *some neighborhood* U *of* $\mathbf{a}$ *such that*

1. $\mathbf{f}$ *maps each pair of distinct points of* U *into two distinct points in* $\mathbb{R}^n$.
2. *The image set* $\mathbf{f}(U)$, *consisting of all points* $\mathbf{f}(\mathbf{x})$ *for which* $\mathbf{x} \in U$, *is open.*
3. *The inverse mapping* $\mathbf{f}^{-1}$, *defined on* $\mathbf{f}(U)$ *by* $\mathbf{f}^{-1}(\mathbf{y}) = \mathbf{x}$ *where* $\mathbf{x}$ *is that unique point of* U *such that* $\mathbf{y} = f(\mathbf{x})$, *is differentiable at each point of* $\mathbf{f}(U)$, *and if we write* $\mathbf{f}'(\mathbf{x}) = T$ *and denote* $\mathbf{f}^{-1}$ *for convenience by* $\mathbf{g}$, *then* $\mathbf{g}'(\mathbf{y}) = T^{-1}$.
4. $\mathbf{f}^{-1}$ *is continuously differentiable at* $\mathbf{f}(\mathbf{a})$ *and if* $\mathbf{f}$ *happens to be continuously differentiable at other points* $\mathbf{x}$ *of* U, *then* $\mathbf{f}^{-1}$ *is continuously differentiable at the corresponding points* $\mathbf{f}(\mathbf{x})$ *in* $\mathbf{f}(U)$.

Proof. We take the proof in steps, corresponding to the assertions (1) to (4) in the theorem.

Step 1. For convenience we write T_0 for $\mathbf{f}'(\mathbf{a})$ and T for $\mathbf{f}'(\mathbf{x})$, bearing in mind, then, that T depends on $\mathbf{x}$. By hypothesis T_0 is invertible and so for each $\mathbf{w}$ in $\mathbb{R}^n$, $\mathbf{w} = T_0^{-1}[T_0(\mathbf{w})]$. Therefore

$$\|\mathbf{w}\| \leqq \|T_0^{-1}\| \, \|T_0[\mathbf{w}]\|, \qquad \text{or} \qquad \frac{\|w\|}{\|T_0^{-1}\|} \leqq \|T_0[\mathbf{w}]\|. \tag{12.6–3}$$

This relation will be used later. To simplify the notation we shall denote $\dfrac{1}{\|T_0^{-1}\|}$ by m.

Because T_0 is invertible we know that T will also be invertible if $\|T - T_0\|$ is sufficiently small. In fact, we know by Theorem IV in §11.10 that T will be invertible if $\|T - T_0\| < m$. Now because $\mathbf{f}$ is continuously differentiable at $\mathbf{a}$, we can insure $\|T - T_0\| = \|\mathbf{f}'(\mathbf{x}) - \mathbf{f}'(\mathbf{a})\| < m$ if we keep $\|\mathbf{x} - \mathbf{a}\|$ sufficiently small. We choose a positive number c small enough so that two things are true:

(i) The open ball of radius c centered at $\mathbf{a}$ is contained in $\mathcal{N}$.

(ii) For each positive number $r < c$, $\underset{\|\mathbf{x}-\mathbf{a}\| \leq r}{\text{l.u.b.}} \|T - T_0\| < m$.

We shall denote by $\phi(r)$ the least upper bound of $\|T - T_0\|$ referred to in (ii).

Now we introduce U as the open ball $B(\mathbf{a}, c) = \{\mathbf{x} : \|\mathbf{x} - \mathbf{a}\| < c\}$; its image set, $\mathbf{f}(U)$, is of course the set of all points $\mathbf{f}(\mathbf{x})$ where $\mathbf{x} \in U$. Let $\mathbf{u}$ and $\mathbf{v}$ be two distinct points of U. We shall apply the law of the mean, formula (12.4–1), which we can do because the line segment from $\mathbf{u}$ to $\mathbf{v}$ lies in U and hence in $\mathcal{N}$, where $\mathbf{f}$ is continuous and differentiable. So we have that for each $\mathbf{w}$ there exists some $\boldsymbol{\xi}$ strictly between $\mathbf{u}$ and $\mathbf{v}$ such that

$$(\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})) \cdot \mathbf{w} = \mathbf{f}'(\boldsymbol{\xi})[\mathbf{v} - \mathbf{u}] \cdot \mathbf{w}. \tag{12.6–4}$$

Since $\mathbf{u}$ and $\mathbf{v}$ both belong to U, $\|\mathbf{u} - \mathbf{a}\|$ and $\|\mathbf{v} - \mathbf{a}\|$ are each less than c. Let r denote the larger of these two numbers. The closed ball $\overline{B(\mathbf{a}, r)}$ is contained in U, and since balls are convex, $\boldsymbol{\xi} \in \overline{B(\mathbf{a}, r)}$. Since $\|\boldsymbol{\xi} - \mathbf{a}\| \leqq r$, $\|\mathbf{f}'(\boldsymbol{\xi}) - \mathbf{f}'(\mathbf{a})\| \leqq \phi(r) < m$. For convenience let us write $\mathbf{f}'(\boldsymbol{\xi}) = T_1$. We have that $\|T_1 - T_0\| \leqq \phi(r)$. Now (12.6–4)

can be written in the form

$$[\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})] \cdot \mathbf{w} = T_1(\mathbf{v} - \mathbf{u}) \cdot \mathbf{w}$$

from which we can get

$$\|(\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})) \cdot \mathbf{w}\| = \|(T_1 - T_0)[\mathbf{v} - \mathbf{u}] \cdot \mathbf{w} + T_0[\mathbf{v} - \mathbf{u}] \cdot \mathbf{w}\|;$$

applying the triangle inequality (see (11.5–2)) to this we find that

$$\|(\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})) \cdot \mathbf{w}\| \geqq \|T_0[\mathbf{v} - \mathbf{u}] \cdot \mathbf{w}\| - \|(T_1 - T_0)[\mathbf{v} - \mathbf{u}] \cdot \mathbf{w}\|$$

But because

$$\|(T_1 - T_0)(\mathbf{v} - \mathbf{u}) \cdot \mathbf{w}\| \leqq \|T_1 - T_0\| \, \|\mathbf{v} - \mathbf{u}\| \, \|\mathbf{w}\| \leq \phi(r)\|\mathbf{v} - \mathbf{u}\| \, \|\mathbf{w}\|,$$

and

$$\|(\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})) \cdot \mathbf{w}\| \leqq \|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})\| \, \|\mathbf{w}\|,$$

we see that

$$\|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})\| \, \|\mathbf{w}\| \geqq \|T_0[\mathbf{v} - \mathbf{u}] \cdot \mathbf{w}\| - \phi(r)\|\mathbf{v} - \mathbf{u}\| \, \|\mathbf{w}\|.$$

In the last part of the foregoing reasoning, we have twice used the Cauchy inequality (10.12–5).

We can now conclude this part of the proof by exploiting the arbitrariness of $\mathbf{w}$ to choose it in such a way that $\|\mathbf{w}\| = 1$ and $\|T_0[\mathbf{v} - \mathbf{u}] \cdot \mathbf{w}\| = \|T_0[\mathbf{v} - \mathbf{u}]\|$. Such a choice is $\mathbf{w} = T_0[\mathbf{v} - \mathbf{u}]/\|T_0[\mathbf{v} - \mathbf{u}]\|$. Then we have

$$\|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})\| \geqq \|T_0[\mathbf{v} - \mathbf{u}]\| - \phi(r)\|\mathbf{v} - \mathbf{u}\|.$$

But we know from (12.6–3) that $\|T_0[\mathbf{v} - \mathbf{u}]\| \geqq m\|\mathbf{v} - \mathbf{u}\|$, and so we obtain the result

$$\|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})\| \geqq (m - \phi(r))\|\mathbf{v} - \mathbf{u}\|. \tag{12.6–5}$$

This implies that $\mathbf{f}(\mathbf{v}) \neq \mathbf{f}(\mathbf{u})$ when $\mathbf{v} \neq \mathbf{u}$ and completes the proof of Step 1.

Step 2. To show that $\mathbf{f}(U)$ is open we shall prove that to each point $\mathbf{x}_0$ in U corresponds some positive number ρ, which will depend on $\mathbf{x}_0$, such that the open ball with center $f(\mathbf{x}_0)$ and radius ρ is contained in $\mathbf{f}(U)$. As we shall see in the course of the proof, once $\mathbf{x}_0$ is chosen a suitable choice of ρ can be made as follows: If $\mathbf{x}_0 \in U$, then $\|\mathbf{x}_0 - \mathbf{a}\| < c$, and so we can choose r so that $\|\mathbf{x}_0 - \mathbf{a}\| < r < c$. We do this and then define

$$\rho = \tfrac{1}{2}\{m - \phi(r)\}(r - \|\mathbf{x}_0 - \mathbf{a}\|), \tag{12.6–6}$$

where m and $\phi(r)$ are as defined in Step 1.

What we shall prove is that if $\mathbf{y}$ is in $\mathbb{R}^n$ but not in $\mathbf{f}(U)$, then $\|\mathbf{y} - \mathbf{f}(\mathbf{x}_0)\| \geqq \rho$, for this is equivalent to showing that, if $\|\mathbf{y} - \mathbf{f}(\mathbf{x}_0)\| < \rho$, then $\mathbf{y}$ must be in $\mathbf{f}(U)$. Suppose, then, that there is a particular $\mathbf{y}$, say $\mathbf{y} = \mathbf{b}$, in $\mathbb{R}^n$ but not in $\mathbf{f}(U)$. We begin by letting

$$d = glb\|\mathbf{f}(\mathbf{x}) - \mathbf{b}\| \text{ for all } \mathbf{x} \text{ such that } \|\mathbf{x} - \mathbf{a}\| \leqq r.$$

Because $\|\mathbf{f}(\mathbf{x}) - \mathbf{b}\|$ is a continuous function of $\mathbf{x}$ on the closed and bounded set of $\mathbf{x}$'s for which $\|\mathbf{x} - \mathbf{a}\| \leq r$, it follows that the lower bound d is attained at some point $\mathbf{x}_1$, that is, that $d = \|\mathbf{f}(\mathbf{x}_1) - \mathbf{b}\|$. This $\mathbf{x}_1$ is clearly in U, whence $\mathbf{f}(\mathbf{x}_1) \in \mathbf{f}(U)$, and therefore $d > 0$, because $\mathbf{b}$ is not in $\mathbf{f}(U)$. We shall prove that $\|\mathbf{x}_1 - \mathbf{a}\| = r$. Our method will be to show that if we assume $\|\mathbf{x}_1 - \mathbf{a}\| < r$ we are led to a contradiction.

So, suppose $\|\mathbf{x}_1 - \mathbf{a}\| < r$. The contradiction to which we shall come is that $d < d$, and we shall arrive at this by an adroit use of the defining property of the derivative of $\mathbf{f}$ at $\mathbf{x}_1$. Because $\mathbf{x}_1 \in U$, $T_1 = \mathbf{f}'(\mathbf{x}_1)$ is invertible; there is, therefore, a unique vector $\mathbf{h}$ such that $T_1\mathbf{h} = \mathbf{b} - \mathbf{f}(\mathbf{x}_1)$. Evidently $\mathbf{h} \neq \mathbf{0}$, because $\mathbf{b} \neq \mathbf{f}(\mathbf{x}_1)$. We know nothing else about the size of $\|\mathbf{h}\|$, but we can make $\|t\mathbf{h}\| = |t|\,\|\mathbf{h}\|$ small by making t small. We wish to make sure that $\|\mathbf{x}_1 + t\mathbf{h} - \mathbf{a}\| < r$. We can do this as follows: Choose t so that $0 < t$ and $t\|\mathbf{h}\| < r - \|\mathbf{x}_1 - \mathbf{a}\|$. (Here we make use of the assumption that $\|\mathbf{x}_1 - \mathbf{a}\| < r$.) Then

$$\|\mathbf{x}_1 + t\mathbf{h} - \mathbf{a}\| \leqq \|\mathbf{x}_1 - \mathbf{a}\| + t\|\mathbf{h}\| < r.$$

Because $\mathbf{f}$ is differentiable at $\mathbf{x}_1$ it follows (see (12.1–4)) that

$$\lim_{t \to 0} \frac{\|\mathbf{f}(\mathbf{x}_1 + t\mathbf{h}) - \mathbf{f}(\mathbf{x}_1) - T_1(t\mathbf{h})\|}{t\|\mathbf{h}\|} = 0.$$

Let us therefore impose on t the additional requirement that it be small enough to make

$$\frac{\|\mathbf{f}(\mathbf{x}_1 + t\mathbf{h}) - \mathbf{f}(\mathbf{x}_1) - T_1(t\mathbf{h})\|}{t\|\mathbf{h}\|} < \frac{d}{2\|\mathbf{h}\|}.$$

Then

$$\|\mathbf{f}(\mathbf{x}_1 + t\mathbf{h}) - \mathbf{f}(\mathbf{x}_1) - T_1(t\mathbf{h})\| < \frac{td}{2}. \tag{12.6–7}$$

The reason for the choice of the magnitude $d/2\|\mathbf{h}\|$ in the inequality before (12.6–7) will appear presently. We place one more restriction on t, namely, that $t < 1$. These three restrictions are compatible and achievable.

From the definition of d we know that

$$d < \|\mathbf{f}(\mathbf{x}_1 + t\mathbf{h}) - \mathbf{b}\|.$$

But $\mathbf{b} = \mathbf{f}(\mathbf{x}_1) + T_1\mathbf{h}$, and so we see that

$$d \leqq \|\mathbf{f}(\mathbf{x}_1 + t\mathbf{h}) - \mathbf{f}(\mathbf{x}_1) - T_1\mathbf{h}\|.$$

From the triangle inequality we see that

$$d \leqq \|\mathbf{f}(\mathbf{x}_1 + t\mathbf{h}) - \mathbf{f}(\mathbf{x}_1) - T_1\mathbf{h}\| \leqq \|\mathbf{f}(\mathbf{x}_1 + t\mathbf{h}) - \mathbf{f}(\mathbf{x}_1) - T_1(t\mathbf{h})\| + \|T_1(t\mathbf{h}) - T_1(\mathbf{h})\|. \tag{12.6–8}$$

But $\|T_1(t\mathbf{h}) - T_1(\mathbf{h})\| = \|(t-1)T_1\mathbf{h}\| = (1-t)\|\mathbf{b} - f(\mathbf{x}_1)\| = (1-t)d$.
Therefore, from (12.6–7) and (12.6–8) we see that

$$d < \frac{td}{2} + (1-t)d = d\left(1 - \frac{t}{2}\right) < d.$$

This is the contradiction we have been seeking, so we must conclude that $\|\mathbf{x}_1 - \mathbf{a}\| = r$.

We now return to the consideration of $\|\mathbf{b} - \mathbf{f}(\mathbf{x}_0)\|$. We are trying to show that it is bounded away from 0. Now

$$\|\mathbf{b} - \mathbf{f}(\mathbf{x}_0)\| = \|\mathbf{b} - \mathbf{f}(\mathbf{x}_1) + \mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_0)\| \geqq \|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_0)\| - \|\mathbf{b} - \mathbf{f}(\mathbf{x}_1)\|.$$

But $\|\mathbf{b} - \mathbf{f}(x_1)\| = d \leqq \|\mathbf{b} - \mathbf{f}(\mathbf{x}_0)\|$, and

$$\|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_0)\| \geqq \{m - \phi(r)\}\|\mathbf{x}_1 - \mathbf{x}_0\|,$$

as we see by applying (12.6–5) with $\mathbf{x}_0$ and $\mathbf{x}_1$ in place of $\mathbf{u}$ and $\mathbf{v}$, respectively.

[This is legitimate, as we can see by reviewing the reasoning leading to (12.6–5), because $\max\{\|\mathbf{x}_1 - \mathbf{a}\|, \|\mathbf{x}_0 - \mathbf{a}\|\} = r$.] Thus we see from the foregoing that

$$\|\mathbf{b} - \mathbf{f}(\mathbf{x}_0)\| \geqq \{m - \phi(r)\}\|\mathbf{x}_1 - \mathbf{x}_0\| - \|\mathbf{b} - \mathbf{f}(\mathbf{x}_0)\|,$$

or

$$\|\mathbf{b} - \mathbf{f}(\mathbf{x}_0)\| \geqq \tfrac{1}{2}\{m - \phi(r)\}\|\mathbf{x}_1 - \mathbf{x}_0\|.$$

Finally,

$$\|\mathbf{x}_1 - \mathbf{x}_0\| = \|\mathbf{x}_1 - \mathbf{a} + \mathbf{a} - \mathbf{x}_0\| \geqq \|\mathbf{x}_1 - \mathbf{a}\| - \|\mathbf{x}_0 - \mathbf{a}\| = r - \|x_0 - a\|,$$

and so

$$\|\mathbf{b} - f(\mathbf{x}_0)\| \geqq \tfrac{1}{2}\{m - \phi(r)\}(r - \|\mathbf{x}_0 - \mathbf{a}\|) = \rho.$$

With this we have done what we set out to do at the beginning of this Step 2 part of the proof of the theorem.

Step 3. It is clear from the already proven conclusions (1) and (2) that the mapping $\mathbf{f}$ from U to $\mathbf{f}(U)$ has an inverse $\mathbf{f}^{-1}$; for convenience we denote it by $\mathbf{g}$. Let $\mathbf{y}$ be any point of $\mathbf{f}(U)$ and let $\mathbf{x} = \mathbf{g}(\mathbf{y})$, $T = \mathbf{f}'(\mathbf{x})$. We know from the details of Step 1 that T is invertible. What we have to prove is that

$$\lim_{\|\mathbf{k}\| \to 0} \frac{\|\mathbf{g}(\mathbf{y} + \mathbf{k}) - \mathbf{g}(\mathbf{y}) - T^{-1}\mathbf{k}\|}{\|\mathbf{k}\|} = 0, \tag{12.6–9}$$

for this is what is meant by the claim that $\mathbf{g}$ is differentiable at $\mathbf{y}$ and that $\mathbf{g}'(\mathbf{y}) = T^{-1}$.

Before undertaking to prove (12.6–9), however, let us explain why it is plausible to suppose that T^{-1} is the correct value for $\mathbf{g}'(\mathbf{y})$. We know that $\mathbf{f}(\mathbf{g}(\mathbf{y})) = \mathbf{y}$, because $\mathbf{g} = \mathbf{f}^{-1}$. If we *knew* that $\mathbf{g}$ were differentiable, we could use the chain rule to conclude that the derivative of $\mathbf{f}(\mathbf{g}(\mathbf{y}))$ would be $\mathbf{f}'(\mathbf{g}(\mathbf{y}))\mathbf{g}'(\mathbf{y}) = T\mathbf{g}'(\mathbf{y})$. On the other hand, the derivative of the function whose value at $\mathbf{y}$ is $\mathbf{y}$ is clearly the identity operator on $\mathbb{R}^n$. (See Exercise 4 at the end of the chapter.)

We now proceed to prove (12.6–9). We assume that $\|\mathbf{k}\|$ is small enough to insure that $\mathbf{y} + \mathbf{k} \in \mathbf{f}(U)$. Let $\mathbf{h} = \mathbf{g}(\mathbf{y} + \mathbf{k}) - \mathbf{g}(\mathbf{y})$. Then $\mathbf{g}(\mathbf{y} + \mathbf{k}) = \mathbf{x} + \mathbf{h}$ and $\mathbf{y} + \mathbf{k} = \mathbf{f}(\mathbf{x} + \mathbf{h})$. Because the mapping is one-to-one, it follows that $\mathbf{h} \neq \mathbf{0}$ if $\mathbf{k} \neq \mathbf{0}$. Observe that

$$\mathbf{g}(\mathbf{y} + \mathbf{k}) - \mathbf{g}(\mathbf{y}) = \mathbf{h} = T^{-1}T[\mathbf{h}]$$

and

$$T^{-1}[\mathbf{k}] = T^{-1}[\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x})],$$

so that the norm in the numerator of (12.6–9) becomes

$$\|T^{-1}\{T[\mathbf{h}] - [\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x})]\}\|,$$

which is less than or equal to

$$\|T^{-1}\|\,\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - T[\mathbf{h}]\|.$$

For the denominator in (12.6–9) we have

$$\|\mathbf{k}\| = \|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x})\| \geqq (m - \phi(r))\|\mathbf{h}\|,$$

by applying (12.6–5) with $\mathbf{v} = \mathbf{x} + \mathbf{h}$, $\mathbf{u} = \mathbf{x}$. The r here is the maximum of $\|\mathbf{x} + \mathbf{h} - \mathbf{a}\|$ and $\|\mathbf{x} - \mathbf{a}\|$; it may change as $\mathbf{h} \to \mathbf{0}$, but we know that $\|\mathbf{x} + \mathbf{h} - \mathbf{a}\| \leq \|\mathbf{x} - \mathbf{a}\| + \|\mathbf{h}\|$, and

so, if r_0 is chosen so that $\|\mathbf{x}-\mathbf{a}\| < r_0 < c$, we can be assured that $r < r_0$ if $\|\mathbf{h}\|$ is sufficiently small. Then, since $\phi(r) \leq \phi(r_0)$ (by the way in which ϕ is defined in Step 1 of this proof), we see that

$$\|\mathbf{k}\| \geqq (m - \phi(r_0))\|\boldsymbol{h}\|. \tag{12.6–10}$$

Thus the quotient on the left side in (12.6–9) does not exceed

$$\frac{\|T^{-1}\|}{m - \phi(r_0)} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - T[\mathbf{h}]\|}{\|\mathbf{h}\|}. \tag{12.6–11}$$

Here $\mathbf{h}$ is variable, depending on $\mathbf{k}$, and we see from (12.6–10) that $\|\mathbf{h}\| \to 0$ as $\|\mathbf{k}\| \to 0$. But then the expression in (12.6–11) converges to 0 as $\|\mathbf{h}\| \to 0$, by virtue of the fact that $T = \mathbf{f}'(\mathbf{x})$. This completes the proof of Step 3.

Step 4. The argument here is very simple. We recall that $\mathbf{g}'(\mathbf{y}) = T^{-1}$ where $T = \mathbf{f}'(\mathbf{g}(\mathbf{y}))$. Now $\mathbf{g}$ is continuous because, as we proved in Step 3, it is differentiable at each point $\mathbf{y}$ of $\mathbf{f}(U)$, and T^{-1} depends continuously on T (by Theorem V of §11.10). Therefore, by application of the composition theorem (Theorem II, §11.7), $\mathbf{g}'$ is continuous at each point $\mathbf{y} = \mathbf{f}(\mathbf{x})$ for which $\mathbf{x} = \mathbf{g}(\mathbf{y})$ is a point in U at which $\mathbf{f}'$ is continuous; of course $\mathbf{a}$ is one such point.

A few comments are in order about Theorem VIII.

1. It is a powerful result showing how the invertibility of a system of nonlinear simultaneous equations can be inferred from the invertibility of a closely related linear system.
2. It is a *local*, rather than a *global*, theorem. That is, it draws conclusions merely about what happens in some sufficiently small neighborhood of a point. By contrast, Cramer's rule is a global theorem about a system of simultaneous *linear* equations.
3. The proof we have given is independent of the particular dimension number n, and does not require us to display the co-ordinates of any of the vectors. It is a good example of the effectiveness of the use of vector space methods. Finally it should be noted that the assumptions used are *sufficient* to conclude invertibility but not necessary. That is, there are instances in which the mapping may be invertible, *even globally*, and yet some of our assumptions may not be valid (e.g., $\mathbf{f}'(\mathbf{a})$ may fail to be invertible). An elementary example is given in Exercise 17.

12.7 / THE IMPLICIT FUNCTION THEOREM

An important introduction to this theorem is given in §8 of Chapter VIII. The student is advised to reread it at this point. The introduction which we give here is somewhat similar but is intended to introduce the implicit function theorem as an extension to nonlinear analysis of a well-known theorem in linear algebra.

If one has three simultaneous homogeneous linear equations in five variables, for example,

$$\begin{aligned} a_1x_1 + a_2x_2 + \cdots + a_5x_5 &= 0 \\ b_1x_1 + b_2x_2 + \cdots + b_5x_5 &= 0 \\ c_1x_1 + c_2x_2 + \cdots + c_5x_5 &= 0, \end{aligned} \tag{12.7–1}$$

it makes sense to ask whether the system implies that some three of the unknowns, say x_3, x_4, and x_5 are functions of the other two. By transposing the terms involving the other two variables to the other side and using Cramer's rule, we see that a necessary and sufficient condition that x_3, x_4, and x_5 be implicitly determined uniquely as functions of x_1 and x_2 is that the determinant

$$\begin{vmatrix} a_3 & a_4 & a_5 \\ b_3 & b_4 & b_5 \\ c_3 & c_4 & c_5 \end{vmatrix}$$

be different from zero. In other words, by using the inverse function theorem for linear functions, we have obtained an implicit function theorem. We shall now obtain a partial extension to nonlinear functions, that is, to systems like

$$\begin{aligned} f_1(x_1, x_2, x_3, x_4, x_5) &= 0 \\ f_2(x_1, x_2, x_3, x_4, x_5) &= 0 \\ f_3(x_1, x_2, x_3, x_4, x_5) &= 0 \end{aligned} \tag{12.7–2}$$

which reduce to (12.7–1) when the functions are linear. Notice that in the linear case, the determinant above, whose nonvanishing is so important, is the same as the Jacobian determinant

$$\begin{vmatrix} \dfrac{\partial f_1}{\partial x_3} & \dfrac{\partial f_1}{\partial x_4} & \dfrac{\partial f_1}{\partial x_5} \\ \dfrac{\partial f_2}{\partial x_3} & \dfrac{\partial f_2}{\partial x_4} & \dfrac{\partial f_2}{\partial x_5} \\ \dfrac{\partial f_3}{\partial x_3} & \dfrac{\partial f_3}{\partial x_4} & \dfrac{\partial f_3}{\partial x_5} \end{vmatrix}$$

Expressed in this way, this same condition will also be of central importance in the nonlinear case.

The equations (12.7–2) can advantageously be written

$$f_i(x_1, x_2, y_1, y_2, y_3) = 0 \qquad (i = 1, 2, 3). \tag{12.7–3}$$

The change in notation, writing y_1, y_2, y_3 instead of x_3, x_4, x_5, is to help us remember that we want to solve for the y's in terms of the x's. Further, the ordered pair (x_1, x_2) will be denoted by the vector $\mathbf{x}$ and we shall let $(y_1, y_2, y_3) = \mathbf{y}$. The three equations then become

$$f_i(\mathbf{x}, \mathbf{y}) = 0 \qquad (i = 1, 2, 3).$$

For each i, f_i is defined on ordered pairs of the form $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x} \in \mathbb{R}^2$ and $\mathbf{y} \in \mathbb{R}^3$, and each of the three functions is real valued. Since an ordered triple of real-valued functions is commonly thought of as a vector-valued function taking its values in $\mathbb{R}^3$, the system of three equations can conveniently be thought of as

$$\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}. \tag{12.7–4}$$

We have to remember of course that $\mathbf{x} \in \mathbb{R}^2$ and $\mathbf{y} \in \mathbb{R}^3$, and that $\mathbf{f}$ takes its values in $\mathbb{R}^3$.

If A and B are sets, then the set of all ordered pairs of the form (a, b), where $a \in A$ and $b \in B$ is called the *Cartesian product* of A and B and is denoted by $A \times B$. Ordered pairs $(\mathbf{x}, \mathbf{y})$ introduced in the preceding paragraph belong to $\mathbb{R}^2 \times \mathbb{R}^3$; thus we say that $\mathbf{f}$ is a function from some subset of $\mathbb{R}^2 \times \mathbb{R}^3$ to $\mathbb{R}^3$. Since the points of $\mathbb{R}^2 \times \mathbb{R}^3$ are ordered pairs of the form $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ is an ordered pair of real numbers and $\mathbf{y}$ is an ordered triple of real numbers, $(\mathbf{x}, \mathbf{y})$ can be identified with an ordered quintuple of real numbers. Hence, it is common to identify $\mathbb{R}^2 \times \mathbb{R}^3$ with $\mathbb{R}^5$.

In order to get clearly in mind what is meant by solving (12.7–4) for $\mathbf{y}$ as a function of $\mathbf{x}$, consider the following very simple special cases.

$$f(x, y) = 3x + 2y - 5 = 0$$

This can be solved at a glance to get

$$y = \phi(x) = \tfrac{1}{2}(5 - 3x).$$

The essential feature of this function $\phi(x)$ is that if we substitute it for y in the equation $f(x, y) = 0$, we get an identity, namely

$$f(x, \phi(x)) = 3x + 2\phi(x) - 5 = 3x + (5 - 3x) - 5 \equiv 0.$$

In this special instance, the identity holds for all x. In more complicated cases, we may have to settle for identities which are valid only over some subset of the vector space to which x belongs. And in the more general case of the equation $\mathbf{f}(\mathbf{x}, \mathbf{y}) = 0$, to solve for $\mathbf{y}$ in terms of $\mathbf{x}$ means to find a function $\phi(\mathbf{x})$ such that $\mathbf{f}[\mathbf{x}, \phi(\mathbf{x})] \equiv 0$, at least for all $\mathbf{x}$ belonging to some set in $\mathbb{R}^2$.

Consider the set of all pairs $(\mathbf{x}, \mathbf{y})$ such that $\mathbf{f}(\mathbf{x}, \mathbf{y}) = 0$. We call this the *solution set* for the given equation. To avoid dealing with a situation which is of no interest, we must assume that the equation does have solutions, that is, that the solution set is not empty. We are interested in knowing whether this set has the property that when $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$ both belong and $\mathbf{x}_1 = \mathbf{x}_2$, then necessarily $\mathbf{y}_1 = \mathbf{y}_2$. If it does have this property, then $\mathbf{y}$ is determined as a function of $\mathbf{x}$.

There are instances in which the solution set does not determine $\mathbf{y}$ uniquely as a function of $\mathbf{x}$. For example, to take a case in which x and y are both in $\mathbb{R}$, suppose $f(x, y) = x^2 + y^2 - 1$. Then $(0, 1)$ and $(0, -1)$ are both in the solution set, so that there are two values of y (instead of only one) corresponding to $x = 0$. But if we start with the pair $(0, 1)$ and confine attention to pairs (x, y) of the solution set for which x is close to 0 and y is close to 1, we find that this restricted portion of the solution set does define y uniquely as a function of x, the formula being $y = \sqrt{1 - x^2}$ (positive square root). This restriction of attention to all points $(\mathbf{x}, \mathbf{y})$ of the solution set close to a particular point $(\mathbf{a}, \mathbf{b})$ of the solution set is a standard feature of implicit function theorems.

Now for the implicit function theorem. We shall state and prove it for the case we have been discussing of three equations with five variables, but only

trivial changes in notation are required to get a proof valid for m equations in n variables, where $m < n$.

THEOREM IX. (AN IMPLICIT FUNCTION THEOREM). *Suppose that $\mathcal{W}$ is an open subset of $\mathbb{R}^5 = \mathbb{R}^2 \times \mathbb{R}^3$ and that $\mathbf{f}$ is a differentiable function of class $C^{(1)}$ from $\mathcal{W}$ to $\mathbb{R}^3$. Assume further that there is a point $(\mathbf{a}, \mathbf{b})$ in $\mathcal{W}$ such that $\mathbf{f}(\mathbf{a}, \mathbf{b}) = \mathbf{0}$, and assume that the derivative of $\mathbf{f}(\mathbf{a}, \mathbf{y})$, as a function of $\mathbf{y}$, is nonsingular at $\mathbf{y} = \mathbf{b}$. Then the equation $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ determines $\mathbf{y}$ uniquely as a $C^{(1)}$ function of $\mathbf{x}$ near $(\mathbf{a}, \mathbf{b})$. More precisely, there is some neighborhood E of $\mathbf{b}$ in $\mathbb{R}^3$ and some neighborhood S of $\mathbf{a}$ in $\mathbb{R}^2$ and a continuously differentiable function ϕ from S to $\mathbb{R}^3$ such that $\phi(\mathbf{x}) \in E$ and $\mathbf{f}(\mathbf{x}, \phi(\mathbf{x})) = \mathbf{0}$ whenever $\mathbf{x}$ is in S. Furthermore, the only points $\mathbf{y}$ of E which satisfy $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ for some $\mathbf{x}$ in S are those for which $\mathbf{y} = \phi(\mathbf{x})$.*

Proof. Our main tool is the inverse function theorem for functions from $\mathbb{R}^n$ to $\mathbb{R}^n$. Since our function $\mathbf{f}$ is from $\mathbb{R}^5$ to $\mathbb{R}^3$, the tool seems ill-suited to the task, but this disparity can easily be remedied. In the special case of the linear problem (12.7–1), we reduced it to a problem about functions from $\mathbb{R}^3$ to $\mathbb{R}^3$ by transposing the x_1 and the x_2 to the other side. In the nonlinear case (12.7–3), there is no way to get the x_1 and x_2 to the other side, so we artificially inflate the problem to one about functions from $\mathbb{R}^5$ to $\mathbb{R}^5$ in the following trivial way. We define functions $F_1, \ldots, F_5$ as follows:

$$
\begin{aligned}
F_1(x_1, x_2, y_1, y_2, y_3) &= x_1 \\
F_2(x_1, x_2, y_1, y_2, y_3) &= x_2 \\
F_3(x_1, x_2, y_1, y_2, y_3) &= f_1(x_1, x_2, y_1, y_2, y_3) \\
F_4(x_1, x_2, y_1, y_2, y_3) &= f_2(x_1, x_2, y_1, y_2, y_3) \\
F_5(x_1, x_2, y_1, y_2, y_3) &= f_3(x_1, x_2, y_1, y_2, y_3)
\end{aligned}
$$

Notice that this vector-valued function $\mathbf{F}$ can be defined more briefly as follows:

$$\mathbf{F}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{f}(\mathbf{x}, \mathbf{y})).$$

Since $\mathbf{f}$ is continuously differentiable, Theorem III tells us that the same is true of $\mathbf{F}$. By Theorem III, the derivative of $\mathbf{F}$ at $(\mathbf{a}, \mathbf{b})$ is represented by the Jacobian matrix

$$
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
\dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \dfrac{\partial f_1}{\partial y_1} & \dfrac{\partial f_1}{\partial y_2} & \dfrac{\partial f_1}{\partial y_3} \\
\dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \dfrac{\partial f_2}{\partial y_1} & \dfrac{\partial f_2}{\partial y_2} & \dfrac{\partial f_2}{\partial y_3} \\
\dfrac{\partial f_3}{\partial x_1} & \dfrac{\partial f_3}{\partial x_2} & \dfrac{\partial f_3}{\partial y_1} & \dfrac{\partial f_3}{\partial y_2} & \dfrac{\partial f_3}{\partial y_3}
\end{pmatrix}
$$

where all the partial derivatives are evaluated at $(\mathbf{a}, \mathbf{b})$. We wish to show that the linear transformation represented by this matrix is nonsingular.

Notice that the 3×3 submatrix in the lower right-hand corner is nonsingular, because it represents the derivative of $\mathbf{f}(\mathbf{a}, \mathbf{y})$ as a function of $\mathbf{y}$ at $\mathbf{y} = \mathbf{b}$; in the upper left-hand corner we have the 2×2 identity matrix having determinant 1. So the determinant of the 5×5 matrix has the same value as that of the determinant of the 3×3 submatrix in the lower right-hand corner. In other words, the derivative of $\mathbf{F}$ at $(\mathbf{a}, \mathbf{b})$ is nonsingular.

We observe that $\mathbf{F}$ maps $(\mathbf{a}, \mathbf{b})$ into the point $(\mathbf{a}, \mathbf{0})$ of $\mathbb{R}^2 \times \mathbb{R}^3$. By applying the inverse function theorem (Theorem VIII) to $\mathbf{F}$, we see that there must exist a neighborhood $\mathcal{U}$ of $(\mathbf{a}, \mathbf{b})$ in $\mathbb{R}^5$ such that $\mathcal{U}$ is contained in $\mathcal{W}$ and such that $\mathbf{F}$ defines a one-to-one mapping of $\mathcal{U}$ onto a neighborhood $\mathcal{V}$ of $(\mathbf{a}, \mathbf{0})$ in $\mathbb{R}^5$. Moreover, the inverse mapping $\mathbf{F}^{-1}$ is of class $C^{(1)}$ on $\mathcal{V}$. It is easy to prove (see Exercise 21) that every neighborhood of $(\mathbf{a}, \mathbf{b})$ contains a neighborhood of this point which is a Cartesian product $D \times E$, where D is a neighborhood of $\mathbf{a}$ in $\mathbb{R}^2$ and E is a neighborhood of $\mathbf{b}$ in $\mathbb{R}^3$; we can assume that $\mathcal{U}$ itself is such a Cartesian product, and shall do this for convenience.

Let us write points of $\mathcal{V}$ in the form $(\mathbf{x}, \mathbf{z})$, where $\mathbf{x}$ is in $\mathbb{R}^2$ and $\mathbf{z}$ belongs to $\mathbb{R}^3$. To each $(\mathbf{x}, \mathbf{z})$ in $\mathcal{V}$, there corresponds a unique point $\mathbf{F}^{-1}(\mathbf{x}, \mathbf{z}) = (\mathbf{x}, \mathbf{y})$ in $\mathcal{U}$ such that $f(\mathbf{x}, \mathbf{y}) = \mathbf{z}$ and $\mathbf{F}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{z})$. Thus, $\mathbf{y}$ is determined uniquely by $\mathbf{x}$ and $\mathbf{z}$, and this defines $\mathbf{y}$ as a function of $\mathbf{x}$ and $\mathbf{z}$, say $\mathbf{y} = \mathbf{g}(\mathbf{x}, \mathbf{z})$. In particular, $\mathbf{g}(\mathbf{a}, \mathbf{0}) = \mathbf{b}$. Then $\mathbf{F}^{-1}(\mathbf{x}, \mathbf{z}) = (\mathbf{x}, \mathbf{g}(\mathbf{x}, \mathbf{z}))$. Since $\mathbf{F}^{-1}$ is of class $C^{(1)}$, it is readily seen that $\mathbf{g}$ is also of class $C^{(1)}$.

Let S be the set of $\mathbf{x}$'s such that $(\mathbf{x}, \mathbf{0})$ is in $\mathcal{V}$. Clearly, S is contained in D, since $(\mathbf{x}, \mathbf{0})$ comes from some point in $\mathcal{U} = D \times E$ by the mapping $\mathbf{F}$. Observe that $(\mathbf{a}, \mathbf{0})$ comes from $(\mathbf{a}, \mathbf{b})$. It is easy to prove (see Exercise 20) that S is an open subset of $\mathbb{R}^2$. Let us define a function ϕ on S by the formula $\phi(\mathbf{x}) = \mathbf{g}(\mathbf{x}, \mathbf{0})$. Observe that $\phi(\mathbf{x})$ is in $\mathbb{R}^3$ and that $(\mathbf{x}, \phi(\mathbf{x}))$ is in $\mathcal{U}$ when $\mathbf{x}$ is in S. In particular, $\phi(\mathbf{a}) = \mathbf{b}$ and $\phi(\mathbf{x})$ is in E. Observe also that $\mathbf{F}^{-1}(\mathbf{x}, \mathbf{0}) = (\mathbf{x}, \phi(\mathbf{x}))$. This means that $(\mathbf{x}, \mathbf{0}) = \mathbf{F}(\mathbf{x}, \phi(\mathbf{x})) = (\mathbf{x}, \mathbf{f}[\mathbf{x}, \phi(\mathbf{x})])$, and hence, that $\mathbf{f}(\mathbf{x}, \phi(\mathbf{x})) = \mathbf{0}$. The function ϕ is of class $C^{(1)}$ on S, because $\phi(\mathbf{x}) = \mathbf{g}(\mathbf{x}, \mathbf{0})$ and we know that $\mathbf{g}$ is of class $C^{(1)}$.

Finally, to prove the last assertion in Theorem IX, assume that $\mathbf{y}$ is a point of E such that $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ for some $\mathbf{x}$ in S. Then $\mathbf{F}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{0})$ is in $\mathcal{V}$, and hence, since the mapping of $\mathcal{U}$ onto $\mathcal{V}$ is one-to-one, we are assured that $\mathbf{y} = \mathbf{g}(\mathbf{x}, \mathbf{0}) = \phi(\mathbf{x})$. This completes the proof.

In conclusion, we shall now state a more general form of the implicit function theorem. Except for minor changes in notation, the proof is the same as that for the special case just treated.

THEOREM X (THE IMPLICIT FUNCTION THEOREM). *Suppose that $\mathcal{W}$ is an open subset of $\mathbb{R}^{p+q}$ (which we shall identify with $\mathbb{R}^p \times \mathbb{R}^q$) and let $\mathbf{f}$ be a continuously differentiable function from $\mathcal{W}$ to $\mathbb{R}^q$. Assume further that there is a point $(\mathbf{a}, \mathbf{b})$ in $\mathcal{W}$ such that $\mathbf{f}(\mathbf{a}, \mathbf{b}) = \mathbf{0}$, and such that the derivative of $\mathbf{f}(\mathbf{a}, \mathbf{y})$ as a function of $\mathbf{y}$ is nonsingular at $\mathbf{y} = \mathbf{b}$. Then the equation $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$*

determines $\mathbf{y}$ *uniquely as a continuously differentiable function of* $\mathbf{x}$ *near* $(\mathbf{a}, \mathbf{b})$. *More precisely, there is some neighborhood* E *of* $\mathbf{b}$ *in* $\mathbb{R}^q$, *some neighborhood* S *of* $\mathbf{a}$ *in* $\mathbb{R}^p$, *and a continuously differentiable function* ϕ *from* S *to* $\mathbb{R}^q$ *such that* $\phi(\mathbf{x}) \in E$ *and* $\mathbf{f}(\mathbf{x}, \phi(\mathbf{x})) = \mathbf{0}$ *whenever* $\mathbf{x}$ *is in* S. *Furthermore, the only points* $\mathbf{y}$ *of* E *which satisfy* $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ *for some* $\mathbf{x}$ *in* S *are those for which* $\mathbf{y} = \phi(\mathbf{x})$.

12.8 / DIFFERENTIATION OF SCALAR PRODUCTS OF VECTOR VALUED FUNCTIONS OF A VECTOR VARIABLE

We recall from elementary calculus that if two real functions of one real variable are differentiable, the product of the functions is also differentiable, and the formula for the derivative of the product is

$$(fg)'(x) = f(x)g'(x) + f'(x)g(x).$$

This formula has a counterpart for the case of the dot product (also called scalar product or inner product) of $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$ where $\mathbf{f}$ and $\mathbf{g}$ are functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. Recall from (10.12–1) the definition of the dot product in $\mathbb{R}^n$.

THEOREM XI. *Let* $\mathbf{f}$ *and* $\mathbf{g}$ *be functions from an open subset of* $\mathbb{R}^n$ *to* $\mathbb{R}^m$, *and define* ϕ *from* $\mathbb{R}^n$ *to* $\mathbb{R}$ *by*

$$\phi(\mathbf{x}) = \mathbf{f}(\mathbf{x}) \cdot \mathbf{g}(\mathbf{x}).$$

If $\mathbf{f}$ *and* $\mathbf{g}$ *are each differentiable at a particular* $\mathbf{x}$, *so is* ϕ, *and*

$$d\phi(\mathbf{x}, \mathbf{h}) = \phi'(\mathbf{x}) \cdot \mathbf{h} = \mathbf{f}(\mathbf{x}) \cdot \mathbf{g}'(\mathbf{x})\mathbf{h} + \mathbf{f}'(\mathbf{x})\mathbf{h} \cdot \mathbf{g}(\mathbf{x}), \tag{12.8–1}$$

for each vector $\mathbf{h}$ *in* $\mathbb{R}^n$.

Proof. The special case in which one of the two factors, $\mathbf{f}$ and $\mathbf{g}$, is a constant vector is treated in Exercise 7(a). This more general result follows by an extension of the reasoning which establishes the special case. The method of proof is to make straightforward use of the definition of differentiability as applied to $\mathbf{f}$ and $\mathbf{g}$, and then recognize from what we get the fact that $\mathbf{f}(\mathbf{x}) \cdot \mathbf{g}'(\mathbf{x})\mathbf{h} + \mathbf{f}'(\mathbf{x})\mathbf{h} \cdot \mathbf{g}(\mathbf{x})$ fulfills the requirement for being the differential of ϕ at $\mathbf{x}$, which means that it is a good approximation to $\phi(\mathbf{x}+\mathbf{h}) - \phi(\mathbf{x})$ in the appropriate sense when $\|\mathbf{h}\|$ is small.

We begin by writing

$$\mathbf{f}(\mathbf{x}+\mathbf{h}) = \mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x})\mathbf{h} + \boldsymbol{\epsilon}(\mathbf{h})\|\mathbf{h}\|, \tag{12.8–2}$$

$$\mathbf{g}(\mathbf{x}+\mathbf{h}) = \mathbf{g}(\mathbf{x}) + \mathbf{g}'(\mathbf{x})\mathbf{h} + \boldsymbol{\eta}(\mathbf{h})\|\mathbf{h}\|, \tag{12.8–3}$$

where $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$ are vector functions of $\mathbf{h}$ which approach zero when $\mathbf{h}$ does. [Compare with the use of ϵ as a function of $h_1, \ldots, h_n$ in (7–2).] Now form

$$\Delta\phi = \mathbf{f}(\mathbf{x}+\mathbf{h}) \cdot \mathbf{g}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) \cdot \mathbf{g}(\mathbf{x}), \tag{12.8–4}$$

and substitute from (12.8–2) and (12.8–3) on the right in (12.8–4). By the

properties of the scalar product in Exercise 8 of §10.12 one can express the first dot product on the right in (12.8–4) as a sum of nine terms, one of which is $\mathbf{f}(\mathbf{x}) \cdot \mathbf{g}(\mathbf{x})$. On looking carefully at the other eight terms, we can see that

$$\begin{aligned}\phi(\mathbf{x}+\mathbf{h}) - \phi(\mathbf{x}) = \mathbf{f}(\mathbf{x}) \cdot \mathbf{g}'(\mathbf{x})\mathbf{h} + \mathbf{f}'(\mathbf{x})\mathbf{h} \cdot \mathbf{g}(\mathbf{x}) \\ + \|\mathbf{h}\| \{\text{a sum of five terms}\} \\ + \mathbf{f}'(\mathbf{x})\mathbf{h} \cdot \mathbf{g}'(\mathbf{x})\mathbf{h}.\end{aligned}$$

The sum of the first two terms here on the right of the equality sign is evidently linear in $\mathbf{h}$, and therefore of the type desired for the differential of ϕ. What is needed, then, is to show that the remaining expressions on the right, taken together, have an absolute value less than or equal to $\|\mathbf{h}\|$ times some function of $\mathbf{h}$ that approaches zero as $\mathbf{h}$ does. This is not very difficult, and we leave it to the student to carry out the steps. (Exercise 28).

For purposes of applications, it is advantageous to translate (12.8–1) from the language of differentials to the language of derivatives. This reformulation can be broken down into several small steps. Since dot multiplication of vectors is commutative, we can write

$$\phi'(\mathbf{x}) \cdot \mathbf{h} = \mathbf{f}(\mathbf{x}) \cdot \mathbf{g}'(\mathbf{x})\mathbf{h} + \mathbf{g}(\mathbf{x}) \cdot \mathbf{f}'(\mathbf{x})\mathbf{h},$$

and expressing dot products in terms of matrix multiplication (§11.4) leads to

$$\phi'(\mathbf{x}) \cdot \mathbf{h} = [\mathbf{f}(\mathbf{x})]^T[\mathbf{g}'(\mathbf{x})\mathbf{h}] + [\mathbf{g}(\mathbf{x})]^T[\mathbf{f}'(\mathbf{x})\mathbf{h}].$$

Using the fact that matrix multiplication is associative,

$$\begin{aligned}\phi'(\mathbf{x}) \cdot \mathbf{h} &= \{[\mathbf{f}(\mathbf{x})]^T\mathbf{g}'(\mathbf{x})\}\mathbf{h} + \{[\mathbf{g}(\mathbf{x})]^T\mathbf{f}'(\mathbf{x})\}\mathbf{h} \\ &= [\mathbf{f}(\mathbf{x})^T\mathbf{g}'(\mathbf{x}) + \mathbf{g}(\mathbf{x})^T\mathbf{f}'(\mathbf{x})]\mathbf{h}.\end{aligned}$$

In Exercise 4, Chapter 11, it is indicated that the transpose of the sum of two matrices is the sum of their transposes, and that the transpose of a product is the product of the transposes in the reverse order. Using these two facts we get

$$\phi'(\mathbf{x}) \cdot \mathbf{h} = [\mathbf{g}'(\mathbf{x})^T\mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x})^T\mathbf{g}(\mathbf{x})]^T\mathbf{h},$$

and from (11.4–2) this can be written as the dot product of the vector in brackets with $\mathbf{h}$, that is,

$$d\phi(\mathbf{x}, \mathbf{h}) = \phi'(\mathbf{x}) \cdot \mathbf{h} = [\mathbf{g}'(\mathbf{x})^T\mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x})^T\mathbf{g}(\mathbf{x})] \cdot \mathbf{h}.$$

Comparing this with (12.2–8) we see that the only way this can hold for all $\mathbf{h}$ is for the vector in brackets to be the gradient of ϕ at $\mathbf{x}$. Therefore

$$\phi'(\mathbf{x}) = \mathbf{grad}\ \phi(\mathbf{x}) = \mathbf{g}'(\mathbf{x})^T\mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x})^T\mathbf{g}(\mathbf{x}),$$

and we have arrived at the following reformulation of THEOREM XI.

THEOREM XI'. *Under the hypotheses of THEOREM XI where* $\phi(\mathbf{x}) = \mathbf{f}(\mathbf{x}) \cdot \mathbf{g}(\mathbf{x})$, *$\phi$ is differentiable at* $\mathbf{x}$ *and*

$$\phi'(\mathbf{x}) = \mathbf{g}'(\mathbf{x})^T\mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x})^T\mathbf{g}(\mathbf{x}). \qquad (12.8\text{–}5)$$

Notice that when $\mathbf{f} = \mathbf{g}$, *this formula reduces to*

$$\phi'(\mathbf{x}) = 2\mathbf{f}'(\mathbf{x})^T\mathbf{f}(\mathbf{x}). \tag{12.8–6}$$

Example 1. Consider the following system of m linear equations in n unknowns,

$$A\mathbf{x} = \mathbf{b} \tag{12.8–7}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, $\mathbf{b} = (b_1, \ldots, b_m)$ is given, and A is a given $m \times n$ matrix.

Perhaps there is no $\mathbf{x}$ which satisfies the equation (12.8–7). If there is such an $\mathbf{x}$, then of course $A\mathbf{x} - \mathbf{b} = 0$, and hence $\|A\mathbf{x} - \mathbf{b}\|^2 = 0$. Otherwise, $A\mathbf{x} - \mathbf{b}$ is different from zero for all $\mathbf{x}$. Even when no solution exists, one sometimes needs to find a vector which comes as close as possible to satisfying (12.8–7). To be more precise, notice that (12.8–7) associates with each vector $\mathbf{x}$ the vector $A\mathbf{x} - \mathbf{b}$, which is called the *residual* of $\mathbf{x}$. We now broaden our point of view regarding (12.8–7); instead of looking at it as an equation to be solved, we take the problem to be that of finding the vector having the smallest possible residual, that is, we try to make $\|A\mathbf{x} - \mathbf{b}\|$ as small as possible, which is equivalent to making $\|A\mathbf{x} - \mathbf{b}\|^2$ an absolute minimum. If (12.8–7) does have solutions, this approach will lead us to one, since the residual of a solution is the zero vector. Otherwise we shall find the nearest thing to a solution, in the sense of a vector whose residual has the least possible norm. Such a vector is called a *least squares solution* of (12.8–7). If its residual is not actually the zero vector, then a least squares solution is not, of course, a solution in the strict sense. It is a good example of what is frequently called a "generalized solution"—which usually refers to something obtained by relaxing somewhat the requirements that a true solution would have to satisfy. Generalized solutions have great importance in some areas of mathematics.

To find a least squares solution of (12.8–7), we begin with the function

$$\phi(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|^2 = (A\mathbf{x} - \mathbf{b}) \cdot (A\mathbf{x} - \mathbf{b}),$$

and try to minimize it. What we have here is a function from $\mathbb{R}^n$ to $\mathbb{R}$ which, by Theorem X, is clearly differentiable everywhere, since $A\mathbf{x} - \mathbf{b}$ is. By §7.6, the critical points of ϕ are the solutions of

$$\mathbf{grad}\ \phi(\mathbf{x}) = \phi'(\mathbf{x}) = 0,$$

and by putting $\mathbf{f}(\mathbf{x}) = A\mathbf{x} - \mathbf{b}$ in (12.8–6), the critical point equation becomes

$$\phi'(\mathbf{x}) = 2A^T(A\mathbf{x} - \mathbf{b}) = 0,$$

since in this case, $\mathbf{f}'(\mathbf{x}) = A$ for all $\mathbf{x}$. The above equation is clearly equivalent to

$$A^TA\mathbf{x} = A^T\mathbf{b}. \tag{12.8–8}$$

Recall that A is an $m \times n$ matrix, making A^TA an $n \times n$ matrix; and since $\mathbf{b}$ can here be thought of as an $m \times 1$ matrix, $A^T\mathbf{b}$ is an $n \times 1$ matrix, or a vector in $\mathbb{R}^n$.

The point now is that the $n \times n$ matrix A^TA may be invertible, even when the matrix A of our original problem is not. (An example in which this possibility is realized is found in Exercise 29.) When this turns out to be the case, (12.8–8) has the unique solution

$$\mathbf{x} = (A^TA)^{-1}A^T\mathbf{b}. \tag{12.8–9}$$

This shows that when A^TA is invertible, the function ϕ has a unique critical point, namely the $\mathbf{x}$ given by (12.8–9).

There are several ways we can see that ϕ has its absolute minimum at this critical point. The function is obviously nonnegative and takes on the value $\|\mathbf{b}\|^2$ at $\mathbf{x} = 0$. It is easily seen that $\phi(\mathbf{x})$ becomes very large when $\|\mathbf{x}\|$ is large, and therefore, if the positive number R is chosen large enough, we can be sure that $\phi(\mathbf{x}) > \|\mathbf{b}\|^2$ if $\|\mathbf{x}\| > R$. It then follows (from what theorem in Chapter 6?) that $\phi(\mathbf{x})$ has an absolute minimum value for some $\mathbf{x}$, and that the minimum occurs for an $\mathbf{x}$ such that $\|\mathbf{x}\| \leq R$. Exercise 30 asks the student to write out the argument in full. It is very much like Exercises 4 and 5 of §6.3.

Another way of identifying the critical point as a minimum point is to find the Hessian matrix of ϕ. This method was developed in §12.41. It is easy to see that this Hessian matrix is just the second derivative of ϕ. Since the first derivative of ϕ is $2(A^TA\mathbf{x} - A^T\mathbf{b})$, we see at once that the Hessian, H, is simply $2A^TA$. We now consider the quadratic form $\mathbf{h}^TH\mathbf{h}$. Since

$$\mathbf{h}^TH\mathbf{h} = 2\mathbf{h}^TA^TA\mathbf{h} = 2(A\mathbf{h})^TA\mathbf{h} = 2(A\mathbf{h}) \cdot (A\mathbf{h}) = 2\|A\mathbf{h}\|^2,$$

we see that this quadratic form can never be negative. But neither can it be zero unless $\mathbf{h} = 0$; because suppose there were an $\mathbf{h} \neq 0$ such that $\mathbf{h}^TA\mathbf{h} = 0$. Then, by the above equations $A\mathbf{h} = 0$ and therefore $A^TA\mathbf{h} = 0$. But A^TA is nonsingular, and therefore maps only the zero vector into zero. This contradiction forces us to conclude that the Hessian is positive definite. From §12.41 we know that the unique $\mathbf{x}$ given by (12.8–9) makes ϕ a minimum, and since this $\mathbf{x}$ is the only critical point, ϕ must have its absolute minimum there.

Since (12.8–9) gives us a generalized solution of (12.8–7), $(A^TA)^{-1}A^T$ is called a *generalized inverse* of A in those cases where A^TA is invertible. There is vastly more to this subject, but we must leave it at this point. One final thing: If $n = m$ and if A is invertible, so is A^T, and in that case

$$(A^TA)^{-1}A^T = A^{-1}(A^T)^{-1}A^T = A^{-1},$$

showing that if A is invertible, then the generalized inverse turns out to be the same as the inverse.

Example 2. If A is an $n \times n$ symmetric matrix (so that $A^T = A$), then $\mathbf{x} \cdot A\mathbf{x}$ is a quadratic form in n variables $x_1, \ldots, x_n$. Consider the problem of seeking the extreme values of $\mathbf{x} \cdot A\mathbf{x}$ subject to the constraint that $\|\mathbf{x}\|^2 = 1$, that is, that $\mathbf{x} \cdot \mathbf{x} = 1$. Since all vectors having unit length are called unit vectors, and since the set of all unit vectors is called the unit sphere, our problem is one of finding the extreme values of a quadratic form on the unit sphere.

Using the Lagrange multiplier method of §6.8, we form the function

$$\phi(\mathbf{x}) = \mathbf{x} \cdot A\mathbf{x} - \lambda(\mathbf{x} \cdot \mathbf{x}) = \mathbf{x} \cdot (A\mathbf{x} - \lambda\mathbf{x}),$$

calculate $\phi'(\mathbf{x})$ and examine the equation $\phi'(\mathbf{x}) = 0$. To calculate the derivative we apply (12.8–5) with $\mathbf{f}(\mathbf{x}) = \mathbf{x}$ and $\mathbf{g}(\mathbf{x}) = A\mathbf{x} - \lambda\mathbf{x}$. Then $\mathbf{f}'(\mathbf{x}) = I$ and $\mathbf{g}'(\mathbf{x}) = A - \lambda I$, where I is the identity matrix (for which $I\mathbf{x} = \mathbf{x}$ for all $\mathbf{x}$). The result is

$$\boldsymbol{\phi}'(\mathbf{x}) = (A - \lambda I)^T\mathbf{x} + I^T(A\mathbf{x} - \lambda\mathbf{x}).$$

Since I is represented by the matrix with 1's along the main diagonal and 0's everywhere else, it is obvious that $I^T = I$. Since A is symmetric and $A - \lambda I$ differs from A only along the main diagonal, $A - \lambda I$ is also symmetric. This gives

$$\phi'(\mathbf{x}) = (A - \lambda I)\mathbf{x} + A\mathbf{x} - \lambda\mathbf{x} = 2(A\mathbf{x} - \lambda\mathbf{x}),$$

so $\phi'(\mathbf{x}) = 0$ is equivalent to the equation

$$A\mathbf{x} = \lambda\mathbf{x}. \tag{12.8–10}$$

This reveals that any vector that solves our original problem must have the interesting property that when it is transformed by the operator A, the result is exactly the same as when it is multiplied by a certain scalar. Such vectors are called *characteristic vectors* of A. For each characteristic vector $\mathbf{v}$, the scalar by which we can multiply $\mathbf{v}$ to get $A\mathbf{v}$ is called the *characteristic value* associated with $\mathbf{v}$.

In matrix algebra we learn that for the system (12.8–10) there are n distinct characteristic vectors $\{\mathbf{u}_k\}_1^n$ of unit length, and n associated characteristic values (which may or may not be distinct). Thus we have

$$A\mathbf{u}_k = \lambda_k\mathbf{u}_k, \qquad k = 1, 2, \ldots, n.$$

Recall that the extreme values that the quadratic form has on the unit sphere can occur only on certain of these $\mathbf{u}_k$'s. Now notice that it is easy to see what values $\mathbf{x} \cdot A\mathbf{x}$ has on these characteristic vectors; we simply take the dot product of both sides of the above equation with $\mathbf{u}_k$. This gives

$$\mathbf{u}_k \cdot A\mathbf{u}_k = \lambda_k\mathbf{u}_k \cdot \mathbf{u}_k = \lambda_k.$$

In other words, on each unit characteristic vector the value which the quadratic form assumes is simply the associated characteristic value. In matrix algebra courses one learns to find these characteristic values. The largest of them is always the absolute maximum of $\mathbf{x} \cdot A\mathbf{x}$ on the unit sphere, and the smallest is the absolute minimum of this quadratic form subject to the constraint $\|\mathbf{x}\| = 1$.

Example 3. In this final example we show how the method of steepest descent can be used to find solutions of systems of nonlinear equations where the number of equations may or may not be the same as the number of unknowns.

Consider the system of m nonlinear equations in n unknowns,

$$f^{(1)}(x_1, \ldots, x_n) = 0$$
$$\ldots$$
$$f^{(m)}(x_1, \ldots, x_n) = 0$$

which we shall henceforth write in vector form,

$$\mathbf{f}(\mathbf{x}) = \mathbf{0},$$

and assume that $\mathbf{f}$ is a differentiable function which maps $\mathbb{R}^n$ (or an open subset of it) into $\mathbb{R}^m$. We may begin our search not even knowing whether any solutions exist. Our goal of course is to find a solution if any exist, and if there are none, to find a vector which comes as close as possible to being one. In either case, what we are trying to do is to make $\|\mathbf{f}(\mathbf{x})\|$ as small as possible. We shall sometimes speak of it as an attempt to make $\|\mathbf{f}(\mathbf{x})\|^2$ as small as possible, bearing in mind that the two ways of saying it come down to the same thing. Even if there is no value of $\mathbf{x}$ which makes $\mathbf{f}(\mathbf{x})$ vanish, we shall call $\mathbf{x}$ a *least square solution* of $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ if it makes $\|\mathbf{f}(\mathbf{x})\|$ a minimum, that is, which makes $\mathbf{f}(\mathbf{x})$ as close to the zero vector as possible.

Since $\|\mathbf{f}(\mathbf{x})\|^2 = \mathbf{f}(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x})$, an application of (12.8–6) gives

$$\mathbf{grad}\, \|\mathbf{f}(\mathbf{x})\|^2 = 2[\mathbf{f}'(\mathbf{x})]^T \mathbf{f}(\mathbf{x})$$

where $\mathbf{f}'(\mathbf{x})$ is the $(m \times n)$ Jacobian matrix of $\mathbf{f}$, so its transpose is an $(n \times m)$ matrix which maps $\mathbf{f}(\mathbf{x})$ into $\mathbb{R}^n$.

Since $[\mathbf{f}'(\mathbf{x})]^T \mathbf{f}(\mathbf{x})$ is a vector in $\mathbb{R}^n$ pointing from $\mathbf{x}$ in the direction of most rapid increase of $\|\mathbf{f}(\mathbf{x})\|$, to decrease $\|\mathbf{f}(\mathbf{x})\|$ most rapidly we should move from $\mathbf{x}$ in the opposite direction. This suggests that we begin our search for a solution of $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ by making an initial guess $\mathbf{x}_0$. Then, for proper choice of some positive number α_0, the next approximation, $\mathbf{x}_1 = \mathbf{x}_0 - \alpha_0[\mathbf{f}'(\mathbf{x}_0)]^T \mathbf{f}(\mathbf{x}_0)$, should be such that $\|\mathbf{f}(\mathbf{x}_1)\|$ is smaller than $\|\mathbf{f}(\mathbf{x}_0)\|$. Repetition of this process leads to the sequence-generating procedure

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \alpha_n[\mathbf{f}'(\mathbf{x}_n)]^T \mathbf{f}(\mathbf{x}_n).$$

For a good choice of the α_n's, $\{\|\mathbf{f}(\mathbf{x}_n)\|\}_0^\infty$ should be a decreasing sequence of positive numbers converging to a minimum value of $\|\mathbf{f}(\mathbf{x})\|$, in which case the sequence $\{\mathbf{x}_n\}_0^\infty$ could be expected to converge to either a solution or a least square solution of $\mathbf{f}(\mathbf{x}) = \mathbf{0}$.

Skill in the use of this method is a matter of making good choices for the α_n's. A bad choice of α_n would be one which would make $\|\mathbf{f}(\mathbf{x}_{n+1})\|$ bigger than $\|\mathbf{f}(\mathbf{x}_n)\|$. This would mean that even though we started out from $\mathbf{x}_n$ in the right direction, we went too far. To remedy this mistake, one might then repeat the determination of $\mathbf{x}_{n+1}$, using a value for α_n only half as big.

As you might have begun to suspect, this method of steepest descent is likely to be tediously slow in converging, and in this respect it is much inferior to Newton's method—if Newton's method converges at all. In cases where New-

ton's method can be used if one starts close enough to the desired solution, a good strategy might be to begin with the method of steepest descent and then switch to Newton's method when indications are that a solution is near.

EXERCISES

1. Prove that, if a function $\mathbf{f}$ from an open set in $\mathbb{R}^n$ to $\mathbb{R}^m$ is differentiable at a point $\mathbf{a}$, the derivative $\mathbf{f}'(\mathbf{a})$ is uniquely determined. HINT: Suppose there were two T's, T_1 and T_2, satisfying (12.1–4). Use the triangle inequality to show that

$$\lim_{\|\mathbf{h}\|\to 0} \frac{\|(T_1 - T_2)\mathbf{h}\|}{\|\mathbf{h}\|} = 0,$$

and from this explain in detail why $\|T_1 - T_2\| = 0$.

2. Prove that if $\mathbf{f}$ (as in Exercise 1) is differentiable at $\mathbf{a}$, it is continuous at $\mathbf{a}$.

3. Suppose the real-valued function f of $\mathbf{x} = (x_1, \ldots, x_n)$, defined in a neighborhood of $\mathbf{a} = (a_1, \ldots, a_n)$, has first partial derivatives $\dfrac{\partial f}{\partial x_i} = f_i$ at each point of the neighborhood, and that these are continuous at $\mathbf{a}$. Show that f is differentiable at $\mathbf{a}$. HINT: The proof rests on n applications of the law of the mean for functions of one real variable, together with the addition and subtraction of $n - 1$ terms, as is done in (7.1–2) for the case $n = 2$. For the general case some good notation is very helpful. We illustrate for the case $n = 3$. Let $\mathbf{h} = (h_1, h_2, h_3)$, $\mathbf{h}^{(0)} = (0, 0, 0)$, $\mathbf{h}^{(1)} = (h_1, 0, 0)$, $\mathbf{h}^{(2)} = (h_1, h_2, 0)$, $\mathbf{h}^{(3)} = (h_1, h_2, h_3)$. Verify that

$$\begin{aligned} f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) &= \sum_{k=1}^{3} [f(\mathbf{a}+\mathbf{h}^{(k)}) - f(\mathbf{a}+\mathbf{h}^{(k-1)})] \\ &= \sum_{k=1}^{3} h_k f_k(\mathbf{a} + (1-\theta_k)\mathbf{h}^{(k-1)} + \theta_k \mathbf{h}^{(k)}) \end{aligned}$$

for certain numbers θ_k, where $0 < \theta_k < 1$. Now complete the argument.

4. Consider the function $\mathbf{G}(\mathbf{y}) = \mathbf{y}$ defined on some open set S in $\mathbb{R}^n$. Prove that G is differentiable on S with $\mathbf{G}'(\mathbf{y}) = I$ (the identity operator on R^n) for each $\mathbf{y}$ in S.

5. Find the flaws in the following. If we let $n = 1$ in Exercise 4, we have that if $G(\mathbf{y}) = \mathbf{y}$ on some open interval (a, b), then G' exists and is equal to I at all points of (a, b). And clearly G is the identity function on (a, b). Therefore $G' = G$ on (a, b). Hence $G(\mathbf{y}) = ce^{\mathbf{y}}$, proving that $\mathbf{y} = ce^{\mathbf{y}}$ for all $\mathbf{y}$ in (a, b).

6. Given $T \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, define $\mathbf{f}(\mathbf{x}) = T\mathbf{x}$ for each $\mathbf{x}$ in an open set U in $\mathbb{R}^n$. Prove that $\mathbf{f}$ is differentiable at each point of U and that $\mathbf{f}'(\mathbf{x}) = T$ for each $\mathbf{x}$.

7. (a) Given a function $\mathbf{f}$ from an open set in $\mathbb{R}^n$ to $\mathbb{R}^m$ and a fixed vector $\mathbf{a}$ in $\mathbb{R}^m$, show that if $\mathbf{f}$ is differentiable at $\mathbf{x}_0$, then so is the function $\mathbf{F}$ from $\mathbb{R}^n$ to $\mathbb{R}$ defined by $\mathbf{F}(\mathbf{x}) = \mathbf{a} \cdot \mathbf{f}(\mathbf{x})$, the differential being given by $d\mathbf{F}(\mathbf{x}_0, \mathbf{h}) = \mathbf{a} \cdot \mathbf{f}'(\mathbf{x}_0)\mathbf{h}$.
(b) Show that $\mathbf{F}$ is continuous at $\mathbf{x}_0$ if $\mathbf{f}$ is, even when $\mathbf{f}$ is not differentiable.

8. Suppose that $\mathbf{f}$ is a differentiable function from an open connected set G in $\mathbb{R}^n$ to $\mathbb{R}^m$, and suppose that its derivative at each point is the zero transformation. Show that $\mathbf{f}$ has the same value at all points of G. HINT: Treat first the case where G is convex.

This is a generalization of Theorem V of §1.2 and also of Theorem VII of §7.4.

9. (a) Let $\mathbf{g}$ be a function from $[0, 1]$ in $\mathbb{R}$ to $\mathbb{R}^n$, defined by $\mathbf{g}(t) = \mathbf{u} + t\mathbf{w}$, where $\mathbf{u}$ and $\mathbf{w}$ are fixed vectors in $\mathbb{R}^n$. Find $d\mathbf{g}(t, h)$ for $t \in [0, 1]$ and $h \in \mathbb{R}$. Explain the difference between $\mathbf{g}'(t)$ as a member of $\mathscr{L}(\mathbb{R}, \mathbb{R}^n)$ and $\mathbf{g}'(t)$ regarded as a vector in $\mathbb{R}^n$.

(b) If $\mathbf{f}$ is a differentiable function from an open set in $\mathbb{R}^n$ to $\mathbb{R}^m$ and $\mathbf{F}(t) = \mathbf{f}(\mathbf{g}(t))$, where $\mathbf{g}$ is as in part (a), find $d\mathbf{F}(t, h)$ and $\mathbf{F}'(t)$.

10. Show that the function f from $\mathbb{R}^2$ to $\mathbb{R}$ defined as follows,

$$f(x, y) = \frac{xy}{x^2 + y^2} \qquad \text{if } (x, y) \neq (0, 0),$$

and

$$f(0, 0) = 0,$$

has partial derivatives with respect to both x and y at $(0, 0)$ but is not continuous there, and hence is certainly not differentiable at the origin.

11. Let $g(x, y) = \frac{xy^2}{x^2 + y^4}$ if $(x, y) \neq (0, 0)$, and

$$g(0, 0) = 0.$$

Show that g has directional derivatives in all directions at the origin, but that it is not differentiable there. Is g continuous at the origin?

12. Define $F(x, y) = (2x^2 - y)(y - x^2)$ if $y \leqq x^2$ or $2x^2 \leqq y$, and define

$$F(x, y) = \frac{(2x^2 - y)(y - x^2)}{xy} \quad \text{if} \quad 0 < x^2 < y < 2x^2.$$

Show that **(a)** F is continuous at each point, including the point $(0, 0)$, and **(b)** F has a directional derivative in every direction at $(0, 0)$, but **(c)** F is not differentiable at $(0, 0)$.

13. With f as in the opening paragraph of §12.21, suppose f is differentiable at $\mathbf{x} = \mathbf{a}$.
(a) Prove that $D_{\mathbf{u}}f(\mathbf{a}) = -D_{-\mathbf{u}}f(\mathbf{a})$.
(b) Assuming that $\mathbf{v} \neq -\mathbf{u}$, express $D_{\mathbf{u}}f(\mathbf{a}) + D_{\mathbf{v}}f(\mathbf{a})$ in terms involving just one directional derivative at $\mathbf{a}$.

14. If $\mathbf{f}$ is a function from some subset A of $\mathbb{R}^n$ to $\mathbb{R}^m$, prove that $\mathbf{f}$ is continuous at the point $\mathbf{a}$ of A if and only if each of its component functions is continuous at $\mathbf{a}$.

15. Assume that $\mathbf{f}$ is a function from an open subset U of $\mathbb{R}^n$ to $\mathbb{R}^m$, and that $\mathbf{a}$ is a point of U at which all mn of the first partial derivatives of the component functions of $\mathbf{f}$ exist. Prove that if T is that element of $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ whose standard representation is the Jacobian matrix of $\mathbf{f}$ at $\mathbf{a}$, then either $\mathbf{f}$ is differentiable at $\mathbf{a}$ with derivative T, or else,

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{a} + \mathbf{h}) - \mathbf{f}(\mathbf{a}) - T\mathbf{h}\|}{\|\mathbf{h}\|}$$

does not exist. HINT: Show that the indicated limit, if it exists, must be 0.

16. Prove that if T is an invertible linear operator on $\mathbb{R}^n$, and $\boldsymbol{\eta}$ is a function from some neighborhood of the origin in $\mathbb{R}^n$ to $\mathbb{R}^n$, and $\lim_{\mathbf{x} \to \mathbf{0}} T\boldsymbol{\eta}[(\mathbf{x})] = 0$, then

$$\lim_{\mathbf{x} \to \mathbf{0}} \boldsymbol{\eta}(\mathbf{x}) = 0.$$

17. Observe that the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$ is continuously differentiable everywhere and that the derivative at 0 is singular (i.e., not invertible). The function f is nevertheless invertible—in every neighborhood of the origin, in fact. Construct functions from $\mathbb{R}^n$ to $\mathbb{R}^n$ exhibiting these properties.

18. Show that it is possible to prove the inverse function theorem in the special case where $n = 1$ without assuming the function to be continuously differentiable. It is sufficient that it merely be differentiable and that its derivative be different from zero at

each point of an open interval (a, b). Show that one can then prove that the function has an inverse on (a, b), that its inverse function is differentiable at each point of its domain, and that the derivative of the inverse can be expressed very simply in terms of the derivative of the given function. HINT: Use Exercise 21 of §3.3.

The authors do not know whether, in Theorem VIII, the hypothesis that $\mathbf{f}$ be continuously differentiable can be replaced by the hypothesis that $\mathbf{f}$ merely have an invertible derivative throughout some neighborhood of $\mathbf{x} = \mathbf{a}$. This hint ought to come at the end of the preceding paragraph.

19. Consider the function $f: \mathbb{R} \to \mathbb{R}$ defined as follows:

$$f(x) = x \qquad \text{if } x \text{ is rational,}$$
$$f(x) = x^2 + x \qquad \text{if } x \text{ is irrational.}$$

Prove that f is not only continuous at 0 but differentiable there with a positive derivative. Then show that f is not invertible in any neighborhood of 0. The fact that f has a positive derivative at 0 does imply that there is some neighborhood of 0 in which $f(x) > f(0)$ if $x > 0$, and $f(x) < f(0)$ if $x < 0$, but there is no neighborhood of 0 throughout which f is either always increasing or always decreasing.

20. Suppose that $\mathcal{W}$ is an open subset of $\mathbb{R}^5$ and that $(a_1, a_2, 0, 0, 0) \in \mathcal{W}$. Show that $\{(x_1, x_2)\colon (x_1, x_2, 0, 0, 0) \in \mathcal{W}\}$ is an open subset of $\mathbb{R}^2$.

21. Suppose that $\mathcal{W}$ is an open subset of $\mathbb{R}^5 = \mathbb{R}^2 \times \mathbb{R}^3$ and that $(\mathbf{a}, \mathbf{b}) \in \mathcal{W}$. Prove that there are open sets $\mathcal{U}$ and $\mathcal{V}$ in $\mathbb{R}^2$ and $\mathbb{R}^3$ respectively such that $\mathbf{a} \in \mathcal{U}$, $\mathbf{b} \in \mathcal{V}$, and $\mathcal{U} \times \mathcal{V} \subset \mathcal{W}$. Notice that this argument can be extended to prove that every open set in $\mathbb{R}^{p+q} = \mathbb{R}^p \times \mathbb{R}^q$ is the union of Cartesian products of open sets in $\mathbb{R}^p$ with open sets in $\mathbb{R}^q$.

How can one distinguish geometrically between those subsets of the plane which are Cartesian products of subsets of the line and those that are not?

22. Write out in complete detail, with accompanying figures, a proof of Theorem X in the special case where $p = q = 1$. Draw a figure for the case which illustrates the possibility that the set S may consist of several disjoint sets (intervals).

23. **(a)** Suppose there exists a function $u = F(x, y)$, continuously differentiable in an open set in $\mathbb{R}^2$, with values in $\mathbb{R}$, and invertible, so that F cannot have the same value at two different points. Show that this leads to a contradiction. HINT: Either $\dfrac{\partial F}{\partial x}$ or $\dfrac{\partial F}{\partial y}$ must be nonzero at some point.

(b) Show that there cannot exist an invertible continuously differentiable mapping from an open subset of $\mathbb{R}^3$ to $\mathbb{R}^2$. HINT: Use an implicit function argument to reduce the situation to one in which part **(a)** can be used.

24. Let $f(T) = T^{-1}$ when $T \in \Omega$. [See §11.10 for definition of Ω and other relevant information, especially the lemma and the fact that Ω is an open set in $\mathscr{L}(\mathbb{R}^n)$.] Thus f is a function from Ω to itself. Prove that f is differentiable at each point of Ω and that its differential is given by $df(T, H) = -T^{-1}HT^{-1}$ for each H in $\mathscr{L}(\mathbb{R}^n)$. HINT: The essential tools are the lemma from §11.10 and the formula in Exercise 21, Chapter 11, for the case $k = 1$. Note that when $\|H\|$ is sufficiently small

$$(T + H)^{-1} = [T(I + T^{-1}H)]^{-1} = (I + T^{-1}H)^{-1}T^{-1}.$$

Now use the result from Exercise 21, Chapter 11 with the T there replaced by $-T^{-1}H$.

25. Using the method of steepest descent as presented in §12.21 as a guide, devise a method of steepest ascent and use it to solve Exercise 11 of §6.3. The calculations should be carried out on a programmable calculator.

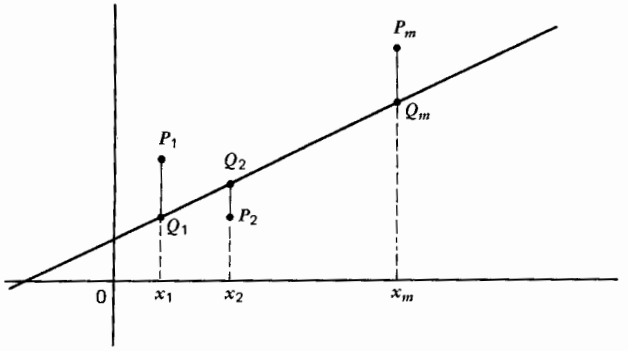


Fig. 89.

26. Use Newton's method to find to four places of decimals a number x such that $\log_e x - e^{-x} = 0$. This does not require a programmable calculator. Any pocket calculator with a few memory locations makes it possible to generate the sequence of x_n's defined by (12.3–1) without writing down any intermediate results. A good starting approximation can be got by sketching the graphs of e^{-x} and $\log_e x$ and estimating the abscissa of their point of intersection.

27. Using a programmable calculator, solve the nonlinear system (12.3–3) by Newton's method.

28. Complete the derivation of (12.8–1).

29. Let points $P_1(x_1, y_1), \ldots, P_m(x_m, y_m)$ be given in xy-plane, not all on the same vertical line. They may or may not lie on some straight line, but sometimes it is important to find a straight line $y = \alpha_1 x + \alpha_2$ that "fits" the points as accurately as possible in the sense that the sum of the squares

$$\phi(\alpha_1, \alpha_2) = (\overline{P_1Q_1})^2 + (\overline{P_2Q_2})^2 + \cdots + (\overline{P_mQ_m})^2$$

is as small as possible, where Q_i is the point in which the line $x = x_i$ intersects the line $y = \alpha_1 x + \alpha_2$. See Fig. 89.

Show that the problem of finding α_1 and α_2 to make $\phi(\alpha_1, \alpha_2)$ a minimum is the same as the problem of finding the least squares solution of the system of m equations

$$\begin{aligned} x_1\alpha_1 + \alpha_2 &= y_1 \\ x_2\alpha_1 + \alpha_2 &= y_2 \\ &\vdots \\ x_m\alpha_1 + \alpha_2 &= y_m \end{aligned}$$

in the two unknowns α_1 and α_2. In this case the $\mathbf{b}$ of (12.8–7) is $(y_1, \ldots, y_m)$, the $\mathbf{x}$ of (12.9–7) is (α_1, α_2) and A is a matrix of m rows and two columns. What is the matrix A^TA in this case, and how do you know that it is nonsingular?

30. Fill in the details of the first argument discussed for showing that the value of $\mathbf{x}$ given by (12.8–9) is that one for which ϕ has its absolute minimum value.

13 / DOUBLE AND TRIPLE INTEGRALS

13 / PRELIMINARY REMARKS

This chapter is designed in such a way that it does not require of the student any previous knowledge of the subject of multiple integrals. Most students of elementary calculus will have had some acquaintance with double and triple integrals before coming to a course in advanced calculus. The extent of this acquaintance will vary considerably with the student, however, and it seems desirable here to start from the beginning.

Multiple integrals have important applications in geometry and the sciences. For these applications the *concept* of the integral (double or triple) is vital, quite apart from the important matter of knowing how to calculate the value of the integral. It is therefore very important for the student to pay attention to the *definitions* of double and triple integrals.

Naturally, the study of multiple integrals builds upon prior knowledge of ordinary definite integrals

$$\int_a^b f(x)\,dx.$$

Such integrals may be called *single* integrals, since they involve functions of a single independent variable, while multiple integrals involve functions of several independent variables.

In the study of integration, some of the salient matters to be considered are: the definition of the integral, the type of function which is integrable (i.e., which has an integral in the sense of the definition), properties of the integral, methods of finding the values of integrals, and applications. Of these matters, the question as to what types of functions are integrable is the most difficult. For most ordinary applications it is sufficient to deal with continuous functions. In Chapter 18, we shall consider the theory of integration, paying special attention to questions of integrability; in particular, we shall prove in that later chapter that continuous functions are integrable. In the present chapter we shall avoid discussions of integrability, taking for granted the existence of the integrals which come to hand.

13.1 / MOTIVATIONS

Consider a thin plane sheet of metal covering a certain region R in the xy-plane. If the sheet is of uniform thickness and texture, the mass of any portion of the sheet will be directly proportional to the area of that portion. The constant of

proportionality may be called the *areal density* of the sheet; it is the mass per unit area. If we denote this density by σ, the mass ΔM of an area ΔA of the sheet is

$$\Delta M = \sigma \Delta A. \tag{13.1–1}$$

How can we locate the center of mass (which is the same as the center of gravity) of the sheet? An intuitive attack on this question may be made as follows. Divide the sheet into a large number, say n, of small pieces. Denote the areas of these pieces by $\Delta A_1, \ldots, \Delta A_n$, and their masses by $\Delta M_1, \ldots, \Delta M_n$. If the maximum dimensions of the pieces are all sufficiently small, it is a reasonable approximation to regard each piece as a particle, all of the mass of the piece being thought of as concentrated at some point within the piece. We thus arrive at the picture of a system of particles, the mass ΔM_k being located at a point $P_k(x_k, y_k)$ in the kth piece, as in Fig. 90. The center of mass of this finite system of particles is at the point $(\bar{x}, \bar{y})$, where

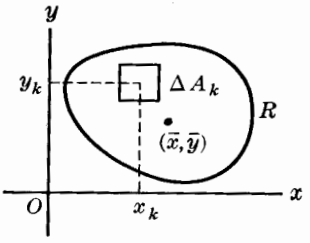


Fig. 90.

$$\begin{aligned} M\bar{x} &= x_1 \Delta M_1 + \cdots + x_n \Delta M_n, \\ M\bar{y} &= y_1 \Delta M_1 + \cdots + y_n \Delta M_n, \\ M &= \Delta M_1 + \cdots + \Delta M_n. \end{aligned} \tag{13.1–2}$$

It is plausible to suppose that, as the number of pieces is increased and the greatest dimension of all the pieces approaches zero, the point $(\bar{x}, \bar{y})$ located by formulas (13.1–2) will approach a limiting position which is the exact center of mass of the plate. If this limiting point has co-ordinates $(\bar{x}, \bar{y})$, we see from (13.1–1) and (13.1–2) that

$$M\bar{x} = \lim(x_1 \sigma \Delta A_1 + \cdots + x_n \sigma \Delta A_n), \tag{13.1–3}$$

with a similar formula for $\bar{y}$. The limit here is to be understood in much the same way as the limit defining a definite single integral (see §1.63). In the present case the limit is a double integral, and we write

$$M\bar{x} = \int\int_R x\sigma \, dA. \tag{13.1–4}$$

The expression on the right here is called the double integral of $x\sigma$ over the region R. There will be a similar formula for $\bar{y}$.

We shall see that double integrals can arise conceptually from many different geometrical or physical problems. For the present we select one further illustration, this time from geometry.

Consider a surface $z = f(x, y)$, where f is defined and continuous in a bounded closed region R of the xy-plane. Suppose that the values of f are

everywhere positive, or perhaps zero, so that the surface never falls below the xy-plane. We now pose the problem: How can we calculate the volume which is under the surface and directly above the region R? This volume will be bounded laterally by a cylindrical surface composed of lines parallel to the z-axis erected at the points of the boundary of R.

The procedure for arriving at a formulation of the volume is very similar to the procedure used in expressing the area under a curve $y = f(x)$ from $x = a$ to $x = b$ as the definite integral $\int_a^b f(x)\,dx$. We divide the region R into a large number (say n) of small pieces (called subregions) of areas $\Delta A_1, \ldots, \Delta A_n$. This divides the volume under consideration into thin columns. Let the portion of the volume directly above the small area ΔA_k be denoted by ΔV_k, so that the total volume is

$$V = \Delta V_1 + \cdots + \Delta V_n.$$

If $P_k(x_k, y_k)$ is any point in the kth subregion of R, let $z_k = f(x_k, y_k)$ be the distance from P_k up to the surface (see Fig. 91). If the dimensions of ΔA_k are sufficiently small, the expression

$$z_k\,\Delta A_k = f(x_k, y_k)\,\Delta A_k$$

is a good approximation to the volume ΔV_k. In fact, denoting by m_k and M_k the minimum and maximum values, respectively, of $f(x, y)$ in the kth subregion, we have

$$m_k\,\Delta A_k \leqq \Delta V_k \leqq M_k\,\Delta A_k. \qquad (13.1\text{–}5)$$

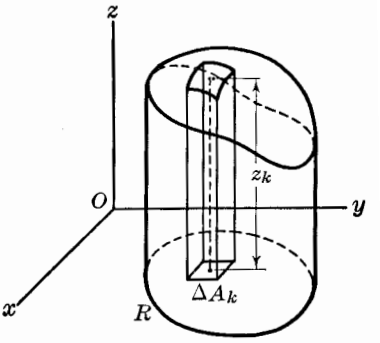


Fig. 91.

Since z_k lies between m_k and M_k, it is clear that ΔV_k and $z_k\,\Delta A_k$ do not differ by more than $(M_k - m_k)\,\Delta A_k$. It thus appears very plausible that the sum

$$f(x_1, y_1)\,\Delta A_1 + \cdots + f(x_n, y_n)\,\Delta A_n \qquad (13.1\text{–}6)$$

is a good approximation to V, and that we have exactly (using a summation symbol to abbreviate the expression (13.1–6)),

$$V = \lim \sum_{k=1}^{n} f(x_k, y_k)\,\Delta A_k, \qquad (13.1\text{–}7)$$

the limit being understood in the sense that the number n is increased indefinitely and the maximum dimension of the subregions $\Delta A_1, \ldots, \Delta A_n$ approaches zero. The choice of the point P_k in the kth subregion is arbitrary, and the exact shape of the subregions is immaterial. The limit on the right in (13.1–7) is equal to a double integral; the notation for the integral is

$$V = \iint_R f(x, y)\,dA. \qquad (13.1\text{–}8)$$

If we compare the limits in (13.1–3) and (13.1–7), we see that they have the same form. In fact, the limit in (13.1–3) is the special case of that in (13.1–7) for which $f(x, y) = x\sigma$ (which happens to be independent of y). Limits of sums having the general form (13.1–6) occur in a variety of contexts, with widely different interpretations. The mathematical properties common to all such limits furnish us with a starting point for the general theory of double integrals.

13.2 / DEFINITION OF A DOUBLE INTEGRAL

Let R be a closed, bounded region in the xy-plane, and let $f(x, y)$ be a function defined in R. In a very general theory of integration, we might seek to place no more restrictions on the function f and the region R than are absolutely necessary for the development of the theory. In the interests of simplicity, however, we shall make rather severe limitations on R, and we shall assume at the outset that the function f is continuous in R. Later it will be possible (and desirable) to broaden the treatment so that certain kinds of discontinuities of f are permitted.

The term "region" was defined in §5.1. We are now concerned with closed, bounded regions. If R is such a region, it has an interior and a boundary. Since R is closed, the boundary is part of the region. The limitations we place on R are in the nature of assumptions about the character of the boundary. We have in mind, roughly speaking, that the boundary of R shall consist of a finite number or arcs of smooth curves joined together to form a closed curve, or possibly several (but a finite number of) such curves. A smooth curve is defined to be a curve with a continuously turning tangent. Circles, parabolas, and straight lines are among the simplest kinds of smooth curves. It is more difficult than one might suppose to be precise in describing the boundary of a region; we shall not attempt to express our assumptions more exactly than in the above statement. Hereafter in this chapter, in speaking of a region R in connection with a double integral, the foregoing assumptions will be taken for granted without explicit mention.

In defining a double integral, we start from approximating sums having the appearance of (13.1–6), but the subregions ΔA_k are chosen in a prescribed manner, and are not arbitrary in shape. Let two sets of lines be drawn, one set parallel to the x-axis, the other set parallel to the y-axis (see Fig. 92). The spacing of the lines need not be regular, but the spacing should be close enough so that the rectangles formed by the intersections of the two sets of lines are small in comparison with R. The network thus formed in the xy-plane is called a rectangular partition; one of the rectangles of the network is called a *cell*. Some of the cells will belong entirely to R; others will contain points which do

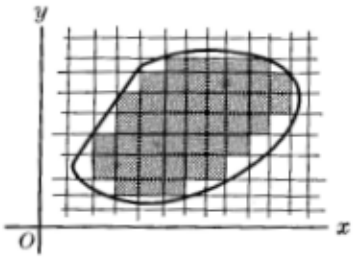


Fig. 92.

not belong to R. For our purposes we discard these latter cells, retaining only those which do not go outside R (these are shaded in Fig. 92). Let the retained cells be numbered (in any order), and denote their areas by $\Delta A_1, \ldots, \Delta A_N$, N being the number of cells retained. It is convenient to refer to the kth cell as the cell ΔA_k. Now let P_k, with co-ordinates (x_k, y_k), be any point in the cell ΔA_k, and form the sum

$$f(x_1, y_1)\,\Delta A_1 + \cdots + f(x_N, y_N)\,\Delta A_N = \sum_{k=1}^{N} f(x_k, y_k)\,\Delta A_k. \tag{13.2–1}$$

The limit of this sum is defined as the double integral of the function f over the region R:

$$\iint_R f(x, y)\,dA = \lim \sum_{k=1}^{N} f(x_k, y_k)\,\Delta A_k. \tag{13.2–2}$$

The integral is the limit of the sum in the following sense: The integral is the real number to which we may approximate as closely as we please by the sums (13.2–1); to get any desired degree of approximation all that is necessary is to make the dimensions of all the relevant cells sufficiently small. The choice of P_k in ΔA_k is arbitrary, and the rectangular partition itself is arbitrary. In choosing partitions so that all the cells have very small dimensions, it is of course apparent that N will become very large.

The precise meaning of (13.2–2) is then as follows: If ϵ is any positive number, there is some corresponding positive number δ which depends upon ϵ (and also upon f and R) such that the inequality

$$\left| \iint_R f(x, y)\,dA - \sum_{k=1}^{N} f(x_k, y_k)\,\Delta A_k \right| < \epsilon \tag{13.2–3}$$

holds for all rectangular partitions in which the maximum cell dimensions are less than δ, and for all choices of the point P_k in the kth cell.

We take it for granted that the continuous function f is integrable, i.e., that the approximating sums (13.2–1) do actually converge to a limit in the sense just specified. This assumption is examined more closely in Chapter 18.

The student may already have speculated upon the fact that in forming the sum (13.2–1) we dealt only with those cells which belong *entirely* to R. We might have proceeded somewhat differently and retained not only those cells just mentioned, but also all those cells which touch the region R in any way (see Fig. 93). This procedure would give us more terms. The limiting value of the approximating sums would be the same as in (13.2–2), however. For, let ΔS denote the area

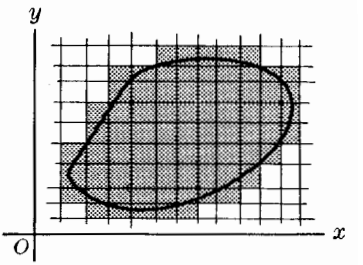


Fig. 93.

of the additional cells, and let M be the maximum of $|f(x, y)|$ in R. Then the contribution of these additional cells to the approximating sum would not exceed $M\,\Delta S$. But ΔS approaches zero as the mesh of the partition is made finer and the maximum cell dimension approaches zero. Hence $M\,\Delta S \to 0$. In a fully detailed treatment of these matters the proof that $\Delta S \to 0$ turns out to depend on the nature of the boundary of R. The assumptions we made concerning this boundary are sufficient to insure that $\Delta S \to 0$.

The notation

$$\iint_R f(x, y)\,dx\,dy \text{ instead of } \int\int_R f(x, y)\,dA \tag{13.2–4}$$

is frequently used for the double integral. The letter which is used for area is of course immaterial. Thus, if we used S for the area of R, and $\Delta S_1, \ldots, \Delta S_N$ for the areas of the cells in the partition, we might denote the double integral by

$$\int\int_R f(x, y)\,dS.$$

13.21 / SOME PROPERTIES OF THE DOUBLE INTEGRAL

A number of important properties of double integrals follow readily from the definition (13.2–2). Two of these properties are embodied in the formulas

$$\int\int_R cf(x, y)\,dA = c\int\int_R f(x, y)\,dA, \tag{13.21–1}$$

$$\iint_R [f(x, y) + g(x, y)]\,dA = \iint_R f(x, y)\,dA + \iint_R g(x, y)\,dA. \tag{13.21–2}$$

In (13.21–1) c is a constant; *a constant factor may be taken outside the integral sign.* In (13.21–2) f and g are any two functions which are integrable in R; *the integral of a sum is the sum of the integrals.* These formulas are immediate consequences of the definition (13.2–2) and the fundamental theorems about limits (see §1.6).

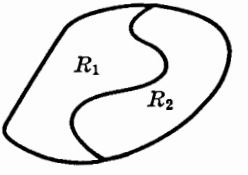


Fig. 94.

Another important property concerns the situation when the region R is composed of two regions R_1, R_2 which have no common points except for parts of their boundaries (see Fig. 94). The formula here is

$$\iint_R f(x, y)\,dA = \iint_{R_1} f(x, y)\,dA + \iint_{R_2} f(x, y)\,dA. \tag{13.21–3}$$

The subregions R_1, R_2 are of course subject to the same assumptions as R as far as their boundaries are concerned.

13.22 / INEQUALITIES. THE MEAN-VALUE THEOREM

It is at once apparent from the definition of the double integral that

$$\iint_R f(x, y)\, dA \geqq 0 \qquad \text{if } f(x, y) \geqq 0 \text{ in } R. \tag{13.22–1}$$

Hence, if $f(x, y) \geqq g(x, y)$ in R, we have

$$\iint_R f(x, y)\, dA \geqq \iint_R g(x, y)\, dA.$$

Now

$$-|f(x, y)| \leqq f(x, y) \leqq |f(x, y)|,$$

and therefore

$$-\iint_R |f(x, y)|\, dA \leqq \iint_R f(x, y)\, dA \leqq \iint_R |f(x, y)|\, dA.$$

This result can be written

$$\left|\iint_R f(x, y)\, dA\right| \leqq \iint_R |f(x, y)|\, dA. \tag{13.22–2}$$

Let A be the area of R. Then, taking $f(x, y) = 1$, we see that

$$\iint_R f(x, y)\, dA = \lim \sum_{k=1}^{n} \Delta A_k = A.$$

Hence, for any constant c,

$$\iint_R c\, dA = cA. \tag{13.22–3}$$

Suppose now (returning to the case of an arbitrary continuous f) that m, M are numbers such that, in R,

$$m \leqq f(x, y) \leqq M.$$

Then

$$mA = \iint_R m\, dA \leqq \iint_R f(x, y)\, dA \leqq \iint_R M\, dA = MA,$$

so that the integral of f has a value between mA and MA. Accordingly, we have the following theorem:

THEOREM I (MEAN-VALUE THEOREM). *If m and M are the minimum and maximum values of $f(x, y)$ in R, there is a number μ such that $m \leqq \mu \leqq M$ and*

$$\iint_R f(x, y)\, dA = \mu A. \tag{13.22–4}$$

The number μ is called the *average* (or mean) *value* of f in R. We cannot as a rule find the value of μ unless we know the value of the integral, so that μ is in fact defined by (13.22–4). Nevertheless, even without exact knowledge of the value of μ, the fact that $m \leqq \mu \leqq M$ makes the formula (13.22–4) useful.

On occasion it is useful to know that there is some point $P(x, y)$ in R at which f takes on its average value μ. When f is continuous, there is always at least one such point P provided R is what is called a *connected* region, i.e., is not composed of two or more closed regions completely separated from each other. For a fuller discussion of what is meant by saying that a region is connected, the student is referred to §17.7. This matter need not be considered any further at present, however.

13.23 / A FUNDAMENTAL THEOREM

In our motivation of the definition of the double integral we used approximating sums which were obtained by decomposing the region R into subregions of arbitrary shape. In the definition (13.2–2), however, we restricted ourselves to subregions which are cells of a rectangular partition. It is important to know that we get the same limit of the approximating sums, no matter how the subdivision of R is made (as long as the pieces are sufficiently regular in shape that we can without ambiguity assign each of them an area). Assurance of this is given by the following theorem:

THEOREM II. *Let the double integral of the continuous function f over R be defined by* (13.2–2), *using rectangular partitions. Let R be divided in any manner into a finite number of subregions, of areas $\Delta A_1, \ldots, \Delta A_n$, the shapes being arbitrary except as qualified in the previous paragraph. Let a point $P_k(x_k, y_k)$ be chosen arbitrarily in ΔA_k. Then*

$$\iint_R f(x, y)\, dA = \lim \sum_{k=1}^{n} f(x_k, y_k)\, \Delta A_k.$$

Furthermore, it is not essential that the areas $\Delta A_1, \ldots, \Delta A_n$ completely fill out the area of R, provided that the amount of area omitted approaches zero in the limiting process.

This theorem appears to be intuitively evident from the geometrical interpretation of the double integral as a volume, as explained in §13.1. A purely analytical proof may be given. In this proof the property (13.21–3) plays an important role. We forego the details.

Among other things, this theorem has the consequence that we are able to define the double integral of a continuous scalar point function over a region R; the integral is independent of co-ordinate systems, and is therefore a scalar invariant. Before reading the following brief remarks on this subject, the student will do well to read the first part of §10.5.

Let R be a plane region of the type assumed in §13.2, and let $f(P)$ be a continuous scalar point function defined in R. With an arbitrary choice of rectangular co-ordinates in the plane, let the representation of $f(P)$ be

$$f(P) = F(x, y),$$

P having co-ordinates (x, y). Consider the integral

$$\iint_R F(x, y)\, dA, \tag{13.23–1}$$

as defined earlier in this chapter. If some other rectangular co-ordinate system is set up in the plane, denote the new co-ordinates of P by (x', y'), and the new representation of $f(P)$ by $\Phi(x', y')$. Then the integral

$$\iint_R \Phi(x', y')\, dA \tag{13.23–2}$$

has the same value as (13.23–1); for, the approximating sums converging to the integral (13.23–2), formed for a rectangular partition of the $x'y'$-coordinate system, will also converge to the integral (13.23–1), by virtue of Theorem II, since $F(x, y) = \Phi(x', y')$ when (x, y) and (x', y') refer to the same point. It follows that if we define the double integral of $f(P)$ over R by

$$\iint_R f(P)\, dA = \iint_R F(x, y)\, dA, \tag{13.23–3}$$

then the integral is a scalar invariant.

13.3 / ITERATED INTEGRALS. CENTROIDS

We shall now learn how to calculate the value of a double integral by performing two successive single integrations. Our initial explanation of this method rests on the geometric interpretation of the double integral as a volume, as in the discussion which culminates in formula (13.1–8).

Our problem is to evaluate the integral

$$\iint_R f(x, y)\, dA, \tag{13.3–1}$$

where f is continuous in R. We shall for convenience assume that $f(x, y)$ is positive at each point of R. The integral (13.3–1) is equal to the volume under the surface $z = f(x, y)$ and directly above the region R. But we can calculate this volume in the following manner:

Let the region R be contained between the lines $y = a$, $y = b$ in the xy-plane (see Fig. 96), and suppose that an arbitrary line $y = y'$, where $a < y' < b$, intersects the boundary of R just twice, as shown in Fig. 96. The co-ordinates X_1, X_2 of these intersections will depend on y'.

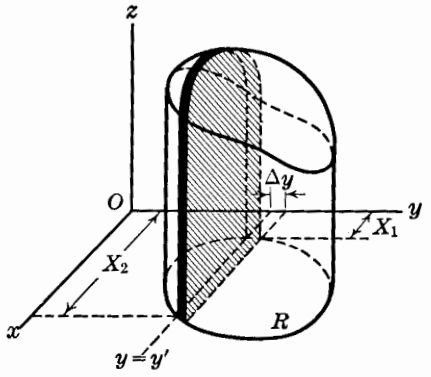


Fig. 95.

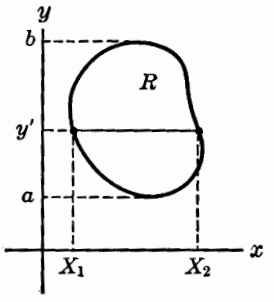


Fig. 96.

Turning now to the three-dimensional picture (Fig. 95), consider the intersection of the plane $y = y'$ with the volume we are seeking. Let the area of this plane section of the volume be denoted by $S(y')$. This area may be expressed as an integral with respect to x, y' being held constant:

$$S(y') = \int_{X_1}^{X_2} f(x, y')\, dx. \tag{13.3–2}$$

If Δy is a small positive number, the parallel plane $y = y' + \Delta y$ will also intersect the solid, and the volume of the slice between these two planes will be approximately $S(y')\,\Delta y$. Hence we should expect the total volume in question to be

$$V = \int_a^b S(y)\, dy. \tag{13.3–3}$$

This is precisely the method which is followed in elementary calculus for finding

the volumes of various kinds of solids, notably solids of revolution, pyramids, and so on. If we now drop the prime on y in (13.3–2), we see from (13.3–3) that the desired volume is

$$V = \int_a^b \left(\int_{X_1}^{X_2} f(x, y)\, dx \right) dy.$$

It is usual to write this expression in the form

$$V = \int_a^b dy \int_{X_1}^{X_2} f(x, y)\, dx. \tag{13.3–4}$$

We have here what is called an *integrated integral.* Two successive integrations are indicated. First we integrate with respect to x, holding y constant in the integrand. The limits of integration X_1, X_2 generally depend on y, and are found by referring to Fig. 96 and taking account of the equations of the curves which form the boundary of R. The second integration is then performed with respect to y, between the limits $y = a$, $y = b$; these limits are the algebraically smallest and largest values, respectively, which y can assume in the region R.

We now have two expressions for the volume under the surface $z = f(x, y)$, one given by a double integral, and the other by an iterated integral. Therefore, we can state a theorem.

THEOREM III. *Let the region R have its extremes in the direction of the y-axis at $y = a$ and $y = b$ respectively $(a < b)$. Let any line between these extremes and parallel to the x-axis cut the boundary of R in exactly two points, so that for $a < y < b$ the boundary of R is formed by two curves*

$$x = X_1(y), x = X_2(y) \qquad (X_1 < X_2).$$

Then a double integral over R can be expressed as an iterated integral:

$$\iint_R f(x, y)\, dA = \int_a^b dy \int_{X_1}^{X_2} f(x, y)\, dx. \tag{13.3–5}$$

It is of course also true that the double integral can be expressed as an iterated integral in which the first integration is with respect to y, provided the appropriate conditions on the boundary of R are fulfilled. See Fig. 97 and formula (13.3–6).

$$\iint_R f(x, y)\, dA = \int_c^d dx \int_{Y_1}^{Y_2} f(x, y)\, dy. \tag{13.3–6}$$

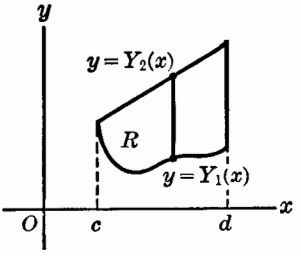


Fig. 97.

We have been led to the discovery of Theorem III by an argument based on geometrical considerations which are highly plausible. The theorem itself is not dependent on the geometrical interpretation, however. Nor is the assumption

that $f(x, y)$ is positive an essential one. The purely analytical proof of the theorem is discussed in Chapter 18, §18.61.

Example 1. Let the region R lie in the first quadrant of the xy-plane, and be bounded by $y = 0$, $y^2 = x$, and $x + 2y = 3$. Let the surface be the plane $3x + 4y + 2z = 12$. Sketch the solid and find the volume for this particular instance of the foregoing discussion.

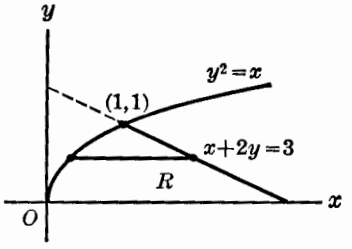


Fig. 98.

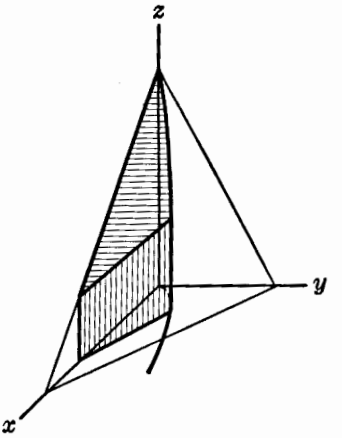


Fig. 99.

The region R is shown in Fig. 98. The intersection of the parabola and line is at (1, 1). The solid in question is shown in Fig. 99. The top surface is

$$z = \tfrac{1}{2}(12 - 3x - 4y).$$

Hence the volume is

$$V = \tfrac{1}{2}\iint_R (12 - 3x - 4y)\, dA.$$

To express this as an iterated integral we read from Fig. 98 that a line parallel to the x-axis cuts the boundary of R at

$$x = y^2 \qquad \text{and} \qquad x = 3 - 2y.$$

Thus $X_1 = y^2$ and $X_2 = 3 - 2y$ in this case. The extreme values of y are $y = 0$, $y = 1$. Hence

$$V = \tfrac{1}{2}\int_0^1 dy \int_{y^2}^{3-2y} (12 - 3x - 4)\, dx.$$

The first integration yields

$$[12x - \tfrac{3}{2}x^2 - 4xy]_{y^2}^{3-2y} = 36 - 24y - \tfrac{3}{2}(9 - 12y + 4y^2) - 12y + 8y^2 - (12y^2 - \tfrac{3}{2}y^4 - 4y^3) = \tfrac{45}{2} - 18y - 10y^2 + 4y^3 + \tfrac{3}{2}y^4.$$

Hence

$$V = \tfrac{1}{2}\int_0^1 (\tfrac{45}{2} - 18y - 10y^2 + 4y^3 + \tfrac{3}{2}y^4)\, dy = \tfrac{1}{2}[\tfrac{45}{2}y - 9y^2 - \tfrac{10}{3}y^3 + y^4 + \tfrac{3}{10}y^5]_0^1 = \tfrac{1}{2}[\tfrac{45}{2} - 9 - \tfrac{10}{3} + 1 + \tfrac{3}{10}] = \tfrac{86}{15},$$

or

$$V = 5.733\ldots.$$

Example 2. Express the double integral of Example 1 as an iterated integral in which the first integration is with respect to y.

When the integration is done in the order requested, we observe that the upper limit Y_2 of the formula (13.3–6) has two different expressions, according as $0 \leqq x \leqq 1$ or $1 \leqq x \leqq 3$ (see Fig. 100).

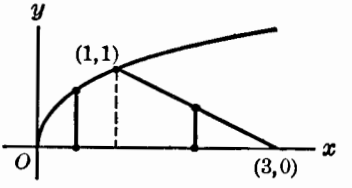


Fig. 100.

$$Y_2 = \sqrt{x}, \quad 0 \leqq x \leqq 1$$

$$Y_2 = \tfrac{1}{2}(3 - x), \quad 1 \leqq x \leqq 3.$$

Under these conditions the iterated integral must be written as the sum of two iterated integrals:

$$\iint_R f(x, y)\, dA = \int_0^1 dx \int_0^{\sqrt{x}} f(x, y)\, dy + \int_1^3 dx \int_0^{\frac{1}{2}(3-x)} f(x, y)\, dy.$$

In the present case $f(x, y) = \frac{1}{2}(12 - 3x - 4y)$.

We leave it for the student to complete the integrations and check with the result of Example 1.

The exercises following this section are intended to give the student practice in evaluating double integrals by means of iterated integrals. In these exercises the applications are limited to the calculation of volumes and the location of centers of gravity of thin sheets of constant areal density. The use of double integrals for this purpose was explained in §13.1. If the thin sheet (often called a lamina) is specified as to shape and position by a region R in the xy-plane, and if the constant areal density is σ, the co-ordinates $(\bar{x}, \bar{y})$ of its center of gravity are given by

$$M\bar{x} = \iint_R x\sigma\, dA, \qquad M\bar{y} = \iint_R y\sigma\, dA, \tag{13.3–7}$$

M being the mass of the sheet. If A is the area of R, we have $M = \sigma A$. Dividing both sides of the formulas (13.3–7) by the constant factor σ, we have

$$A\bar{x} = \iint_R x\,dA, \qquad A\bar{y} = \iint_R y\,dA. \tag{13.3–8}$$

It thus appears that $\bar{x}$ and $\bar{y}$ are independent of σ. The point $(\bar{x}, \bar{y})$ is thus a geometric characteristic of the region R, and is not affected by the material composing the sheet. This point is often called the *centroid* of R. It will be observed (compare Theorem I, §13.22) that $\bar{x}$ is the mean value of x in R; a similar statement holds for $\bar{y}$.

EXERCISES

1. Compute the volume of each of the following regions by the use of a double integral. All literal constants a, b, c, etc. are assumed to be positive.

(a) The tetrahedron cut from the first octant by the plane $3x + 4y + 2z = 12$.

(b) The first octant section cut from the region inside the cylinder $x^2 + z^2 = a^2$ by the planes $z = 0$, $y = 0$, $x = y$.

(c) The hemisphere $x^2 + y^2 + z^2 \leqq a^2$, $z \geqq 0$.

(d) The region between the paraboloid $a^2 z = H(a^2 - x^2 - y^2)$ and the xy-plane.

(e) The region bounded by the ellipsoid $(x^2/a^2) + (y^2/b^2) + (z^2/c^2) = 1$.

(f) The first octant portion of the region inside the cone $a^2y^2 = h^2(x^2 + z^2)$ and between $y = 0$ and $y = h$.

(g) The first octant region bounded by the co-ordinate planes and the cylinders $a^2 y = b(a^2 - x^2)$, $a^2 z = c(a^2 - x^2)$.

2. Find by double integrals the volumes of the tetrahedrons described:

(a) With plane faces $x = 0$, $z = 0$, $x + y = 5$, $8x - 12y + 15z = 0$;

(b) With vertices $(0, 0, 0)$, $(3, 0, 0)$, $(2, 1, 0)$, $(3, 0, 4)$;

(c) Cut from the first octant by the plane $(x/a) + (y/b) + (z/c) = 1$.

3. Interpret the iterated integral

$$\int_0^a dy \int_0^{\sqrt{a^2-y^2}} \frac{2x + 4y}{3}\,dx$$

as the volume of a certain solid, and describe the solid geometrically. Calculate the volume.

4. Locate the centroids of the following plane regions, using double integrals. All literal constants are assumed to be positive.

(a) The triangle with vertices at $(0, 0)$, $(a, 0)$, (a, b).

(b) The triangle with vertices at $(0, 0)$, $(a, 0)$, (b, c), where $a > b$.

(c) The semicircular region $x^2 + y^2 \leqq a^2$, $x \geqq 0$.

(d) The region in the first quadrant and inside the ellipse $(x^2/a^2) + (y^2/b^2) = 1$.

(e) The first quadrant region bounded by $By^2 = H^2 x$, $x = B$, $y = 0$.

(f) The region between the curve $x^3 = y^2$ and the line $x = 1$.

(g) The region bounded by $bx^2 = a^2 y$ and the line $ay = bx$.

5. Locate the centroid of the region described by $x^n \leqq y \leqq 1$, $0 \leqq x \leqq 1$. What is the limiting position of the centroid as $n \to \infty$?

6. Solve Exercise 1(b) by use of a double integral over a suitable region in the xz-plane.

7. Interpret the integral

$$\int_{-a}^{a} dy \int_{a-\sqrt{a^2-y^2}}^{a+\sqrt{a^2-y^2}} x\, dx$$

as a double integral arising in the location of the centroid of a certain plane region, and hence write down the value of the integral without actually carrying out the integration.

8. Locate the centroids of each of the following plane regions:

(a) The region in the first quadrant between $x = 0$, $x = 1$, and between $y = x - x^2$, $y^2 = 2x$.

(b) The region bounded by the two parabolas $y = x^2 + x$, $y = 2x^2 - 2$.

(c) The region defined by $x^2 \leqq y \leqq 2 - x$, $0 \leqq x \leqq 1$.

(d) The region bounded by $y = 0$, $x + y = 2$, and the first quadrant part of $y = x^2$.

13.4 / USE OF POLAR CO-ORDINATES

In suitable situations the evaluation of double integrals is greatly simplified by the use of polar co-ordinates. We shall explain the details of such use.

Suppose we wish to find the value of a double integral

$$\iint_R f(x, y)\, dA. \tag{13.4–1}$$

If we make the change to polar co-ordinates

$$x = r\cos\theta,\ y = r\sin\theta,$$

the function $f(x, y)$ becomes a function of r and θ, say

$$f(x, y) = F(r, \theta). \tag{13.4–2}$$

We are going to appeal to Theorem II (§13.23). Instead of decomposing R by a rectangular partition, we make a subdivision based on a series of circles $r =$ constant and a series of rays $\theta =$ constant. These two series of curves are the two one-parameter families associated with the curvilinear co-ordinates r, θ (see §9.4). We consider the cells of this subdivision which belong entirely to the region R, and number them consecutively in any order (see Fig. 101). Suppose that there are n cells and that ΔA_k is the area of the kth cell. According to Theorem II, referred to above,

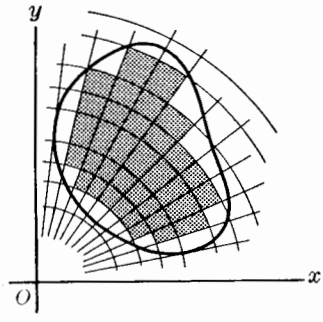


Fig. 101.

$$\iint_R f(x, y)\, dA = \lim \sum_{k=1}^{n} f(x_k, y_k)\, \Delta A_k, \tag{13.4–3}$$

where (x_k, y_k) is a point which may be chosen arbitrarily in the kth cell. Let us agree to choose the point midway between the two circular arcs bounding the cell, and also midway between the two rays bounding the cell (see Fig. 102). If the radii of the circular arcs differ by Δr_k, and if the rays make an angle $\Delta\theta_k$ with each other, the area ΔA_k is easily computed in terms of $\Delta\theta_k$, Δr_k and the polar co-ordinates (r_k, θ_k) of the chosen point. The two circles are respectively

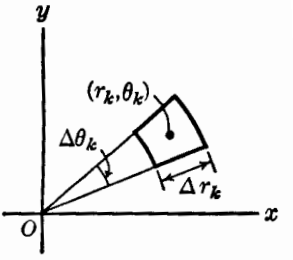


Fig. 102.

$$r = r_k - \tfrac{1}{2}\Delta r_k,\ r = r_k + \tfrac{1}{2}\Delta r_k.$$

Using the formula for the area of a sector of a circle, we have

$$\Delta A_k = \tfrac{1}{2}[(r_k + \tfrac{1}{2}\Delta r_k)^2 - (r_k - \tfrac{1}{2}\Delta r_k)^2]\,\Delta\theta_k,$$

or

$$\Delta A_k = r_k\,\Delta r_k\,\Delta\theta_k.$$

Hence in view of (13.4–2), (13.4–3) may be rewritten

$$\iint_R f(x, y)\,dA = \lim \sum_{k=1}^{n} F(r_k, \theta_k) r_k\,\Delta r_k\,\Delta\theta_k. \tag{13.4–4}$$

Our next task is to work out a method of calculating the right member of (13.4–4). We do this with the aid of the concept of a point transformation, or mapping, from one plane to another, as developed in §9.2. We regard (r, θ) as rectangular co-ordinates in one plane, and (x, y) as rectangular co-ordinates in another plane. The equations $x = r\cos\theta$, $y = r\sin\theta$ define a mapping from one plane to the other. Under this mapping, each of the cells of the subdivision of R shown in Fig. 101 corresponds to a *rectangular cell* in the $r\theta$-plane. The region R itself maps into a certain region in the $r\theta$-plane. Let us denote this region by T. Corresponding to the partition of R by the cells of the polar co-ordinate net, we have a rectangular partition of T. We may consider the function $F(r, \theta)$ as being defined in T, its value at (r, θ) being the same as the value of $f(x, y)$ at the corresponding point of R. If we number the cells in T in correspondence with

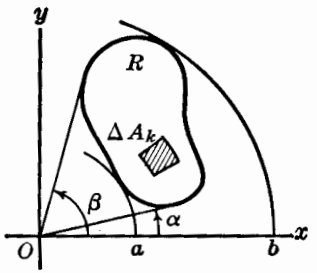


Fig. 103.

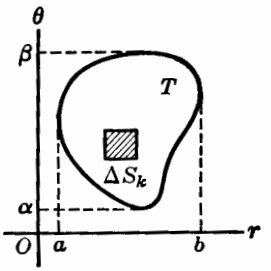


Fig. 104.

their counterparts in R let ΔS_k be the area of the kth cell (see Fig. 103 and Fig. 104) then, in accordance with the notation of Fig. 102 the dimensions of this kth cell are Δr_k by $\Delta\theta_k$, so that $\Delta S_k = \Delta r_k\, \Delta\theta_k$. Accordingly, we see that the limit on the right in (13.4–4) is precisely the limit which defines the double integral

$$\iint_T F(r, \theta) r\, dS$$

of the function $F(r, \theta)r$ over the region T in the $r\theta$-plane. In the notation of (13.2–4) we may write

$$\iint_R f(x, y)\, dx\, dy = \iint_T F(r, \theta) r\, dr\, d\theta. \tag{13.4–5}$$

Observe the factor r in the integrand on the right. This formula is the fundamental result we have been seeking. The last step is to express the integral on the right as an iterated integral with respect to r and θ. The limits of integration will of course depend on the particular region T and may be found by the method explained in §13.3. In practice, however, these limits of integration are usually found directly by examining R. One must know the equations in polar co-ordinates for the curves forming the boundary of R. If we regard $f(x, y)$ and $F(r, \theta)$ as defining the same scalar point function in R, we may write

$$\iint_R f(x, y)\, dA \quad \text{and} \quad \iint_R F(r, \theta)\, dA$$

interchangeably. The use of polar co-ordinates in evaluating the double integral (13.4–1) is then summed up as follows:

$$\begin{aligned} \iint_R F(r, \theta)\, dA &= \int_\alpha^\beta d\theta \int_{R_1}^{R_2} F(r, \theta) r\, dr \\ &= \int_a^b dr \int_{\Theta_1}^{\Theta_2} F(r, \theta) r\, d\theta. \end{aligned} \tag{13.4–6}$$

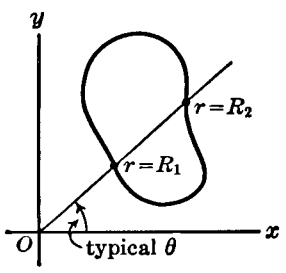


Fig. 105a.

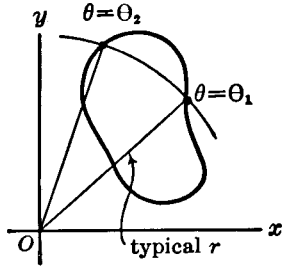


Fig. 105b.

In these iterated integrals α, β are the extreme values of θ, and a, b are the extreme values of r, in the region R. The inner limits R_1, R_2, Θ_1, Θ_2 are read off from the appropriate one of the two figures, as shown (Fig. 105*a* or Fig. 105*b*).

The use of polar co-ordinates may prove advantageous either by simplification of the integrand, or by simplification of the limits of integration in dealing with the iterated integrals. Experience and discernment are required to be able to judge whether or not to use polar co-ordinates. The student's first task is to practice the use of polar co-ordinates.

Example 1. Locate the centroid of the plane region R shown in Fig. 106 (above the x-axis and between the circles of radii a, b).

The centroid is obviously on the y-axis, so $\bar{x} = 0$. The area A of R is $(\pi/2)(b^2 - a^2)$. Hence

$$\frac{\pi}{2}(b^2 - a^2)\bar{y} = \iint_R y\, dA.$$

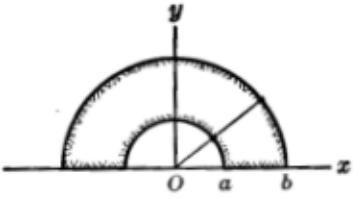


Fig. 106.

The boundaries of R have very simple equations in polar co-ordinates. Therefore, we evaluate the double integral by an iterated integral in polar co-ordinates. Here

$$f(x, y) = y = r \sin\theta = F(r, \theta).$$

Also, $\alpha = 0$, $\beta = \pi$, $R_1 = a$, $R_2 = b$. Therefore, supplying the extra factor r in the integrand, we have

$$\iint_R y\, dA = \int_0^{\pi} d\theta \int_a^b r^2 \sin\theta\, dr$$

$$= \frac{b^3 - a^3}{3} \int_0^{\pi} \sin\theta\, d\theta = \tfrac{2}{3}(b^3 - a^3).$$

Then

$$\bar{y} = \frac{4}{3\pi} \frac{b^3 - a^3}{b^2 - a^2} = \frac{4}{3\pi} \frac{b^2 + ba + a^2}{b + a}.$$

For a semicircular region we put $a = 0$. In this case $\bar{y} = \dfrac{4}{3\pi} b$.

Example 2. Find the volume inside the cylinder $x^2 + (y - a)^2 = a^2$ and between the plane $z = 0$ and the paraboloid $4az = x^2 + y^2$.

The volume in question is given by

$$V = \iint_R \frac{1}{4a}(x^2 + y^2)\, dA,$$

where R is the region in the xy-plane bounded by the circle $x^2 + (y - a)^2 = a^2$. Half of the volume is shown in Fig. 107. Polar co-ordinates are convenient for

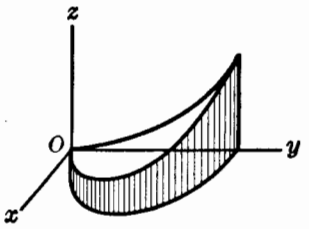


Fig. 107.

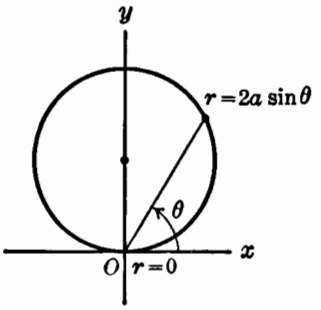


Fig. 108.

this problem, because of both the form of the integrand and the equation of the boundary of R in polar co-ordinates. The polar equation of the boundary of R is $r = 2a \sin\theta$ (see Fig. 108). Hence we have

$$V = \frac{1}{4a}\iint_R r^2\,dA = \frac{1}{4a}\int_0^{\pi} d\theta \int_0^{2a\sin\theta} r^3\,dr.$$

$$V = a^3\int_0^{\pi} \sin^4\theta\,d\theta = 2a^3\int_0^{\pi/2} \sin^4\theta\,d\theta.$$

$$V = 2a^3\frac{3\cdot 1}{4\cdot 2}\frac{\pi}{2} = \frac{3\pi}{8}a^3.$$

In evaluating the last integral we have used one of the standard tabulated formulas for the definite integrals

$$\int_0^{\pi/2} \sin^n\theta\,d\theta,\ n = 1, 2, \ldots.$$

These formulas are very convenient, and the student should be familiar with them.

EXERCISES

1. Calculate the volumes of the solids here described, using double integrals and polar co-ordinates.

(a) Inside the cylinder $x^2 + y^2 = a^2$, between the planes $z = 0$, $z = x$, and in which $x \geqq 0$.

(b) Inside the sphere $x^2 + y^2 + z^2 = a^2$. Deal with the first octant only, and use symmetry.

(c) Inside the cylinder $x^2 + y^2 = a^2$ and between $z = 0$ and $a^2z = h(x^2 + y^2)$.

(d) Between the cone $c^2(z - h)^2 = h^2(x^2 + y^2)$ and the plane $z = 0$.

(e) Inside both the sphere $x^2 + y^2 + z^2 = 4a^2$ and the cylinder $x^2 + (y - a)^2 = a^2$.

(f) Inside the cylinder $x^2 + y^2 = 2ax$ and between the plane $z = 0$ and the cone $z^2 = x^2 + y^2$.

(g) Inside the prism bounded by the planes $y = x$, $y = 0$, $x = a/\sqrt{2}$, and between the plane $z = 0$ and the cone $az = h(x^2 + y^2)^{1/2}$.

2. Locate the centroids of the plane regions described as follows, using double integrals and polar co-ordinates:

(a) In the first quadrant, between $x^2 + y^2 = 2ax$ and $y = 0$.

(b) Between $r = 2a \cos \theta$ and $r \cos \theta = a$, and on the side of the latter curve away from the origin.

(c) Inside the cardioid $r = a(1 + \sin \theta)$.

(d) Inside the first quadrant loop of $r = a \sin 2\theta$.

(e) In the first quadrant, inside $r = 2a \cos \theta$ and outside $r = a$.

(f) Inside the loop of $r^2 = 2a \cos 2\theta$ which is bisected by the ray $\theta = 0$.

3. Find each of the two volumes into which the volume in Exercise 1(e) is divided by the cylinder $x^2 + y^2 = a^2$.

4. Find the volume inside the sphere $x^2 + (y - a)^2 + z^2 = a^2$ and between the planes $x = 0$, $y = x$.

5. Find the volume between the paraboloid $z = x^2 + y^2$ and the plane $z = x$.

13.5 / APPLICATIONS OF DOUBLE INTEGRALS

In §13.1 we introduced the concept of a thin sheet of material substance. The concept of a distribution of matter without thickness is a very useful one. A plane region which carries such a mass distribution is called a lamina. A lamina is a mathematical idealization of a thin sheet, just as a particle is a mathematical idealization of a small, concentrated bit of matter. One may also speak of laminas which are curved surfaces, but here we shall deal only with plane laminas.

We wish to introduce the concept of a lamina of variable density. In the case of constant density, the density of lamina is the ratio of mass to area:

$$\sigma = \frac{M}{A}. \tag{13.5–1}$$

But we may imagine a lamina in which the mass is so distributed that various pieces of the lamina, although of equal area, will have different masses. For the general case, the density of a lamina is, by definition, an integrable function $\sigma(x, y)$ such that when it is integrated over any subregion ΔR of the lamina, it gives the mass of that portion:

$$\Delta M = \iint_{\Delta R} \sigma \, dA. \tag{13.5–2}$$

In particular, the total mass is

$$M = \iint_{R} \sigma \, dA. \tag{13.5–3}$$

We shall consider only the case of continuous densities. If the area of ΔR is

ΔA we see from (13.5–2) by the mean-value theorem (§13.2) that

$$\sigma' \leqq \frac{\Delta M}{\Delta A} \leqq \sigma'', \tag{13.5–4}$$

where σ' and σ'' are respectively the minimum and maximum values of the density in ΔR. Hence, if (x, y) is a fixed point of ΔR, and if we shrink the subregion ΔR so that its maximum diameter approaches zero, we see by the continuity of σ that

$$\lim \frac{\Delta M}{\Delta A} = \sigma(x, y). \tag{13.5–5}$$

This relation replaces (13.5–1) when we deal with laminas of variable density.

Example 1. The density of a square lamina of side b varies in direct proportion to the distance from a particular corner of the square. Find the mass of the lamina.

Let us take the square as shown in Fig. 109 with the particular corner at the origin.

We are given that

$$\sigma = k\sqrt{x^2 + y^2},$$

k being a constant of proportionality. Hence,

$$M = k \iint_R \sqrt{x^2 + y^2}\, dx\, dy,$$

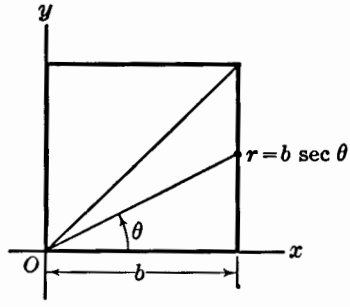


Fig. 109.

R being the square region. We meet difficulties if we proceed to the iterated integral in rectangular co-ordinates. It is better to use polar co-ordinates. The line $y = x$ divides the square into two triangles which are evidently of equal mass. The line $x = b$ has the polar equation $r = b \sec \theta$. Hence, if R_1 denotes the triangle below the line $y = x$,

$$M = 2k \iint_{R_1} r\, dA = 2k \int_0^{\pi/4} d\theta \int_0^{b \sec \theta} r^2\, dr.$$

$$M = \frac{2kb^3}{3} \int_0^{\pi/4} \sec^3 \theta\, d\theta$$

$$= \frac{2kb^3}{3} \left[\tfrac{1}{2} \tan \theta \sec \theta + \tfrac{1}{2} \log \tan\left(\frac{\pi}{4} + \frac{\theta}{2}\right) \right]_0^{\pi/4}$$

$$M = \frac{kb^3}{3} \left[\sqrt{2} + \log \tan \frac{3\pi}{8} \right].$$

Many physical concepts are first formulated for systems of a finite number of discrete particles, and then extended to continuous distributions of matter by

the use of integrals. The guiding principle is that of subdividing the continuous distribution into small parts. Each part is then replaced by a particle, which is obtained by concentrating all the mass of the part at some point within the part. The resulting system of particles is then regarded as an approximation to the continuous distribution, and physical attributes of the continuous mass are assumed to be obtained as limits of the corresponding physical attributes of the approximating system of particles. Where the physical attribute of the system of particles is expressed by a sum, that of the continuous mass will be expressed as the limit of a sum, and this limit will normally be a definite integral. Illustrations are furnished by such concepts as center of mass, moment of inertia, and gravitational or electrostatic attraction. In the case of a lamina of continuous density, when it is subdivided into small parts, of areas $\Delta A_1, \ldots, \Delta A_n$, the mass ΔM_k of the kth part will be expressed by

$$\Delta M_k = \sigma(x_k, y_k)\, \Delta A_k, \tag{13.5–6}$$

where (x_k, y_k) is a suitably chosen point in the part [see (13.5–4) and the remarks at the end of §13.22].

Using the general principle described in the foregoing paragraph we find that the center of mass (center of gravity) of a lamina is given by formulas (13.3–7). When σ is variable it cannot be taken from under the integral sign, of course. In this case we cannot use formulas (13.3–8), and M must be found by integration.

If L is a straight line in space, and if we have a system of particles of masses $m_1, \ldots, m_n$, the perpendicular distance from m_k to L being r_k, then the moment of inertia of the system about L as an axis is defined to be

$$I = m_1 r_1^2 + \cdots + m_n r_n^2.$$

To extend this definition to laminas in the xy-plane, let P_k denote the point (x_k, y_k) in (13.5–6), and let Q_k be the foot of the perpendicular drawn from P_k to L. Then for the approximating system of particles we have $r_k = P_kQ_k$, $m_k = \sigma(x_k, y_k)\, \Delta A_k$; therefore, the moment of inertia of the lamina about L is

$$I = \lim \sum_{k=1}^{n} (P_kQ_k)^2 \sigma(x_k, y_k)\, \Delta A_k.$$

This limit is clearly a double integral. If $D(x, y)$ denotes the perpendicular distance PQ from a typical point $P(x, y)$ of the lamina to the axis L, the double integral is

$$I = \iint_R \sigma D^2\, dA. \tag{13.5–7}$$

In each particular problem D must be expressed as a function of the coordinates. For instance, if L is the y-axis, $D^2 = x^2$, while if L is the z-axis, $D^2 = x^2 + y^2$.

Moments of inertia are often expressed in the form $I = Mk^2$, M being the mass. The constant k is called the *radius of gyration.*

Example 2. Prove the following proposition: *Let a lamina occupy the region R, and let L be a line in the plane of the lamina. Let I denote the moment of inertia of the lamina about L, and let I_0 denote the moment of inertia about an axis L_0 parallel to L through the center of mass of the lamina. Then*

$$I = I_0 + Mh^2 \tag{13.5–8}$$

where h is the distance between L and L_0.

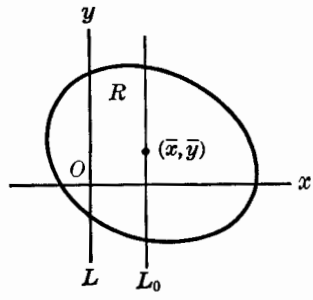


Fig. 110.

It is convenient to locate our co-ordinate system so that L coincides with the y-axis. See Fig. 110. This is permissible, since the physical quantities are independent of the position of the co-ordinate axes. The distance from $P(x, y)$ to L is $D = |x|$. Hence

$$I = \iint_R \sigma x^2 \, dA.$$

The equation of L_0 is $x = \bar{x}$, and so

$$I_0 = \iint_R \sigma(x - \bar{x})^2 \, dA = \iint_R \sigma(x - h)^2 \, dA,$$

for $\bar{x} = h$.

Thus

$$I - I_0 = \iint_R \sigma[x^2 - (x - h)^2] \, dA.$$

Now $x^2 - (x - h)^2 = 2xh - h^2$. Therefore

$$I - I_0 = 2h \iint_R \sigma x \, dA - h^2 \iint_R \sigma \, dA = 2hM\bar{x} - h^2M = Mh^2.$$

Thus (13.5–8) is proved.

If we set $\sigma = 1$ in (13.5–7), the resulting integral

$$\iint_R D^2 \, dA \tag{13.5–9}$$

is called the *second moment* of the region R with respect to the axis L. This is a purely geometric characteristic of R in relation to L. Second moments are used in the theory of elasticity and strength of materials, particularly in the theory of the bending of beams. In the literature these second moments are often called moments of inertia. This is a misnomer, since no mass concept is involved. The physical dimensions of a moment of inertia are mass $\times$ (length)^2, while those of a second moment of a plane region are (length)^4. Second moments are also

important in statistics and elsewhere. The integrals

$$\iint_R x\,dA, \iint_R y\,dA$$

occurring in the formulas (13.3–8) for the centroid are, by contrast, called *first moments* (about the y-axis and x-axis, respectively).

EXERCISES

1. In each of the parts of this exercise a lamina of a certain shape is described, and the manner in which its density varies is defined. Find the mass and locate the center of mass of each lamina. Wherever it occurs in this exercise, k denotes a constant of proportionality.

(a) Triangular lamina with vertices at $(0, 0)$, $(a, 0)$, (a, b); $\sigma = kx$.

(b) The same lamina as in **(a)**, but with $\sigma = ky$.

(c) The lamina occupying the region defined by $x^2 + y^2 \leqq a^2$, $x \geqq 0$, $y \geqq 0$, with $\sigma = kx$.

(d) The lamina of **(c)**, but with $\sigma = kxy$.

(e) The lamina of **(c)**, but with $\sigma = k(x^2 + y^2)^{1/2}$.

(f) The lamina in the first quadrant, bounded by $bx^2 = a^2y$, $x = 0$, $y = b$, with $\sigma = kx$.

(g) The lamina of **(f)**, but with $\sigma = k(b - y)$.

(h) The triangular lamina cut from the first quadrant by the line $x + y = a$, with σ directly proportional to the product of the distances from (x, y) to the sides of the triangle.

(i) The lamina in the first quadrant, bounded by $r = 2a \cos\theta$ and $\theta = 0$, with $\sigma = kr$.

(j) The lamina of **(i)**, but with $\sigma = kr \sin 2\theta$.

2. For any distribution of mass, let I_x and I_y denote the moments of inertia of the distribution about the x-axis and the y-axis, respectively, and let J_0 denote the moment of inertia about the axis perpendicular to the xy-plane at the origin. Show that $J_0 = I_x + I_y$.

3. In each part of this exercise, a homogeneous lamina is described. Find I_x, I_y, and J_0 in each case (see Exercise 2).

(a) The circular lamina bounded by $x^2 + y^2 = a^2$.

(b) The annulus bounded by the two circles $x^2 + y^2 = r_i^2$ $(i = 1, 2, r_1 < r_2)$.

(c) The rectangular lamina bounded by $x = \pm a$, $y = \pm b$.

(d) The triangular lamina bounded by $y = 0$, $x = a$, $ay = bx$.

(e) The elliptical lamina bounded by $b^2x^2 + a^2y^2 = a^2b^2$.

(f) The lamina bounded by $y^2 = 2ax$ and $x = 2a$.

(g) The lamina occupying the circular segment $x^2 + y^2 \leqq b^2$, $x \geqq b \cos\alpha$ where $0 < \alpha < \pi/2$.

(h) The lamina occupying the circular sector $0 \leqq r \leqq b$, $-\beta \leqq \theta \leqq \beta$, where $0 < \beta \leqq \pi/2$.

4. For a lamina occupying a region R in the xy-plane, the double integral

$$U_{xy} = \iint_R \sigma xy\,dA$$

is called the *product of inertia* of the lamina with respect to the co-ordinate axes.

Calculate this product of inertia for each of the following laminas, assuming constant density:

(a) The rectangular lamina bounded by $x = 0$, $x = a$, $y = 0$, $y = b$.

(b) The first-quadrant quarter of the circular lamina bounded by $x^2 + y^2 = a^2$.

(c) The triangular lamina with vertices at $(0, 0)$, $(a, 0)$, (a, b).

(d) The square lamina bounded by $x = -a$, $x = 2a$, $y = -2a$, $y = a$.

(e) The lamina composed of all except the third-quadrant portion of the region inside the circle $x^2 + y^2 = a^2$.

(f) The lamina bounded by $y^2 = 2a(x + a)$ and $x = a$.

5. Products of inertia, as defined in Exercise 4, play an important role in the discussion of what happens to moments of inertia when the co-ordinate axes are rotated. Suppose the xy-system can be rotated into the $x'y'$-system by turning counterclockwise through an angle α. The equations relating the two systems are

$$x' = x \cos \alpha + y \sin \alpha,$$

$$y' = -x \sin \alpha + y \cos \alpha,$$

and a similar pair of equations giving x, y in terms of x', y'. For a given lamina, let $A = I_x$, $B = U_{xy}$, $C = I_y$ (using the notation defined in Exercises 2 and 4), and let A', B', C' denote the corresponding moments and products of inertia relative to the axes of the $x'y'$-system.

Show that

$$A' = A \cos^2 \alpha - 2B \sin \alpha \cos \alpha + C \sin^2 \alpha,$$

$$A = A' \cos^2 \alpha + 2B' \sin \alpha \cos \alpha + C' \sin^2 \alpha,$$

$$B' = B(\cos^2 \alpha - \sin^2 \alpha) + (A - C) \sin \alpha \cos \alpha,$$

and find three more formulas of this type. Hence prove that the curve

$$Ax^2 - 2Bxy + Cy^2 = 1$$

has the equation

$$A'x'^2 - 2B'x'y' + C'y'^2 = 1$$

when referred to the $x'y'$-axes.

This curve is called the *ellipse of inertia* for the given lamina, relative to the origin O. If the xy-axes coincide with the axes of symmetry of the ellipse, the xy-term in the equation must disappear, so that $B = 0$. In this position, the co-ordinate axes are called *principal axes of inertia* for the lamina relative to O. If $B \neq 0$, the position of the principal axes may be found by the method of choosing α so that $B' = 0$, i.e.,

$$B \cos 2\alpha + \tfrac{1}{2}(A - C) \sin 2\alpha = 0.$$

6. Show that, if one regards A' as a function of α, $\dfrac{dA'}{d\alpha} = -2B'$; hence show that A' is either a maximum of a minimum when the $x'y'$-axes are principal axes of inertia.

7. Use the formula for A' in Exercise 5 to compute the following moments of inertia:

(a) The lamina of Exercise 4(a), about its diagonal.

(b) The lamina of Exercise 4(c), about the line $2ay = bx$.

(c) The lamina of Exercise 4(f), about the line $y = 2x$.

8. Find the principal axes of inertia for the following laminas:

(a) The lamina of Exercise 4(a), if $a = 2$, $b = 1$.

(b) The lamina of Exercise 4(b).

(c) The lamina of Exercise 4(d).

(d) The lamina of Exercise 4(e).

(e) The lamina of Exercise 4(f).

9. A lamina in the shape of the circle $x^2 + y^2 \leq a^2$ has density $\sigma = (x + y)^2$. Find its principal axes of inertia relative to its center, and the moments of inertia about these axes.

13.51 / POTENTIALS AND FORCE FIELDS

In the theory of electrostatics, the concepts of charge and charge density are entirely analogous to the concepts of mass and mass density, with this exception: Charges may be either positive or negative, while we habitually think of masses as positive. A particle of electric charge e exerts an electrostatic force on another particle of charge e' according to the inverse-square law of Coulomb: The magnitude of the force is inversely proportional to the square of the distance between the charges, and directly proportional to the product of the charges. The force is directed along the line joining the charges, and like charges repel each other, while unlike charges attract. With proper choice of units (electrostatic units) the constant of proportionality may be taken as unity.

The vector form of Coulomb's law is as follows: Let e be at P, e' at P', and r be the distance PP'. Then the force exerted by e on e' is

$$\mathbf{F} = \frac{ee'}{r^3}\overrightarrow{PP'}. \tag{13.51–1}$$

This should be compared with the analogous formula for gravitational attraction between two particles (see (10.51–4)).

Next we consider how to deal with the notion of electrostatic force produced by a continuous distribution of charge on a plane lamina. Consider a particle of unit positive charge at a fixed point Q, anywhere in space, but not on the lamina. Let σ be the charge density on the lamina, which we assume occupies a region R in the xy-plane. In the usual manner, we subdivide R and consider the force exerted on Q by the system of point charges which is obtained when we concentrate the charge Δe of each part ΔR of the lamina at a point P within the part. The contribution of this part to the total is a force

$$\Delta\mathbf{F} = \frac{\Delta e}{r^3}\overrightarrow{PQ},$$

where r is the distance PQ (see Fig. 111). All such vectors must be added, and then we must carry out the limiting process. Since Δe is approximately $\sigma\,\Delta A$, the total force exerted by the lamina is

$$\mathbf{F} = \iint_R \frac{\sigma}{r^3}\overrightarrow{PQ}\,dA. \tag{13.51–2}$$

This is a *vector* double integral; i.e., the integrand is a vector function. We have not formally defined such integrals, but the work is entirely like that for scalar double integrals.

When it comes to actual computation of **F**, we work with components of the vectors. It frequently occurs that because of symmetry we know in advance the direction of **F**, and hence need only to deal with components of **F** in that one direction. If L is a directed line and if ψ is the angle which $\overrightarrow{PQ}$ makes with L, the component of **F** in the direction of L is

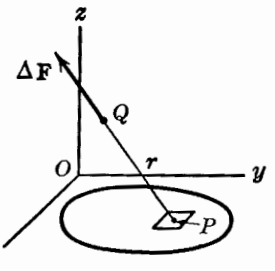


Fig. 111.

$$F_L = \iint_R \frac{\sigma}{r^2} \cos\psi \, dA. \tag{13.51–3}$$

To evaluate, we must express the integrand in suitable co-ordinates and then pass to an iterated integral. *It must be kept in mind that the r in* (13.51–3) *is not necessarily the r of polar co-ordinates.*

The force exerted by a system of charges on a unit positive charge at Q is called the *field* at Q due to the system.

Example 1. Find the field due to a uniformly charged circular lamina of radius b, at a point Q a distance c from the center of the lamina along the perpendicular to it (Fig. 112).

The phrase "uniformly charged" means that the density is constant. We take the origin and the z-axis as shown in Fig. 112. Using polar co-ordinates in the xy-plane, we have the distance PQ given by

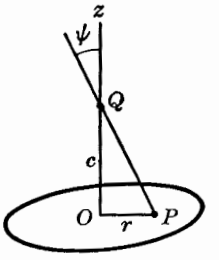


Fig. 112.

$$(PQ)^2 = r^2 + c^2.$$

By symmetry the force is evidently in the direction OQ, so that

$$\cos\psi = \frac{c}{\sqrt{r^2 + c^2}}.$$

Hence, denoting the x, y, and z components of **F** by F_1, F_2, F_3, we have $F_1 = F_2 = 0$, and

$$F_3 = \iint_R \frac{\sigma c}{(r^2 + c^2)^{3/2}} \, dA = \sigma c \int_0^{2\pi} d\theta \int_0^b \frac{r \, dr}{(r^2 + c^2)^{3/2}},$$

$$F_3 = 2\pi\sigma c \left(\frac{1}{c} - \frac{1}{\sqrt{b^2 + c^2}}\right) = 2\pi\sigma \left(1 - \frac{c}{\sqrt{b^2 + c^2}}\right).$$

Observe that the force very near O is almost of amount $2\pi\sigma$.

A systematic study of the theory of electrostatic fields is greatly simplified by introducing the concept of the *potential* of the field. The potential at a point Q, produced by a charge e at the point P, is defined to be

$$\frac{e}{r}, \qquad \text{where } r = PQ.$$

For the potential of several particles, the principle of superposition is used, and for continuous distributions of charge, the standard integral calculus procedure is employed. For a lamina on the region R, with charge density σ at P, the potential at Q is defined to be

$$u(Q) = \iint_R \frac{\sigma}{PQ}\, dA. \tag{13.51–4}$$

The potential is a *scalar* point function. The electrostatic field is a *vector* point function. The relation between the two functions is shown in the fact that the gradient of the potential gives the negative of the field vector:

$$\nabla_Q u(Q) = -\mathbf{F}. \tag{13.51–5}$$

The Q on the gradient symbol is to remind one that we must differentiate with respect to the co-ordinates of Q. If P is (x, y, z), and Q is (ξ, η, ζ), we have

$$r^2 = (PQ)^2 = (\xi - x)^2 + (\eta - y)^2 + (\zeta - z)^2, \tag{13.51–6}$$

and

$$u(Q) = \iint_R \frac{\sigma(x, y)}{r}\, dx\, dy, \tag{13.51–7}$$

$$\mathbf{F} = \iint_R \frac{\sigma(x, y)}{r^3} [(\xi - x)\mathbf{i} + (\eta - y)\mathbf{j} + (\zeta - z)\mathbf{k}]\, dx\, dy. \tag{13.51–8}$$

Formula (13.51–5) is then equivalent to

$$\mathbf{F} \cdot \mathbf{i} = -\frac{\partial u}{\partial \xi} = \iint_R \frac{\sigma(x, y)}{r^3} (\xi - x)\, dx\, dy, \tag{13.51–9}$$

and two similar formulas for the other components of $\mathbf{F}$.

It is only in very special instances that the potential can be computed in elementary form by integration. Usually the work leads to elliptic or other nonelementary integrals. Nevertheless, the study of the potential is very fruitful. Extensive consideration of the theory of potential functions is outside the scope of the present book.

Example 2. The lamina bounded by the lines $x = 0$, $x = a$, $y = 0$, $y = b$ in the xy-plane carries a charge of density $\sigma = xy$. Find the potential at the point $Q(0, 0, \zeta)$ on the z-axis.

The potential is

$$u = \iint_R \frac{xy}{(x^2+y^2+\zeta^2)^{1/2}}\,dx\,dy = \int_0^b y\,dy \int_0^a \frac{x\,dx}{(x^2+y^2+\zeta^2)^{1/2}}.$$

The first integration gives

$$(a^2+y^2+\zeta^2)^{1/2} - (y^2+\zeta^2)^{1/2},$$

so

$$u = \int_0^b [y(a^2+y^2+\zeta^2)^{1/2} - y(y^2+\zeta^2)^{1/2}]\,dy,$$

$$u = \tfrac{1}{3}(a^2+b^2+\zeta^2)^{3/2} - \tfrac{1}{3}(a^2+\zeta^2)^{3/2} - \tfrac{1}{3}(b^2+\zeta^2)^{3/2} + \tfrac{1}{3}\zeta^2.$$

EXERCISES

1. Find the potential at Q in Example 1, and verify that $F_3 = -\dfrac{\partial u}{\partial c}$ (assuming $c > 0$).

2. Find $F_3 = \mathbf{F}\cdot\mathbf{k}$ directly in Example 2, and then verify that $F_3 = -\dfrac{\partial u}{\partial \zeta}$ from the answer found in Example 2.

3. Find u and F_3 at the point Q in Example 1 if, instead of constant density, we have $\sigma = r$.

4. Find the potential at a corner of a uniformly charged square lamina of side b.

5. Find the potential at a point on the edge of a uniformly charged circular lamina of radius b. It is most convenient to take the point in question at the origin.

6. Find the potential at $Q(0, 0, b)$, where $b > 0$, due to a uniformly charged square lamina with corners at $(0, 0, 0)$, $(a, 0, 0)$, $(a, a, 0)$, $(0, a, 0)$. Set up the integral in polar co-ordinates, using the fact that the square can be divided by a diagonal so that each half contributes the same amount to the potential. The integral formula

$$\int \frac{\sqrt{a^2+b^2\cos^2\theta}}{\cos\theta}\,d\theta = b\tan^{-1}\left(\frac{b\sin\theta}{\sqrt{a^2+b^2\cos^2\theta}}\right) + \frac{a}{2}\log\frac{\sqrt{a^2+b^2\cos^2\theta}+a\sin\theta}{\sqrt{a^2+b^2\cos^2\theta}-a\sin\theta} + C$$

will be useful. The z-component of the field at Q may be computed from $F_3 = -\partial u/\partial b$ but it is perhaps easier to compute F_3 directly by integration.

7. It can be shown that $\partial u/\partial\xi$ can be computed from (13.51–7) by doing the differentiation under the integral sign, provided Q is a point not in the region R or on its boundary. Proceed from this to verify (13.51–9), using (13.51–6). Thus (13.51–5) is proved.

13.6 / TRIPLE INTEGRALS

We shall deal with the definition of a triple integral somewhat more briefly than we did with the definition of a double integral. We begin with a closed bounded region R in three dimensions, and let $f(x, y, z)$ be a function defined and continuous in R. As in §13.2 we must make some assumptions about the

character of the boundary of R. The precise nature of these assumptions need not be made explicit as long as we do not go carefully into questions of integrability. We shall for simplicity think of the boundary R as consisting of a finite number of surfaces, each of which is smooth except possibly at certain isolated points (e.g., the vertex of a cone) or along certain curves (e.g., the edges of a cube or the rims of a solid right circular cylinder).

We take three sets of planes, parallel respectively to the x-, y-, and z-axes. The mesh of rectangular blocks which these planes form in space is called a rectangular partition. Those blocks, or cells, which belong entirely to R are numbered consecutively in any order. Let ΔV_k be the volume of the kth cell, and let its x-, y-, and z-dimensions be Δx_k, Δy_k, Δz_k respectively, so that $\Delta V_k = \Delta x_k \, \Delta y_k \, \Delta z_k$. Finally, let (x_k, y_k, z_k) be an arbitrarily selected point in the kth cell. Then we define the triple integral of the function f over R by following limit, as the maximum dimensions of all the cells approach zero:

$$\iiint_R f(x, y, z)\, dV = \lim \sum_k f(x_k, y_k, z_k)\, \Delta V_k, \tag{13.6–1}$$

or, in another notation,

$$\iiint_R f(x, y, z)\, dx\, dy\, dz = \lim \sum_k f(x_k, y_k, z_k)\, \Delta x_k \, \Delta y_k \, \Delta z_k. \tag{13.6–2}$$

We take for granted that this limit exists and is independent of the particular method of forming the partitions and choosing the points (x_k, y_k, z_k).

The analogue of Theorem II, §13.23, is true for triple integrals; that is, the integral is given by (13.6–1) when the subregions, instead of being rectangular blocks, are formed in any manner (as long as they are sufficiently regular in shape). They need not completely fill out the region R, provided that the amount of volume omitted approaches zero in the limit. These remarks are of importance for the understanding of what happens when we use cylindrical or spherical co-ordinates.

The properties of double integrals explained in §13.21 extend at once to triple integrals. The same is true of the inequalities of §13.22, and the mean-value theorem.

When it comes to devising an explanation of the evaluation of triple integrals by iterated integrals, we must proceed differently than in the case of double integrals, for no intuitive geometric procedure analogous to that of §13.3 is available to us (a four-dimensional space would be required). There is a direct analytical method, however. This method could have been used for double integrals as well. We shall give a heuristic account of the method, thus making its plausibility clear. A fully rigorous account is rather long, and it seems advisable to leave the details for later study.

Let us first state the result. The letters x, y, z can be written in six possible orders. Corresponding to each such order there is an iterated integral evaluation

of the triple integral, calling for three successive single integrations. The main problem of technique is that of learning how to write the limits of integration for the iterated integrals. The notation for an iterated integral is illustrated by

$$\int_0^1 dy \int_0^{1-y} dx \int_{x+y}^1 (x^2 + y^2)\, dz. \tag{13.6–3}$$

The integrations in (13.6–3) are to be performed in the order z, x, y.

It will be enough to explain the transition from the triple integral to an iterated integral for one particular order of integration. Suppose this order is first with respect to z, then with respect to x, and finally with respect to y. Choosing a typical value of y, consider the cross section of R by a plane $y =$ constant, parallel to the xz-plane. We assume that R is of such a shape that all the foregoing cross sections are plane regions of the type dealt with in our discussion of iterated integrals in two dimensions. As shown in Fig. 113, let the

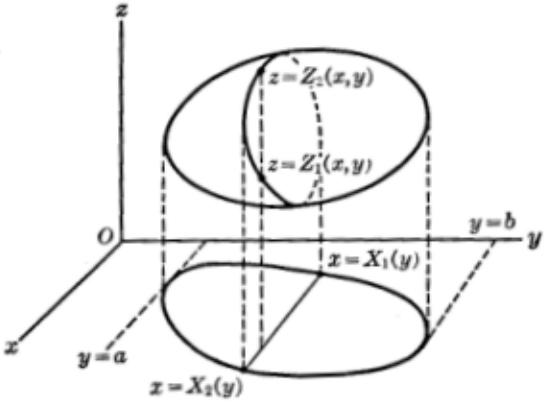


Fig. 113.

largest and smallest values of x in the cross section be respectively $X_1(y)$ and $X_2(y)$, and let $Z_1(x, y)$, $Z_2(x, y)$ be the values of z for which a typical line parallel to the z-axis in the cross section cuts the boundary of R. Finally, let $y = a$ and $y = b$ be the extreme values of y in the region R. Then

$$\iiint_R f(x, y, z)\, dV = \int_a^b dy \int_{X_1}^{X_2} dx \int_{Z_1}^{Z_2} f(x, y, z)\, dz. \tag{13.6–4}$$

The formula (13.6–4) is the fundamental theorem about evaluating triple integrals by iterated integrals in rectangular co-ordinates. Before giving a heuristic justification of the formula we give an illustrative example.

Example. Find the centroid of an octant of a solid sphere.

Let $x^2 + y^2 + z^2 = a^2$ be the equation of the surface of the sphere. We consider

the first octant. Evidently $\bar{x} = \bar{y} = \bar{z}$, so we find $\bar{x}$ only. Analogous to (13.3–8) we have

$$V\bar{x} = \iiint_R x\, dV,$$

where V is the volume of R. In the present case $V = (\pi/6)a^3$. In the notation of (13.6–4) we see from Fig. 114 that

$$Z_1 = 0,\ Z_2 = \sqrt{a^2 - x^2 - y^2},\ X_1 = 0,\ X_2 = \sqrt{a^2 - y^2}.$$

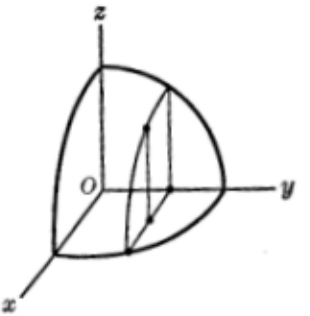


Fig. 114.

Hence

$$\frac{\pi}{6} a^3 \bar{x} = \int_0^a dy \int_0^{\sqrt{a^2-y^2}} dx \int_0^{\sqrt{a^2-x^2-y^2}} x\, dz$$
$$= \int_0^a dy \int_0^{\sqrt{a^2-y^2}} x\sqrt{a^2 - x^2 - y^2}\, dx.$$

The x-integration yields

$$-\tfrac{1}{3}[a^2 - y^2 - x^2]^{3/2}\Big|_0^{\sqrt{a^2-y^2}} = \tfrac{1}{3}(a^2 - y^2)^{3/2}.$$

Hence

$$\frac{\pi}{6} a^3 \bar{x} = \tfrac{1}{3}\int_0^a (a^2 - y^2)^{3/2}\, dy = \frac{a^4}{8}\frac{\pi}{2};$$

we omit the details of the last integration. Finally, then $\bar{x} = \frac{3}{8}a$.

Now to explain (13.6–4). We go back to the definition (13.6–2). Let us single out all the cells of the partition which belong to R and lie in a particular column parallel to the z-axis (see Fig. 115). We may choose the points (x_k, y_k, z_k) so that the co-ordinates x_k, y_k are the same for all the points belonging to cells in the

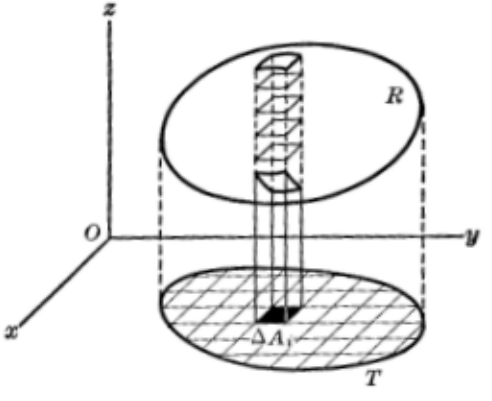


Fig. 115.

same vertical column. The values Δx_k and Δy_k will also be the same for all the cells in one column, and the area of the base of the column will be $\Delta x_k \, \Delta y_k$. Let us number the columns, say from 1 to N. Suppose ΔA_i is the area of the base of the ith column, and suppose the number of cells in the ith column is m_i. Let the points associated with these cells be (x_i', y_i', z_{ij}), $j = 1, \ldots, m_i$, and let their z-dimensions be Δz_{ij}. Then the sum in (13.6–2) can be written in the form

$$\sum_{i=1}^{N} \left(\sum_{j=1}^{m_i} f(x_i', y_i', z_{ij}) \, \Delta z_{ij} \right) \Delta A_i. \tag{13.6–5}$$

The inside sum here is of the type occurring in the definition of a definite integral with respect to z. The interval of z-values that is being subdivided is approximately from the lower to the upper bounding surface of R, that is, from $Z_1(x_i', y_i')$ to $Z_2(x_i', y_i')$. Hence the inner sum is an approximation to

$$\int_{Z_1(x_i', y_i')}^{Z_2(x_i', y_i')} f(x_i, y_i, z) \, dz. \tag{13.6–6}$$

For convenience let us write

$$g(x, y) = \int_{Z_1(x, y)}^{Z_2(x, y)} f(x, y, z) \, dz. \tag{13.6–7}$$

Then the expression (13.6–6) is $g(x_i', y_i')$, and (13.6–5) is seen to be approximately equal to

$$\sum_{i=1}^{N} g(x_i', y_i') \, \Delta A_i \tag{13.6–8}$$

if the cell dimensions in the z-direction are all sufficiently small. This sum, in turn, is of the type occurring in the definition of a double integral. If T is the plane region obtained by projecting the points of R perpendicularly on the xy-plane, the bases of the columns form a rectangular partition of T. When the dimensions of the cells of this partition are small enough, the sum (13.6–8) is very nearly equal to the double integral

$$\iint_T g(x, y) \, dA,$$

which in turn is equal to the iterated integral

$$\int_a^b dy \int_{X_1}^{X_2} g(x, y) \, dx, \tag{13.6–9}$$

as we see from Fig. 113. We see, therefore, on combining (13.6–7) and (13.6–9), that the sum (13.6–5) is an approximation to the iterated integral

$$\int_a^b dy \int_{X_1}^{X_2} dx \int_{Z_1}^{Z_2} f(x, y, z) \, dz.$$

It may be shown in more detail that the approximation becomes better and better as we take the limit defining the triple integral, so that (13.6–4) is exactly true.

13.7 / APPLICATIONS OF TRIPLE INTEGRALS

Triple integrals may be used to calculate the locations of centers of gravity, the masses of solids of variable density, moments of inertia, and other quantities of physical or geometrical significance. The fundamental principles of such applications are the same as those set forth in connection with double integrals (§13.5).

We shall use the Greek letter μ for volume density. The mass of a solid of variable density $\mu(x, y, z)$ occupying a region R is then

$$M = \iiint_R \mu \, dV.$$

The center of gravity $(\bar{x}, \bar{y}, \bar{z})$ is found from the formula

$$M\bar{x} = \iiint_R x\mu \, dV$$

and two other similar formulas. The moment of inertia about the z-axis is

$$I_z = \iiint_R (x^2 + y^2)\mu \, dV.$$

The product of inertia relative to the planes $x = 0$ and $y = 0$ is

$$U_{xy} = \iiint_R xy\mu \, dV.$$

Other moments of inertia I_x, I_y, and other products of inertia U_{yz}, U_{zx} are defined by analogous formulas.

Problems in gravitational attraction are mathematically almost identical with problems of electrostatic forces, since Newton's law and Coulomb's law are both inverse-square laws. There is a difference in sign, since two masses attract each other, whereas two positive charges repel. Newton's law for mass particles m and m' at P and P', a distance r apart, states that m exerts on m' a force

$$\mathbf{F} = k \frac{mm'}{r^3} \overrightarrow{P'P},$$

where k is a universal constant depending only on the units of mass, distance, and force. In theoretical work it is customary to choose units such that $k = 1$. We shall do this. The force of attraction on a unit mass at Q, produced by a solid of density μ occupying a region R, is

$$\mathbf{F} = \iiint_R \frac{\mu}{r^3} \overrightarrow{QP} \, dV,$$

where $r = QP$, μ is evaluated at P, and integration is carried out with respect to the co-ordinates of P.

The concept of potential is useful in the theory of gravitational attraction. The potential at Q is defined to be

$$u(Q) = \iiint_R \frac{\mu}{r}\, dV.$$

The relation between the potential u and the gravitational field force $\mathbf{F}$ is expressed by the equation

$$\mathbf{F} = \nabla_Q u;$$

i.e., the field is the gradient of the potential. The situation is comparable to that in electrostatics (see §13.51); there, however, the field is the *negative* of the gradient of the potential. The difference in sign arises from the difference in sign between Newton's and Coulomb's laws.

Example 1. The first octant portion of the solid inside the cylinder $x^2 + y^2 = a^2$ and between the planes $z = 0$, $z = h$ has density $\sigma = x$. Find its mass. We have

$$M = \iiint_R x\, dV = \int_0^h dz \int_0^a dy \int_0^{\sqrt{a^2-y^2}} x\, dx;$$

$$M = \int_0^h dz \int_0^a \tfrac{1}{2}(a^2 - y^2)\, dy = \int_0^h \frac{a^3}{3}\, dz = \tfrac{1}{3}a^3 h.$$

The finding of the limits of integration is illustrated in Fig. 116.

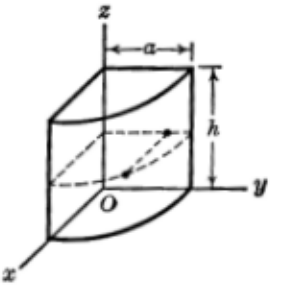


Fig. 116.

Example 2. Find the moment of inertia about the z-axis of the homogeneous tetrahedron bounded by the planes $z = x + y$, $x = 0$, $y = 0$, $z = 1$. The integral in this case is

$$I_z = \iiint_R \mu(x^2 + y^2)\, dV$$

$$= \mu \int_0^1 dy \int_0^{1-y} dx \int_{x+y}^1 (x^2 + y^2)\, dz.$$

The limits of integration are found by an examination of Fig. 117. Completion of the integration is left as an exercise for the student. The result is

$$I_z = \frac{\mu}{30}.$$

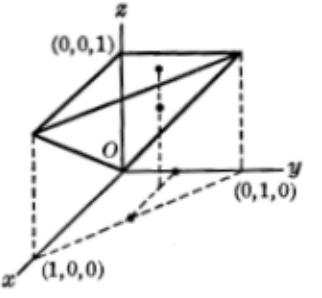


Fig. 117.

Since the volume of the tetrahedron is $\frac{1}{6}$, the mass is $M = \mu/6$, whence $\mu = 6M$ and

$I_z = M/5$. This means that the radius of gyration of the tetrahedron about the z-axis is $k_z = 1/\sqrt{5}$.

EXERCISES

1. A homogeneous solid block is bounded by the planes $x = 0$, $x = a$, $y = 0$, $y = b$, $z = 0$, $z = c$. Show that $I_x = (M/3)(b^2 + c^2)$.

2. Locate the center of gravity of the tetrahedron in Example 2. Consider some of the other orders of integration, and select a convenient order for each part of the work.

3. **(a)** Locate the center of gravity of the solid tetrahedron cut from the first octant by the plane $(x/a) + (y/b) + (z/c) = 1$. **(b)** Find the moment of inertia of this tetrahedron about the y-axis.

4. Find $\bar{x}$ and $\bar{y}$ for the solid of Example 1.

5. The unit cube bounded by the planes $x = 0$, $x = 1$, $y = 0$, $y = 1$, $z = 0$, $z = 1$ has density $\mu = xz$. **(a)** Find the mass and locate the center of gravity of the cube. **(b)** Find the y-component of the attraction which the cube exerts on a unit mass at the origin.

6. A homogeneous solid is bounded by the plane $z = 0$ and the paraboloid $(x^2/a^2) + (y^2/b^2) + (z/c) = 1$. **(a)** Locate its center of gravity. **(b)** Find I_x, I_y, and I_z for the solid.

7. Consider the homogeneous solid bounded by the ellipsoid

$$(x^2/a^2) + (y^2/b^2) + (z^2/c^2) = 1.$$

(a) Locate the center of gravity of the first octant portion of this solid.

(b) Calculate the moment of inertia of the entire solid about the x-axis.

8. The density of a cube of edge $2a$ is proportional to the square of the distance from its center, the density being unity at the center of each face. **(a)** Find the total mass and the average density. **(b)** Find I_z for this cube. **(c)** Find the ratio of the I_z in **(b)** to the value which I_z would have if the cube had the same total mass distributed uniformly throughout the volume.

9. If a solid has density at (x, y, z) equal to the distance from (x, y, z) to the origin, show that the potential at the origin is equal to the volume of the solid.

10. Suppose a solid has density at (x, y, z) equal to the cube of the distance from (x, y, z) to the origin. Show that the attraction of the solid on a unit mass at the origin is a force directed toward the centroid of the volume occupied by the solid, and equal in magnitude to the product of the volume and the distance from the origin to the centroid.

11. Let I denote the moment of inertia of a solid about a certain axis, and let I_0 be the moment of inertia about a parallel axis through the center of gravity. If h is the distance between the two axes, prove that $I = I_0 + Mh^2$, where M is the total mass of the body.

12. Consider two rectangular co-ordinate systems with the same origin, the relations between the xyz-system and the $x'y'z'$-system being expressed in the notation of §10.3. For a given body, show that

$$I_{x'} = I_x l_1^2 + I_y m_1^2 + I_z n_1^2 - 2U_{yz} m_1 n_1 - 2U_{zx} n_1 l_1 - 2U_{xy} l_1 m_1.$$

This shows how products of inertia enter into consideration of the effect of rotation of axes upon moments of inertia.

The equation

$$I_x x^2 + I_y y^2 + I_z z^2 - 2U_{yz} yz - 2U_{zx} zx - 2U_{xy} xy = 1$$

defines what is called the *ellipsoid of inertia* for the body relative to the origin O. A set of axes such that the products of inertia all vanish is called a set of *principal axes of inertia* for the body.

13.8 / CYLINDRICAL CO-ORDINATES

If we use polar co-ordinates in a plane, and a rectangular co-ordinate along an axis perpendicular to the plane at the origin of the polar system, the combination is called a cylindrical co-ordinate system. Most commonly the polar co-ordinates are taken in the xy-plane (see Fig. 118), but there is no logical necessity for this choice. It is often convenient to evaluate a triple integral by an iterated integral in cylindrical co-ordinates. As we saw in §13.6,

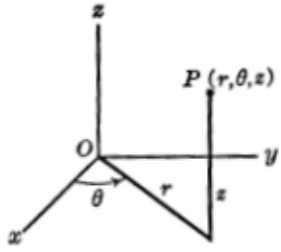


Fig. 118.

$$\iiint_R f(x, y, z)\, dV = \iint_T dA \int_{Z_1}^{Z_2} f(x, y, z)\, dz. \tag{13.8–1}$$

If we express the integrand in cylindrical co-ordinates, say $f(x, y, z) = F(r, \theta, z)$, the double integral in (13.8–1) may be evaluated as an iterated integral in polar co-ordinates. This leads to the result

$$\iiint_R F(r, \theta, z)\, dV = \int_\alpha^\beta d\theta \int_{R_1}^{R_2} r\, dr \int_{Z_1}^{Z_2} F(r, \theta, z)\, dz. \tag{13.8–2}$$

Do not fail to observe the factor r which is introduced into the integrand of the iterated integral. The limits Z_1, Z_2 must be expressed in terms of r and θ; the r and θ limits are found by inspection of the plane region T, the "shadow" of R on the xy-plane (see Fig. 115). The result (13.8–2) and others like it may also be obtained by an argument similar to that beginning after the Example in §13.6. There are five other possible orders of integration. A systematic method for determining the limits of integration for any given order is illustrated in the following example:

Example. Find the moment of inertia of a homogeneous right circular cone about its axis.

Let the radius of the base be b, the altitude be h. The density μ is constant, so $M = \mu(\pi/3)b^2h$. We place the cone as shown in Fig. 119 (we draw only one-fourth the cone). Let the integration order be r, z, θ. We must first set up the triple integral:

$$I = \iiint_R \mu(x^2 + y^2)\, dV = \mu \iiint_R r^2\, dV.$$

Now picture a section of the region R made by holding the last integration variable (here θ) constant. In the present case this is the triangle OAB. Next assign the second integration variable z a typical value, and determine the range of freedom left to the first integration variable r. This process is indicated in Fig. 119 by the line CD. Since $OC = z$, the value of r at D is given by

$$\frac{r}{z} = \frac{b}{h}.$$

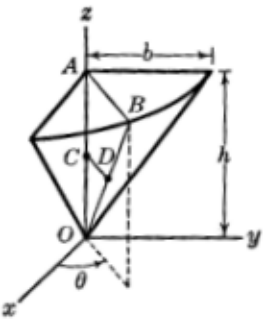


Fig. 119.

The r limits of integration are therefore 0 and zb/h. Now let the line CD range in the z-direction as much as it may (from 0 to h); these are the z-limits of integration. Finally, let θ vary through all values necessary to have the θ-sections sweep out the entire region R. We see that the θ-limits of integration are 0 and 2π. Therefore (remembering the additional factor r),

$$I = \mu \int_0^{2\pi} d\theta \int_0^h dz \int_0^{bz/h} r^3\, dr = \frac{\mu\pi}{10} b^4 h.$$

This may be written $I = \frac{3}{10}Mb^2$.

EXERCISES

1. For the solid cone of the illustrative example find **(a)** I_y; **(b)** the location of the center of gravity of the first octant portion; **(c)** the attraction exerted on a unit mass at the origin; **(d)** the potential at a point $(0, 0, \zeta)$, where $\zeta < 0$.

2. Find the moment of inertia of a homogeneous solid sphere of radius a, about a diameter.

3. For a homogeneous solid right circular cylinder of height h and radius of base a, find the moments of inertia **(a)** about the axis of the cylinder; **(b)** about a line through the center of gravity of the cylinder, perpendicular to the axis of the cylinder; **(c)** the attraction exerted by the cylinder on a unit mass at the center of one end; **(d)** the potential at a point $(0, 0, \zeta)$, assuming the cylinder defined by $x^2 + y^2 \leqq a^2$, $0 \leqq z \leqq h$, and assuming $\zeta \geqq h$.

13.9 / SPHERICAL CO-ORDINATES

To form a spherical co-ordinate system we start from an origin O and a fixed ray issuing from O. We shall take the ray as the positive z-axis; there is, however, no necessity for any one special relation between spherical and rectangular co-ordinates. The spherical co-ordinates are the distance $\rho = OP$ and the two angles θ, ϕ (see Fig. 120). The angle θ, sometimes called the *azimuth* of P, is the same as that used in plane polar co-ordinates. The angle ϕ is the *colatitude* of P. We always choose ϕ in the range $0 \leqq \phi \leqq \pi$. For most work ρ is taken nonnegative.

The student should be aware that in some books the roles of θ and ϕ are

reversed, so that θ denotes the colatitude. This is particularly true of European texts, and books on mathematical physics. With due caution the student should have no trouble accommodating himself to such differences in notation.

The standard formulas relating spherical to rectangular co-ordinates are

$$z = \rho \sin \phi \cos \theta,\ y = \rho \sin \phi \sin \theta,\ z = \rho \cos \phi. \tag{13.9–1}$$

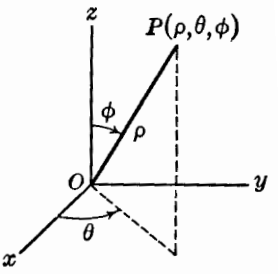


Fig. 120.

To explain the use of spherical co-ordinates in evaluating triple integrals, we employ a method like that of §13.4. The curvilinear co-ordinates ρ, θ, ϕ give rise to a partition of R into cells which resemble cubes. A typical such cell is bounded by three pairs of surfaces, two each of the types ρ = constant (spheres), θ = constant (half-planes through the z-axis), ϕ = constant (cones). Such a cell is generated by taking a shaded area as shown in Fig. 121 and turning it about the z-axis through an angle $\Delta\theta$. By formulas of elementary solid geometry the volume of this cell is found to be

$$\Delta V = \tfrac{1}{3}[(\rho + \Delta\rho)^3 - \rho^3][\cos \phi - \cos(\phi + \Delta\phi)]\,\Delta\theta. \tag{13.9–2}$$

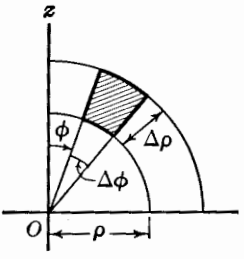


Fig. 121.

Now, by the law of the mean, we have

$$(\rho + \Delta\rho)^3 - \rho^3 = 3\rho'^2\,\Delta\rho,$$
$$\cos(\phi + \Delta\phi) - \cos\phi = -\sin\phi'\,\Delta\phi,$$

where ρ' is between ρ and $\rho + \Delta\rho$ and ϕ' is between ϕ and $\phi + \Delta\phi$.

Thus (13.9–2) becomes

$$\Delta V = \rho'^2 \sin\phi'\,\Delta\rho\,\Delta\theta\,\Delta\phi.$$

This may be expressed by saying that there is in each cell, say the kth one, a point $(\rho_k, \theta_k, \phi_k)$ such that the volume of the cell is

$$\Delta V_k = \rho_k^2 \sin\phi_k\,\Delta\rho_k\,\Delta\theta_k\,\Delta\phi_k. \tag{13.9–3}$$

It is worth while observing the expression $\rho^2 \sin\phi\,\Delta\rho\,\Delta\theta\,\Delta\phi$ is evidently a good approximation to ΔV in (13.9–2) when the dimensions of the cell are small, for three concurrent edges of the cell have the lengths

$$\Delta\rho,\ \rho\,\Delta\phi,\ \rho \sin\phi\,\Delta\theta.$$

Now consider a triple integral

$$\iiint_R F(\rho, \theta, \phi)\,dV,$$

the integrand being expressed in terms of spherical co-ordinates. We express this integral as the limit of a sum

$$\sum_k F(\rho_k, \theta_k, \phi_k)\,\Delta V_k,$$

using the formula (13.9–3) for ΔV_k, so that

$$\iiint_R F(\rho, \theta, \phi)\, dV = \lim \sum_k F(\rho_k, \theta_k, \phi_k)\rho_k^2 \sin \phi_k\, \Delta\rho_k\, \Delta\theta_k\, \Delta\phi_k. \quad (13.9\text{–}4)$$

By looking upon the equations (13.9–1) as defining a mapping between an xyz-space and a $\rho\theta\phi$-space, we interpret the limit on the right in (13.9–4) as defining a triple integral over the image of R in the $\rho\theta\phi$-space (compare with (13.4–4), and so arrive at the iterated integral

$$\iiint_R F(\rho, \theta, \phi)\, dV = \int_\alpha^\beta d\theta \int_{\Phi_1}^{\Phi_2} d\phi \int_{R_1}^{R_2} F(\rho, \theta, \phi)\rho^2 \sin \phi\, d\rho, \quad (13.9\text{–}5)$$

or any one of five other iterated integrals. The essential thing to observe is the factor $\rho^2 \sin \phi$ on the right. The finding of the limits of integration follows the same principles illustrated in §13.8 and earlier work.

Spherical co-ordinates seem to be particularly convenient for many gravitational-attraction problems.

Example. The smaller volume bounded by the sphere $x^2 + y^2 + z^2 = 4a^2$ and the plane $z = a$ is filled with a homogeneous solid. Find the gravitational field which this solid produces at the origin.

The density μ is constant; the force is in the z-direction, so $F_1 = F_2 = 0$. In the integral for F_3 we have $\psi = \phi$; hence

$$F_3 = \mu \iiint_R \frac{\cos \phi}{\rho^2}\, dV.$$

We integrate in this order: first ρ, then θ, and finally ϕ. Finding the limits is illustrated in Fig. 122 and Fig. 123. The equation of the plane $z = a$ is written as $\rho = a \sec \phi$; that of the sphere is $\rho = 2a$. Putting in the factor $\rho^2 \sin \phi$, we have

$$F_3 = \mu \int_0^{\pi/3} d\phi \int_0^{2\pi} d\theta \int_{a \sec \phi}^{2a} \sin \phi \cos \phi\, d\rho.$$

$$F_3 = 2\pi\mu \int_0^{\pi/3} (2a \sin \phi \cos \phi - a \sin \phi)\, d\phi$$

$$= 2\pi\mu a[\sin^2 \phi + \cos \phi]_0^{\pi/3} = \frac{\pi\mu a}{2}.$$

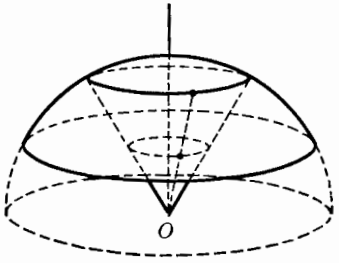


Fig. 122.

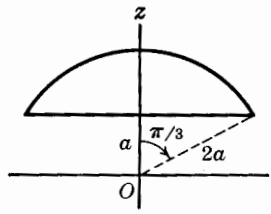


Fig. 123.

EXERCISES

1. Find the moment of inertia of a homogeneous spherical solid about a diameter, using spherical co-ordinates.

2. Locate the centroid of an octant of a solid sphere, using spherical co-ordinates.

3. Consider the homogeneous solid cone defined by $h^2(x^2+y^2) \leqq b^2z^2$, $0 \leqq z \leqq h$. Find **(a)** the attraction it exerts on a unit mass at the origin; **(b)** the moment of inertia about the z-axis; **(c)** the moment of inertia about the y-axis.

4. Consider the homogeneous solid sphere whose surface is defined by $\rho = 2a\cos\phi$. It is divided into two parts by the plane $z = a$. **(a)** Find the attraction which each part exerts on a unit mass at the origin. **(b)** What is the ratio of the magnitudes of these attractions? **(c)** What is the combined attraction of both parts?

5. Suppose the sphere $x^2+y^2+(z-h)^2 = a^2$, where $0 < a \leqq h$, is filled with matter of constant density μ. Show that its attraction on a unit mass particle at the origin is $\frac{4}{3}\pi\mu a^3/h^2$, and hence that the attraction is the same as though all the mass were concentrated at the center of the sphere. Begin by showing that a typical ray from O pierces the sphere at the two points given by $\rho = h\cos\phi \pm (h^2\cos^2\phi + a^2 - h^2)^{1/2}$, and that the range of ϕ values to be considered is $0 \leqq \phi \leqq \sin^{-1}(a/h)$.

6. In Exercise 5, suppose that $0 \leqq h < a$, so that the origin is inside the sphere. In this case show that the attraction is $\frac{4}{3}\pi\mu h$. This shows that, if the sphere is divided up into a smaller concentric sphere of radius h, and a shell between the spheres of radii a and h, the net force of attraction due to the matter in the shell is zero, the total force being just that produced by the matter in the sphere of radius h, on whose surface O lies.

7. Show that the value of the integral

$$\int_0^\pi \frac{(c - p\cos\phi)\sin\phi}{(c^2+\rho^2-2c\rho\cos\phi)^{3/2}}\,d\phi$$

is $2/c^2$ if $c > \rho \geqq 0$, and is equal to 0 if $\rho > c \geqq 0$. Also show that it is equal to $1/c^2$ if $\rho = c > 0$. Some of these results are useful in Exercises 8. If $c \neq \rho$ and $c\rho \neq 0$, the integral may be evaluated by making the substitution $t^2 = c^2+\rho^2-2c\rho\cos\phi$, $t\,dt = c\rho\sin\phi\,d\phi$.

8. Consider a solid sphere of radius a, center at the origin, not necessarily of constant density, but with the mass distributed symmetrically about the center, so that μ depends only on ρ: $\mu = \mu(\rho)$. Show that the mass of the sphere is

$$M = 4\pi\int_0^a \rho^2\mu(\rho)\,d\rho.$$

If a unit mass particle is h units from the center of the sphere, show that it is attracted with a force M/h^2 if $h \geqq a$, and with a force

$$\frac{4\pi}{h^2}\int_0^h \rho\mu(\rho)\,d\rho$$

if $0 < h < a$. If the density is constant, these results are the same as those obtained in a different way in Exercises 5, 6.

14 / CURVES AND SURFACES

14 / INTRODUCTION

Curves and surfaces are geometric entities with which the student is to some extent familiar. The simplest examples of these entities, such as the conic curves in the plane, and spheres, cylinders, cones, and other quadric surfaces in space, have been encountered repeatedly from analytic geometry through calculus. Geometrical interpretations of functions of one or two independent variables have led the student to think of curves and surfaces in quite general terms. In this chapter we propose to make a careful study of the means by which we render our intuitive notions about curves and surfaces amenable to precise mathematical treatment. This is done partly as an introduction to a branch of geometry—what is known as *differential geometry*—and partly as preparation for the following chapter on line and surface integrals.

A point to be emphasized is this: Our intuitive notions about curves and surfaces are all derived from relatively simple examples of these things. The general concepts of curves and surfaces are very inclusive, however, and in our studies we must remember that when we wish to *prove* something, we must appeal to the definitions and previously established theorems, not solely to our intuitions, which may present us with an oversimplified picture. Direct geometric visualization of the subjects of our discussion is, however, of great value, both for the suggestions we can derive and for the better understanding and retention of what we learn.

14.1 / REPRESENTATIONS OF CURVES

Intuitively we think of a curve as a one-dimensional configuration, like the path of a moving particle, or as something we might obtain by bending and twisting a straight line. We shall define a curve by saying that it is an ordered configuration of points (x, y, z) given by three continuous functions of a parameter:

$$x = f(t), y = g(t), z = h(t); \tag{14.1–1}$$

the range of the parameter is to be some interval (finite or infinite) of the real axis. We speak of (14.1–1) as a parametric representation of the curve. A curve may have more than one parametric representation. If we interpret t as time, (14.1–1) may be regarded as defining the path of a moving point. The point may pass through the same position in space several times; in this case the curve intersects itself. Evidently a curve in the above sense of the word is very general, and may not be very smooth. Imagine, for instance, the track of a tiny particle in Brownian movement over a long period of time.

To avoid some of the complexities which may occur in dealing with curves in general, we introduce some restrictions. By an *arc* we mean a curve which does not intersect itself, which has two distinct ends, and which is represented in the form (14.1–1) with a finite range of the parameter t, say $a \leqq t \leqq b$, where $a < b$. A semicircle is an arc, but an entire circle is not. If a curve is defined by (14.1–1) with a finite range $a \leqq t \leqq b$ of the parameter, and if the points corresponding to $t = a$ and $t = b$ are coincident, we say that the curve is closed (it has no free ends). Such a closed curve without self-intersections is called a *simple closed curve*, or a closed *Jordan* curve (after a French mathematician). The prototype of an arc is the unit interval $0 \leqq x \leqq 1$, while the prototype of a simple closed curve is the unit circle $x^2 + y^2 = 1$. Continuous deformation (bending, twisting, stretching, shrinking) of an arc leaves it still an arc, provided no points are brought together which were originally distinct. The like can be said about closed Jordan curves. The boundary of a square is a closed Jordan curve. A parabola is neither an arc nor a closed curve. It may be regarded, however, as an infinite number of arcs joined end to end.

A curve is called *smooth* if two conditions are satisfied: (1) it does not intersect itself, and (2) it has a tangent line at each point, whose direction varies continuously as the point moves along the curve. The second of these conditions is satisfied if the functions f, g, h in (14.1–1) have continuous derivatives which do not all vanish together for any value of t. The direction of the tangent line is specified by the ratios

$$f'(t) : g'(t) : h'(t), \tag{14.1–2}$$

and we can find direction cosines by normalization (i.e., division by the square root of the sum of the squares of the three quantities in (14.1–2)).

The curves with which we normally deal either are smooth or each bounded portion of the curve is *sectionally smooth*, that is, composed of a finite number of smooth arcs joined end to end. Such a curve may have corners at the junction points. The periphery of a square is an example of a closed curve with corners.

Frequently we encounter curves as the intersections of two surfaces. A parametric representation of the form (14.1–1) may not immediately present itself, but may theoretically be derived from the analytical representations of the surfaces by implicit-function arguments.

14.2 / ARC LENGTH

The length of a smooth arc from $t = a$ to $t = b$ is given by an integral:

$$L = \int_a^b \left[\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2 \right]^{1/2} dt. \tag{14.2–1}$$

This formula, or the special form of it for plane curves, is familiar from elementary calculus. The derivation will now be given.

Consider any curve C, closed or not, without self-intersections, and defined

by (14.1–1) with a finite parameter interval $a \leqq t \leqq b$. Consider a subdivision

$$a = t_0 < t_1 < t_2 < \cdots < t_n = b$$

of the interval (a, b) into n subintervals, and let P_k be the point of C corresponding to $t = t_k$. Joining the points $P_0, P_1, \ldots, P_n$ in order, we obtain a polygonal line inscribed on C. The length of the polygonal line is

$$\overline{P_0P_1} + \cdots + \overline{P_{n-1}P_n}. \tag{14.2–2}$$

Now consider what happens if we allow n to increase, and make the length of the longest of the segments $P_{k-1}P_k$ approach zero. If the sum (14.2–2) approaches a finite limit, we call this limit the length of the curve C, and we say that C is *rectifiable.* (For a curve that is *not* rectifiable, see Exercise 11.) A curve may be rectifiable without being smooth or even sectionally smooth; we shall confine ourselves to showing that a smooth arc is rectifiable, with length given by (14.2–1). The length of a sectionally smooth curve is found by adding the lengths of its component arcs.

Consider a segment $P_{k-1}P_k$ of the polygonal line. Its length is

$$[(\Delta x_k)^2 + (\Delta y_k)^2 + (\Delta z_k)^2]^{1/2}, \tag{14.2–3}$$

where

$$\Delta x_k = f(t_k) - f(t_{k-1}),$$

with similar formulas for Δy_k and Δz_k. By the law of the mean we find

$$\Delta x_k = f'(\alpha_k)\,\Delta t_k, \Delta y_k = g'(\beta_k)\,\Delta t_k, \Delta z_k = h'(\gamma_k)\,\Delta t_k,$$

where $\Delta t_k = t_k - t_{k-1}$ and the points $\alpha_k, \beta_k, \gamma_k$ are between t_{k-1} and t_k. Thus (14.2–2) becomes

$$\sum_{k=1}^{n} [(f'(\alpha_k))^2 + (g'(\beta_k))^2 + (h'(\gamma_k))^2]^{1/2}\,\Delta t_k. \tag{14.2–4}$$

If the points $\alpha_k, \beta_k, \gamma_k$ were all the same, the limit of the sum (14.2–4) would be, by definition, the integral

$$\int_a^b [(f'(t))^2 + (g'(t))^2 + (h'(t))^2]^{1/2}\,dt,$$

which is the same as (14.2–1). Because the three points need not be the same, more argument is needed. The matter is covered, however, by a general theorem (Duhamel's principle) in the theory of integration, and we refer the student to Example 2, §18.21, for the final discussion of the issue.

Let C be a smooth arc and let s be the length measured along C from $t = a$ to a variable point. It follows from (14.2–1), with the upper limit b replaced by a variable, that

$$\frac{ds}{dt} = \left[\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2\right]^{1/2}. \tag{14.2–5}$$

Hence

$$ds^2 = dx^2 + dy^2 + dz^2. \tag{14.2–6}$$

The integral giving the arc length is frequently such that no evaluation of the integral can be made in terms of elementary functions. This is the case even

with plane curves. Sometimes the arc length is expressible in terms of tabulated standard integrals, such as elliptic integrals.

Example. Consider the first octant portion of the curve of intersection of the sphere and cylinder

$$x^2 + y^2 + z^2 = 4a^2, \qquad x^2 + (y - a)^2 = a^2, \tag{14.2–7}$$

as shown in Fig. 124.

It is convenient to use z as a parameter for this curve. If we eliminate x by subtracting the two equations in (14.2–7), we find

$$y = \frac{4a^2 - z^2}{2a}.$$

Substituting this result in the second of the equations (14.2–7), we find

$$x = \frac{z}{2a}\sqrt{4a^2 - z^2}.$$

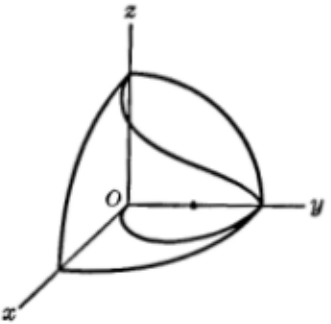


Fig. 124.

As parametric equations of the curve, we have

$$x = \frac{z}{2a}\sqrt{4a^2 - z^2}, \qquad y = \frac{4a^2 - z^2}{2a}, \qquad z = z.$$

A direct calculation shows that

$$ds^2 = \frac{8a^2 - z^2}{4a^2 - z^2}\,dz^2.$$

The range of z is from 0 to $2a$, so the length of the first-octant portion of the curve is

$$L = \int_0^{2a} \left(\frac{8a^2 - z^2}{4a^2 - z^2}\right)^{1/2} dz. \tag{14.2–8}$$

This integral is improper at the limit $z = 2a$, but it is convergent. It can be put in the form of a standard elliptic integral of the second kind. For further discussion of this problem see Exercise 8.

EXERCISES

The standard elliptic integral of the second kind is defined as

$$E(k, \phi) = \int_0^{\phi} \sqrt{1 - k^2 \sin^2 t}\,dt,$$

where $0 < k < 1$. If $\phi = \pi/2$, the integral is called *complete*. Values of this integral for various values of the parameters k, ϕ are given in many books of tables. In some of the exercises it is required that the arc length be expressed in the form of such a standard integral.

1. Consider the cylindrical helix $x = a\cos\theta$, $y = a\sin\theta$, $z = k\theta$. Show that the arc length from $\theta = \theta_0$ to $\theta = \theta_1$ is directly proportional to $\theta_1 - \theta_0$.

2. A point moves on a conical helix according to the formulas $x = t\cos\omega t$, $y = t\sin\omega t$, $z = t$, where t is the time. Find the speed of the point at $t = 0$ and at $t = 1$, and its average speed during this time interval.

3. Find the length of the curve $x = a(\theta - \sin\theta)$, $y = a(1 - \cos\theta)$, $z = 4a\sin(\theta/2)$, $0 \leqq \theta \leqq 2\pi$.

4. If $0 < b < 4a$, show that the length of the curve $x = a(\theta - \sin\theta)$, $y = a(1 - \cos\theta)$, $z = b\sin(\theta/2)$ from $\theta = 0$ to $\theta = 2\pi$ is

$$\int_0^{2\pi} \left[4a^2 - \left(4a^2 - \frac{b^2}{4}\right)\cos^2\frac{\theta}{2}\right]^{1/2} d\theta$$

and that this is equal to $8aE(k, \pi/2)$, where $k^2 = 1 - (b^2/16a^2)$.

5. If $0 < 4a < b$, show that the arc length sought in Exercise 4 is $2bE(k, \pi/2)$, where $k^2 = 1 - (16a^2/b^2)$.

6. Show that the total circumference of an ellipse of major axis $2a$ and eccentricity e is $4aE(e, \pi/2)$. Start with the parametric form $x = a\sin t$, $y = b\cos t$, $b^2 = a^2(1 - e^2)$.

7. Consider the intersection of the ellipsoid $(x^2/a^2) + (y^2/b^2) + (z^2/c^2) = 1$ and the cylinder $x^2 + y^2 = b^2$, where $a > b$. Show that the first-octant portion of this curve has arc length

$$\frac{c\sqrt{a^2 - b^2}}{ak} E\left(k, \frac{\pi}{2}\right), \qquad \text{where } k^2 = \frac{c^2(a^2 - b^2)}{c^2(a^2 - b^2) + a^2b^2}.$$

8. Show that the integral (14.2–8) is equal to $2\sqrt{2}\,aE(\sqrt{2}/2, \pi/2)$. This may be done either by making a suitable change of variable in (14.2–8) or by using the parametric representation of the original curve: $x = a\sin 2t$, $y = a + a\cos 2t$, $z = 2a\sin t$.

9. Find the length of the curve $x = a\cosh t$, $y = a\sinh t$, $z = at$, from $t = 0$ to $t = t_1$. This curve lies on the hyperbolic cylinder $x^2 - y^2 = a^2$.

10. A point starts at $(0, a, 0)$ and moves into the first octant along the curve $y = a\cosh(x/a)$, $z = a\sinh(x/a)$. Show that when it is at (x, y, z) it has traveled a distance $\sqrt{2}z$.

11. Consider the plane curve defined by $y = f(x)$, $0 \leqq x \leqq \frac{2}{\pi}$, where $f(x) = x\sin\frac{1}{x}$ if $x \neq 0$ and $f(0) = 0$. Show by direct use of the definition of a rectifiable curve that this curve is not rectifiable. HINT: Consider the points of the curve with x-coordinates 0 and $\frac{2}{k\pi}$, $k = 1, 2, \ldots, 2n$.

14.3 / THE TANGENT VECTOR

Let C be a smooth arc. Let us choose a certain direction along the arc as the positive direction. At any point of C we shall define what we call the *tangent vector*. This is a vector of unit length along the tangent line in the positive direction. See Fig. 125. It is denoted by $\mathbf{T}$; the vector $\mathbf{T}$ is a vector point function (in the sense of §10.51) defined along C. If the angles which $\mathbf{T}$ makes with the positive co-ordinate axes are α, β, γ we can

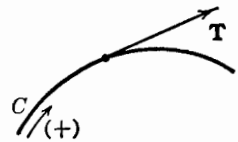


Fig. 125.

write

$$\mathbf{T} = \cos\alpha\,\mathbf{i} + \cos\beta\,\mathbf{j} + \cos\gamma\,\mathbf{k}.$$

If s is arc length measured along C in the positive direction from some fixed point, the direction cosines are given by $\cos\alpha = \dfrac{dx}{ds}$, and so on, so that

$$\mathbf{T} = \frac{dx}{ds}\,\mathbf{i} + \frac{dy}{ds}\,\mathbf{j} + \frac{dz}{ds}\,\mathbf{k}. \tag{14.3–1}$$

If x, y, z are functions of t with continuous derivatives which are never all simultaneously zero,

$$\frac{ds}{dt} = \pm\left[\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2\right]^{1/2}, \tag{14.3–2}$$

with the plus sign chosen if s and t increase in the same sense along C; otherwise the minus sign is chosen. Since

$$\frac{dx}{dt} = \frac{dx}{ds}\frac{ds}{dt},$$

and so on, we see that $\mathbf{T}$ may be found from the formula

$$\mathbf{T} = \frac{1}{\left(\dfrac{ds}{dt}\right)}\left(\frac{dx}{dt}\,\mathbf{i} + \frac{dy}{dt}\,\mathbf{j} + \frac{dz}{dt}\,\mathbf{k}\right), \tag{14.3–3}$$

together with (14.3–2).

For many purposes it is convenient to discuss $\mathbf{T}$ entirely in vector notation, without the use of a co-ordinate system. If $P(x, y, z)$ is a variable point on C, we denote the vector $\overrightarrow{OP}$ by $\mathbf{R}$. In the co-ordinate notation

$$\mathbf{R} = x\,\mathbf{i} + y\,\mathbf{j} + z\,\mathbf{k}. \tag{14.3–4}$$

Equation (14.3–1) may be written

$$\frac{d\mathbf{R}}{ds} = \mathbf{T}. \tag{14.3–5}$$

This formula for the tangent vector can be derived directly by vector methods; we give the derivation, because it provides an instructive exercise in learning how to differentiate a vector function of a scalar variable.

We consider R as a function of s. Let the points P, P' on C correspond to s and $s + \Delta s$, respectively. Write

$$\overrightarrow{OP} = \mathbf{R},\ \overrightarrow{OP'} = \mathbf{R} + \Delta\mathbf{R},$$

so that

$$\Delta\mathbf{R} = \overrightarrow{PP'}.$$

The quotient $\Delta\mathbf{R}/\Delta s$ is a vector along the line of the chord PP'. See Fig. 126. Since the length of $\Delta\mathbf{R}$ is the length of the chord PP', we see that when P' approaches P the limit of the length of $\Delta\mathbf{R}/\Delta s$ is unity. Furthermore, the limiting direction of PP' is that of the tangent at P. Therefore

$$\frac{d\mathbf{R}}{ds} = \lim_{\Delta s \to 0} \frac{\Delta \mathbf{R}}{\Delta s} = \mathbf{T}.$$

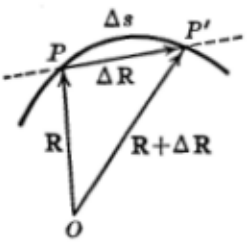


Fig. 126.

Differentiation of $\mathbf{R}$ with respect to t gives

$$\frac{d\mathbf{R}}{dt} = \frac{d\mathbf{R}}{ds}\frac{ds}{dt} = \frac{ds}{dt}\mathbf{T}. \tag{14.3–6}$$

This is equivalent to formula (14.3–3). If t is the time variable, $\dfrac{d\mathbf{R}}{dt}$ is the *vector velocity* of the point P moving on C.

EXERCISES

1. Let $\mathbf{F}$ and $\mathbf{G}$ denote vector functions of a scalar variable t. Assuming that $\mathbf{F}$ and $\mathbf{G}$ are differentiable, prove the formulas

(a) $\dfrac{d}{dt}(\mathbf{F}\cdot\mathbf{G}) = \mathbf{F}\cdot\dfrac{d\mathbf{G}}{dt} + \dfrac{d\mathbf{F}}{dt}\cdot\mathbf{G}$,

(b) $\dfrac{d}{dt}(\mathbf{F}\times\mathbf{G}) = \mathbf{F}\times\dfrac{d\mathbf{G}}{dt} + \dfrac{d\mathbf{F}}{dt}\times\mathbf{G}$,

using the same method by which the rule for differentiating products is derived in elementary calculus.

2. If $\mathbf{F}$ in Exercise 1 is a vector of constant length, prove that $\mathbf{F}\cdot\dfrac{d\mathbf{F}}{dt} = 0$, and thus conclude that $\dfrac{d\mathbf{F}}{dt}$ is perpendicular to $\mathbf{F}$ unless $\dfrac{d\mathbf{F}}{dt} = 0$.

14.31 / PRINCIPAL NORMAL. CURVATURE

In this section and the next we continue with the notations used in §14.3. We shall define two more unit vectors, the principal normal $\mathbf{N}$ and the binormal $\mathbf{B}$, which, along with the tangent vector $\mathbf{T}$, form an orthonormal set of vectors associated with the point (x, y, z) on the curve C. These vectors, especially $\mathbf{T}$ and $\mathbf{N}$, are important in the study of the motion of the point (x, y, z) along the curve. The acceleration vector lies in the plane of $\mathbf{T}$ and $\mathbf{N}$. The component of the acceleration along the line of $\mathbf{N}$ will be shown to depend on the curvature of C. We shall assume that x, y, z have second derivatives with respect to s. From (14.3–1) we have

$$\frac{d\mathbf{T}}{ds} = \frac{d^2x}{ds^2}\mathbf{i} + \frac{d^2y}{ds^2}\mathbf{j} + \frac{d^2z}{ds^2}\mathbf{k}. \tag{14.31–1}$$

We shall prove that

$$\mathbf{T} \cdot \frac{d\mathbf{T}}{ds} = 0. \tag{14.31–2}$$

This means that $\frac{d\mathbf{T}}{ds}$ is either $\mathbf{0}$ or perpendicular to $\mathbf{T}$. The demonstration may be given in several ways. From (14.2–6) we have

$$1 = \left(\frac{dx}{ds}\right)^2 + \left(\frac{dy}{ds}\right)^2 + \left(\frac{dz}{ds}\right)^2.$$

Therefore, differentiating with respect to s,

$$0 = 2\left(\frac{dx}{ds}\frac{d^2x}{ds^2} + \frac{dy}{ds}\frac{d^2y}{ds^2} + \frac{dz}{ds}\frac{d^2z}{ds^2}\right).$$

The right member of this formula is exactly $2\mathbf{T} \cdot \frac{d\mathbf{T}}{ds}$ as we see from (14.3–1) and (14.31–1) and the formula for the dot product in terms of components. Thus (14.31–2) is proved. Alternatively (14.31–2) is a special case of the result proved in Exercise 2, §14.3.

When $\frac{d\mathbf{T}}{ds}$ is not $\mathbf{0}$, its direction is perpendicular to $\mathbf{T}$. A unit vector in the direction of $\frac{d\mathbf{T}}{ds}$ is called the *principal normal* to C at the point in question. The principal normal is denoted by $\mathbf{N}$, and the length of $\frac{d\mathbf{T}}{ds}$ is denoted by κ. Therefore

$$\frac{d\mathbf{T}}{ds} = \kappa\mathbf{N}. \tag{14.31–3}$$

The scalar κ is called the *curvature* of C; it is given by

$$\kappa = \left[\left(\frac{d^2x}{ds^2}\right)^2 + \left(\frac{d^2y}{ds^2}\right)^2 + \left(\frac{d^2z}{ds^2}\right)^2\right]^{1/2}. \tag{14.31–4}$$

This definition makes the curvature always either positive or zero. In the case of plane curves it is easy to see, by sketching a figure, that the principal normal lies in the plane of the curve and points toward the concave side of the curve. In the theory of plane curves, the principal normal is called simply the *normal.*

If $\frac{d\mathbf{T}}{ds} = \mathbf{0}$, our definition of $\mathbf{N}$ breaks down. This happens at all points of C if C is a straight line, and in this case there is no basis for calling any particular normal the principal one. Otherwise the vanishing of $\frac{d\mathbf{T}}{ds}$ is exceptional.

The principal normal and the curvature play an important role in connection with the acceleration of a particle moving in a curve.

We saw earlier that the vector velocity is

$$\mathbf{v} = \frac{d\mathbf{R}}{dt}.$$

The vector acceleration is therefore

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} = \frac{d^2\mathbf{R}}{dt^2}.$$

Now, from (14.3–6), we see that

$$\frac{d^2\mathbf{R}}{dt^2} = \frac{ds}{dt}\frac{d\mathbf{T}}{dt} + \frac{d^2s}{dt^2}\mathbf{T}.$$

Also,

$$\frac{d\mathbf{T}}{dt} = \frac{d\mathbf{T}}{ds}\frac{ds}{dt}.$$

Therefore, using (14.31–3), we see that

$$\frac{d^2\mathbf{R}}{dt^2} = \frac{d^2s}{dt^2}\mathbf{T} + \kappa\left(\frac{ds}{dt}\right)^2\mathbf{N}. \tag{14.31–5}$$

The acceleration vector is made up of a component of magnitude $\frac{d^2s}{dt^2}$ along the tangent vector, and one of magnitude $\kappa\left(\frac{ds}{dt}\right)^2$ along the principal normal.

The reciprocal of the curvature is called the *radius of curvature* ρ:

$$\rho = \frac{1}{\kappa}. \tag{14.31–6}$$

If we write $v = \left|\frac{ds}{dt}\right|$ for the speed of the particle moving along C, the component $\kappa\left(\frac{ds}{dt}\right)^2$ can be written $\frac{v^2}{\rho}$. For motion in a circular path this is the familiar form of the centripetal acceleration.

The tip of the vector $\rho\mathbf{N}$, drawn from the point P as initial point, is called the *center of curvature* of C corresponding to P.

The plane of $\mathbf{T}$ and $\mathbf{N}$ is called the *osculating plane* of C at P.

14.32 / BINORMAL. TORSION

The vector

$$\mathbf{B} = \mathbf{T} \times \mathbf{N} \tag{14.32–1}$$

is called the *binormal* vector of C at P. Observe that $\mathbf{T}$, $\mathbf{N}$, and $\mathbf{B}$ are mutually perpendicular unit vectors; moreover, they form a right-handed system, just as $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ do. The binormal $\mathbf{B}$ is perpendicular to the osculating plane. See Fig. 127.

There is a scalar quantity called *torsion* associated with C at each point. To define torsion we consider $\frac{d\mathbf{B}}{ds}$. From (14.32–1) we have (see Exercise 1(**b**), §14.3)

$$\frac{d\mathbf{B}}{ds} = \mathbf{T} \times \frac{d\mathbf{N}}{ds} + \frac{d\mathbf{T}}{ds} \times \mathbf{N}.$$

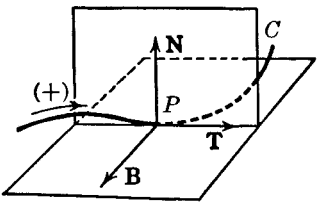


Fig. 127.

We know by (14.31–3) that

$$\frac{d\mathbf{T}}{ds} \times \mathbf{N} = \kappa \mathbf{N} \times \mathbf{N} = \mathbf{0},$$

since the cross product of any vector with itself is zero. Therefore

$$\frac{d\mathbf{B}}{ds} = \mathbf{T} \times \frac{d\mathbf{N}}{ds}. \tag{14.32–2}$$

Also, $\mathbf{B} \cdot \mathbf{B} = 1$, since $\mathbf{B}$ has unit length. Therefore, by Exercise 2, §14.3,

$$\mathbf{B} \cdot \frac{d\mathbf{B}}{ds} = 0. \tag{14.32–3}$$

Now, if $\frac{d\mathbf{B}}{ds} \neq \mathbf{0}$, the two preceding equations show that $\frac{d\mathbf{B}}{ds}$ is perpendicular both to $\mathbf{B}$ and to $\mathbf{T}$; it is therefore a multiple of $\mathbf{N}$, for $\mathbf{N}$ is also perpendicular to both $\mathbf{B}$ and $\mathbf{T}$. Thus we can write

$$\frac{d\mathbf{B}}{ds} = -\tau \mathbf{N}, \tag{14.32–4}$$

where $-\tau$ is simply the proper multiple of $\mathbf{N}$ to give $\frac{d\mathbf{B}}{ds}$. The quantity τ thus defined is called the torsion of C at the point under consideration. If $\frac{d\mathbf{B}}{ds} = \mathbf{0}$, we define $\tau = 0$, so that (14.32–4) still holds.

For plane curves, $\mathbf{B}$ is constant, being a unit vector perpendicular to the plane. For such curves $\tau = 0$ at all points. A curve which does not lie in a single plane is called a *twisted* curve; its torsion measures, to some extent, the amount by which the curve is twisted.

For methods of finding $\mathbf{T}$, $\mathbf{N}$, $\mathbf{B}$, κ, and τ from the parametric equations of the curve, see Exercises 3, 4.

EXERCISES

1. If P and P' correspond to s and $s + \Delta s$, and if $\mathbf{T}$ and $\mathbf{T} + \Delta\mathbf{T}$ are the corresponding tangents to C, let $\Delta\theta$ be the angle between these tangents. Show that

$$\lim_{\Delta s \to 0} \frac{\Delta\theta}{\Delta s} = \kappa.$$

2. Prove that

$$\frac{d\mathbf{N}}{ds} = \tau \mathbf{B} - \kappa \mathbf{T}.$$

Formulas (14.31–3), (14.32–4), and the foregoing formula are called the Frenet formulas, or the Serret-Frenet formulas.

3. (**a**) Starting from (14.31–5), show that

$$\frac{d^3\mathbf{R}}{dt^3} = \left(\frac{d^3s}{dt^3} - \left(\frac{ds}{dt}\right)^3\kappa^2\right)\mathbf{T} + \left(3\kappa\frac{ds}{dt}\frac{d^2s}{dt^2} + \left(\frac{ds}{dt}\right)^2\frac{d\kappa}{dt}\right)\mathbf{N} + \kappa\tau\left(\frac{ds}{dt}\right)^3\mathbf{B}.$$

Use the result in Exercise 2.

(**b**) Show that

$$\frac{d\mathbf{R}}{dt}\times\frac{d^2\mathbf{R}}{dt^2} = \kappa\left(\frac{ds}{dt}\right)^3\mathbf{B}, \qquad \left(\frac{d\mathbf{R}}{dt}\times\frac{d^2\mathbf{R}}{dt^2}\right)\cdot\left(\frac{d^3\mathbf{R}}{dt^3}\right) = \kappa^2\tau\left(\frac{ds}{dt}\right)^6.$$

(**c**) Show that

$$\kappa = \frac{|\mathbf{R}'\times\mathbf{R}''|}{|\mathbf{R}'|^3}, \qquad \tau = \frac{(\mathbf{R}'\times\mathbf{R}'')\cdot\mathbf{R}'''}{|\mathbf{R}'\times\mathbf{R}''|^2},$$

where primes denote differentiation with respect to t. In terms of x, y, z as functions of t, these formulas become

$$\kappa = \frac{[(y'z''-y''z')^2+(z'x''-z''x')^2+(x'y''-x''y')^2]^{1/2}}{(x'^2+y'^2+z'^2)^{3/2}},$$

$$\tau = \frac{\begin{vmatrix} x' & y' & z' \\ x'' & y'' & z'' \\ x''' & y''' & z''' \end{vmatrix}}{(y'z''-y''z')^2+(z'x''-z''x')^2+(x'y''-x''y')^2}.$$

4. To find **T**, **N**, and **B** from the parametric equations of the curve, observe first of all that **T** is a positive multiple of $\frac{d\mathbf{R}}{dt}$, provided $\frac{ds}{dt} > 0$, i.e., provided the positive sense along the curve is the same as that in which t increases. Since **T** is a unit vector, it follows that

$$\mathbf{T} = \frac{\mathbf{R}'}{|\mathbf{R}'|}.$$

Show with the help of Exercise 3(**b**) that

$$\mathbf{B} = \frac{\mathbf{R}'\times\mathbf{R}''}{|\mathbf{R}'\times\mathbf{R}''|}, \qquad \mathbf{N} = \frac{(\mathbf{R}'\times\mathbf{R}'')\times\mathbf{R}'}{|(\mathbf{R}'\times\mathbf{R}'')\times\mathbf{R}'|}.$$

If $\frac{ds}{dt} < 0$, the formulas for **T** and **B** require a minus sign on the right, but the formula for **N** is unchanged. If **T** and **B** have already been found, **N** is easily found from $\mathbf{N} = \mathbf{B}\times\mathbf{T}$.

5. (**a**) Find **T**, **N**, **B**, κ, and τ at $t = 0$ on the curve $x = t$, $y = t^2$, $z = t^3$.

(**b**) Find κ at $z = 1$ on the curve $x = \sqrt{z}$, $y = 2\sqrt{z}$.

(**c**) Find κ and τ at $x = 3a$ on the curve $3ay = x^2$, $2xz = a^2$.

(**d**) Find **T**, **N**, **B**, κ, and τ at $t = \pi$ on the curve $x = t - \sin t$, $y = 1 - \cos t$, $z = 4\sin(t/2)$.

(**e**) Find κ and τ in terms of t for the curve $x = \cos t$, $y = \sin t$, $z = e^t$.

(**f**) For $x = 3t - t^3$, $y = 3t^2$, $z = 3t + t^3$, show that $\kappa = \tau = \frac{1}{3(1+t^2)^2}$.

(**g**) Find **T**, **N**, **B**, κ, and τ for the conical helix $x = 3t\cos t$, $y = 3t\sin t$, $z = 4t$, at $t = 0$.

6. Consider the helix $x = a\cos t, y = a\sin t,\ z = at\tan\alpha$. It lies on the cylinder $x^2 + y^2 = a^2$. It is a *screw curve*, with pitch angle α. As t increases, the point describes a right-handed screw path if $0 < \alpha < \pi/2$, and a left-handed screw path if $-\pi/2 < \alpha < 0$.

(a) Find the vectors **T**, **N**, **B** in terms of t. Show that **T** and **B** make constant angles with the z-axis, and that **N** is perpendicular to the z-axis and directed toward it.

(b) Show that

$$\kappa = \frac{\cos^2\alpha}{a}, \qquad \tau = \frac{\sin\alpha\cos\alpha}{a}.$$

(*c*) Show that the center of curvature of the helix, corresponding to the point on the helix at the tip of **R**, is at $\mathbf{R}_1 = \mathbf{R} + \rho\mathbf{N}$, and hence show that the center of curvature moves on a helix whose pitch angle is $(\pi/2) - \alpha$ and which lies on a coaxial cylinder of radius $a\tan^2\alpha$. Show also that the center of curvature for this second helix is the original point on the original helix.

7. Show that, if a point moves so that its velocity and acceleration vectors always have unit length, $\kappa = 1$ at all points of the path.

8. For the curve $x = a\sin 2t,\ y = a + a\cos 2t,\ z = 2a\sin t$ (see Fig. 124) find **T**, **N**, **B**, κ, and τ **(a)** at $t = \pi/2$; **(b)** at $t = \pi/4$; **(c)** at $t = 0$.

9. Suppose $|\mathbf{R}| = a$, so that the tip of **R** moves on a sphere of radius a with center at O. **(a)** Show that $\mathbf{R} = \alpha\mathbf{N} + \beta\mathbf{B}$, where α and β are related to ρ and τ by the equations $\alpha = -\rho,\ \beta\tau = -\dfrac{d\rho}{ds}$. **(b)** If $\tau \equiv 0$, show that the tip of **R** moves on a circle about the line of **B** as an axis. **(c)** If $\tau \neq 0$, show that $\rho^2 + \dfrac{1}{\tau^2}\left(\dfrac{d\rho}{ds}\right)^2 = a^2$.

14.4 / SURFACES

Speaking roughly, a surface is a configuration of points having a two-dimensional character; that is, a point moving on the surface, but otherwise unrestricted, has two degrees of freedom. This description of a surface is not a strict mathematical definition, however, and we need much greater precision in the work which is to follow.

Considerations about surfaces are of two kinds, which we may designate respectively as *local* and *in the large*. By a local property of a surface at a point we mean a property which can be completely described and analyzed by considering only the part of the surface in the immediate vicinity of the point. A property "in the large" is a characteristic of the surface as a whole, and cannot be defined merely in terms of the features of the surface near a single point. The property of *having a tangent plane at a given point* is a local property. A surface may have this property at some points but not at others. *Being a closed surface* is a property in the large (a spherical surface is closed, while a circular disk is not). There is clearly a difference in the large between the surface of a basketball and the surface of an inner tube. The mathematical name for the latter surface is *torus*. There is a classification of surfaces according to certain properties in the large and ignoring local properties. In this classification a spherical and an ellipsoidal surface are of the same type, while the torus is of a different type. Properties in the large are in many ways harder to study than local properties.

There are three common methods for the analytical representation of a surface. A locus in space defined by

$$z = f(x, y), \tag{14.4–1}$$

where f is a single-valued continuous function defined on a region R of the xy-plane, is called a surface. In order to have the surface consist of one piece we assume that R is a connected region. A point (x, y, z) is called an interior point of the surface if the corresponding point (x, y) of R is interior to R. The points (x, y, z) corresponding to the boundary of R form the boundary or edge of the surface.

Often we have surfaces represented by equations of the form

$$F(x, y, z) = 0. \tag{14.4–2}$$

If (x_0, y_0, z_0) is a point of such a surface, we can in many cases represent the portion of the surface near (x_0, y_0, z_0) in a form analogous to (14.4–1) by solving (14.4–2) for x, y, or z in terms of the two other variables.

For many purposes, especially in theoretical work, parametric representation of a surface is the most convenient. A locus defined by

$$x = f(u, v), \qquad y = g(u, v), \qquad z = h(u, v), \tag{14.4–3}$$

where f, g, h are continuous functions defined on a connected region R of the uv-plane, is called a *parametric surface.* There may be other parametric representations of the same surface. This definition is so general that a parametric surface may not have any resemblance to what we would consider a surface to be, intuitively. It might, for example, intersect itself along infinitely many distinct curves in the vicinity of a single point. Hence, just as in dealing with curves in §14.1, we find it necessary to introduce more restrictive definitions.

Suppose R is a closed rectangular region of the uv-plane: $a \leqq u \leqq b$, $c \leqq v \leqq d$, and suppose the equation (14.4–3) define a continuous one-to-one mapping of R onto a point set S in xyz-space. Then S is called a *simple surface element.* The one-to-one condition means that distinct points of R are not mapped into the same point of S. A simple surface element may be thought of as any configuration which may be obtained from a rectangular plane region by continuous deformation (bending, twisting, stretching, shrinking) without tearing and without bringing together any points which were originally distinct. In particular, a plane circular region is a simple surface element; so is a hemispherical surface.

If S is a simple surface element corresponding to a rectangular region R in the uv-plane, the points of S which correspond to the boundary of R form what is called the boundary of S. Other points of S are called interior points of the surface element.

The lateral surface of a right circular cylinder of finite length is not a simple surface element. It can, however, be regarded as composed of two simple

surface elements joined together along the lines AB and CD (see Fig. 128). One of the elements in AA_1CDB_1B; the other is AA_2CDB_2B.

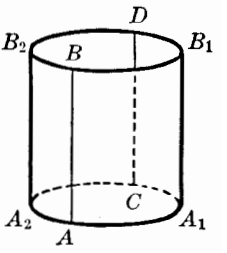


Fig. 128.

All surfaces which we shall subsequently consider may be thought of as being built up out of simple surface elements by matching together portions of the edges of the latter. This matching is always done in such a way that never more than two surface elements have an arc of edge in common. An infinite number of surface elements will be required to form a surface of infinite extent, such as a paraboloid or a complete cylinder, but, in considering surfaces in bounded portions of space, we shall assume that such surfaces can be formed from a finite number of simple surface elements. The exact manner of drawing the curves which divide a surface into elements is not unique; the curves can always be chosen so as to avoid the neighborhood of any particular point. Thus, in studying a local property of a surface, we may always assume that we are dealing with an interior point of a single surface element.

The boundary of a surface consists of the unmatched edges of its surface elements. If there are no unmatched edges, there is no boundary. A surface which has no boundary, and which lies in a bounded portion of space, is called a *closed surface.* For example, the surface of a sphere is closed, but the cylinder surface in Fig. 128 is not closed.

A surface is called *smooth* at a point P if it has a tangent plane at each point in the neighborhood of P, and if the direction of the normal to this plane varies continuously from point to point. The whole surface is called smooth if it is smooth at every point.

The surface represented by (14.4–1) is smooth if f has continuous first partial derivatives. The direction of the normal in this case is specified by the ratios

$$\frac{\partial z}{\partial x} : \frac{\partial z}{\partial y} : -1, \tag{14.4–4}$$

as we saw in §6.2. In the representation (14.4–2) the surface is smooth at a point where the three partial derivatives $\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}, \frac{\partial F}{\partial z}$ do not all vanish, provided they are continuous in the neighborhood of the point. The normal direction is specified by the ratios

$$F_1 : F_2 : F_3$$

(see Example 1, §6.6).

In the parametric case (14.4–3) let us suppose that the functions f, g, h have continuous first partial derivatives. We introduce the notations

$$j_1 = \frac{\partial(y, z)}{\partial(u, v)}, \qquad j_2 = \frac{\partial(z, x)}{\partial(u, v)}, \qquad j_3 = \frac{\partial(x, y)}{\partial(u, v)}. \tag{14.4–5}$$

We assume that these three Jacobians do not vanish simultaneously. Under these conditions we can show that the surface is smooth, and that the normal direction is given by the ratios

$$j_1 : j_2 : j_3. \tag{14.4–6}$$

For the proof, suppose $j_3 \neq 0$. Then, by the implicit-function theorem (Theorem III, §8.3), the equations $x = f(u, v)$, $y = g(u, v)$ can be solved for u, v in terms of x, y. Substituting the solutions into $z = h(u, v)$ we obtain z as a function of x, y. This means that the portion of the surface near a point where $j_3 \neq 0$ can be represented in the form (14.4–1). By the rules for differentiating composite functions we shall have

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial x}, \tag{14.4–7}$$

with a similar formula for $\dfrac{\partial z}{\partial y}$. If we regard u, v as dependent and x, y as independent in the equations $x = f(u, v)$, $y = g(u, v)$, we have

$$1 = \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x},$$

$$0 = \frac{\partial g}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial g}{\partial v}\frac{\partial v}{\partial x}.$$

Solving for $\dfrac{\partial u}{\partial x}$ and $\dfrac{\partial v}{\partial x}$, we find

$$\frac{\partial u}{\partial x} = \frac{\begin{vmatrix} 1 & f_2 \\ 0 & g_2 \end{vmatrix}}{\begin{vmatrix} f_1 & f_2 \\ g_1 & g_2 \end{vmatrix}}, \quad \frac{\partial v}{\partial x} = \frac{\begin{vmatrix} f_1 & 1 \\ g_1 & 0 \end{vmatrix}}{\begin{vmatrix} f_1 & f_2 \\ g_1 & g_2 \end{vmatrix}},$$

or

$$\frac{\partial u}{\partial x} = \frac{1}{j_3}\frac{\partial y}{\partial v}, \quad \frac{\partial v}{\partial x} = -\frac{1}{j_3}\frac{\partial y}{\partial u}.$$

The foregoing formulas are the same as some of those in (9.1–5), except for minor changes in notation. Hence, from (14.4–7)

$$\frac{\partial z}{\partial x} = \frac{1}{j_3}\left(\frac{\partial z}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial z}{\partial v}\frac{\partial y}{\partial u}\right) = -\frac{j_1}{j_3}.$$

Similarly,

$$0 = \frac{\partial f}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial y},$$

$$1 = \frac{\partial g}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial g}{\partial v}\frac{\partial v}{\partial y},$$

whence

$$\frac{\partial u}{\partial y} = -\frac{1}{j_3}\frac{\partial x}{\partial v}, \frac{\partial v}{\partial y} = \frac{1}{j_3}\frac{\partial x}{\partial u},$$

and

$$\frac{\partial z}{\partial y} = \frac{1}{j_3}\left(-\frac{\partial z}{\partial u}\frac{\partial x}{\partial v} + \frac{\partial z}{\partial v}\frac{\partial x}{\partial u}\right) = -\frac{j_2}{j_3}.$$

It follows that

$$\frac{\partial z}{\partial x} : \frac{\partial z}{\partial y} : -1 = -\frac{j_1}{j_3} : -\frac{j_2}{j_3} : -1.$$

Thus (14.4–4) and (14.4–6) define the same direction; this is what we set out to prove.

Once the direction of the normal is known, it is of course an easy matter to write out the equation of the tangent plane.

EXERCISES

1. Show that the parametric surface defined by $x = a \sin\phi \cos\theta$, $y = b \sin\phi \sin\theta$, $z = c \cos\phi$, $0 \leqq \theta \leqq 2\pi$, $0 \leqq \phi \leqq \pi$, is an ellipsoid, and that it is a sphere if $a = b = c$. The surface is not a simple surface element, however. Which part of the definition of a simple surface element is not satisfied in this case?

2. Explain how to divide the surface of a sphere into simple surface elements in several ways. In particular, if the sphere is $x^2 + y^2 + z^2 = 1$, describe a mode of division such that the points $(0, 0, \pm 1)$ are interior points of elements on which they lie. Describe a mode of division such that the points $(\pm 1, 0, 0)$ and $(0, \pm 1, 0)$ are interior points of the elements on which they lie.

3. For the case of each of the following parametric surfaces, obtain an equation of the surface in rectangular co-ordinates.

(a) $x = au \cos v$, $y = bu \sin v$, $z = u$ (elliptic cone).

(b) $x = u \cos v$, $y = u \sin v$, $z = ku^2$ (paraboloid of revolution).

(c) $x = a \sin u \cosh v$, $y = b \cos u \cosh v$, $z = c \sinh v$ (hyperboloid of one sheet).

(d) $x = r \cos\theta$, $y = r \sin\theta$, $z = (r^2/2) \sin 2\theta$ (hyperbolic paraboloid).

(e) $x = au \cos v$, $y = bu \sin v$, $z = u^2 \cos 2v$ (hyperbolic paraboloid).

(f) $x = a \cosh v$, $y = b \cosh v \cos u$, $z = c \cosh v \sin u$.

4. Show that the tangent plane to the ellipsoid $(x^2/a^2) + (y^2/b^2) + (z^2/c^2) = 1$ at (x_0, y_0, z_0) is $(x_0x/a^2) + (y_0y/b^2) + (z_0z/c^2) = 1$.

5. Show that the direction of the normal to the surface in Exercise **3(e)** is $-2bu \cos v : 2au \sin v : ab$.

6. Describe the parametric surface $x = a \cos u$, $y = a \sin u$, $z = v$, and find its equation in rectangular co-ordinates.

7. Describe the parametric surface $x = 2u + v$, $y = u - v$, $z = 3u$, and find its equation in rectangular co-ordinates.

8. Show that the parametric surface $x = u + v$, $y = u - v$, $z = 4v^2$ is the parabolic cylinder $z = (x - y)^2$. Show that the tangent plane at the point corresponding to (u, v) is $4vx - 4vy - z = 4v^2$.

9. If the curve $y = f(x)$ in the xy-plane is revolved around the x-axis, show that the resulting surface can be represented parametrically in the form $x = u$, $y = f(u) \cos v$, $z = f(u) \sin v$. Assuming that $f'(u)$ is continuous and $f(u) > 0$, show that the direction of the normal is $f'(u) : -\cos v : -\sin v$.

10. (a) Consider the parametric surface $x = u + v$, $y = u^2 + v^2$, $z = u^3 + v^3$. Show that it is part of the surface $x^3 - 3xy + 2z = 0$, and, in fact, just the part for which $x^2 \leqq 2y$ (u and v are required to be real, or course). (b) Find the direction of the normal to the surface at $(0, 8, 0)$, using (14.4–4), and also using (14.4–6). (c) Show that the direction of the normals is $6uv : -3(u + v) : 2$. Is this correct even when $u = v$? (Check with the result obtained from the rectangular equation when $u = v$, i.e., when $2y = x^2$.)

14.5 / CURVES ON A SURFACE

Consider a fixed surface. We are going to study local properties of curves on the surface, so we shall find it convenient to deal with a single simple surface element. We suppose equations (14.4–3) give a parametric representation of the element, and that the functions f, g, h have continuous first partial derivatives. If u and v are made to depend continuously on a parameter t, then x, y, z will become continuous functions of t, and we shall have a curve defined on the surface. If we suppose u and v are continuously differentiable functions of t, any arc of the curve will be rectifiable. We wish to obtain a formula for the differential of the arc-length in terms of u, v, du, and dv. No matter what the parameter is, we have the following formulas:

$$ds^2 = dx^2 + dy^2 + dz^2,$$

$$dx = \frac{\partial x}{\partial u}\,du + \frac{\partial x}{\partial v}\,dv,$$

and similar formulas for dy and dz. Hence

$$ds^2 = \left(\frac{\partial x}{\partial u}\,du + \frac{\partial x}{\partial v}\,dv\right)^2 + \cdots.$$

On carrying out the details, we obtain the result

$$ds^2 = E\,du^2 + 2F\,du\,dv + G\,dv^2, \tag{14.5–1}$$

where

$$E = \left(\frac{\partial x}{\partial u}\right)^2 + \left(\frac{\partial y}{\partial u}\right)^2 + \left(\frac{\partial z}{\partial u}\right)^2,$$

$$F = \frac{\partial x}{\partial u}\frac{\partial x}{\partial v} + \frac{\partial y}{\partial u}\frac{\partial y}{\partial v} + \frac{\partial z}{\partial u}\frac{\partial z}{\partial v}, \tag{14.5–2}$$

$$G = \left(\frac{\partial x}{\partial v}\right)^2 + \left(\frac{\partial y}{\partial v}\right)^2 + \left(\frac{\partial z}{\partial v}\right)^2.$$

We call attention to the structure of (14.5–1). It is a quadratic form in du and dv, with coefficients which are functions of u, v; these coefficients may be found directly from the parametric equations of the surface, as we see from (14.5–2). The quadratic form is called the *first fundamental form* of the surface. This form is of prime importance for what is called the metric differential geometry of the surface.

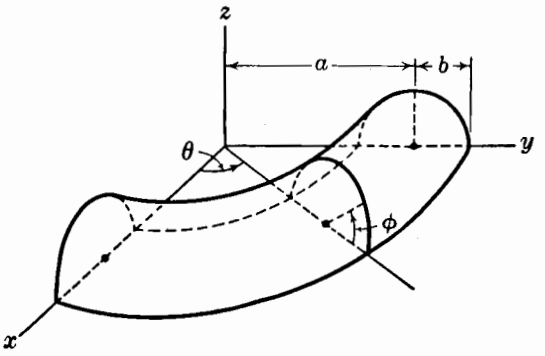


Fig. 129.

Example. The parametric equations

$$x = (a + b\cos\phi)\cos\theta$$
$$y = (a + b\cos\phi)\sin\theta \qquad (14.5\text{–}3)$$
$$z = b\sin\phi$$

in which $0 < b < a$, and θ and ϕ have the ranges $0 \leqq \theta \leqq 2\pi$, $0 \leqq \phi \leqq 2\pi$, define a torus. The part of the surface of the torus in the first octant is shown in Fig. 129. This part is a simple surface element, corresponding to the ranges $0 \leqq \theta \leqq \pi/2$, $0 \leqq \phi \leqq \pi$ of the parameters θ, ϕ. The entire torus is a closed surface, which may be divided into four simple surface elements, corresponding to the four squares in the $\theta\phi$-plane shown in Fig. 130.

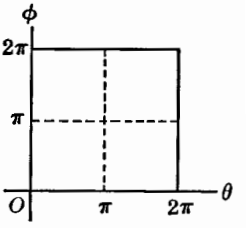


Fig. 130

Let us calculate the first fundamental form for the torus. We do this by direct calculation from (14.5–3), rather than by substitution in (14.5–1) and (14.5–2). Here θ and ϕ play the roles of u and v. We have

$$dx = -(a + b\cos\phi)\sin\theta\, d\theta - b\sin\phi\cos\theta\, d\phi,$$
$$dy = (a + b\cos\phi)\cos\theta\, d\theta - b\sin\phi\sin\theta\, d\phi,$$
$$dz = b\cos\phi\, d\phi.$$

Hence, using the formula (14.2–6) for ds^2, we have

$$ds^2 = (a + b\cos\phi)^2(\sin^2\theta + \cos^2\theta)\, d\theta^2$$
$$+ b^2\sin^2\phi(\sin^2\theta + \cos^2\theta)\, d\phi^2 + b^2\cos^2\phi\, d\phi^2,$$
$$ds^2 = (a + b\cos\phi)^2\, d\theta^2 + b^2\, d\phi^2. \qquad (14.5\text{–}4)$$

Observe that there is no term involving $d\theta\, d\phi$.

When a surface is defined in terms of parameters u, v, the curves $u =$ constant and $v =$ constant are called *co-ordinate curves*. Along a curve $v =$ constant (called a v-curve) we may consider u as a parameter of the curve. The direction of the tangent to this curve is given by the ratios

$$\frac{\partial x}{\partial u} : \frac{\partial y}{\partial u} : \frac{\partial z}{\partial u}.$$

Likewise the direction of the tangent to a u-curve ($u =$ constant) is

$$\frac{\partial x}{\partial v} : \frac{\partial y}{\partial v} : \frac{\partial z}{\partial v}.$$

If the two sets of curves intersect orthogonally at each point, we see from (14.5–2) that $F \equiv 0$. In this case, then, the first fundamental form becomes

$$ds^2 = E\,du^2 + G\,dv^2,$$

and no term in $du\,dv$ is present.

Along a v-curve $ds = \pm\sqrt{E}\,du$, and along a u-curve $ds = \pm\sqrt{G}\,dv$. Hence, if by direct geometric observation we can determine the values of $\dfrac{ds}{du}$ along a v-curve, we can find E without resorting to the work of calculating from (14.5–2). Similar remarks apply to G. In the case of a number of familiar surfaces, these short cuts are quite convenient.

These short cuts are available in the case of the torus of the illustrative Example. In that example the ϕ-curves are circles in horizontal planes, with centers on the z-axis. The radius of a ϕ-circle is $a + b\cos\phi$, and hence $ds = \pm(a + b\cos\phi)\,d\theta$ along the circle. The coefficient of $d\theta^2$ in the general formula for ds^2 is therefore $(a + b\cos\phi)^2$. Discussion of the θ-curves is left to the student. In this example the co-ordinate curves intersect each other at right angles. This accounts for the absence of the $d\theta\,d\phi$ term in the quadratic form.

From the directions of the tangents to the co-ordinate curves we may determine the direction of their common perpendicular, which is normal to the surface. Use of the usual scheme of three two-row determinants shows us that the direction of the normal is $j_1 : j_2 : j_3$. Thus we obtain the result (14.4–6) in a different way.

For certain purposes it is very convenient to use sets of variables with indices. Thus, we might write x_1, x_2, x_3 instead of x, y, z, and u_1, u_2 instead of u, v. If this is done the formulas for E, F, G can be written compactly in summation notation; for example,

$$F = \sum_{i=1}^{3} \frac{\partial x_i}{\partial u_1}\frac{\partial x_i}{\partial u_2}.$$

The formula for ds^2 can be written

$$ds^2 = \sum_{\alpha,\beta=1}^{2} \sum_{i=1}^{3} \frac{\partial x_i}{\partial u_\alpha}\frac{\partial x_i}{\partial u_\beta}\,du_\alpha\,du_\beta.$$

In practice, the index notation is normally used when *tensor* methods are being employed. The conventions of tensor notation require the use of two kinds of indices, *upper* and *lower*, and for the present case the convention calls for the notations x^1, x^2, x^3, u^1, u^2. The upper indices are not exponents, but merely distinguishing marks. The coefficients in the formula for ds^2 are now denoted by $g_{11}, g_{12}, g_{21}, g_{22}$; g_{11} being the coefficient of $du^1\,du^1$; g_{12} that of $du^1\,du^2$; and so on. Thus

$$ds^2 = \sum_{\alpha,\beta=1}^{2} g_{\alpha\beta}\,du^\alpha\,du^\beta,$$

where

$$g_{\alpha\beta} = \sum_{i=1}^{3} \frac{\partial x^i}{\partial u^\alpha}\frac{\partial x^i}{\partial u^\beta}.$$

Observe that $g_{12} = g_{21}$. In our previous notation, $g_{11} = E$, $g_{12} = F$, $g_{22} = G$. The four quantities $g_{11}, g_{12}, g_{21}, g_{22}$ are the components of what is called the *fundamental metric tensor* of the surface. In all this discussion we have not defined the term *tensor*. To do so is outside the scope of this book. Suffice it to say that the concept of a tensor is an extension of the concept of a vector.

EXERCISES

1. Describing the θ-curves in the torus of Fig. 129, and, by direct inspection, find the formula for ds along a θ-curve.

2. Describe the θ-curves and ϕ-curves on the sphere $x = a\sin\phi\cos\theta$, $y = a\sin\phi\sin\theta$, $z = a\cos\phi$, and find the formula for ds along each curve, thus obtaining the first fundamental form $ds^2 = a^2(\sin^2\phi\,d\theta^2 + d\phi^2)$.

3. Describe the u-curves and v-curves on the cylinder $x = u$, $y = a\sin v$, $z = a\cos v$, and find the general formula for ds^2.

4. Proceed as in Exercise 3 with the cone $x = u\sin v$, $y = mu$, $z = u\cos v$.

5. Calculate the first fundamental form for each of the following surfaces:

(a) $x = u\cos v$, $y = u\sin v$, $z = ku^2$.

(b) $x = a\sin u\cosh v$, $y = b\cos u\cosh v$, $z = c\sinh v$.

(c) $x = r\cos\theta$, $y = r\sin\theta$, $z = (r^2/2)\sin 2\theta$.

(d) $x = u + v$, $y = u - v$, $z = 4v^2$.

(e) $x = u + v$, $y = u^2 + v^2$, $z = u^3 + v^3$.

6. **(a)** Let $\lambda = (\pi/2) - \phi$ be the latitude of a point on the sphere of Exercise 2. Let a curve on the sphere be given. If ω is the angle at which the curve crosses the circle of λ latitude, show that

$$\frac{d\lambda}{\cos\lambda\,d\theta} = \tan\omega$$

if a suitable convention is made about the sign of ω.

(b) If ω is constant, the curve is called a *rhumb line*, or a *loxodrome*. Find the equation in terms of λ and θ of the loxodrome through $\lambda = 0$, $\theta = 0$ with ω given $(0 < \omega < \pi/2)$. Show that, although the curve spirals infinitely often about both poles of the sphere, its length is finite and equal to $\pi a\csc\omega$.

14.6 / SURFACE AREA

One of the first objectives in the study of surfaces with the aid of calculus is the derivation of formulas for calculating the area of a surface. There are various ways of getting at such formulas. We begin with a method which is frankly heuristic; that is, we formulate an argument which is plausible and which leads to a formula for the area as a double integral. This argument will not do, however, as the last word on the subject, for it is not based on an explicit definition of the area of a surface, but rather on our intuition of what ought to be true of such areas.

Consider a simple surface element, as given by parametric equations (14.4–3), with the parameters (u, v) representing a point varying over the rectangle $R: a \leqq u \leqq b,\ c \leqq v \leqq d$ in the uv-plane. Suppose that this rectangle is divided up into small cells by a rectangular partition after the fashion used in defining a double integral (§13.2). We shall suppose that the functions f, g, h entering in the parametric representation have continuous first partial derivatives in the closed region R. We shall use vector notation in what follows. We think of (14.4–3) as a vector mapping from the rectangle R into $\mathbb{R}^3$:

$$\mathbf{F}(u, v) = (f(u, v), g(u, v), h(u, v)). \tag{14.6–1}$$

By Theorem VII of Chapter 12, we know that $\mathbf{F}$ is continuously differentiable because we have assumed that the coordinate functions f, g, h have continuous first partial derivatives. Consider a typical cell in the interior of R and the corresponding cell of the surface element (Fig. 131*a* and Fig. 131*b*). We first seek to find an approximate expression for the area of this small cell of the surface element. The guiding idea is to think of the cell on the curved surface element as being approximately a plane parallelogram. The plane of this parallelogram is the plane tangent to the surface element at one corner of the cell under consideration. The parallelogram itself is the figure into which the cell in R is mapped by a certain affine transformation which we obtain by using the differential of $\mathbf{F}$. To explain this more clearly we must now introduce some additional notation.

In Fig. 132*a* we designate the lower left corner A of the typical cell in R by (u_i, v_j). Denote the other corners by B, C, D as indicated, so that B is $(u_i + \Delta u, v_j)$ and C is $(u_i, v_j + \Delta v)$. In Fig. 132*b* is shown the cell $A'B'C'D'$ on the surface element,

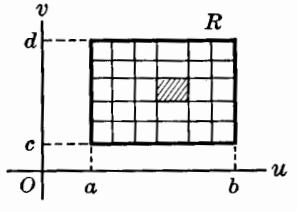


Fig. 131a.

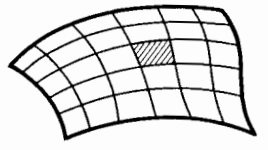


Fig. 131b.

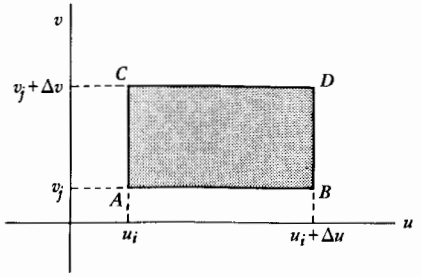


Fig. 132a.

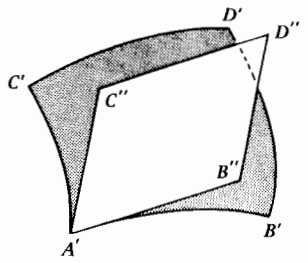


Fig. 132b.

the image of the cell $ABCD$ under the mapping function $\mathbf{F}$. In particular, A' is the vector $\mathbf{F}(u_i, v_j)$.

To locate a point in the cell $ABCD$ of Fig. 132*a*, we write its coordinates as $(u_i + \delta u, v_j + \delta v)$, where $0 \leqq \delta u \leqq \Delta u$, $0 \leqq \delta v \leqq \Delta v$. Now consider the affine mapping from the uv-plane into $\mathbb{R}^3$ that carries $(u_i + \delta u, v_j + \delta v)$ into

$$\mathbf{F}(u_i, v_j) + \mathbf{F}'(u_i, v_j)(\delta u, \delta v). \tag{14.6–2}$$

Here the derivative $\mathbf{F}'(u_i, v_j)$, as we know from Theorem III in Chapter 12, is represented by the 3×2 matrix

$$\begin{pmatrix} \dfrac{\partial f}{\partial u} & \dfrac{\partial f}{\partial v} \\ \dfrac{\partial g}{\partial u} & \dfrac{\partial g}{\partial v} \\ \dfrac{\partial h}{\partial u} & \dfrac{\partial h}{\partial v} \end{pmatrix}$$

in which the partial derivatives are evaluated at (u_i, v_j). This matrix transforms the vector $(\delta u, \delta v)$ into the vector

$$\mathbf{F}'(u_i, v_j)(\delta u, \delta v) = \left(\frac{\partial f}{\partial u}\,\delta u + \frac{\partial f}{\partial v}\,\delta v, \frac{\partial g}{\partial u}\,\delta u + \frac{\partial g}{\partial v}\,\delta v, \frac{\partial h}{\partial u}\,\delta u + \frac{\partial h}{\partial v}\,\delta v\right). \tag{14.6–3}$$

The affine mapping (14.6–2) maps the portion of the uv-plane near A into a portion of the plane tangent at A' to the surface element in $\mathbb{R}^3$, and maps the rectangle $ABCD$ into a parallelogram in the tangent plane, with AB mapping into $A'B''$, AC mapping into $A'C''$, and so on. (Lines are mapped into lines because the mapping is affine.) That the rectangle $ABCD$ maps into a piece of the tangent plane is because of the basic property of the differential $\mathbf{F}'(u_i, v_j)(\delta u, \delta v)$, which is an approximation to the difference $\mathbf{F}(u_i + \delta u, v_j + \delta v) - \mathbf{F}(u_i, v_j)$.

We shall need the following relationships:

$$\mathbf{F}'(u_i, v_j)(\Delta u, 0) = \overrightarrow{A'B''}, \qquad \mathbf{F}'(u_i, v_j)(0, \Delta v) = \overrightarrow{A'C''}.$$

The first of these follows from the fact that, when $\delta u = \Delta u$ and $\delta v = 0$ in

(14.6–2), we obtain the affine mapping image of B, which is B'' while the affine mapping image of A is A'. The vector difference is therefore $\overrightarrow{A'B''}$. The other relation is obtained in a similar way. Therefore we have, from (14.6–3),

$$\overrightarrow{A'B''} = \left(\frac{\partial f}{\partial u}\Delta u, \frac{\partial g}{\partial u}\Delta u, \frac{\partial h}{\partial u}\Delta u\right),$$

$$\overrightarrow{A'C''} = \left(\frac{\partial f}{\partial v}\Delta v, \frac{\partial g}{\partial v}\Delta v, \frac{\partial h}{\partial v}\Delta v\right).$$

Now, the area of the parallelogram $A'B''C''D''$ is equal to the magnitude of the cross product of the vectors $\overrightarrow{A'B''}$ and $\overrightarrow{A'C''}$. This cross product is, [see (10.2–7)],

$$\overrightarrow{A'B''} \times \overrightarrow{A'C''} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial f}{\partial u}\Delta u & \dfrac{\partial g}{\partial u}\Delta u & \dfrac{\partial h}{\partial u}\Delta u \\ \dfrac{\partial f}{\partial v}\Delta v & \dfrac{\partial g}{\partial v}\Delta v & \dfrac{\partial h}{\partial v}\Delta v \end{vmatrix},$$

or, in the notation of (14.4–5)

$$\overrightarrow{A'B''} \times \overrightarrow{A'C''} = j_1\,\Delta u\,\Delta v\mathbf{i} + j_2\,\Delta u\,\Delta v\mathbf{j} + j_3\,\Delta u\,\Delta v\mathbf{k},$$

so that the area of the parallelogram $A'B''C''D''$ is

$$\sqrt{j_1^2 + j_2^2 + j_3^2}\,\Delta u\,\Delta v. \tag{14.6–4}$$

Here the Jacobians j_1, j_2, j_3 are evaluated at (u_i, v_j).

Our plausibility argument now continues by using the expression in (14.6–4) as our approximation to the area of the cell $A'B'C'D'$ on the surface element. The next step in the process is to form the sum of all the terms of the form (14.6–4), corresponding to all of the cells into which R is divided in Fig. 131*a*. When we form this sum and then find its limit as the partition of R is refined, we obtain a double integral as the limit. This double integral is

$$S = \iint_R \sqrt{j_1^2 + j_2^2 + j_3^2}\,du\,dv. \tag{14.6–5}$$

We are led in this manner to define by this double integral the area S of the surface element into which R is mapped by (14.6–1).

This formula can be given another appearance, using the coefficients E, F, G from the quadratic form expressing ds on the surface in terms of du and dv (see (14.5–2)). The alternative expression for the double integral is

$$S = \iint_R \sqrt{EG - F^2}\,du\,dv. \tag{14.6–6}$$

One way of passing from (14.6–5) to (14.6–6) is to make the straightforward

calculation showing that

$$EG - F^2 = j_1^2 + j_2^2 + j_3^2. \tag{14.6–7}$$

In working problems it will sometimes be found to be easier to compute $j_1^2 + j_2^2 + j_3^2$ than $EG - F^2$ and vice versa.

Example 1. Compute the total area of the torus $x = (a + b\cos\phi)\cos\theta$, $y = (a + b\cos\phi)\sin\theta$, $z = b\sin\phi$, $0 < b < a$.

This torus was discussed in §14.5 (see Fig. 129). The part in the first octant is a simple surface element corresponding to $0 \leqq \theta \leqq \pi/2$, $0 \leqq \phi \leqq \pi$, and the area of this part is one eighth of the total. From (14.5–4) we see that, if $u = \theta$ and $v = \phi$,

$$E = (a + b\cos\phi)^2, \qquad F = 0, \quad G = b^2.$$

Thus $\sqrt{EG - F^2} = b(a + b\cos\phi)$, and the total area is

$$S = 8\int_0^{\pi} d\phi \int_0^{\pi/2} b(a + b\cos\phi)\, d\theta = 4\pi^2 ab.$$

This is in accord with the theorem of Pappus.

If we use (14.6–5) instead of (14.6–6) to calculate the first octant portion of the torus, we find that

$$j_1 = b(a + b\cos\phi)\cos\theta\cos\phi,$$

$$j_2 = b(a + b\cos\phi)\sin\theta\cos\phi,$$

$$j_3 = b(a + b\cos\phi)\sin\phi,$$

from which

$$j_1^2 + j_2^2 + j_3^2 = b^2(a + b\cos\phi)^2.$$

Thus, the identity (14.6–7) is verified in this particular case, and the integral for the area is calculated as before.

In the argument leading up to (14.6–5), it will be seen that the fact that the region R was a rectangle in the uv-plane was not essential. If R is any bounded closed region of the uv-plane of the type described in the discussion of double integrals in §13.2, the discussion leading up to (14.6–4) applies to any cell in a rectangular partition of the type shown in Fig. 131*a* and (14.6–4) gives the area of the parallelogram which is the image of this cell under the affine mapping (14.6–2). Hence, the formulas (14.6–5) and (14.6–6) can be used to find the area of any portion of a smooth surface which is obtained by a one-to-one and continuously differentiable mapping from the region R in the uv-plane.

We now consider the special case of a surface defined by an equation $z = f(x, y)$ for all (x, y) belonging to some region R in the xy-plane. It will be assumed that f has continuous first partial derivatives in R. We can think of x and y

as being the parameters u, v. This leads us to the following very special case of (14.4–3).

$$x = x, \qquad y = y, \qquad z = f(x, y). \tag{14.6–8}$$

By simple calculations we see from (14.4–5) that

$$j_1 = -\frac{\partial f}{\partial x}, \qquad j_2 = -\frac{\partial f}{\partial y}, \qquad j_3 = 1.$$

Consequently, the area of the portion of the surface corresponding to the plane region R is

$$S = \iint_R \left[1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2\right]^{1/2} dx\, dy. \tag{14.6–9}$$

Alternatively, if we wish to use (14.6–6), we can calculate as follows:

$$dz = \frac{\partial z}{\partial x}\, dx + \frac{\partial z}{\partial y}\, dy,$$

$$ds^2 = dx^2 + dy^2 + \left(\frac{\partial z}{\partial x}\, dx + \frac{\partial z}{\partial y}\, dy\right)^2$$

$$= \left[1 + \left(\frac{\partial z}{\partial x}\right)^2\right] dx^2 + 2\frac{\partial z}{\partial x}\frac{\partial z}{\partial y}\, dx\, dy + \left[1 + \left(\frac{\partial z}{\partial y}\right)^2\right] dy^2.$$

Interpreting x as u and y as v, we have

$$E = 1 + \left(\frac{\partial z}{\partial x}\right)^2, \qquad F = \frac{\partial z}{\partial x}\frac{\partial z}{\partial y}, \qquad G = 1 + \left(\frac{\partial z}{\partial y}\right)^2.$$

Hence the area of the surface is

$$S = \iint_R \left[1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2\right]^{1/2} dx\, dy. \tag{14.6–10}$$

Example 2. Find the area of the upper half of the sphere $x^2 + y^2 + z^2 = a^2$ by using formula (14.6–10).
Here

$$z = \sqrt{a^2 - x^2 - y^2}, \qquad \frac{\partial z}{\partial x} = \frac{-x}{\sqrt{a^2 - x^2 - y^2}},$$

with a similar formula for $\frac{\partial z}{\partial y}$. Thus the integrand in (14.6–10) becomes

$$\left[1 + \frac{x^2}{a^2 - x^2 - y^2} + \frac{y^2}{a^2 - x^2 - y^2}\right]^{1/2} = \frac{a}{\sqrt{a^2 - x^2 - y^2}}.$$

There is one difficulty. The hemisphere lies above the region R bounded by the circle $x^2 + y^2 = a^2$ in the xy-plane, and we see that the partial derivatives of z

with respect to x and y are not defined on the boundary of this region. Moreover, the integral (14.6–10) in this case is

$$\iint_R \frac{a}{\sqrt{a^2-x^2-y^2}}\,dx\,dy;$$

this is an improper integral, since the integrand becomes infinite at the boundary of R. We can get around the difficulty by considering the area above a smaller concentric circle of radius b, and then making $b \to a$ afterward. It is convenient to evaluate the double integral by an iterated integral in polar co-ordinates. The result is

$$S = \lim_{b\to a} \int_0^b dr \int_0^{2\pi} \frac{ar\,d\theta}{\sqrt{a^2-r^2}} = 2\pi a \int_0^a \frac{r\,dr}{\sqrt{a^2-r^2}},$$

or $S = 2\pi a^2$. This is correct for the hemisphere.

We may put formula (14.6–10) in a different form. Let α, β, γ be the angles which the normal to the surface $z = f(x, y)$ makes with the positive axes, the positive direction of the normal being chosen so that γ is acute. Now

$$\cos\alpha : \cos\beta : \cos\gamma = \frac{\partial z}{\partial x} : \frac{\partial z}{\partial y} : -1.$$

Hence

$$\sec^2\gamma = \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1, \tag{14.6–11}$$

and so from (14.6–10) we have

$$S = \iint_R \sec\gamma\,dA. \tag{14.6–12}$$

Let R be a connected region. Applying the mean-value theorem to formula (14.6–12), we conclude that there is a point P' on the surface such that, if γ' is the value of γ at P' and A is the area of R, then $S = A \sec\gamma'$. If a cylindrical surface parallel to the z-axis is constructed around the boundary of R, and if the plane tangent at P' to the original surface is constructed, then the area of this plane which is cut off within the cylinder is exactly $A \sec\gamma'$ (see Fig. 133). Hence, the area of the original surface $z = f(x, y)$ is the same as the area cut from a certain one of its tangent planes by the cylinder parallel to the z-axis and intersecting the xy-plane in the boundary of the region R. If the region R is very small, it follows that $A \sec\gamma$ is a good approximation to S, no matter at which point P of the surface we evaluate γ. This remark is sometimes taken as the intuitive foundation for a derivation of formula (14.6–2), the procedure being to subdivide R into small parts, the area S

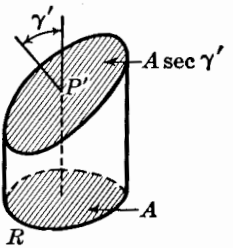


Fig. 133

being obtained as the limit of the sum of areas $\Delta A \sec \gamma$. It is this derivation which is usually found in elementary calculus textbooks.

Formulas (14.6–6) and (14.6–12) are the standard formulas of calculus for dealing with surface area. Where, however, is the *definition* of surface area? Are there surfaces which have area, and yet which are such that the area cannot be found by the integrals mentioned above, perhaps because of lack of sufficient smoothness? It is logically and aesthetically desirable to have a definition of surface area which is directly geometric, and which does not put too many restrictions on the surface. A good definition ought not to depend upon the method of representing the surface analytically, and should not be limited to smooth surfaces. The demand for such a definition poses a very difficult problem, however. It may surprise the student to know that the problem has occupied the attention of many able mathematicians over the last fifty years, and that the end of research on the question is not yet in sight.

To present the concept of surface area to the student at the advanced calculus level, the most satisfactory logical approach seems to be the following: for a smooth simple surface element, with appropriate conditions on its parametric representation, formula (14.6–6) is to be taken as a *definition*; the discussion leading up to the formula is by way of motivation. It can be shown that the area so defined is independent of the particular parametrization, and is therefore an intrinsic characteristic of the surface. This demonstration requires the theory of transformation of double integrals, and is discussed in Chapter 15. In the particular case of surfaces $z = f(x, y)$, the formula (14.6–12) makes it clear that the area does not depend on the parametrization of the surface, but it must still be shown that the orientation of the z-axis is inessential, since the direction of this axis plays a role in the formula.

EXERCISES

1. Find the area of a sphere, using the parametric representation

$$x = a \sin \phi \cos \theta, \qquad y = a \sin \phi \sin \theta, \qquad z = a \cos \phi.$$

2. Find the area of the part of the cylinder $x^2 + z^2 = a^2$ inside the cylinder $y^2 = a(x + a)$.

3. Find the area of the part of the cone $x^2 + y^2 = z^2$ inside the cylinder $x^2 + y^2 = 2ax$.

4. Find the area of the part of the surface $z = xy$ inside the cylinder $x^2 + y^2 = a^2$.

5. Find the area of the surface element $x = au \cos v$, $y = bu \sin v$, $z = \frac{1}{2}(u^2 a \cos^2 v + b \sin^2 v)$, $0 \leqq u \leqq 1$, $0 \leqq v \leqq 2\pi$. Identify the surface and the portion of it whose area is found.

6. A part of the surface $z^2 = 2xy$ can be parametrized by $x = u^2$, $y = v^2$, $z = \sqrt{2}uv$. **(a)** Find the area of the part of the surface above the rectangle $0 \leqq x \leqq a$, $0 \leqq y \leqq b$. **(b)** Find the area of the part of the surface above the region in the xy-plane between the xy-axes and the curve $x^{1/2} + y^{1/2} = 1$. Compare the solutions by (14.6–12) and (14.6–5).

7. Find the area defined by $x = r \cos \theta$, $y = r \sin \theta$, $z = \theta$, $0 \leqq r \leqq 1$, $0 \leqq \theta \leqq 2\pi$. Describe the surface.

8. Find the area of the spheroid $x = a \sin \phi \cos \theta$, $y = a \sin \phi \sin \theta$, $z = c \cos \phi$, distinguishing the cases $a \geqq c > 0$, $c > a > 0$.

9. Find the total area of the part of the sphere $x^2 + y^2 + z^2 = 4a^2$ inside the cylinder $x^2 + y^2 = 2ay$, **(a)** using (14.6–12) and polar co-ordinates, **(b)** using the type of parametric representation occurring in Exercise 1. In **(b)** the main problem is to find the proper region in the $\theta\phi$-plane to correspond to the first octant portion of the required area.

10. Find the total area of the part of the cylinder $x^2 + z^2 = a^2$ inside the cylinder $x^2 + y^2 = ax$.

11. Find the area of the portion of the surface $y^2 + z^2 = 4ax$ in the first octant, between $x = 0$ and $x = 3a$, and inside the cylinder $y^2 = ax$. Solve **(a)** by using (14.6–12), and **(b)** by using the parametric representation $x = r^2(4/a)$, $y = r \cos \theta$, $z = r \sin \theta$. Show that the latter method is equivalent to using the counterpart of (14.6–12) for projection on the yz-plane, and then introducing polar co-ordinates to do the integration.

12. Show that the area on the sphere $x^2 + y^2 + z^2 = c^2$ and inside the paraboloid $(x^2/a) + (y^2/b) = 2(z + c)$ is $4\pi c\sqrt{ab}$, provided that $0 < b \leqq a \leqq c$.

13. Prove the equality in (14.6–7).

14. On the surface $z = f(x, y)$ consider the locus of points where the angle γ is constant. Suppose the projection of this locus on the xy-plane is a closed curve bounding an area $A = \phi(\gamma)$. Show that it is plausible to think that the area of the part of the surface on which $\gamma_0 \leqq \gamma \leqq \gamma_1$ (where $0 \leqq \gamma_0 \leqq \gamma_1 \leqq \pi/2$) is $\int_{\gamma_0}^{\gamma_1} \sec \gamma \phi'(\gamma)\, d\gamma$.

Check by applying to the hemisphere $z = \sqrt{a^2 - x^2 - y^2}$.

15. Find the area of the portion of the cylinder surface $y^2 + z^2 = a^2$ which is in the first octant and inside the cylinder $(x - a)^2 + y^2 = a^2$. SUGGESTION: As one convenient possible parametrization use $u = \theta$, $v = x$, where $y = a \cos \theta$, $z = a \sin \theta$. From symmetry it may be seen that one half of the desired area comes from the part of the cylindrical surface corresponding to the part of the θx-plane defined by $a(1 - \sin \theta) \leq x \leq a$, $0 \leq \theta \leq \pi/2$.

15 / LINE AND SURFACE INTEGRALS

15 / INTRODUCTION

In this chapter we consider some new concepts, the concepts of line integrals and surface integrals. These new kinds of integrals will be defined as limits of sums in the same general way that single and double integrals are defined. An ordinary single integral

$$\int_a^b f(x)\,dx$$

is an integral of a function which is defined along a line segment (an interval of a co-ordinate axis). There is a corresponding kind of integral for a function which is defined along a curve. Such an integral might well be called a *curvilinear* integral; the usual name is *line integral*, where *line* means, in general, a *curved line*. Likewise, the concept of a surface integral is a generalization of the concept of a double integral. A double integral

$$\iint_R f(x, y)\,dx\,dy$$

is an integral of a function which is defined on a region R in the xy-plane. A surface integral is an integral of a function which is defined on a surface. The double integral is a particular case, for the plane region R is a flat surface.

These new kinds of integrals have important applications to geometry and physics. They are also tools of great usefulness in analytical reasoning.

15.1 / POINT FUNCTIONS ON CURVES AND SURFACES

In §10.5 we introduced the concept of a scalar point function, the essential idea of which is that of considering the function values as associated with a point rather than with the co-ordinates of that point. A notation such as $f(P)$ signifies the value of the function f at the point P. Now a point function may be defined throughout some region of space, or merely on a curve or a surface. Thus, for instance, the curvature of a curve is a scalar point function defined only at points on the curve, while a unit vector normal to a smooth surface is a vector point function defined only on the surface.

We shall need the concept of continuity for point functions defined merely on curves or surfaces. Let $f(P)$ be defined on a curve C, and let P_0 be a fixed

point of C. We say that f is continuous at P_0 if

$$\lim_{P \to P_0} f(P) = f(P_0),$$

it being understood, of course, that P can approach P_0 in any manner consistent with its being on C. The arithmetical form of this definition is as follows: f is continuous at P_0 if to each positive number ϵ corresponds some positive number δ (depending in general on P_0 and ϵ) such that $|f(P) - f(P_0)| < \epsilon$ whenever P is on C and at distance less than δ from P_0. The function is called continuous on C if it is continuous at each point of C. A similar definition is made for continuity on a surface.

A point function may, of course, be expressed as a function of co-ordinates when a co-ordinate system is introduced.

15.12 / LINE INTEGRALS

Let C be a curve in xyz-space. It may in particular be a plane curve, as for instance a curve in the xy-plane. There are two directions along a curve; we may arbitrarily choose to call one of these directions the positive direction, and the opposite direction the negative direction. When this choice has been made we say the curve is *oriented*, or given an orientation. If an arc is oriented, it has an initial point A and a terminal point B. An oriented simple closed curve has no initial or terminal point, but it is often convenient to select some point of the curve and regard it as both the initial and terminal point of the curve (see Fig. 134).

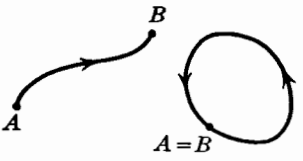


Fig. 134.

In what follows we shall limit our discussion to curves which are formed by joining a finite number of arcs end to end. Such a curve may intersect itself a great deal. If it does not intersect itself at all, it is either a simple closed curve or else it may be regarded as a single arc.

Let C be an oriented curve with initial point A and terminal point B (A and B *may* coincide, as in Fig. 134). Let F denote a scalar point function which is defined and continuous along C. Let points $P_0, P_1, \ldots, P_n$ be chosen in order along C, with $A = P_0$, $P_n = B$. Let Q_k be any point of C between $\mathbf{P}_{k-1}$ and P_k (see Fig. 135). In a given rectangular co-ordinate system let the co-ordinates of P_k be (x_k, y_k, z_k), and let $\Delta x_k = x_k - x_{k-1}$. Form the sum

$$\sum_{k=1}^{n} F(Q_k)\, \Delta x_k; \tag{15.12-1}$$

if the sums of this sort have a limiting value as $n \to \infty$ and the greatest of the chord lengths $P_0P_1, \ldots, P_{n-1}P_n$ approaches zero, we denote the limit by

$$\int_C F(P)\, dx,$$

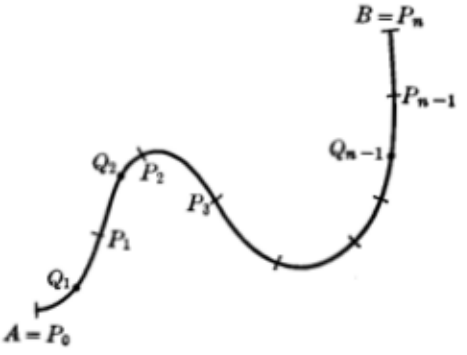


Fig. 135.

and call it the line integral of F with respect to x along C. If P is the point (x, y, z), and if $F(P)$ is denoted by $f(x, y, z)$, an alternative notation for the line integral is

$$\int_C f(x, y, z)\, dx.$$

Line integrals with respect to y or z are defined in the same way, with Δy_k or Δz_k replacing Δx_k in (15.12–1).

To compute the value of a line integral, we use some parametric representation of the curve C. Suppose the parametric equations of C are

$$x = \lambda(t), \quad y = \mu(t), \quad z = \nu(t), \quad a \leqq t \leqq b,$$

and suppose that x, y, and z have continuous derivatives with respect to t. We further suppose that the points A and B correspond to $t = a$ and $t = b$, respectively, and that (x, y, z) traces out C from A to B as t goes from a to b. Let the points P_k on C correspond to points t_k such that $a = t_0 < t_1 < \cdots < t_n = b$; let $\Delta t_k = t_k - t_{k-1}$, and let Q_k correspond to t_k', where $t_{k-1} \leqq t_k' \leqq t_k$. The sum (15.12–1) now takes the form

$$\sum_{k=1}^{n} f(\lambda(t_k'), \mu(t_k'), \nu(t_k'))[\lambda(t_k) - \lambda(t_{k-1})]. \tag{15.12–2}$$

By the law of the mean,

$$\lambda(t_k) - \lambda(t_{k-1}) = \lambda'(\tau_k)\, \Delta t_k,$$

where τ_k is some number between t_{k-1} and t_k. Therefore

$$\int_C f(x, y, z)\, dx = \int_a^b f(\lambda(t), \mu(t), \nu(t))\lambda'(t)\, dt.$$

In drawing this conclusion we use a standard theorem about definite integrals; this theorem appears as (18.21–4), §18.21. It is a special case of Duhamel's

principle. This argument shows that the limit defining the line integral exists provided C has a parametric representation in which x, y, z have continuous derivatives with respect to t.

The result may be written

$$\int_C f(x, y, z)\, dx = \int_a^b f(x, y, z) \left(\frac{dx}{dt}\right) dt, \tag{15.12–3}$$

where the integral on the right is an ordinary definite integral of a function of t, whose integrand is found by expressing x, y, z in terms of t from the parametric equations of C.

Example 1. Find the values of

$$(a)\ \int_C (xy + y^2 - xyz)\, dx \qquad \text{and} \qquad (b)\ \int_C (x^2 - xy)\, dy$$

if C is the arc of the parabola $y = x^2$, $z = 0$ from $(-1, 1, 0)$ to $(2, 4, 0)$, shown in Fig. 136.

Here we use x as the parameter. The integral (a) becomes

$$\int_C (xy + y^2 - xyz)\, dx = \int_{-1}^{2} (x^3 + x^4)\, dx = \tfrac{15}{4} + \tfrac{33}{5} = \tfrac{207}{20}.$$

For (b) we have $dy/dx = 2x$, and so

$$\int_C (x^2 - xy)\, dy = \int_{-1}^{2} (x^2 - x^3)2x\, dx = \tfrac{15}{2} - \tfrac{66}{5} = -\tfrac{57}{10}.$$

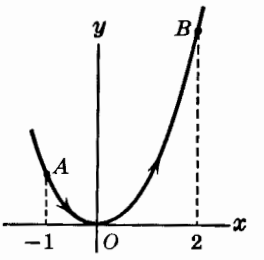


Fig. 136.

It is not essential that the parameter be increasing as we go along the curve in the positive direction.

Example 2. Consider the first quadrant arc C of the circle $x^2 + y^2 = 1$ in the xy-plane, oriented positively in the direction from $(0, 1)$ to $(1, 0)$ (see Fig. 137). With the parametric representation $x = \cos\theta$, $y = \sin\theta$, the initial point of C corresponds to $\theta = \pi/2$ and the terminal point to $\theta = 0$. In evaluating a line integral, the lower limit of integration will be $\theta = \pi/2$, and the upper limit $\theta = 0$. For instance, $dy = \cos\theta$, and so

$$\int_C x^2 y\, dy = \int_{\pi/2}^{0} \cos^2\theta \sin\theta \cos\theta\, d\theta = -\frac{\cos^4\theta}{4}\Bigg|_{\pi/2}^{0}$$
$$= -\tfrac{1}{4}.$$

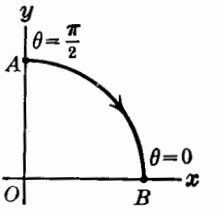


Fig. 137.

The value of a line integral depends on the orientation which is assigned to the curve; if the orientation is reversed, the value of the integral is replaced by the negative of its former value. This is because the limits of integration are reversed in (15.12–3).

The value of a line integral does not depend upon the particular parametric

representation of the curve which is used to calculate the value of the integral.

A sum of line integrals with respect to x, y, and z is often written with just one integral sign. Thus,

$$\int_C f(x, y, z)\,dx + g(x, y, z)\,dy + h(x, y, z)\,dz$$

means

$$\int_C f(x, y, z)\,dx + \int_C g(x, y, z)\,dy + \int_C h(x, y, z)\,dz.$$

Example 3. Compute the value of

$$\int_C xz\,dx + x\,dy - yz\,dz \tag{15.12–4}$$

along the oriented curve shown in Fig. 138, consisting of a quarter circle in the xz-plane, and line segments in the xy-plane and yz-plane, respectively. Denote the three parts of C by C_1, C_2, C_3, respectively. On C_1 we choose x as parameter. Then $z = \sqrt{1-x^2}$, $y = 0$, so $dy = 0$, and

$$\int_{C_1} xz\,dx + x\,dy - yz\,dz = \int_{C_1} xz\,dx + x\cdot 0 - 0\cdot dz$$

$$= \int_0^1 x\sqrt{1-x^2}\,dx = \tfrac{1}{3}.$$

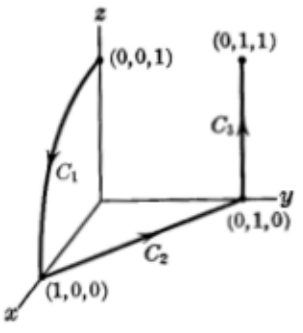


Fig. 138.

For C_2 we use y as parameter; the equations are $x = 1 - y$, $z = 0$; so $dz = 0$, and

$$\int_{C_2} xz\,dx + x\,dy - yz\,dz = \int_{C_2} 0\cdot dx + x\,dy - 0\cdot 0$$

$$= \int_0^1 (1-y)\,dy = \tfrac{1}{2}.$$

Finally, using z as parameter on C_3, we have $x = 0$, $y = 1$, $dx = 0$, $dy = 0$, and the integral over C_3 is just

$$\int_{C_3} -yz\,dz = \int_0^1 -z\,dz = -\tfrac{1}{2}.$$

Thus the line integral (15.12–4) has the value

$$\tfrac{1}{3} + \tfrac{1}{2} - \tfrac{1}{2} = \tfrac{1}{3}.$$

EXERCISES

1. Find the values of the following line integrals. All the curves are in the xy-plane

(a) $\int_C y^2\,dx - x\,dy$, along $y^2 = 4x$ from $(0, 0)$ to $(1, 2)$.

(b) $\int_C -y\,dx + x\,dy$, along $y^2 = 4x$ from $(4, 4)$ to $(0, 0)$.

(c) $\int_C (4x - y)\,dx$, along $y = 8x - 2x^2$ from $(4, 0)$ to $(0, 0)$.

(d) $\int_C (4x - y)\,dy$ along the curve of part (c).

(e) $\int_C x^2\,dy$, along the curve $y = x^3 - 3x^2 + 2x$ from $(0, 0)$ to $(2, 0)$.

(f) $\int_C x\,dy - y\,dx$, along $2y = 3x + 2$ from $(2, 4)$ to $(4, 7)$.

(g) $\int_C (y - x)\,dx + x^2 y\,dy$, along $y^2 = x^3$ from $(1, -1)$ to $(1, 1)$.

(h) $\int_C (x^2 - y^2)\,dx + x\,dy$, along the first quadrant arc of $x^2 + y^2 = 4$, from $(0, 2)$ to $(2, 0)$.

2. Find the values of the following line integrals:

(a) $\int_C x^2y^2\,dx + xy^2\,dy$, counterclockwise around the closed curve formed by parts of the line $x = 1$ and the parabola $y^2 = x$.

(b) $\int_C (x^2 - y^2)\,dx + 2xy\,dy$, counterclockwise around the square formed by the lines $x = 0$, $x = 2$, $y = 0$, $y = 2$.

(c) $\int_C \frac{xy^2}{x^2 + y^2}\,dy$, counterclockwise around the circle $x^2 + y^2 = a^2$.

(d) $\int_C x^2y\,dx$, counterclockwise around the circle $x^2 + y^2 = a^2$.

(e) $\int_C -3y\,dx + 2x\,dy + 4z\,dz$, around the circle $x^2 + y^2 = 1$, $z = 1$, in the counterclockwise sense as viewed from $(0, 0, 2)$.

(f) $\int_C (x^2 - y^2)\,dx + x\,dy$, in the counterclockwise sense around the circle $x^2 + y^2 = 4$.

(g) $\int_C (x\sqrt{3} - y)\,dx + (y\sqrt{3} + x)\,dy$, counterclockwise around the circle $x^2 + y^2 = 1$.

(h) $\int_C -y^2\,dx + x^2\,dy$, counterclockwise around the closed curve formed by the upper half of the ellipse $(x^2/a^2) + (y^2/b^2) = 1$ and the x-axis from $x = -a$ to $x = a$.

3. Find the value of $\int_C \frac{-y}{x^2 + y^2}\,dx + \frac{x}{x^2 + y^2}\,dy$ in each of the following cases:

(a) If C is the counterclockwise arc of $x^2 + y^2 = 2$ from $(1, 1)$ to $(-\sqrt{2}, 0)$.

(b) If C is the line $x = 1$ from $(1, 0)$ to $(1, \sqrt{3})$.

(c) If C is the line $x + y = 1$ from $(0, 1)$ to $(1, 0)$.

4. Calculate $\int_C y\,dx - x\,dy + dz$, where C is the arc of the helix $x = a \sin t$, $y = a \cos t$, $z = t$, from $t = 0$ to $t = \pi/2$.

5. Calculate $\int_C \sqrt{y}\,dx + 2x\,dy + 3y\,dz$, where C is the arc of $x = t$, $y = t^2$, $z = t^3$ from $t = 1$ to $t = 2$.

6. Calculate $\int_C x\left(\frac{1 - y^2}{y^2 + z^2}\right)^{1/2} dx$ from $(0, 0, 1)$ to $(\sqrt{2}/2, \sqrt{2}/2, 0)$ along the first-octant part of the curve of intersection of the plane $x = y$ and the cylinder $2y^2 + z^2 = 1$.

7. Calculate $\int_C (z/y)\,dx + (x^2 + y^2 + z^2)\,dz$ from $(0, 1, 4)$ to $(1, 0, 6)$ along the first-octant part of the intersection of $x^2 + y^2 = 1$ and $z = 2x + 4$.

8. Calculate $\int_C y\,dx - y(x - 1)\,dy + y^2z\,dz$ along the first-octant part of the curve $x^2 + y^2 + z^2 = 4$, $(x - 1)^2 + y^2 = 1$ from $(2, 0, 0)$ to $(0, 0, 2)$.

9. Calculate each of the integrals

(a) $\int_C z^2\,dx$, (b) $\int_C x^2\,dy$, (c) $\int_C y^2\,dz$ from $(2, 0, 0)$ to $(0, 4/\sqrt{3}, 2/\sqrt{3})$ along the first-octant part of the ellipse defined by $x^2 + y^2 - z^2 = 4$, $2z = y$.

10. Consider the integral $\int_C (z + y)\,dx + (x + z)\,dy + (y + x)\,dz$, where C is

(a) The broken line joining $(0, 0, 0)$, $(1, 0, 0)$, $(1, 1, 0)$, and $(1, 1, 1)$ in that order.

(b) The straight line from $(0, 0, 0)$ to $(1, 1, 1)$.

(c) The broken line joining $(0, 0, 0)$, $(0, 0, 1)$, and $(1, 1, 1)$ in that order.

Calculate the line integral in each case and show that the values are all equal.

11. Prove that the line integral in Exercise 10 has the same value for all curves C with initial point at $(0, 0, 0)$ and terminal point at $(1, 1, 1)$. HINT: If the curve is expressed in terms of a parameter t, consider $F'(t)$, where $F(t) = xy + yz + zx$ when x, y, z are expressed in terms of t.

12. Let C be the clockwise closed curve bounded by the lines $x = a$, $x = b$, the x-axis, and a curve $y = f(x)$, $a \leqq x \leqq b$, assuming that $a < b$ and that $f(x)$ is continuous and never negative. Using results from elementary calculus, show (**a**) that $\int_C y\,dx$ is the area enclosed by C; (**b**) that $\int_C \pi y^2\,dx$ is the volume generated when this area is revolved around the x-axis; (**c**) that $\int_C xy\,dx$ is the first moment of this area with respect to the y-axis; (**d**) that $\int_C \frac{1}{2}y^2\,dx$ is the first moment of the area with respect to the x-axis; (**e**) that $\int_C x^2 y\,dx$ is the second moment of the area with respect to the y-axis. It can be shown later, after we have learned more about line integrals, that these same interpretations may be made for the foregoing line integrals if C is any sectionally smooth, simple closed curve in the xy-plane (except that in (**b**) we must require that the curve lie entirely on one side of the x-axis).

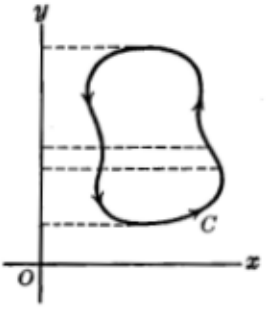


Fig. 139.

13. Using Fig. 139 explain why it appears correct to say that, if C is a simple closed curve oriented counterclockwise, $\int_C x\,dy$ is equal to the area enclosed by C.

14. Using Fig. 139 as a guide, set up a line integral with respect to y, giving the volume of the solid generated when the area enclosed by C is revolved around the x-axis. Assume, as in the figure, that the curve lies entirely above the x-axis.

15.13 / VECTOR FUNCTIONS AND LINE INTEGRALS. WORK

Consider a line integral of the form

$$\int_C P\,dx + Q\,dy + R\,dz, \tag{15.13–1}$$

where P, Q, R are continuous functions defined along a certain oriented curve C. Such integrals often occur in connection with vector point functions, and we shall now indicate how the integral (15.13–1) can be expressed in a different notation by the use of vectors.

Let

$$\mathbf{F}(x, y, z) = P\mathbf{i} + Q\mathbf{j} + R\mathbf{k}$$

be the vector function defined at each point of C in such a way that P, Q, R are its components in the xyz-co-ordinate system. Let s denote arc length along C, with $s = 0$ at the initial point of C and $s = l$ at the terminal point. We assume that C is smooth. Then the unit vector tangent to C in the positive direction at a given point is

$$\mathbf{T} = \frac{dx}{ds}\mathbf{i} + \frac{dy}{ds}\mathbf{j} + \frac{dz}{ds}\mathbf{k}.$$

Therefore, by the formula for dot products,

$$\mathbf{F} \cdot \mathbf{T} = P\frac{dx}{ds} + Q\frac{dy}{ds} + R\frac{dz}{ds}.$$

But, if we use s as the parameter along C, we know by (15.12–3) that

$$\int_C P\,dx = \int_0^l P\left(\frac{dx}{ds}\right) ds,$$

with similar formulas involving Q and R.

Therefore, we see that

$$\int_C P\,dx + Q\,dy + R\,dz = \int_0^l (\mathbf{F} \cdot \mathbf{T})\,ds. \tag{15.13–2}$$

Let ψ be an angle between the vectors $\mathbf{F}$ and $\mathbf{T}$, with the understanding that $0 \leqq \psi \leqq \pi$; see Fig. 140. Since $\mathbf{T}$ has unit length and the length of $\mathbf{F}$ is

$$(P^2 + Q^2 + R^2)^{1/2},$$

we know that

$$\mathbf{F} \cdot \mathbf{T} = (P^2 + Q^2 + R^2)^{1/2} \cos\psi.$$

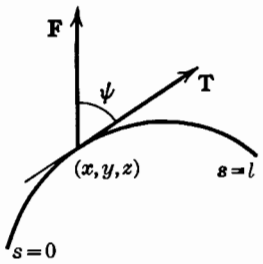


Fig. 140.

Therefore

$$\int_C P\,dx + Q\,dy + R\,dz = \int_0^l (P^2 + Q^2 + R^2)^{1/2} \cos\psi\,ds. \tag{15.13–3}$$

From this formula we can get a useful inequality concerning the value of the line integral. Let M be the maximum length of the vector $\mathbf{F}$. Then

$$|(P^2 + Q^2 + R^2)^{1/2} \cos\psi| \leqq M$$

at all points of C, and so

$$\left|\int_C P\,dx + Q\,dy + R\,dz\right| \leqq Ml. \tag{15.13–4}$$

All these considerations apply to integrals of the form

$$\int_C P\,dx + Q\,dy$$

where P and Q are continuous functions defined along a curve C in the xy-plane. Here we can think of P and Q as components of a vector $\mathbf{F}$ lying in the xy-plane.

The line integral notation is often used for integrals with respect to s, if s is arc length in the positive direction along C. If g is any function of s, a commonly

used notation is to write

$$\int_C g\,ds \qquad \text{instead of} \qquad \int_0^l g\,ds.$$

Example 1. Let C be the semicircle $(x-a)^2+y^2=a^2$, $y \geqq 0$, from $(2a, 0)$ to $(0, 0)$. Let $\mathbf{F}$ be a vector of constant length ($\|\mathbf{F}\| = c$) directed from (x, y) toward $(0, 0)$. Calculate

$$\int_C \mathbf{F}\cdot\mathbf{T}\,ds.$$

In this case $\psi = (\pi/2)-(\phi/2)$ (see Fig. 141) and $\mathbf{F}\cdot\mathbf{T} = c\cos\psi = c\sin(\phi/2)$. Also, $s = a\phi$, so

$$\int_C \mathbf{F}\cdot\mathbf{T}\,ds = \int_0^{\pi} c\sin\frac{\phi}{2}\,a\,d\phi = 2ac.$$

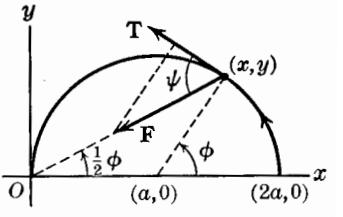


Fig. 141.

One of the important physical applications of line integrals is to the concept of work in analytic mechanics. The most elementary definition of work is that which is given when a constant force acts on a particle while the particle moves in a straight path along the line of action of the force. In this case the work is defined as the product of the magnitude of the force and the distance traversed. In elementary calculus this definition is generalized to cover the situation of a force of variable magnitude, the motion still being in a straight line. The work is then given by a definite integral.

A general definition of work can be made in terms of a line integral. Suppose a particle moves along a curve C, and that while so moving it is acted on by a force $\mathbf{F}$ which may vary both in magnitude and direction. The work done by this force is *defined* to be

$$W = \int_C \mathbf{F}\cdot\mathbf{T}\,ds. \tag{15.13–5}$$

It is seen from Fig. 140 that this definition is in accord with physical intuition. Among other things we see that the component of $\mathbf{F}$ perpendicular to C contributes nothing to the work, which is all done by the tangential component. Also, (15.13–5) agrees with the more elementary definitions already referred to in the appropriate special cases.

If the x, y, and z components of $\mathbf{F}$ are F_1, F_2, F_3, respectively, the work can be expressed as the line integral

$$W = \int_C F_1\,dx + F_2\,dy + F_3\,dz.$$

If time t is introduced as the parameter, we know that the vector velocity of the moving particle is

$$\mathbf{v} = \frac{ds}{dt}\mathbf{T}.$$

Thus

$$\mathbf{T}\,ds = \mathbf{v}\,dt, \quad \mathbf{F}\cdot\mathbf{T}\,ds = \mathbf{F}\cdot\mathbf{v}\,dt,$$

and the formula for work becomes

$$W = \int_{t_0}^{t_1} \mathbf{F}\cdot\mathbf{v}\,dt, \tag{15.13–6}$$

where t_0 and t_1 are the initial and final values of t. This formula is convenient if the path of the particle is defined by giving x, y, z as functions of t. It should be noted, however, that the work does not depend on the particular law of motion, but only on the force and the path which is followed by the particle.

Example 2. A particle moves in the xy-plane according to the law

$$x = 64\sqrt{3}t, \qquad y = 64t - 16t^2,$$

and is acted on by a force $\mathbf{F}$ which is directly proportional to the velocity in magnitude, but opposite in sense to the velocity. Find the work done by $\mathbf{F}$ from $t = 0$ to $t = 4$.

The components of velocity are

$$\frac{dx}{dt} = 64\sqrt{3}, \qquad \frac{dy}{dt} = 64 - 32t.$$

Hence the components of $\mathbf{F}$ are

$$F_1 = -64\sqrt{3}c, \qquad F_2 = -(64 - 32t)c,$$

where c is a positive constant of proportionality. Hence

$$\mathbf{F}\cdot\mathbf{v} = -c\{(64\sqrt{3})^2 + (64 - 32t)^2\},$$

and the work is

$$W = \int_0^4 \mathbf{F}\cdot\mathbf{v}\,dt = -1024c\int_0^4 (16 - 4t + t^2)\,dt,$$

$$W = -\frac{c}{3}(163{,}840).$$

EXERCISES

1. Find the value of the line integral in Example 1 if $\mathbf{F}$, instead of having constant length, always reaches just to the origin.

2. A point moves from $(0, 0)$ to $(2a, 0)$ along the semicircle in Fig. 141. It is acted on by a force of constant magnitude 2, whose direction makes constant angles of 45° with both the positive co-ordinate axes. Find the work done by $\mathbf{F}$.

3. A particle of weight w descends from $(0, 2)$ to $(4, 0)$ along the parabola $8y = (x - 4)^2$. It is acted on by gravity and also by a horizontal force of magnitude equal to the y-coordinate of the point, acting in the positive x-direction. Find the total work done by these two forces.

4. The cycloid $x = a(\theta - \sin\theta)$, $y = a(1 - \cos\theta)$ is generated by a point fixed in the circumference of a rolling circle. Let the point be acted on by a force of unit magnitude directed toward the center of the rolling circle. (**a**) Find the work done by the force as the particle moves from $\theta = 0$ to $\theta = \pi$. (**b**) How much of the work in (**a**) is done by the vertical component of the force?

5. A weight is dragged along the x-axis from $x = 0$ to $x = 7$ (units in feet) by a string which passes over a pulley located at (16, 12) in the xy-plane. If the tension in the string is constantly 10 pounds, find the work done by the pulling force.

15.2 / PARTIAL DERIVATIVES AT THE BOUNDARY OF A REGION

In this section we take up a few matters concerning the meaning and behavior of partial derivatives at the boundary of a region. The discussion is relevant to an exact understanding of some later parts of this chapter. The student may, if he wishes, go directly on to §15.3, and read this section later.

Let us recall the definition of a partial derivative. For the partial derivative of $f(x, y)$ with respect to x at (a, b) we define

$$f_1(a, b) = \lim_{x \to a} \frac{f(x, b) - f(a, b)}{x - a}.$$

Ordinarily we require that the limit be the same when $x \to a$ from the right as when $x \to a$ from the left. This presupposes that $f(x, y)$ is defined along the line $y = b$ for some distance on either side of $x = a$. If, however, $f(x, y)$ is not defined when $y = b$ and $x < a$, we require only that the limit exist as $x \to a$ from the right. The restriction to this kind of one-sided limit is typically necessary in considering the partial derivatives of f at a point on the boundary of the region in which the function is defined.

Example 1. Let R be the region shown in Fig. 142, defined by $x^2 + y^2 \leqq 1$, and let $f(x, y) = (1 - x^2 - y^2)^{3/2}$. This function is not defined if $x^2 + y^2 > 1$. If (a, b) is a point on the boundary of R in the second quadrant, $f_1(a, b)$ must be understood as a limit in which $x \to a$ from the right, and $f_2(a, b)$ must be understood as a limit in which $y \to b$ from below. At most other boundary points the situation is similar. At (0, 1), however, there is an even more unusual situation. Along the line $y = 1$, there is only the single point $x = 0$ which belongs to R. Therefore $f_1(0, 1)$ is not defined, because the quotient

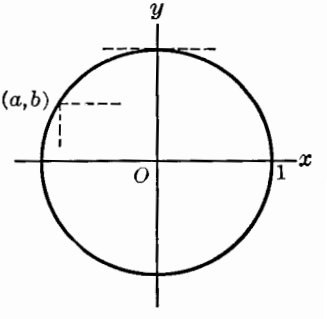


Fig. 142.

$$\frac{f(x, 1) - f(0, 1)}{x - 0}$$

is not defined; *we cannot let* $x \to 0$ *either from the right or the left*! Similar considerations show that $f_1(0, -1)$ is undefined, and that $f_2(\pm 1, 0)$ are undefined.

Now let us look at the formula for $\frac{\partial f}{\partial x}$ when it is defined. The usual rules give

$$\frac{\partial f}{\partial x} = -3x(1 - x^2 - y^2)^{1/2}.$$

Let us define

$$g(x, y) = -3x(1 - x^2 - y^2)^{1/2}$$

at each point of R. Then $f_1(x, y) = g(x, y)$ at all points where $f_1(x, y)$ is defined. But we observe that $g(x, y)$ is even defined at the points $(0, \pm 1)$, where $f_1(x, y)$ is *not* defined. Moreover, g is continuous in R, and $g(0, \pm 1) = 0$. Therefore, even though $f_1(0, 1)$ is not defined, it *is* true that

$$\lim_{(x, y)\to(0, 1)} f_1(x, y) = 0.$$

In situations like this example, it is customary to consider $f_1(0, 1)$ *as being defined* by the limiting value which $f_1(x, y)$ approaches as $(x, y) \to (0, 1)$. This is a conventional agreement which proves to be useful in practice. In the present case, this convention permits us to say that $\partial f/\partial x$ is continuous throughout the entire closed region R.

A general statement of this convention may be made as follows: If R is a closed region, and f is defined in R, we agree to say that $\frac{\partial f}{\partial x}$ is continuous in R if there is some function g which is defined and continuous in R and such that $\frac{\partial f}{\partial x} = g$ at all interior points of R. Similar conventions are made pertaining to $\frac{\partial f}{\partial y}$ and to partial derivatives of higher order. Also, similar conventions are made for functions $f(x, y, z)$ defined in a closed region of three-dimensional space.

The foregoing convention is useful when it comes to considering *normal derivatives* at the boundary of a region. Such considerations are quite important in the physical applications of line and surface integrals (in potential theory, for example).

Suppose a certain part of the boundary of R consists of a smooth arc. Let P_0 be a point of this arc, and let $\mathbf{n}$ be a vector of unit length perpendicular to the boundary at P_0, and pointing outward from R. We call $\mathbf{n}$ the *outer normal* at P_0 (see Fig. 143). Let α be the angle counterclockwise from the positive x-direction to the direction of $\mathbf{n}$. If f is a function defined in R, with first partial derivatives which are continuous in R, the outer normal derivative of f at P_0 is defined to be

$$\frac{\partial f}{\partial n} = \frac{\partial f}{\partial x}\cos\alpha + \frac{\partial f}{\partial y}\sin\alpha, \qquad (15.2\text{–}1)$$

where $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ are evaluated at P_0.

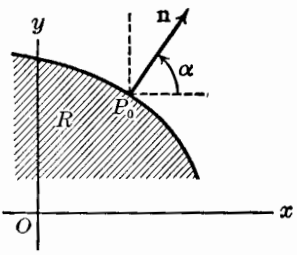


Fig. 143.

Observe that $\frac{\partial f}{\partial n}$ is not, strictly speaking, the rate of change of f at P_0 in the direction of $\mathbf{n}$, because f is not defined outside of R, and therefore such a rate of change is meaningless. It *is* true, however, that $-\frac{\partial f}{\partial n}$ is the rate of change in the *inward* direction at P_0 (i.e., in the direction opposite to that of $\mathbf{n}$).

15.3 / GREEN'S THEOREM IN THE PLANE

The subject of the present section is a very important theorem relating to line integrals around closed curves in the plane. More precisely, the theorem exhibits an exact relation between a line integral taken around the curve (or curves) forming the boundary of a region and a certain double integral taken over the region. With the aid of the theorem we can transform certain double integrals into line integrals, and vice versa. Such transformations are of the highest usefulness in many mathematical arguments, as we shall subsequently see in some instances. Many other instances abound in the literature dealing with the partial differential equations of applied mathematics. A widespread usage sanctions the attachment to the theorem of the name of G. Green, an English mathematician of the early 19th century. There is also justification for calling it Gauss's theorem, after a great German mathematician of the same period.

THEOREM I. *Let R be a closed and bounded region of the xy-plane. Let the boundary of R consist of a finite number of simple closed curves which do not intersect each other, and each of which is sectionally smooth. Let $P(x, y)$ and $Q(x, y)$ be functions which are continuous and have continuous first partial derivatives in R. Let C denote the aggregate of curves forming the boundary of R, each oriented in such a way that the region is on the left as one advances along the curve in the positive direction. Then*

$$\int_C P\,dx + Q\,dy = \iint_R \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dx\,dy. \tag{15.3-1}$$

Definition. *A region having the properties specified in Theorem I will be called a* regular *region.*

There are certain difficulties in proving Green's theorem in the full generality of its statement. However, for regions of sufficiently simple shape the proof is quite easy. We shall begin by giving the proof for such easy cases, and then extending it somewhat. Then we shall proceed to illustrate the theorem in some special cases, and give some applications. We do not give a fully detailed proof of the theorem, but we give indications of such a proof. For further comments on the proof see §15.31.

Before going further, we observe that the functions P and Q are independent of one another, and hence formula (15.3–1) may be broken down into

two separate formulas, namely

$$\int_C P\,dx = -\iint_R \frac{\partial P}{\partial y}\,dx\,dy, \tag{15.3–2}$$

$$\int_C Q\,dy = \iint_R \frac{\partial Q}{\partial x}\,dx\,dy. \tag{15.3–3}$$

The proof of (15.3–1) is equivalent to proofs of (15.3–2) and (15.3–3) separately.

It should be mentioned that we assume the x and y axes have their usual relation to each other, i.e., that a counterclockwise rotation of 90° is needed to carry the positive x-axis into the position occupied by the positive y-axis. This arrangement is responsible for the minus sign in (15.3–2) and the lack of it in (15.3–3).

Let us assume that R has the simple form suggested by Fig. 144. That is, suppose that there is an interval $c \leqq x \leqq d$ such that for an x' outside the interval the line $x = x'$ does not intersect R, while for $c < x' < d$ the line $x = x'$ intersects R in an interval. The lines $x = c$, $x = d$ may intersect R either in an interval or a single point. The boundary of R then consists of a lower curve $y = Y_1(x)$, an upper curve $y = Y_2(x)$, and certain portions (either a segment or a point) of each of the lines $x = c$, d. The positive orientation of the boundary is shown in Fig. 144. Let C_1 and C_2 denote the oriented lower and upper curves, respectively.

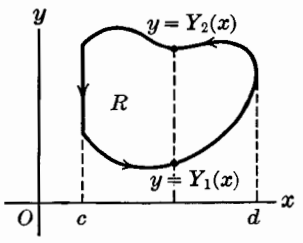


Fig. 144.

Definition. *If a regular region R has the simple form just described, let us call it an* x-simple *region. Likewise we may define a* y-simple *region.*

Now consider formula (15.3–2) for an x-simple region R. There are no contributions to the line integral from the portions of C on the lines $x = c$, $x = d$, since $dx = 0$ along a segment of either line. Hence

$$\int_C P\,dx = \int_{C_1} P\,dx + \int_{C_2} P\,dx.$$

But on C_1 we may take x as the parameter, and put $y = Y_1(x)$. Hence

$$\int_{C_1} P(x, y)\,dx = \int_c^d P(x, Y_1(x))\,dx.$$

Likewise,

$$\int_{C_2} P(x, y)\,dx = \int_d^c P(x, Y_2(x))\,dx = -\int_c^d P(x, Y_2(x))\,dx.$$

(Bear in mind the co-ordinates of the initial and terminal points in determining the limits of integration.) Thus

$$\int_C P\,dx = -\int_c^d \{P(x, Y_2) - P(x, Y_1)\}\,dx. \qquad (15.3\text{–}4)$$

Next consider the double integral, and use the iterated integral formula (13.3–6):

$$\iint_R \frac{\partial P}{\partial y}\,dx\,dy = \int_c^d dx \int_{Y_1}^{Y_2} \frac{\partial P}{\partial y}\,dy. \qquad (15.3\text{–}5)$$

The y integration may now be performed with x held constant. The result, by Theorem VIII, §1.53, is

$$\int_{Y_1}^{Y_2} \frac{\partial P}{\partial y}\,dy = P(x, Y_2) - P(x, Y_1). \qquad (15.3\text{–}6)$$

On combining (15.3–6) with (15.3–5) and comparing with (15.3–4), we see the truth of (15.3–2) for an x-simple region R.

An entirely similar proof may be given for formula (15.3–3) if we assume that R is y-simple. The figure for this case would resemble Fig. 96 (§13.3). Finally, if R is both x-simple and y-simple, we combine (15.3–2) and (15.3–3) to give (15.3–1). Green's theorem is thus easily proved for regions which are both x-simple and y-simple. In particular, a bounded region R is both x-simple and y-simple if its boundary consists of a single sectionally smooth convex curve. A rectangle is such a region.

There are x-simple regions which are not y-simple, and regions which are neither x-simple nor y-simple. On the other hand, many regions may be divided into a finite number of subregions, each of which is both x-simple and y-simple. For such a region it is easy to prove Green's theorem. For instance, suppose R is the region bounded between the circle and the large triangle in Fig. 145, with axes as shown. This region is neither x-simple nor y-simple, but we can divide it into four subregions, each of which is both x-simple and y-simple. The formula of Green's theorem therefore holds for each of the subregions. If we add corresponding parts of the four formulas, the double integrals combine to give the correct double integral over the whole of R. Now consider the line integrals. In dividing R into parts, we introduced four interior connecting lines. Each of these lines occurs twice, but with opposite orientations in the two occurrences, since each line belongs to the boundary of two neighbouring subregions. Hence, when all the line integrals are added, the contributions from these interior lines cancel out in pairs, leaving only the line integral around the total oriented boundary of R, that is, counterclockwise around the triangle and clockwise

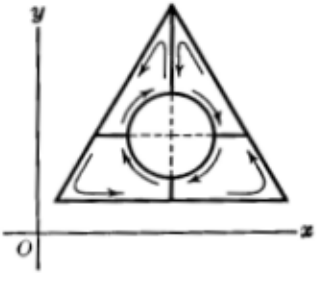


Fig. 145.

around the circle. In this way we obtain the proof of Green's theorem for this region. The idea of the proof may clearly be extended to any region which may be decomposed into a finite number of subregions which are both x-simple and y-simple. There are regions of the sort mentioned in Theorem I which cannot be thus decomposed; Green's theorem for such regions is discussed in §15.31.

Example 1. Show that the area of a regular region R is given by any one of the formulas

$$A = -\int_C y\,dx, \qquad A = \int_C x\,dy, \qquad A = \tfrac{1}{2}\int_C -y\,dx + x\,dy, \tag{15.3–7}$$

where C is the positively oriented boundary of R.

We verify the third formula only, leaving the others to the student. Putting $P = -y$, $Q = x$ in Green's theorem, we get

$$\tfrac{1}{2}\int_C -y\,dx + x\,dy = \tfrac{1}{2}\iint_R (1+1)\,dx\,dy = A.$$

Example 2. Calculate $\int_C (x^2 - y^2)\,dx + 2xy\,dy$, where C is the counterclockwise boundary of the square formed by the lines $x = 0$, $x = 2$, $y = 0$, $y = 2$.

Here we let R be the square region, and put $P = x^2 - y^2$, $Q = 2xy$ in Green's theorem. The line integral is equal to

$$\iint_R \{2y - (-2y)\}\,dx\,dy = 4\iint_R y\,dx\,dy.$$

We could easily calculate the value of this double integral, but it is also possible to recognize its value by the following argument: If A is the area of R, and if $(\bar{x}, \bar{y})$ is the centroid of R, then, as we know,

$$\iint_R y\,dx\,dy = A\bar{y}.$$

In this case $A = 4$ and $\bar{y} = 1$. Hence

$$\int_C (x^2 - y^2)\,dx + 2xy\,dy = 16.$$

Example 3. Green's theorem can be used to prove the following theorem of Pappus: *If R is a regular region lying entirely on one side of the x-axis, and if R is revolved about the x-axis, the volume of the solid so generated is equal to $2\pi A\bar{y}$, where A and $\bar{y}$ have the same significance as in Example 2.*

For brevity, let us consider merely the case of an x-simple region, as pictured in Fig. 144. The reasoning can be extended to the most general regular region. The volume in question is, by elementary calculus,

$$V = \pi\int_c^d \{Y_2(x)\}^2\,dx - \pi\int_c^d \{Y_1(x)\}^2\,dx,$$

in the notation of Fig. 144. This is readily seen to be the same as the line integral

$$V = -\pi \int_C y^2\, dx,$$

with C oriented counterclockwise. We now apply Green's theorem, with $P = -\pi y^2$, $Q = 0$. Thus

$$V = \iint_R 2\pi y\, dx\, dy = 2\pi A\bar{y}.$$

This is what we wanted to prove.

Example 4. Use Green's theorem to deduce the integral formula

$$\iint_R \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) dx\, dy = \int_C \frac{\partial u}{\partial n}\, ds, \tag{15.3–8}$$

where s refers to arc length along C and n refers to the outer normal to C.

Hence it is assumed that R is a regular region and that u and its first and second partial derivatives are continuous in R.

To derive (15.3–8) we start by applying Green's theorem with

$$Q = \frac{\partial u}{\partial x}, \qquad P = -\frac{\partial u}{\partial y}.$$

Then we obtain

$$\iint_R \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) dx\, dy = \int_C -\frac{\partial u}{\partial y}\, dx + \frac{\partial u}{\partial x}\, dy.$$

To complete the derivation we have only to show that on C

$$-\frac{\partial u}{\partial y}\frac{dx}{ds} + \frac{\partial u}{\partial x}\frac{dy}{ds} = \frac{\partial u}{\partial n}. \tag{15.3–9}$$

We know by (15.2–1) that

$$\frac{\partial u}{\partial n} = \frac{\partial u}{\partial x}\cos\alpha + \frac{\partial u}{\partial y}\sin\alpha.$$

We also know that

$$\frac{dx}{ds} = \cos\phi, \qquad \frac{dy}{ds} = \sin\phi, \tag{15.3–10}$$

where ϕ is the angle which the positive tangent to C makes with the positive x-axis. It is easy to show by sketching a figure that as one traverses a curve in such a way that the interior is always on the left, the outward normal lags behind the tangent by $\pi/2$ radians. It follows that $\alpha = \phi - (\pi/2)$ (see Fig. 146) and

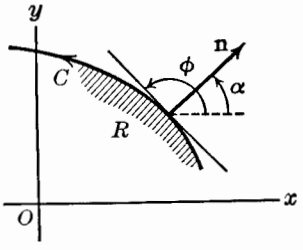


Fig. 146.

therefore

$$\cos\alpha = \frac{dy}{ds}, \qquad \sin\alpha = -\frac{dx}{ds}. \tag{15.3–11}$$

The correctness of (15.3–9) is now apparent, and the proof of (15.3–8) is complete.

EXERCISES

1. Use Green's theorem to evaluate the following line integrals:

(a) $\int_C 2xy\,dx - 3xy\,dy$, clockwise around the square bounded by $x = 3$, $x = 5$, $y = 1$, $y = 3$.

(b) $\int_C xy^2\,dx + 2x^2y\,dy$, counterclockwise around the ellipse $4x^2 + 9y^2 = 36$.

(c) $\int_C (x^2 + 2y^2)\,dy$, counterclockwise around the circle $(x-2)^2 + y^2 = 1$.

(d) $\int_C e^x \sin y\,dx + e^x \cos y\,dy$, around the boundary of any regular region.

(e) $\int_C x^2y\,dx - y^2x\,dy$, counterclockwise around the region bounded by $y = \sqrt{a^2 - x^2}$ and $y = 0$. Use polar co-ordinates to evaluate the double integral.

(f) $\int_C \dfrac{-y\,dx + x\,dy}{x^2 + y^2}$, around the boundary of any regular region not containing the origin.

2. Calculate the line integrals of Exercise 2, §15.12, parts **(a)**, **(d)**, **(f)**, **(g)** and **(h)**, using Green's theorem.

3. Let C be any sectionally smooth simple closed curve in the xy-plane, oriented counterclockwise. Let R be the region bounded by C, and let R have area A and centroid $(\bar{x}, \bar{y})$. Show that

$$\tfrac{1}{2}\int_C x^2\,dy = A\bar{x}, \qquad \int_C xy\,dy = A\bar{y},$$

and, if R is a lamina of constant unit density, interpret

$$\int_C -x^2y\,dx \qquad \text{and} \qquad \int_C -x^2y\,dx + xy^2\,dy$$

as moments of inertia, specifying the axis of rotation in each case.

4. Let R be the region bounded by the rays $\theta = \alpha$, $\theta = \beta$ and the curve $r = f(\theta)$, as shown in Fig. 147. Use the third formula in (15.3–7) to show that the area of R is

$$A = \tfrac{1}{2}\int_\alpha^\beta \{f(\theta)\}^2\,d\theta,$$

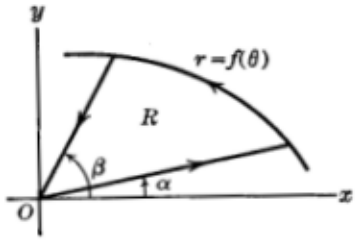


Fig. 147.

by actually calculating the line integral, using r as parameter on the rays, and θ as parameter on the curved side of R. This gives a new derivation of a formula which is familiar from elementary calculus.

5. Derive the formulas

$$\iint_R \left[u\,\Delta v + \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} \right] dx\,dy = \int_C u \frac{\partial v}{\partial n}\,ds,$$

$$\iint_R (u\,\Delta v - v\,\Delta u)\,dx\,dy = \int_C \left(u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n} \right) ds,$$

where the notation

$$\Delta v = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}$$

is employed. Model your work somewhat after that of Example 4.

6. If v satisfies the condition $\Delta v = 0$ (Laplace's equation) in R, show that

$$\int_C \frac{\partial v}{\partial n}\,ds = 0$$

and

$$\iint_R \left[\left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2 \right] dx\,dy = \int_C v \frac{\partial v}{\partial n}\,ds.$$

7. Green's theorem can be expressed in vector notation in two different ways. Consider the plane region R of the xy-plane as being part of three-dimensional space. If $F_1(x, y)$ and $F_2(x, y)$ are functions having continuous derivatives in R, let

$$\mathbf{F} = F_1(x, y)\mathbf{i} + F_2(x, y)\mathbf{j}$$

be a vector field defined in space. It has the special property that the z-component is zero and that the x- and y-components are independent of z. Show that

$$\textbf{(a)} \quad \iint_R (\nabla \cdot \mathbf{F})\,dx\,dy = \int_C \mathbf{F} \cdot \mathbf{n}\,ds$$

and

$$\textbf{(b)} \quad \iint_R (\nabla \times \mathbf{F}) \cdot \mathbf{k}\,dx\,dy = \int_C \mathbf{F} \cdot \mathbf{T}\,ds,$$

where $\nabla \cdot \mathbf{F}$ and $\nabla \times \mathbf{F}$ are as defined in §§10.7 and 10.8, and $\mathbf{n}$ and $\mathbf{T}$ are the unit outer normal and the unit positive tangent, respectively, on C. Formulas (15.3–10) and (15.3–11) will be found useful.

15.31 / COMMENTS ON THE PROOF OF GREEN'S THEOREM

In the statement of Green's theorem we required only that the region R be regular, whereas in the proof given in §15.3 we assumed that R could be decomposed into a finite number of subregions, each both x-simple and y-simple. Now, not every regular region can be so divided, and so the proof of

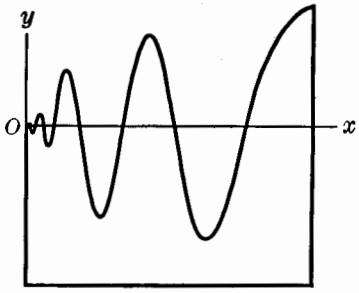


Fig. 148.

Theorem I, §15.3, is not complete. An example of such a region is suggested by Fig. 148; here three sides of the region are formed by straight lines, but the fourth side (the top) is a smooth curve which oscillates more and more rapidly as it approaches the origin, and crosses the x-axis an infinite number of times in the interval shown. The equation $y = x^3 \sin(1/x)$ defines such a curve. Such a region may be x-simple, but is not y-simple, and it cannot be divided into a finite number of y-simple subregions. It is complications of this sort at the boundary of R that make it somewhat difficult to give a complete proof of Theorem I.

One method for completing the proof may be described in outline as follows: Let $R_1, R_2, \ldots, R_n, \ldots$ be a sequence of regions, with oriented boundaries $C_1, C_2, \ldots$, having the following properties:

1. Each R_n is a regular region lying in R.
2. Each R_n can be divided into a finite number of subregions, each of which is both x-simple and y-simple.
3. As $n \to \infty$, R_n approaches R, and C_n approaches C in such a way that

$$\iint_{R_n} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dx\, dy \to \iint_{R} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dx\, dy$$

and

$$\int_{C_n} P\, dx + Q\, dy \to \int_{C} P\, dx + Q\, dy.$$

Green's theorem holds for each R_n, by (1) and (2) and what has already been proved in §15.3. It therefore holds for R, by (3). The crux of the problem now is the construction of the approximating regions R_n so that conditions (1), (2), (3) are satisfied. The details of proving such a construction possible are too intricate for consideration in this book.

There are also other approaches to the problem of a complete proof. One important fact is that any regular region may be divided into a finite number of subregions each of which is *either* x-simple or y-simple.

We conclude this section with the remark that, when a more advanced point of view is taken, and the integral concepts are suitably generalized, Green's

theorem, that is, formula (15.3–1), remains true under hypotheses much less restrictive than those stated in Theorem I of §15.3. The relaxation of restrictions applies not only to the character of the region R, but to the continuity requirements on the functions P, Q and their derivatives.

15.32 / TRANSFORMATION OF DOUBLE INTEGRALS

In this section we consider the effect on the double integral

$$\iint_R F(x, y)\, dx\, dy \tag{15.32–1}$$

when we make a change of variable by equations

$$x = f(u, v), \quad y = g(u, v). \tag{15.32–2}$$

As one important result of our work, we shall discover the significance of the sign of the Jacobian of the mapping (15.32–2) from the uv-plane to the xy-plane.

The discussion is based on the assumption that the student is familiar with Chapter 9, and with §9.2 in particular. We assume that equations (15.32–2) establish a one-to-one mapping of a part of the uv-plane onto a part of the xy-plane. Let the inverse mapping be defined by the equations

$$u = h(x, y), \quad v = k(x, y). \tag{15.32–3}$$

We suppose the functions f, g to be continuous, together with their first and second partial derivatives; we further assume that the Jacobian

$$J(u, v) = \frac{\partial(f, g)}{\partial(u, v)}$$

is always of the same sign—either always positive or always negative, and hence never zero. Let R be a regular region all of whose points are interior to the region of the xy-plane which is being mapped, and let R' be the corresponding region in the uv-plane. Then R' is also a regular region, as one may show with the aid of equations (15.32–3).

THEOREM II. *Let A be the area of R. Then, under the foregoing assumptions,*

$$A = \iint_{R'} |J(u, v)|\, du\, dv. \tag{15.32–4}$$

Proof. We start off with the fact (see Example 1 of §15.3) that

$$A = \int_C x\, dy, \tag{15.32–5}$$

where C is the complete boundary of R, oriented in the usual positive sense, i.e., so that the region R is on the left as one advances along the curve in the positive

direction. Let C' denote the boundary of R'. We orient C' by taking the positive sense along C' to be that which corresponds, under the mapping, to the positive sense along C; thus, as (x, y) moves along C in the positive sense, its image point (u, v) moves along C' in the positive sense. With this agreement we have

$$\int_C x\,dy = \int_{C'} f(u, v)\left[\frac{\partial g}{\partial u}\,du + \frac{\partial g}{\partial v}\,dv\right], \tag{15.32–6}$$

since

$$x = f(u, v) \quad \text{and} \quad dy = \frac{\partial g}{\partial u}\,du + \frac{\partial g}{\partial v}\,dv$$

hold for corresponding points of C and C'.

Next we apply Green's theorem in the uv-plane to the line integral in (15.32–6). Instead of $P\,dx + Q\,dy$ we have

$$f\frac{\partial g}{\partial u}\,du + f\frac{\partial g}{\partial v}\,dv.$$

Therefore, corresponding to

$$\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \quad \text{we have} \quad \frac{\partial}{\partial u}\left(f\frac{\partial g}{\partial v}\right) - \frac{\partial}{\partial v}\left(f\frac{\partial g}{\partial u}\right).$$

On carrying out the indicated differentiations, this latter expression is found to be precisely $J(u, v)$. Consequently,

$$\int_{C'} f\frac{\partial g}{\partial u}\,du + f\frac{\partial g}{\partial v}\,dv = \pm\iint_{R'} J(u, v)\,du\,dv. \tag{15.32–7}$$

The choice of sign on the right is determined by the orientation of C'. If the orientation which we have given to C' coincides with the usual positive orientation of the boundary of R', the plus sign is correct; in the contrary case we must choose the minus sign. Combining (15.32–5), (15.32–6), and (15.32–7), we see that

$$A = \iint_{R'} \pm J(u, v)\,du\,dv.$$

Since A is positive and J is always of the same sign, it follows that the sign chosen in (15.32–7) must be the same as the sign of J. Whichever the sign, formula (15.32–4) is correct.

The last remarks enable us to justify the answer given to the question posed at the end of §9.2. Suppose R is a circular region. The positive orientation of its circumference C is counterclockwise. The image of C will be a simple closed curve C', and R' will consist of the interior of C' and C' itself. Hence, the usual positive orientation of C' will also be counterclockwise. But the mapping of R onto R' induces a certain orientation of C'. *From the discussion in the foregoing paragraph we see that the induced orientation of C' is counterclockwise if and only if the Jacobian of the mapping is positive.*

If the regions R, R' are connected (if one is connected, so is the other) we can apply the law of the mean to (15.32–4), and obtain the following result: Let A' be the area of R'. *There is some point* (u, v) *in* R' *such that*

$$A = |J(u, v)|A'. \tag{15.32–8}$$

The magnitude of the Jacobian is therefore a measure of the distortion of areas by the mapping. If $|J| = 1$ we say that the mapping is *equiareal.*

Let us now turn to the double integral (15.32–1).

THEOREM III. *If* F *is continuous in* R *and if the mapping meets the conditions stated prior to Theorem II, we have*

$$\iint_R F(x, y)\, dx\, dy = \iint_{R'} F[f(u, v), g(u, v)]|J(u, v)|\, du\, dv. \tag{15.32–9}$$

Proof. Let R be divided into a finite number of connected regular subregions $R_1, \ldots, R_n$ of areas $\Delta A_1, \ldots, \Delta A_n$. Let the corresponding regions in the uv-plane be $R'_1, \ldots, R'_n$, of areas $\Delta A'_1, \ldots, \Delta A'_n$. Choose a point (u_k, v_k) in R'_k such that (by (15.32–8))

$$\Delta A_k = |J(u_k, v_k)|\, \Delta A'_k,$$

and let (x_k, y_k) be the corresponding point of R_k. For convenience write

$$G(u, v) = F[f(u, v), g(u, v)].$$

Then

$$\sum_{k=1}^{n} F(x_k, y_k)\, \Delta A_k = \sum_{k=1}^{n} G(u_k, v_k)|J(u_k, v_k)|\, \Delta A'_k.$$

We now pass to the limit as $n \to \infty$ and the maximum dimensions of the subregions approach zero. By Theorem II, §13.23, we obtain formula (15.32–9) as the result.

If we like we may interpret the variables u, v as curvilinear co-ordinates in the xy-plane; this point of view was discussed in §9.5. If we consider a small cell in the mesh of co-ordinate curves, we see from (15.32–8) that the area of the cell with opposite vertices (u, v), $(u + \Delta u,\ v + \Delta v)$ is approximately $|J(u, v)|\, \Delta u\, \Delta v$. By expressing the uv-integral in (15.32–9) as an iterated integral, we obtain a formula for the evaluation of the double integral (15.32–1) by an iterated integral in curvilinear co-ordinates. For polar co-ordinates such a result is already known to us. If

$$x = r \cos \theta, \qquad y = r \sin \theta$$

we readily compute

$$\frac{\partial(x, y)}{\partial(r, \theta)} = r.$$

Thus, for the transformation to polar co-ordinates, the Jacobian is the familiar factor r.

EXERCISES

1. The transformation $x = u^2 - v^2$, $y = 2uv$ maps the rectangle $1 \leqq u \leqq 2$, $1 \leqq v \leqq 3$ of the uv-plane into a certain region R of the xy-plane. Make a sketch of R and find its area.

2. Consider the mapping $x = au + bv$, $y = -bu + av$, (a, b constant). If a certain region of unit area in the xy-plane corresponds to a region R' in the uv-plane, find the area of R'.

3. The equations $u = x^2y - y^3$, $v = 2xy^2$ map points near $(2, 1)$ in the xy-plane into points near $(3, 4)$ in the uv-plane. If R and R' are corresponding small regions containing these respective points, find the approximate ratio of the area of R to that of R'.

4. Calculate $\iint_R x\,dx\,dy$ and the area of R, and so find $\bar{x}$ for R, where R is the region bounded by the lines $x = 3y$, $2x + y = 0$, $x - 3y = 14$, $2x + y = 21$. Use the transformation $x = u + 3v$, $y = -2u + v$.

5. Use the transformation $x = au$, $y = bv$ to map the region R defined by $(x^2/a^2) + (y^2/b^2) \leqq 1$ onto the uv-plane. Evaluate the integral

$$\iint_R \left(\frac{x^2}{a^2} + \frac{y^2}{b^2}\right) dx\,dy$$

with the aid of this transformation and polar co-ordinates in the uv-plane.

6. Calculate $\iint_R \frac{dx\,dy}{x+y}$, where R is the region in the xy-plane bounded by the lines $x + y = 1$, $x + y = 4$, $y = 0$, $x = 0$. Use the transformation $x = u - uv$, $y = uv$.

7. Change the iterated integral $\int_0^1 dx \int_0^{1-x} e^{(y-x)/(y+x)}\,dy$ to a double integral and evaluate it with the aid of the transformation $u = x + y$, $v = x - y$.

8. Use the transformation $x = \sqrt{v - u}$, $y = u + v$ to evaluate the integral $\int_1^{\sqrt{2}} dx \int_{x^2}^{4-x^2} \frac{x\,dy}{x^2+y}$.

9. Let R be the first-quadrant region bounded by the curves $x^2 - y^2 = 1$, $x^2 - y^2 = 4$, $x^2 + y^2 = 9$, $x^2 + y^2 = 16$. Calculate $\iint_R xy\,dx\,dy$ with the aid of the transformation $u = x^2 - y^2$, $v = x^2 + y^2$.

10. Find the area of the first quadrant region bounded by $xy = 4$, $xy = 16$, $y = x + 3$, $y = x - 3$. Use the transformation $2u = x - y$, $v = \sqrt{xy}$.

11. Consider the mapping

$$u = \frac{2x}{x^2+y^2}, \qquad v = \frac{2y}{x^2+y^2}.$$

Draw the circles $u = \frac{1}{3}$, $u = \frac{1}{2}$, $v = \frac{1}{4}$, $v = 1$ in the xy-plane, and let R be the region bounded by them. Calculate the value of $\iint_R \frac{dx\,dy}{(x^2+y^2)^2}$.

12. Find the region in the $r\theta$-plane corresponding to the region R in the xy-plane inside the circle $x^2 + y^2 = 2x$ and to the right of $x = 1$, assuming $x = r\cos\theta$, $y = r\sin\theta$,

$r > 0$, $-\frac{\pi}{4} \leqq \theta \leqq \frac{\pi}{4}$. Transform $\iint_R \frac{x\,dx\,dy}{x^2+y^2}$ to an integral in the $r\theta$-plane, and evaluate it.

13. Consider the families of curves $x^3 = uy$, $y^3 = vx$ in the xy-plane. Show that the first-quadrant area bounded by the curves $u = a^2$, $u = b^2$, $v = \alpha^2$, $v = \beta^2$ is $\frac{1}{2}(a-b)(\alpha-\beta)$, if $a > b > 0$, $\alpha > \beta > 0$.

14. Use the transformation $x = u + uv$, $y = v + uv$ to calculate the integral $\iint_R \frac{dx\,dy}{[(x-y)^2+2(x+y)+1]^{1/2}}$, where R is the triangle with vertices at $(0,0)$, $(2,0)$, $(2,2)$.

15.4 / EXACT DIFFERENTIALS

An expression such as

$$M(x, y)\,dx + N(x, y)\,dy \tag{15.4–1}$$

is called a *first-order differential form in two variables.* As examples we list

$$(y \sin x - 1)\,dx + \cos x\,dy,$$

$$(x^2 + y^2)\,dx - 2xy\,dy.$$

The functions M and N occurring in the form are assumed to be defined in some region of the plane. In practice certain continuity and differentiability restrictions will be imposed on the functions.

The purpose of this section is to study those particular kinds of differential forms which are the differentials of functions of two variables. The following forms are of this kind:

$$x\,dy + y\,dx = d(xy),$$

$$\frac{-y}{x^2+y^2}\,dx + \frac{x}{x^2+y^2}\,dy = d\tan^{-1}\left(\frac{y}{x}\right),$$

$$\frac{x}{x^2+y^2}\,dx + \frac{y}{x^2+y^2}\,dy = d\log(x^2+y^2)^{1/2}.$$

A differential form is called an *exact differential* if it is the differential of some function u at all points of some region in the xy-plane. Since

$$du = \frac{\partial u}{\partial x}\,dx + \frac{\partial u}{\partial y}\,dy,$$

this means that $M\,dx + N\,dy$ is exact if there is some differentiable function u such that

$$M = \frac{\partial u}{\partial x}, \qquad N = \frac{\partial u}{\partial y} \tag{15.4–2}$$

at each point of some region.

There are many nonexact differential forms, as we shall see later.

In order to have enough precision for the statement of our theorems, we make the following formal definition:

Definition. *The differential form $M\,dx + N\,dy$ is said to be exact* at the point (a, b) *if there is some single-valued differentiable function u defined in some neighborhood of (a, b), such that $du = M\,dx + N\,dy$ at all points of the neighborhood.*

Observe that, in defining exactness of a form at a point, we actually place a condition on the form at all points of some neighborhood of the point, not merely at the point itself.

The fundamental problem relating to exact differentials may be stated in two parts:

1. How can one tell by examining M and N whether or not the differential form is exact at a particular point?
2. If the differential form is known to be exact, how can one actually find a function of which the form is the differential?

Both of these questions can be answered with the aid of line integrals, provided we assume that M and N have continuous first partial derivatives.

THEOREM IV. *If M and N have continuous first partial derivatives at all points of some open rectangle, the differential form is exact at each point of the rectangle if and only if the condition*

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x} \tag{15.4–3}$$

is satisfied throughout the rectangle. When this condition is satisfied, a function u such that $du = M\,dx + N\,dy$ is furnished by the line integral

$$u(x, y) = \int_C M(s, t)\,ds + N(s, t)\,dt \tag{15.4–4}$$

along the path from (a, b) to (x, y) shown in Fig. 149, where (a, b) is the center of the rectangle.

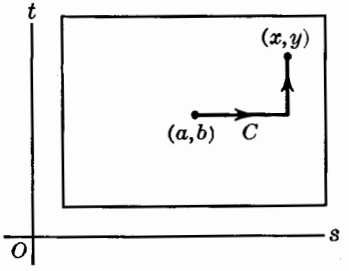


Fig. 149.

Remark. *s and t are used in place of x and y as variables in the line integral, since (x, y) is a fixed point during the computation of the line integral.*

Proof of the Theorem. The necessity of condition (15.4–3) is proved as follows: If the form is exact, there is some function u such that equations (15.4–2) hold. Then

$$\frac{\partial M}{\partial y} = \frac{\partial^2 u}{\partial y\, \partial x} \quad \text{and} \quad \frac{\partial N}{\partial x} = \frac{\partial^2 u}{\partial x\, \partial y}.$$

The second derivatives of u are continuous, since the first derivatives of M and N are continuous, by hypothesis. Therefore

$$\frac{\partial^2 u}{\partial y\, \partial x} = \frac{\partial^2 u}{\partial x\, \partial y},$$

and so the condition (15.4–3) is satisfied.

The sufficiency part of the proof consists in showing that if (15.4–3) holds and $u(x, y)$ is defined by (15.4–4), then $\frac{\partial u}{\partial x} = M$ and $\frac{\partial u}{\partial y} = N$. For this purpose we express the line integral in terms of ordinary integrals. Along the horizontal part of C, $t = b$ and s varies from a to x. Hence $dt = 0$, and the line integral over this part of C is equal to

$$\int_a^x M(s, b)\, ds.$$

On the vertical part of C, $s = x$ and t varies from b to y. Since s is constant during the integration, $ds = 0$, and the line integral over this part of C is equal to

$$\int_b^y N(x, t)\, dt.$$

Thus

$$u(x, y) = \int_a^x M(s, b)\, ds + \int_b^y N(x, t)\, dt. \tag{15.4–5}$$

We may now think of x and y as variables. Doing this, we see that the first integral does not depend on y, while the second one involves y merely as a limit of integration. Therefore, by Theorem VII, §1.52,

$$\frac{\partial u}{\partial y} = N(x, y).$$

All that remains to complete the proof is to show that $\frac{\partial u}{\partial x} = M$. To show this we shall prove that

$$u(x, y) = \int_a^x M(s, y)\, ds + \int_b^y N(a, t)\, dt. \tag{15.4–6}$$

The required result will then follow when we differentiate with respect to x. The

key to this part of the proof is the consideration of the line integral over the alternative path C_2 shown in Fig. 150. In this diagram the former path C is denoted by C_1. The line integral over C_2 is precisely the expression on the right in (15.4–6). If R is the rectangle enclosed by C_1 and C_2, the counterclockwise boundary of R consists of C_1 and the reversal of C_2. Therefore, by Green's theorem,

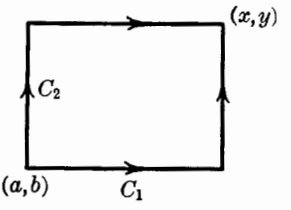


Fig. 150.

$$\int_{C_1} M\,ds + N\,dt - \int_{C_2} M\,ds + N\,dt = \iint_R \left(\frac{\partial N}{\partial s} - \frac{\partial M}{\partial t}\right) ds\,dt.$$

But condition (15.4–3) in our present notation means that $\frac{\partial N}{\partial s} = \frac{\partial M}{\partial t}$. Therefore, the line integrals over C_1 and C_2 are equal, and their common value is $u(x, y)$. Actual calculation of the integral over C_2, using $s = a$, $ds = 0$ on the first part, and $t = y$, $dt = 0$ on the second part, gives us formula (15.4–6). The proof of Theorem IV is then complete.

The fact is that the line integral defining u can be taken over any path from (a, b) to (x, y), so long as the path remains in the originally specified rectangle. This sort of thing is discussed further in §15.41.

Example 1. Show that

$$(y - x^2)\,dx + (x + y^2)\,dy$$

is an exact differential in the whole xy-plane, and find a function of which it is the differential.

Since

$$\frac{\partial}{\partial y}(y - x^2) = \frac{\partial}{\partial x}(x + y^2),$$

we know that we have an exact differential. If (a, b) is an arbitrary point, formula (15.4–5) becomes

$$u(x, y) = \int_a^x (b - s^2)\,ds + \int_b^y (x + t^2)\,dt,$$

or, after a simple calculation,

$$u = -\frac{x^3}{3} + xy + \frac{y^3}{3} + \frac{a^3}{3} - ab - \frac{b^3}{3}.$$

The point (a, b) is arbitrary, so the last three terms may be lumped into a single arbitrary constant C:

$$u = -\frac{x^3}{3} + xy + \frac{y^3}{3} + C.$$

In practice the function u is often found by a formal procedure based on (15.4–5), but not bringing in the point (a, b) explicitly. Let us write

$$u = \phi(x) + F(x, y),$$

where $\phi(x)$ and $F(x, y)$ represent the first and second terms, respectively, in (15.4–5). Since $\dfrac{\partial F}{\partial y} = N(x, y)$, we may think of F as an indefinite integral of N with respect to y, x being held constant. The condition $\dfrac{\partial u}{\partial x} = M$ means that

$$\phi'(x) + \frac{\partial F}{\partial x} = M.$$

Once F has been found we solve for $\phi(x)$ by integrating this last equation. The variable y is not involved in the equation, for

$$\frac{\partial}{\partial y}\left(\frac{\partial F}{\partial x} - M\right) = \frac{\partial^2 F}{\partial x\, \partial y} - \frac{\partial M}{\partial y} = \frac{\partial N}{\partial x} - \frac{\partial M}{\partial y} = 0.$$

Example 2. We illustrate the method on the previous example.

$$N = x + y^2, \quad F(x, y) = xy + \tfrac{1}{3}y^3;$$

$$u = \phi(x) + xy + \tfrac{1}{3}y^3,$$

$$\phi'(x) + y = M = y - x^2, \quad \text{or} \quad \phi'(x) = -x^2.$$

Thus

$$\phi(x) = -\tfrac{1}{3}x^3 + C,$$

$$u = -\tfrac{1}{3}x^3 + xy + \tfrac{1}{3}y^3 + C.$$

Once we have found a function u such that $du = M\,dx + N\,dy$, it is easy to evaluate the line integral of the differential form, as the next theorem shows.

THEOREM V. *Let $u = f(x, y)$ be a single-valued function with the differential $du = M\,dx + N\,dy$ at all points of some region. Let C be a curve lying in this region, and let C have initial point (x_0, y_0) and terminal point (x_1, y_1). We assume M and N continuous, and C sectionally smooth. Then*

$$\int_C M\,dx + N\,dy = f(x_1, y_1) - f(x_0, y_0). \tag{15.4–7}$$

Proof. We shall give the proof for a smooth arc. For a sectionally smooth curve the proof then follows by adding the results for constituent arcs. In terms of a parameter t, going from t_0 to t_1 as (x, y) goes from (x_0, y_0) to (x_1, y_1) along C,

$$\int_C M\,dx + N\,dy = \int_{t_0}^{t_1} \left(M\frac{dx}{dt} + N\frac{dy}{dt}\right) dt = \int_{t_0}^{t_1} \left(\frac{du}{dt}\right) dt$$

$$= u\Big|_{t=t_0}^{t=t_1} = f(x_1, y_1) - f(x_0, y_0).$$

Example 3. Evaluate the integral

$$\int_C \frac{x}{x^2-y^2}\,dx + \frac{y}{y^2-x^2}\,dy,$$

where C is a curve from $(1, 0)$ to $(5, 3)$ and lies between the lines $y = x$, $y = -x$.

We observe by inspection that

$$\frac{x\,dx}{x^2-y^2} + \frac{y\,dy}{y^2-x^2} = du,$$

where $u = \frac{1}{2}\log(x^2 - y^2)$. This function behaves properly as long as $x^2 > y^2$. Hence the integral in question is equal to

$$\tfrac{1}{2}\log(25-9) - \tfrac{1}{2}\log(1-0) = \log 4.$$

15.41 / LINE INTEGRALS INDEPENDENT OF THE PATH

Theorem V in the foregoing section shows that the line integral along C from (x_0, y_0) to (x_1, y_1) has the same value for all curves C starting at (x_0, y_0) and ending at (x_1, y_1), provided that C lies in the region where $u = f(x, y)$ is defined and $du = M\,dx + N\,dy$. Note, however, that we have emphasized the requirement that $f(x, y)$ be single-valued. The reason for insistence on this matter will be more apparent after we have considered the following example.

Example 1. Consider the line integral

$$\int_C \frac{-y}{x^2+y^2}\,dx + \frac{x}{x^2+y^2}\,dy \tag{15.41–1}$$

along various curves from $(1, 0)$ to $(-1, 0)$.

In this case the differential form is exact at each point of the plane except $(0, 0)$, as the student should verify, using (15.4–3). It follows from Theorem IV that in any open rectangle which does not contain the origin there is some function defined whose differential is precisely the differential form appearing in (15.41–1). As a matter of fact, if r, θ are polar co-ordinates, it is easy to verify that

$$d\theta = \frac{-y\,dx + x\,dy}{x^2+y^2}. \tag{15.41–2}$$

We leave it for the student to verify this by calculating dx and dy in terms of dr and $d\theta$ from the equations $x = r\cos\theta$, $y = r\sin\theta$.

If we now attempt to find the value of the line integral (15.41–1) by applying Theorem V, we take $\theta = f(x, y)$. The question then arises: In what region of the plane may we regard θ as a single-valued differentiable function of x and y? Some standard procedure must be adopted so that a unique value of θ is assigned to each point. One possibility is to require $0 \leqq \theta < 2\pi$. Then θ is a single-valued function defined at each point (x, y) except $(0, 0)$. The function is discontinuous (and hence not differentiable) at points on the positive x-axis,

however, for the value of θ experiences a sudden jump as (x, y) crosses the positive x-axis. The value of θ might be standardized in a different way by requiring that $-\pi < \theta \leqq \pi$; this would make θ discontinuous along the negative x-axis. There are infinitely many other ways of standardizing the definition of θ as a single-valued function, but there is no way in which θ can be defined so as to be single-valued and differentiable simultaneously at all points other than $(0, 0)$. Whatever method is used, θ must have at least one point of discontinuity on any closed curve which encircles the origin.

Now consider the line integral (15.41–1). If C goes from $(1, 0)$ to $(-1, 0)$ along a route which does not go through the origin or cross the negative y-axis, we may standardize θ by the requirement $-\pi/2 < \theta \leqq 3\pi/2$. The value of the integral is then found by Theorem V; it is

$$\int_0^{\pi} d\theta = \pi.$$

For a path from $(1, 0)$ to $(-1, 0)$ in the lower half plane, we can define θ so that the discontinuities occur in the upper half plane, say by requiring $-3\pi/2 < \theta \leqq \pi/2$. Then the value of the line integral is

$$\int_0^{-\pi} d\theta = -\pi.$$

Of course, the values of the line integrals for these two paths do not depend on the particular method which is chosen to standardize θ.

The foregoing example shows the following: If $M\,dx + M\,dy$ *is a differential form such that* $\dfrac{\partial M}{\partial y} = \dfrac{\partial N}{\partial x}$ *at all points of a region* D, *it is not necessarily true that* $M\,dx + N\,dy$ *is the differential of a function which is single-valued and differentiable at all points of* D. This raises the question: Is there some kind of a condition which is sufficient to guarantee that if $\dfrac{\partial M}{\partial y} = \dfrac{\partial N}{\partial x}$ in D, then there exists a single-valued function $u = f(x, y)$ such that $du = M\,dx + N\,dy$ in D?

There is such a condition. To explain it we must introduce the concept of a simply connected region.

Definition. *A connected open set* D *is called a* simply connected *region if it has the property that whenever a simple closed curve* C *lies in* D, *all points inside* C *are also in* D. *If* D *is not simply connected it is called* multiply connected.

The property of being simply connected is a property "in the large." The interior of a circle or rectangle is simply connected. The region between two concentric circles is multiply connected. So is the region consisting of the entire plane with the exception of one or any finite number of points.

Using this new concept, we can state the following theorem:

THEOREM VI. *Let D be a simply connected region. Let M and N have continuous first partial derivatives in D, such that*

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x} \tag{15.41–3}$$

at each point. Then there is a function $u = f(x, y)$, single-valued and differentiable in D, such that $du = M\,dx + N\,dy$.

Proof. Let (x_0, y_0) and (x_1, y_1) be any points of D. Consider paths C from (x_0, y_0) to (x_1, y_1), restricted as follows: C consists of a finite number of straight line segments joined end to end, each segment being parallel to either the x or y axis (see Fig. 151). Further, C is to lie in D, and must not intersect itself. Such a curve C will be called an *elementary path*. Since D is connected, it can be shown that any two points of D can be joined by an elementary path in D. We omit the details of this.

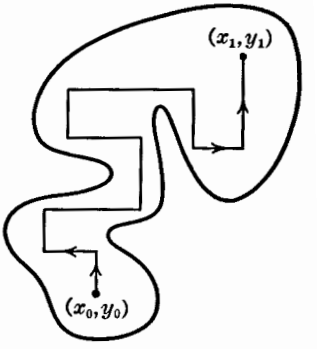


Fig. 151.

Now, the idea of the proof is to show that

$$\int_C M\,dx + N\,dy$$

has the same value for any two elementary paths in D joining (x_0, y_0) to (x_1, y_1). Then, keeping (x_0, y_0) fixed, we define $f(x_1, y_1)$ as the value of the line integral. Since the value depends only on (x_1, y_1), and not on the particular elementary path, we get a single-valued function defined throughout D.

Suppose then that C_1 and C_2 are two such elementary paths in D from (x_0, y_0) to (x_1, y_1). They may coincide along certain line segments, but the noncoincident portions will have just a finite number of intersections, and these portions will form the boundaries of a finite number of regions. Each of these regions will lie entirely in D because of the assumption that D is simply connected. The situation is like that suggested in Fig. 152, where these latter regions are shaded. If R is any one of these regions, the line integral around its complete boundary in the counterclockwise sense is zero, for the line integral is equal to

$$\iint_C \left(\frac{\partial N}{\partial x} - \frac{\partial M}{\partial y}\right) dx\,dy = 0$$

by Green's theorem and the hypothesis (15.41–3). From this it is readily seen that the integrals over C_1 and C_2 are equal.

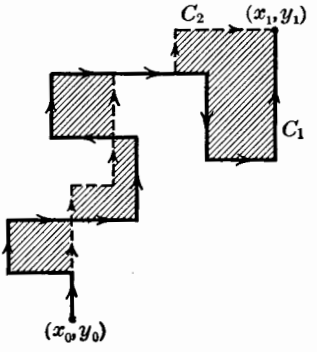


Fig. 152.

We now drop the subscripts on (x_1, y_1), and define $f(x, y)$ as the line integral from (x_0, y_0) to (x, y) over any elementary path in D. This function has partial derivatives

$$\frac{\partial f}{\partial x} = M, \qquad \frac{\partial f}{\partial y} = N. \tag{15.41–4}$$

In fact, if we choose the path so that the last segment of C is horizontal, say from (a, y) to (x, y), then

$$f(x, y) = f(a, y) + \int_a^x M(s, y)\, ds,$$

and so the first equation (15.41–4) must hold. The second equation also holds, by a similar argument in which the last segment of the path is parallel to the y-axis. This completes the proof of Theorem VI.

Make sure that you see where the proof would break down if D were not simply connected.

The restriction to elementary paths is purely for the purposes of the proof of Theorem VI. Once the theorem is proved, we see that the line integral has the same value over any path from (x_0, y_0) to (x_1, y_1). The proof is by Theorem V, where in this case $f(x_0, y_0) = 0$.

A non-simply connected region can often be made simply connected by introducing certain barriers in the form of curves whose points are no longer considered as belonging to the region. Such curves are called *cuts*. For example, let D be the whole plane except for the origin. This is a non-simply connected region. But if we introduce the positive x-axis as a cut, the modified region (call it D_1) is simply connected. No curve enclosing the origin can lie wholly in D_1, and it is the fact that such curves do lie in D that makes D non-simply connected. As another example, consider the region D shown in Fig. 153, lying inside the large curve and outside the two small ones. It is multiply connected, but it may be rendered into a simply connected region by making two cuts, as shown in Fig. 154. The student will readily see that the choice of cuts is not unique. The number of cuts needed in a given case is always the same, however, no matter how they are made.

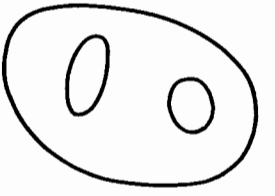


Fig. 153.

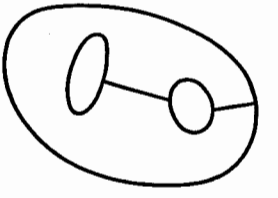


Fig. 154

EXERCISES

1. Test each form for exactness, and if it is exact, find a function of which the form is the differential. Practice the two methods illustrated in Examples 1 and 2, §15.4. The point (a, b) may be taken as $(0, 0)$.

(a) $x\,dy + (y - 7)\,dx$.

(b) $(2y^2 - 3x)\,dx - 4xy\,dy$.

(c) $(2x + 5y - 7)\,dx + (5x - 8y + 3)\,dy$.

(d) $(2xy + x^2)\,dx + x^2\,dy$.

(e) $xe^{xy}\sin y\,dx + (e^{xy}\cos y + y)\,dy$.

(f) $(xy\cos xy + \sin xy)\,dx + x^2\cos xy\,dy$.

(g) $(4x^3 + 10xy^3 - 3y^4)\,dx + (15x^2y^2 - 12xy^3 + 5y^4)\,dy$.

(h) $(e^x\sin y - y)\,dx + (e^x\cos y - x - 2)\,dy$.

2. (a) Find a function u such that

$$du = \frac{1 + y^2}{x^3}\,dx - \frac{y + x^2y}{x^2}\,dy,$$

and describe the region or regions in which u is differentiable.

(b) Find the value of the line integral

$$\int_C \frac{1 + y^2}{x^3}\,dx - \frac{y + x^2y}{x^2}\,dy$$

from $(1, 0)$ to $(5, 2)$; from $(-3, 0)$ to $(-1, 4)$. In each case specify any essential limitations on the path C.

3. (a) Find a function u such that

$$du = \frac{dy}{\sqrt{y^2 - x^2}} - \frac{x\,dx}{y\sqrt{y^2 - x^2} + y^2 - x^2},$$

and describe the region or regions in which u is differentiable.

(b) Find the line integral of the differential form in (a) from $(3, 5)$ to $(5, 13)$, and specify any necessary limitations on the path.

4. Find a function of x alone, $w = \phi(x)$, which makes the differential form $w(x\sin y + y\cos y)\,dx + w(x\cos y - y\sin y)\,dy$ exact; then find the function of which it is the differential, if this function is equal to 0 at $(0, 0)$.

5. Let P_1 be $(1, 0)$, P_2 be $(-1, 0)$, and P be (x, y). Let θ_1 and θ_2 be the angles between the positive x-axis and P_1P and P_2P respectively. Let $u = \theta_1 + \theta_2$. Show that

$$du = -\left(\frac{y}{r_1^2} + \frac{y}{r_2^2}\right)dx + \left(\frac{x - 1}{r_1^2} + \frac{x + 1}{r_2^2}\right)dy,$$

where $r_1 = P_1P$, $r_2 = P_2P$. To make u a single-valued function it is necessary to make some definite agreement about the values of θ_1 and θ_2 at all points except P_1 and P_2.

(a) If it is agreed that $-\pi < \theta_1 \leqq \pi$ and $0 \leqq \theta_2 < 2\pi$, show that u is discontinuous if $y = 0$ and $x^2 > 1$. By making cuts along the lines of discontinuity, we get a simply connected region in which u is differentiable.

(b) If it is agreed that $0 \leqq \theta_1 < 2\pi$ and $0 \leqq \theta_2 < 2\pi$, where are the discontinuities of u?

(c) If $v = \theta_1 - \theta_2$ and the angles are chosen as in (b), where is v discontinuous?

15.5 / FURTHER DISCUSSION OF SURFACE AREA

In §14.6 we arrived at the formula

$$A = \iint_R \sqrt{EG - F^2}\,du\,dv \tag{15.5–1}$$

for the area of a surface represented parametrically, the parameters (u, v) ranging over a region R in the uv-plane. In the earlier discussion of this formula we pointed out that it is logically desirable to show that the number A found by the formula is really not dependent on the particular parametric representation of the surface which is used in calculating A. By using Theorem III, §15.32, it is possible to give an analytic proof that A is independent of the parametrization.

Let another parametrization be given, in terms of parameters (r, t) ranging over a region R' in the rt-plane. Then, there is a one-to-one correspondence between the points of R' and those of R, since both of these regions are mapped in one-to-one fashion onto the same surface. Thus we may consider u and v as functions of r and t. There are now two formulas for ds^2 on the surface:

$$ds^2 = E\,du^2 + 2F\,du\,dv + G\,dv^2, \tag{15.5–2}$$

$$ds^2 = E'\,dr^2 + 2F'\,dr\,dt + G'\,dt^2. \tag{15.5–3}$$

The prime notation here has nothing to do with differentiation; we use it simply to indicate that E', F', G' are related to r and t in the same way that E, F, G are related to u and v.

What we wish to prove is that

$$\iint_R \sqrt{EG - F^2}\,du\,dv = \iint_{R'} \sqrt{E'G' - F'^2}\,dr\,dt. \tag{15.5–4}$$

Now, regarding u and v as functions of r and t, we have

$$du = \frac{\partial u}{\partial r}\,dr + \frac{\partial u}{\partial t}\,dt,$$

and a similar formula for dv. We substitute these expressions for du and dv into (15.5–2), whereupon we get an expression for ds^2 in terms of dr and dt. If this expression is compared with (15.5–3), we see that

$$\begin{aligned} E' &= E\left(\frac{\partial u}{\partial r}\right)^2 + 2F\,\frac{\partial u}{\partial r}\,\frac{\partial v}{\partial r} + G\left(\frac{\partial v}{\partial r}\right)^2 \\ F' &= E\,\frac{\partial u}{\partial r}\,\frac{\partial u}{\partial t} + F\left(\frac{\partial u}{\partial r}\,\frac{\partial v}{\partial t} + \frac{\partial u}{\partial t}\,\frac{\partial v}{\partial r}\right) + G\,\frac{\partial v}{\partial r}\,\frac{\partial v}{\partial t}, \\ G' &= E\left(\frac{\partial u}{\partial t}\right)^2 + 2F\,\frac{\partial u}{\partial t}\,\frac{\partial v}{\partial t} + G\left(\frac{\partial v}{\partial t}\right)^2. \end{aligned} \tag{15.5–5}$$

Now let us consider the expression $E'G' - F'^2$. For convenience, we write

$$u_1 = \frac{\partial u}{\partial r}, \qquad u_2 = \frac{\partial u}{\partial t}, \qquad v_1 = \frac{\partial v}{\partial r}, \qquad v_2 = \frac{\partial v}{\partial t}.$$

Then, after carefully writing out the products involved in $E'G'$ and F'^2 from (15.5–5), we arrive at the final result

$$E'G' - F'^2 = (EG - F^2)(u_1v_2 - u_2v_1)^2,$$

or

$$\sqrt{E'G' - F'^2} = \sqrt{EG - F^2}\left|\frac{\partial(u, v)}{\partial(r, t)}\right|. \tag{15.5–6}$$

Let us now transform the integral (15.5–1) by changing variables from (u, v) to (r, t). By Theorem III, §15.32, the result is

$$\iint_R \sqrt{EG - F^2}\, du\, dv = \iint_{R'} \sqrt{EG - F^2}\left|\frac{\partial(u, v)}{\partial(r, t)}\right| dr\, dt.$$

This formula and (15.5–6) give us (15.5–4), and our proof is thus complete.

15.51 / SURFACE INTEGRALS

A surface integral is a natural generalization of a double integral, and may be used for applications to such things as finding centers of gravity and moments of inertia of curved laminas, the potentials and components of force due to distributions of electrostatic charge on surfaces, and other quantities having physical or geometrical significance.

The direct intuitive formulation of the concept of a surface integral is as follows: Let S be a surface, and let $\phi(P)$ be a point function defined on S. Divide S into a number of surface elements $\Delta S_1, \ldots, \Delta S_n$, of areas $\Delta A_1, \ldots, \Delta A_n$. Let P_k be some point in the element ΔS_k. Then the surface integral of ϕ over S is defined as

$$\iint_S \phi\, dA = \lim \sum_{k=1}^{n} \phi(P_k)\, \Delta A_k, \tag{15.51–1}$$

where the limit is taken as the maximum of the dimensions of the elements ΔS_k approaches zero.

In case the surface S is flat and lies in the xy-plane, this definition takes the same form as the expression of a double integral in Theorem II, §13.23.

If $\phi(P)$ is expressed as a function $F(x, y, z)$, where (x, y, z) are the co-ordinates of P, the surface integral is denoted by

$$\iint_S F(x, y, z)\, dA. \tag{15.51–2}$$

The conception of a surface integral is independent of all co-ordinate systems. It is also independent of the choice of parametric representation for the surface. However, to work with surface integrals, we must learn how to express surface integrals as ordinary double integrals. This is done with the aid of equations for the surface, whether in parametric form or otherwise.

Consider first the case in which S is defined by an equation $z = f(x, y)$,

where (x, y) ranges over a region R in the xy-plane. In this case

$$\iint_S F(x, y, z)\, dA = \iint_R F(x, y, z) \sec \gamma \, dx\, dy, \tag{15.51–3}$$

where, in the integral on the right,

$$z = f(x, y), \sec \gamma = \left[1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2\right]^{1/2};$$

γ is the acute angle between the normal to S at (x, y, z) and the positive z-axis. The derivation of (15.51–3) rests on the fact that if ΔS is a small piece of S which projects into a small rectangular cell of dimensions Δx by Δy in the region R, then the area ΔA of ΔS is given by

$$\Delta A = \sec \gamma \, \Delta x \, \Delta y,$$

where γ is evaluated at some point in ΔS (see the discussion following (14.6–12) in §14.6).

Example 1. Find the moment of inertia of the hemispherical surface $z = \sqrt{a^2 - x^2 - y^2}$ about the x-axis, assuming the surface to be a homogeneous lamina of mass M. See Fig. 155.

The required moment of inertia is

$$I = \iint_S \sigma(y^2 + z^2)\, dA,$$

where σ is the constant density. Thus, by (15.51–3),

$$I = \iint_R \sigma(y^2 + z^2) \sec \gamma \, dx\, dy,$$

where R is the circular region $x^2 + y^2 \leqq a^2$. Now

$$\sec \gamma = \frac{a}{z} = \frac{a}{\sqrt{a^2 - x^2 - y^2}}$$

and $y^2 + z^2 = a^2 - x^2$. Thus

$$I = \sigma a \iint_R \frac{a^2 - x^2}{\sqrt{a^2 - x^2 - y^2}}\, dx\, dy.$$

It is convenient to use polar co-ordinates to evaluate this integral. The iterated integral is

$$I = \sigma a \int_0^{2\pi} d\theta \int_0^a \frac{a^2 - r^2 \cos^2 \theta}{\sqrt{a^2 - r^2}}\, r\, dr.$$

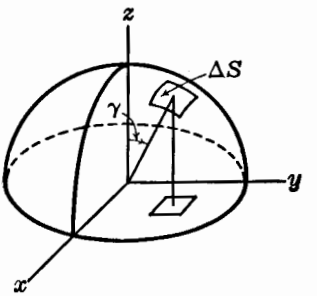


Fig. 155.

We leave most of the calculations to the student.

$$I = \sigma a \int_0^{2\pi} \left(a^3 - \frac{2a^3}{3}\cos^2\theta\right) d\theta = \frac{4\pi\sigma a^4}{3}.$$

Since $M = 2\pi a^2\sigma$, we have $I = \frac{2}{3}Ma^2$.

If the surface is represented parametrically, with parameters (u, v), the appropriate expression for ΔA is

$$\Delta A = \sqrt{EG - F^2}\,\Delta u\,\Delta v$$

(see the discussion leading up to (14.6–6)), and, if $\phi(P)$ is expressed as some function $f(u, v)$, we have

$$\iint_S f(u, v)\,dA = \iint_R f(u, v)\sqrt{EG - F^2}\,du\,dv, \qquad (15.51\text{–}4)$$

where R is the region in the uv-plane which corresponds to the surface S. Here one should also remember that

$$EG - F^2 = j_1^2 + j_2^2 + j_3^2$$

(see (14.6–7)).

Example 2. Find the electrostatic potential at $(0, 0, -a)$ of a uniformly distributed total charge e on the hemisphere in Example 1.

If σ is the constant density of charge, the potential in question is

$$u = \iint_S \frac{\sigma}{[x^2 + y^2 + (z + a)^2]^{1/2}}\,dA.$$

The student may wish to consult §13.51 to refresh his memory on the concept of potential. See (13.51–4), in particular.

We use the parametric representation $x = a\sin\phi\cos\theta$, $y = a\sin\phi\sin\theta$, $z = a\cos\phi$. In this case, with $u = \phi$, $v = \theta$, it is readily found that

$$ds^2 = a^2\,d\phi^2 + a^2\sin^2\phi\,d\theta^2,$$

$$E = a^2,\ F = 0,\ G = a^2\sin^2\phi.$$

Also

$$[x^2 + y^2 + (z + a)^2]^{1/2} = 2a\cos\frac{\phi}{2}.$$

This result may be read directly from Fig. 156, or it may be worked out analytically.

The potential is thus

$$u = \iint_S \frac{\sigma}{2a\cos\frac{\phi}{2}}\,dA = \sigma\int_0^{2\pi} d\theta \int_0^{\pi/2} \frac{a^2\sin\phi}{2a\cos\frac{\phi}{2}}\,d\phi.$$

Since $\sin\phi = 2\sin\frac{\phi}{2}\cos\frac{\phi}{2}$, we have

$$u = 2\pi\sigma a\int_0^{\pi/2}\sin\frac{\phi}{2}\,d\phi = 2\pi\sigma a(2-\sqrt{2}).$$

Since $e = 2\pi a^2\sigma$, the result can be written

$$u = \frac{(2-\sqrt{2})e}{a}.$$

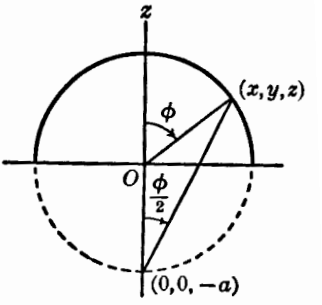


Fig. 156.

EXERCISES

1. Locate the centroid of the hemispherical surface $z = (a^2 - x^2 - y^2)^{1/2}$, using a surface integral.

2. Let S be the surface defined by $x^2 + y^2 + z^2 = 4$, $z \geqq 1$.

(a) Find the value of $\iint_S (x^2+y^2)z\,dA$, using (15.51–3).

(b) Find the value of $\iint_S [x^2+y^2+(z-2)^2]\,dA$, using (15.51–4) and the parameters θ, ϕ as in Example 2.

(c) Find the value of $\iint_S (x^2+y^2)\,dA$.

3. Calculate $\iint_S \frac{dA}{\sqrt{z-y+1}}$ where S is defined by $2z = x^2 + 2y$, $0 \leqq x \leqq 1$, $0 \leqq y \leqq 1$.

4. Compute the value of $\iint_S (x^2+y^2-2z^2)\,dA$ where S is the entire surface $x^2+y^2+z^2 = a^2$.

5. Compute the value of $\iint_S \frac{dA}{\sqrt{2az-z^2}}$ where S is the part of the surface $x^2+y^2+(z-a)^2 = a^2$ which is inside the cylinder $x^2+y^2 = ay$ and underneath the plane $z = a$.

6. Find the value of $\iint_S xyz\,dA$, where S is the part of $x^2+z^2 = 4$ in the first octant and between $y = 0$, $y = 1$.

7. Locate the centroid of the area on the surface of the sphere $x^2+y^2+z^2 = 4a^2$, inside the cylinder $x^2+y^2 = 2ax$, and in the first octant.

8. If the surface in Exercise 7 is a homogeneous lamina, find its moment of inertia about the z-axis.

9. (a) If the cone $az = b\sqrt{x^2+y^2}$ is parametrized by setting $x = r\cos\theta$, $y = r\sin\theta$, $z = br/a$, show that $EG - F^2 = (a^2+b^2)r^2/a^2$.

(b) If S is the part of the cone for which $0 \leqq z \leqq b$, and if S is a homogeneous lamina, show that its moment of inertia about the z-axis is $\frac{1}{2}Ma^2$, using a surface integral and (15.51–4). How does this compare with using (15.51–3) and polar co-ordinates in the xy-plane?

10. On the cylinder $x^2+y^2 = 1$, $0 \leqq z \leqq 2$ use the θ and z of cylindrical co-ordinates as parameters, and calculate the value of the surface integral giving the electrostatic field strength produced at $(0, 0, 0)$ by a uniform density of charge on the cylinder.

11. Let S be a part of the surface of the sphere $x^2+y^2+z^2=a^2$, and let it carry a uniformly distributed total mass M. Show that the gravitational field produced at the origin by this mass is $\mathbf{F}=(M/a^3)\mathbf{R}$, where $\mathbf{R}$ is the vector from 0 to the center of mass of S.

12. Suppose that S in Exercise 11 lies entirely on the hemisphere $z \geqq 0$. Show that the z-component of the field at 0 is $\sigma A/a^2$, where σ is the density and A is the area of the projection of S on the xy-plane.

13. If the torus of Fig. 129 (§14.5) carries a homogeneously distributed total mass M, show that its moment of inertia about the z-axis is $\frac{1}{2}M(2a^2+3b^2)$.

14. Let S be the sphere $x^2+y^2+z^2=a^2$, and let it carry a uniform distribution of charge. Show that the field produced at $(0,0,c)$ is of magnitude 0 if $|c|<a$, e/c^2 if $c>a$, and $\frac{1}{2}e/c^2$ if $c=a$. The substitution $t^2=a^2+c^2-2ac\cos\phi$ will be found useful in this exercise, as well as in Exercises 15, 16.

15. Solve the problem corresponding to Exercise 14 if the density of charge is $\sigma=\cos\phi$, where ϕ is the angle which OP makes with the positive z-axis.

16. Find the potential u at $(0,0,c)$ of a uniform distribution of charge on the hemisphere $z=(a^2-x^2-y^2)^{1/2}$. Then calculate $-\dfrac{\partial u}{\partial c}$ to get the force.

15.6 / THE DIVERGENCE THEOREM

The divergence theorem is a three-dimensional analogue of Green's theorem in the plane (§15.3). The main content of the theorem is the formula

$$\iiint_T \left(\frac{\partial P}{\partial x}+\frac{\partial Q}{\partial y}+\frac{\partial R}{\partial z}\right) dV = \iint_S (P\cos\alpha + Q\cos\beta + R\cos\gamma)\, dA. \qquad (15.6\text{–}1)$$

Here T is a region in three-dimensional space; S is the surface bounding T; P, Q, R are functions of x, y, z which are continuous and have continuous first partial derivatives in T; and $\cos\alpha$, $\cos\beta$, $\cos\gamma$ are the direction cosines of the line normal to S, directed *outward* from T.

The hypotheses covering T and S need to be made more precise. In the usual applications of formula (15.6–1) we may wish to take T to be one of the common solids, e.g., a cube, a sphere, or a right circular cylinder. Or, T might be the region contained between two concentric spheres, or the region which is left when a doughnut-shaped region is removed from the interior of a large ellipsoid. Thus, the surface S of T may actually consist of several detached pieces; in the last-mentioned example, S consists of the surface of the ellipsoid and the surface of the doughnut (a torus).

We saw in §15.31 that there are not inconsiderable difficulties in the way of proving Green's theorem in the generality with which it is stated in §15.3. There are difficulties of a corresponding kind, but even greater, in connection with the divergence theorem. It is not even an easy or brief matter to formulate reasonably general conditions on the region T such that (15.6–1) may be shown to hold true. However, if T is a sufficiently simple type of region, it is quite easy to prove formula (15.6–1). To describe the kind of region we have in mind let us

begin with a description of the boundary. Let G be a regular region in the xy-plane. Let $Z_1(x, y)$ and $Z_2(x, y)$ be continuous functions defined in G and having continuous first partial derivatives there, and such that $Z_1 < Z_2$ at each interior point of G. Let S_1 and S_2 be the surfaces defined by $z_k = Z_k(x, y)$, $k = 1, 2$. Consider the cylindrical surface formed by erecting lines parallel to the z-axis at points of the boundary of G. Let S_3 be the portion of this cylindrical surface which is cut off between the surfaces S_1, S_2. It may happen that $Z_1 = Z_2$ on the boundary of G; in this event there is no surface S_3. Now let T be the region bounded above by S_2, below by S_1, and laterally by S_3. (See Fig. 157, page 486.) A region T formed in this manner will be called *xy-simple.* In a similar manner we may define what we mean by a yz-simple region or a zx-simple region. The region defined by $1 \leqq x^2 + y^2 \leqq 4$, $0 \leqq z \leqq 1$ is xy-simple, but not yz-simple or zx-simple. The region defined by $x^2 + z^2 \leqq y$, $0 \leqq y \leqq 4$ is zx-simple.

We now prove the following lemma:

LEMMA. *Let T be an xy-simple region, and let S be the entire surface of T. Suppose that $F(x, y, z)$ and $\dfrac{\partial F}{\partial z}$ are continuous in T. Let γ be the angle between the positive z-axis and the outward drawn normal to S. Then*

$$\iiint_T \frac{\partial F}{\partial z}\, dV = \iint_S F \cos\gamma\, dA. \tag{15.6–2}$$

Proof. We begin by expressing the triple integral in (15.6–2) as an iterated integral in which the first integration is with respect to z. We have

$$\iiint_T \frac{\partial F}{\partial z}\, dV = \iint_G dx\, dy \int_{Z_1}^{Z_2} \frac{\partial F}{\partial z}\, dz.$$

In the z-integration, x and y are held constant. Hence

$$\int_{Z_1}^{Z_2} \frac{\partial F}{\partial z}\, dz = F(x, y, Z_2) - F(x, y, Z_1),$$

and the triple integral becomes

$$\iiint_T \frac{\partial F}{\partial z}\, dV = \iint_G [F(x, y, Z_2) - F(x, y, Z_1)\, dx\, dy. \tag{15.6–3}$$

Now consider the surface integral in (15.6–2). On the lateral surface S_3 of T we see that $\cos\gamma = 0$, for $\gamma = \pi/2$. Hence

$$\iint_S F \cos\gamma\, dA = \iint_{S_1} F \cos\gamma\, dA + \iint_{S_2} F \cos\gamma\, dA.$$

At a point on S_1 the outer normal extends downward, so that γ is obtuse and

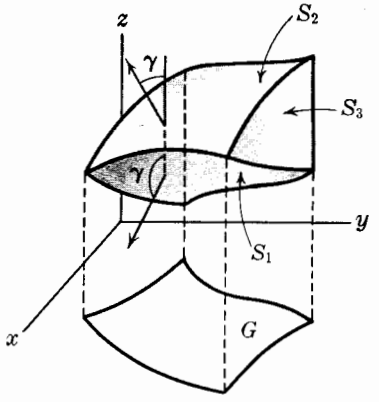


Fig. 157.

$\cos\gamma < 0$. On S_2, however, γ is acute (see Fig. 157). The surface integrals over S_1 and S_2 can be transformed into double integrals over G by formula (15.51–3). In the latter formula γ represents the *acute* angle between the positive z-axis and the *undirected* normal to the surface. Hence since $\sec(\pi - \gamma) = -\sec\gamma$, we have

$$\iint_{S_1} F(x, y, z)\cos\gamma\, dA = \iint_G F(x, y, Z_1)\cos\gamma\sec(\pi - \gamma)\, dx\, dy$$

$$= -\iint_G F(x, y, Z_1)\, dx\, dy.$$

Likewise

$$\iint_{S_2} F(x, y, z)\cos\gamma\, dA = \iint_G F(x, y, Z_2)\cos\gamma\sec\gamma\, dx\, dy$$

$$= \iint_G F(x, y, Z_2)\, dx\, dy.$$

Thus we have shown that

$$\iint_S F\cos\gamma\, dA = \iint_G [F(x, y, Z_2) - F(x, y, Z_1)]\, dx\, dy.$$

On comparing this result with (15.6–3) we see that the lemma is proved.

Clearly, if T is yz-simple, we have a result analogous to (15.6–2) with $\dfrac{\partial F}{\partial x}$

instead of $\frac{\partial F}{\partial z}$, and $\cos\alpha$ instead of $\cos\gamma$; a corresponding result also holds for zx-simple regions. No additional proofs are needed, since the results differ from (15.6–2) in notation only, and the labeling of the axes is purely a matter of notation.

Now suppose that T is a region which is at once xy-simple, yz-simple, and zx-simple, and let S be its surface. Suppose that P, Q, R are functions which are continuous and have continuous first partial derivatives in T. At a point where S is smooth let $\mathbf{n}$ be a unit vector normal to S and extend outward from T, and let $\mathbf{n}$ make angles α, β, γ respectively with the positive x-, y-, and z-axis. Then by the lemma

$$\iiint_T \frac{\partial P}{\partial x}\,dV = \iint_S P\cos\alpha\,dA,$$

$$\iiint_T \frac{\partial Q}{\partial y}\,dV = \iint_S Q\cos\beta\,dA,$$

$$\iiint_T \frac{\partial R}{\partial z}\,dV = \iint_S R\cos\gamma\,dA.$$

Adding, we obtain the divergence theorem (15.6–1) for a region T of this restricted type.

Next we proceed to remove some of the restrictions on T. Let us call a region xyz-simple if it is at once xy-simple, yz-simple, and zx-simple. The region between two concentric spheres (say with centers at O) is not xyz-simple. But the three co-ordinate planes divide this region into eight parts, each of which is xyz-simple. In this subdivision process, certain additional surfaces are introduced as "interior partitions" in T. Each surface element of such an interior partition is on the boundary of two xyz-simple subregions. Let us say that a region T is *xyz-standard* if, by the introduction of a finite number of simple surface elements as interior partitions, we can divide T into a finite number of xyz simple subregions.

THEOREM VII. *Under the stated assumptions on P, Q, R, formula* (15.6–1) *holds when T is an xyz-standard region.*

Proof. Let $T_1, \ldots, T_n$ be the xyz-simple regions composing T, and let S_k be the entire surface of T_k The surface S_k may consist partly of pieces of S and partly of interior partitions. By what we have already proved,

$$\iiint_{T_k} \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z}\right) dV = \iint_{S_k} (P\cos\alpha + Q\cos\beta + R\cos\gamma)\,dA.$$

Here we have n formulas. We add them; the sum of the triple integrals is just the triple integral over over T. Let us see what we get when we add the surface integrals. Suppose that T_j and T_k are adjacent regions, and let S_{jk} denote a surface element which is an interior partition between T_j and T_k. If $\mathbf{n}_j$ is the unit vector normal to S_{jk} outward from T_j, and $\mathbf{n}_k$ is the unit vector normal to S_{jk} outward from T_k, then $\mathbf{n}_j = -\mathbf{n}_k$, and so the direction cosines of $\mathbf{n}_j$ are the negatives of those of $\mathbf{n}_k$. The result is that, when the surface integrals over S_j and S_k are added, the integrals over S_{jk} cancel each other out. Thus, when all the surface integrals are added, the contributions from all the interior partitions cancel in pairs, and we are left with just the integral over the original boundary surface S. This shows that formula (15.6–1) holds true.

The theorem can be written in vector form. Indeed, it is from the vector form that the theorem derives its name.

THEOREM VIII. (THE DIVERGENCE THEOREM.) *Let* $\mathbf{F}$ *be a vector point function which is defined and continuously differentiable in a bounded closed region* T. *Suppose that, for some choice of rectangular co-ordinate system,* T *is an* xyz*-standard region. Let* S *be the surface of* T, *and* $\mathbf{n}$ *the unit outer normal vector to* S. *Then*

$$\iiint_T \operatorname{div} \mathbf{F}\, dV = \iint_S \mathbf{F} \cdot \mathbf{n}\, dA. \tag{15.6–4}$$

This form of the theorem requires a few additional words of proof. Let P, Q, and R be the x, y, and z components, respectively, of $\mathbf{F}$. Then

$$\operatorname{div} \mathbf{F} = \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z}, \tag{15.6–5}$$

$$\mathbf{F} \cdot \mathbf{n} = P \cos \alpha + Q \cos \beta + R \cos \gamma, \tag{15.6–4}$$

so that (15.6–4) and (15.6–1) are equivalent. Now div $\mathbf{F}$ and $\mathbf{F} \cdot \mathbf{n}$ are independent of the way in which the rectangular co-ordinate system is chosen. Hence, if (15.6–4) is true for one choice of the co-ordinate system, it is true for all choices. This shows us that T need not be an xyz-standard region for *all* orientations of the axes; if it is an xyz-standard region for *some* orientation, that is sufficient.

The restriction to xyz-standard regions is not absolutely necessary; it is imposed in order to make the proof simple. Considerations similar to those mentioned in §15.31 will enable one to prove Theorem VIII for certain regions which are never xyz-standard for any choice of axes. We shall not discuss these generalizations, however.

Example 1. Let S be the surface of a region T for which the divergence theorem is applicable. Let O be any fixed point in space, and let $P(x, y, z)$ be a variable point of S. Show that the volume of T is given by

$$V = \tfrac{1}{3} \iint_S r \cos \psi\, dA, \tag{15.6–7}$$

where r is the distance OP and ψ is the angle between the directed line OP and the outer normal to S at P.

To show this, take $P = x$, $Q = y$, $R = z$ in (15.6–5). Then from (15.6–4) we have

$$\iiint_T 3\,dV = \iint_S r\cos\psi\,dA,$$

which gives (15.6–7) at once.

The divergence theorem is useful in connection with the consideration of solid angles. For an exposition of this matter we begin with a definition. Let S be a surface element, and O a fixed point not on S. Assume that S is not intersected more than once by any ray from O, and that no such ray is tangent to S. As a point P varies over S, consider the point Q in which the ray OP (extended, if necessary) intersects the unit sphere with center at O. These points Q fill out a certain portion of the surface of the unit sphere. The area ω of this portion is defined to be *the solid angle subtended by S at O.*

Example 2. Show that the solid angle is given by

$$\omega = \iint_S \frac{\cos\psi}{r^2}\,dA = \iint_S \frac{\mathbf{n}\cdot\overrightarrow{OP}}{r^3}\,dA, \tag{15.6–8}$$

where r is the distance OP and $\mathbf{n}$ is the unit normal to S at P in such a direction that the angle ψ between $\mathbf{n}$ and $\overrightarrow{OP}$ is acute.

To derive (15.6–8) we use the divergence theorem as follows: Construct the sphere with radius a and center O. Choose a so small that S lies entirely outside this sphere. Draw all the rays joining O to S, and let Σ be the portion of the sphere pierced by these rays. Let T be the solid region formed by the bundle of rays cut off between Σ and S (see Fig. 158). For $\mathbf{F}$ we take the vector function $\frac{\overrightarrow{OP}}{r^3}$ with components $\frac{x}{r^3}, \frac{y}{r^3}, \frac{z}{r^3}$. We apply the divergence theorem. Since $r^2 = x^2 + y^2 + z^2$, it is readily calculated that

$$\frac{\partial}{\partial x}\left(\frac{x}{r^3}\right) = \frac{r^2 - 3x^2}{r^5} \quad \text{and} \quad \operatorname{div}\mathbf{F} = 0.$$

thus

$$\iiint_T (\operatorname{div}\mathbf{F})\,dV = 0.$$

Therefore, by (15.6–4), the integral of $\mathbf{F}\cdot\mathbf{n}$ over the entire surface of T is equal to zero. Here $\mathbf{n}$ refers to the outer normal. The surface of T consists of Σ, S, and the lateral surface formed by the rays joining O to the edge of S.

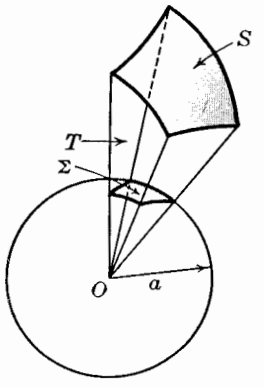


Fig. 158.

Since $\mathbf{F}$ has the same direction as $\overrightarrow{OP}$, $\mathbf{F} \cdot \mathbf{n} = 0$ on this lateral portion, so that the surface integral over this portion is zero. Thus

$$\iint_{\Sigma} \mathbf{F} \cdot \mathbf{n}\, dA + \iint_{S} \mathbf{F} \cdot \mathbf{n}\, dA = 0. \tag{15.6–9}$$

The integral over S in this equation is the same as the surface integral in (15.6–8). Thus we have to show that

$$\iint_{\Sigma} \mathbf{F} \cdot \mathbf{n}\, dA = -\omega. \tag{15.6–10}$$

This is easy, however. On Σ the outer normal points away from T, which is toward O. But $\mathbf{F}$ points away from O, and has magnitude $1/r^2$. Thus, since $r = a$ on Σ, $\mathbf{F} \cdot \mathbf{n} = -1/a^2$, and the integral in (15.6–10) is

$$-\frac{1}{a^2} \iint_{\Sigma} dA = -\frac{1}{a^2} (\text{area of } \Sigma).$$

If $a = 1$, the area of Σ is ω, by the definition of the solid angle. But, in any case,

$$\frac{1}{a^2} (\text{area of } \Sigma) = \omega,$$

since the area of Σ is proportional to a^2 when we vary a. This completes the proof of (15.6–10) and (15.6–8)

EXERCISES

1. Use (15.6–7) to show that the volume of any cone or pyramid is $\frac{1}{3}Bh$, where h is the altitude and B is the area of the base. Take O at the vertex.

2. If S is the hemispherical surface $\bar{z} = (a^2 - x^2 - y^2)^{1/2}$, show that the centroid of the solid hemisphere is given by

$$\bar{z} = \frac{3}{4\pi a^3} \iint_{S} z^2 \cos \gamma\, dA,$$

where γ is the acute angle between the normal and the positive z-axis.

3. Use the divergence theorem to show that

$$\iint_{S} (x^2 \cos \alpha + y^2 \cos \beta + z^2 \cos \gamma)\, dA = \frac{8\pi a^4}{3},$$

if S is the surface $x^2 + y^2 + z^2 = 2az$ and α, β, γ are direction angles of the outer normal.

4. What is the value of the integral in Exercise 3 if S is the surface of the cube bounded by $x = 0$, $x = a$, $y = 0$, $y = a$, $z = 0$, $z = a$?

5. Let $\mathbf{F} = r\overrightarrow{OP}$, where r is the distance OP. Use the divergence theorem to show

that

$$\iint_S \mathbf{F} \cdot \mathbf{n}\, dA = 4 \iiint_T r\, dV.$$

By evaluating the surface integral, show that

$$\iiint_T (x^2 + y^2 + z^2)^{1/2}\, dV = \pi b^4$$

if T is the spherical region $x^2 + y^2 + z^2 \leqq b^2$. Check by direct evaluation of the triple integral.

6. Let $\mathbf{F}$ be the vector function $(x^2 + y^2 + z^2)(x\mathbf{i} + y\mathbf{j} + z\mathbf{k})$. If T is the region $x^2 + y^2 + z^2 \leqq b^2$, calculate $\iiint_T (\nabla \cdot \mathbf{F})\, dV$ directly, and also by the divergence theorem. Observe that the value of the surface integral can be written down by inspection, since $\mathbf{F} \cdot \mathbf{n} = (x^2 + y^2 + z^2)^{3/2}$ on S.

7. Show that, in the notation of (15.6–1),

$$\iint_S (x^2 + y^2)(x \cos \alpha + y \cos \beta)\, dA = 4I_z,$$

where T is regarded as a homogeneous mass of unit density and I_z is its moment of inertia about the z-axis.

8. If $\mathbf{F} = x\mathbf{i} - y\mathbf{j} + z^2\mathbf{k}$, calculate $\iint_S \mathbf{F} \cdot \mathbf{n}\, dA$, where S is the entire surface of any right circular cylinder of radius b with one base in the plane $z = 1$ and the other base in the plane $z = 3$.

9. Let S be the ellipsoidal surface

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

If (x, y, z) is a point on the ellipsoid, and D is the distance from the origin to the plane tangent to the ellipsoid at (x, y, z), show that $D^{-1} = \mathbf{F} \cdot \mathbf{n}$, where $\mathbf{n}$ is the unit outer normal to S at (x, y, z), and $\mathbf{F}$ is the vector with components x/a^2, y/b^2, z/c^2. Hence, show that

$$\iint_S \frac{1}{D}\, dA = \tfrac{4}{3}\pi \left(\frac{bc}{a} + \frac{ca}{b} + \frac{ab}{c} \right).$$

10. Consider a limiting process in which a region T shrinks down onto a certain fixed point. If V is the volume of T, show that the divergence of a vector function at this point is given by

$$\operatorname{div} \mathbf{F} = \lim \frac{1}{V} \iint_S \mathbf{F} \cdot \mathbf{n}\, dA.$$

This formula is sometimes used as an invariant definition of the divergence. It is useful for calculating div $\mathbf{F}$ in terms of various systems of curvilinear co-ordinates. See Exercise 11.

11. Let ρ, θ, ϕ be the usual spherical co-ordinates, and let F_ρ, F_θ, F_ϕ denote the components of a vector function $\mathbf{F}$ in the ρ, θ, ϕ directions respectively. Show that

$$\operatorname{div} \mathbf{F} = \frac{1}{\rho^2 \sin \phi}\left[\frac{\partial}{\partial \rho}(\rho^2 \sin \phi F_\rho) + \frac{\partial}{\partial \theta}(\rho F_\theta) + \frac{\partial}{\partial \phi}(\rho \sin \phi F_\phi)\right].$$

Obtain this by using the result in Exercise 10, using for T a region of the "volume element" type in spherical co-ordinates, with pairs of opposite faces corresponding to values ρ and $\rho + \Delta\rho$, θ and $\theta + \Delta\theta$, ϕ and $\phi + \Delta\phi$. Then let $\Delta\rho$, $\Delta\theta$, and $\Delta\phi$ approach zero. HINT: The volume of T is approximately $\rho^2 \sin \phi\, \Delta\rho\, \Delta\theta\, \Delta\phi$. The surface integral over the pair of faces corresponding to ρ and $\rho + \Delta\rho$ is approximately equal to

$$[F_\rho(\rho + \Delta\rho, \theta, \phi)(\rho + \Delta\rho)^2 \sin \phi - F_\rho(\rho, \theta, \phi)\rho^2 \sin \phi]\, \Delta\theta\, \Delta\phi,$$

and analogous results can be written down for the other pairs of faces.

This procedure can be adapted to any system of orthogonal curvilinear co-ordinates. The reasoning can be made quite precise with the aid of mean-value theorems.

15.61 / GREEN'S IDENTITIES

Among the important consequences of the divergence theorem are certain integral formulas known as Green's identities. These formulas play a significant role in connection with many of the partial differential equations of applied mathematics. The first identity is

$$\iiint_T \left(u\, \Delta v + \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} + \frac{\partial u}{\partial z}\frac{\partial v}{\partial z}\right) dV = \iint_S u \frac{\partial v}{\partial n}\, dA,$$

where we have used the notation

$$\Delta v = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2}.$$

The expression Δv is called the Laplacian of v. In the surface integral, $\partial/\partial n$ refers to the directional derivative in the direction of $\mathbf{n}$, the outer normal to S. It is assumed that T is a region to which the divergence theorem applies, and that u and v, together with the first derivatives of u and the first and second derivatives of v, are continuous in T.

To prove the correctness of (15.61–1), choose

$$P = u\frac{\partial v}{\partial x}, \qquad Q = u\frac{\partial v}{\partial y}, \qquad R = u\frac{\partial v}{\partial z}$$

in the divergence theorem. Then

$$\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z}$$

is the expression under the triple integral in (15.61–1). On the other hand,

$$P \cos \alpha + Q \cos \beta + R \cos \gamma = u\frac{\partial v}{\partial n}$$

because

$$\frac{\partial v}{\partial x}\cos\alpha + \frac{\partial v}{\partial y}\cos\beta + \frac{\partial v}{\partial z}\cos\gamma = \frac{\partial v}{\partial n}.$$

Thus (15.61–1) is just a particular case of the divergence theorem.

The identity can be written in a more compact form by using vector notation. If v is a scalar function, and ∇v is the gradient of v, we recall that the component of ∇v in any particular direction is the directional derivative of v in that direction. Hence, at a point on S

$$\nabla v \cdot \mathbf{n} = \frac{\partial v}{\partial n}.$$

Likewise, we observe that the divergence of ∇v is the Laplacian of v:

$$\nabla \cdot (\nabla v) = \Delta v.$$

Since $\nabla \cdot (\nabla v)$ is conventionally written as $\nabla^2 v$, this latter notation is often used in place of Δv for the Laplacian of v. We may now write (15.61–1) in the form

$$\iiint_T (u\nabla^2 v + \nabla u \cdot \nabla v)\, dV = \iint_S u\nabla v \cdot \mathbf{n}\, dA. \tag{15.61–2}$$

This is the particular case of (15.6–4) in which $\mathbf{F} = u\nabla v$.

We refer to (15.61–1) or (15.61–2) as *Green's first identity.*

Green's second identity is

$$\iiint_T (u\,\Delta v - v\,\Delta u)\, dV = \iint_S \left(u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n}\right) dA, \tag{15.61–3}$$

or, in a different notation,

$$\iiint_T (u\nabla^2 v - v\nabla^2 u)\, dV = \iint_S (u\nabla v - v\nabla u) \cdot \mathbf{n}\, dA. \tag{15.61–4}$$

This is deduced from the first identity by exchanging u and v and subtracting. The second identity presumes that both u and v have continuous second derivatives in T.

EXERCISES

1. Show that

$$\iiint_T \nabla^2 u\, dV = \iint_S \frac{\partial u}{\partial n}\, dA.$$

2. If u is a function such that $\nabla^2 u = 0$ in T, show that

$$\iiint_T \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial u}{\partial z}\right)^2\right] dV = \iint_S u\frac{\partial u}{\partial n}\, dA.$$

3. By putting $\mathbf{F} = \nabla u$ in Exercise 11, §15.6, we obtain the expression for the Laplacian in spherical co-ordinates:

$$\nabla^2 u = \frac{1}{\rho^2 \sin\phi}\left[\sin\phi \frac{\partial}{\partial\rho}\left(\rho^2 \frac{\partial u}{\partial\rho}\right) + \frac{1}{\sin\phi}\frac{\partial^2 u}{\partial\theta^2} + \frac{\partial}{\partial\phi}\left(\sin\phi \frac{\partial u}{\partial\phi}\right)\right].$$

4. Show that the Laplacian can be expressed independently of all co-ordinate systems in the form

$$\nabla^2 u = \lim \frac{1}{V}\iint_S \frac{\partial u}{\partial n}\, dA,$$

where the limit is taken as the surface S enclosing the volume V shrinks down on the point at which the Laplacian is evaluated.

15.62 / TRANSFORMATION OF TRIPLE INTEGRALS

We shall now take up the problem for triple integrals which corresponds to the problem of §15.32 for double integrals. We suppose that the two sets of variables (x, y, z), (u, v, w) are connected by certain equations

$$x = f(u, v, w), \qquad y = g(u, v, w), \qquad z = h(u, v, w), \tag{15.62–1}$$

and that these equations establish a one-to-one mapping of a region in uvw-space onto a region in xyz-space. The functions f, g, h are assumed to be continuous, with continuous first and second partial derivatives. We write

$$J(u, v, w) = \frac{\partial(f, g, h)}{\partial(u, v, w)},$$

and assume that J has a constant sign. Let T be a closed and bounded region which is in the region of xyz-space which is being mapped, and let T' be the corresponding region of uvw-space. We assume that T and T' are regions to which the divergence theorem applies.

THEOREM IX. *Let V be the volume of T. Then*

$$V = \iiint_{T'} |J(u, v, w)|\, du\, dv\, dw. \tag{15.62–2}$$

Proof. The proof is similar to that of Theorem II, §15.32, but the calculations are somewhat more complicated. Let us begin by considering the surfaces S, S' of T and T' respectively. Suppose S_0 is a smooth simple surface element on S, and that S_0' is the corresponding element of S'. We assume that S_0 is represented parametrically in terms of certain parameters s, t. The correspondence set up by the mapping (15.62–1) gives us a representation of S_0' in terms of these same

parameters. Let us write

$$j_1 = \frac{\partial(y, z)}{\partial(s, t)}, \qquad j_2 = \frac{\partial(z, x)}{\partial(s, t)}, \qquad j_3 = \frac{\partial(x, y)}{\partial(s, t)},$$

$$j_1' = \frac{\partial(v, w)}{\partial(s, t)}, \qquad j_2' = \frac{\partial(w, u)}{\partial(s, t)}, \quad j_3' = \frac{\partial(u, v)}{\partial(s, t)}.$$

Also, let

$$D = \sqrt{j_1^2 + j_2^2 + j_3^2}, \quad D' = \sqrt{j_1'^2 + j_2'^2 + j_3'^2}.$$

We first show that

$$j_3 = \frac{\partial(f, g)}{\partial(v, w)} j_1' + \frac{\partial(f, g)}{\partial(w, u)} j_2' + \frac{\partial(f, g)}{\partial(u, v)} j_3'. \tag{15.62–3}$$

There are similar formulas for j_1, j_2, but we shall not need them.

By (15.62–1) we see that

$$\frac{\partial x}{\partial s} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial s} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial s} + \frac{\partial f}{\partial w}\frac{\partial w}{\partial s},$$

with similar formulas for $\frac{\partial x}{\partial t}$, $\frac{\partial y}{\partial s}$, $\frac{\partial y}{\partial t}$.

In the interests of compactness in display, let us write

$$\frac{\partial f}{\partial u} = f_1, \qquad \frac{\partial f}{\partial v} = f_2, \qquad \frac{\partial u}{\partial s} = u_1, \qquad \frac{\partial u}{\partial t} = u_2,$$

and so on. Then we have

$$j_3 = \begin{vmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{vmatrix} = \begin{vmatrix} f_1u_1 + f_2v_1 + f_3w_1 & f_1u_2 + f_2v_2 + f_3w_2 \\ g_1u_1 + g_2v_1 + g_3w_1 & g_1u_2 + g_2v_2 + g_3w_2 \end{vmatrix}.$$

There are theorems on determinants by which the foregoing determinant may be expanded in the form

$$j_3 = \begin{vmatrix} f_2 & f_3 \\ g_2 & g_3 \end{vmatrix} \cdot \begin{vmatrix} v_1 & v_2 \\ w_1 & w_2 \end{vmatrix} + \begin{vmatrix} f_3 & f_1 \\ g_3 & g_1 \end{vmatrix} \cdot \begin{vmatrix} w_1 & w_2 \\ u_1 & u_2 \end{vmatrix} + \begin{vmatrix} f_1 & f_2 \\ g_1 & g_2 \end{vmatrix} \cdot \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix}.$$

This latter form is equivalent to (15.62–3).

It was shown in §14.4 that the direction cosines of the normal to S_0 are given by

$$\cos\alpha = \frac{j_1}{D}, \qquad \cos\beta = \frac{j_2}{D}, \qquad \cos\gamma \frac{j_3}{D}. \tag{15.62–4}$$

In an arbitrary parametrization these may or may not refer to the outward direction of the normal, but since an exchange of the parameters s, t changes the sign of each of the Jacobians j_1, j_2, j_3, we may (and shall) assume that the

parameters have been chosen so that (15.62–4) give the direction of the outer normal to S_0. The direction cosines of the outer normal to S_0' are

$$\cos\alpha' = \pm\frac{j_1'}{D'}, \qquad \cos\beta' = \pm\frac{j_2'}{D'}, \qquad \cos\gamma' = \pm\frac{j_3'}{D'}, \tag{15.62–5}$$

where the sign may be + or −, but is the same in all three equations. As we shall see, the sign depends upon the sign of $J(u, v, w)$.

Now, by the divergence theorem with $P = Q = 0$, $R = z$, we have

$$V = \iiint_T dx\,dy\,dz = \iint_S z\cos\gamma\,dA.$$

We show that

$$\iint_S z\cos\gamma\,dA =$$

$$\pm\iint_{S'} h(u, v, w)\left[\frac{\partial(f, g)}{\partial(v, w)}\cos\alpha' + \frac{\partial(f, g)}{\partial(w, u)}\cos\beta' + \frac{\partial(f, g)}{\partial(u, v)}\cos\gamma'\right] dA, \tag{15.62–6}$$

where the choice of sign is the same as in (15.62–5). It is enough to carry out the demonstration for each pair of corresponding surface elements S_0, S_0'. By (15.51–4) and (15.62–4) we see that

$$\iint_{S_0} z\cos\gamma\,dA = \iint_R (z\cos\gamma)D\,ds\,dt = \iint_R zj_3\,ds\,dt,$$

where R is the region in the st-plane corresponding to the element S_0. The last integral may be transformed into a surface integral over S_0' (again by (15.51–4)):

$$\iint_R zj_3\,ds\,dt = \iint_{S_0'} \frac{zj_3}{D'}\,dA.$$

We now express the integrand in terms of u, v, w, using $z = h(u, v, w)$, (15.62–3) and (15.62–5). The result is the integrand on the right in (15.62–6), and thus the latter formula is established.

The next step is the application of the divergence theorem in th uvw-space, with

$$P = h\frac{\partial(f, g)}{\partial(v, w)}, \quad Q = h\frac{\partial(f, g)}{\partial(w, u)}, \quad R = h\frac{\partial(f, g)}{\partial(u, v)}.$$

A calculation shows that

$$\frac{\partial P}{\partial u} + \frac{\partial Q}{\partial v} + \frac{\partial R}{\partial w} = \frac{\partial(f, g, h)}{\partial(u, v, w)} = J(u, v, w);$$

there are certain second-derivative terms, but they cancel each other out. Thus, by (15.62–6) and the divergence theorem,

$$V = \pm \iiint_{T'} J(u, v, w)\, du\, dv\, dw.$$

This is equivalent to (15.62–2), for J is of constant sign, and V is of course positive.

It follows from the foregoing that the positive signs are to be chosen in (15.62–5) when $J > 0$, while the negative signs are to be chosen if $J < 0$.

With the aid of Theorem IX we can prove the following three-dimensional analogue of Theorem III, §15.32:

THEOREM X. *If $F(x, y, z)$ is continuous in T and if the mapping meets the requirements stated prior to Theorem IX, we have*

$$\iiint_{T} F(x, y, z)\, dx\, dy\, dz = \iiint_{T'} F(x, y, z) \left|\frac{\partial(x, y, z)}{\partial(u, v, w)}\right| du\, dv\, dw,$$

where, on the right, $F(x, y, z)$ is to be expressed as a function of u, v, w by means of (15.62–1).

The proof is entirely analogous to that of Theorem III.

EXERCISES

1. Compute the Jacobian $J(u, v, w)$ for the transformation

$$x = \rho \sin \phi \cos \theta, \qquad y = \rho \sin \phi \sin \theta, \qquad z = \rho \cos \phi,$$

where $u = \rho$, $v = \phi$, $w = \theta$, and compare Theorem X for this case with the known results about evaluating triple integrals in terms of spherical co-ordinates.

2. Consider the transformation

$$x = u + 2v + 3w, \qquad y = 4u + 5v, \qquad z = 6u.$$

(a) If R is a region in xyz-space, and R' is the corresponding region in uvw-space, what is the volume of R if the volume of R' is 10?

(b) If the centroid of R is at $(9, 45, 30)$, find the centroid of R'.

3. Use the transformation $x = au$, $y = bv$, $z = cw$ to calculate the triple integral $\iiint_R (x^2 + y^2)\, dx\, dy\, dz$, where R is defined by $(x^2/a^2) + (y^2/b^2) + (z^2/c^2) \leqq 1$. The transformed integral may be evaluated by use of spherical co-ordinates in uvw-space.

4. Consider the transformation

$$x = \frac{v}{u} \cos w, \qquad y = \frac{v}{u} \sin w, \qquad z = v^2.$$

Let R be the region between the paraboloids $z = x^2 + y^2$, $z = 4(x^2 + y^2)$ and also between the

planes $z = 1$, $z = 4$. Use the transformation to calculate the integral

$$\iiint_R (x^2 + y^2)\, dx\, dy\, dz.$$

5. Consider the transformation

$$x = \frac{v}{1+w}, \qquad y = \frac{vw}{1+w}, \qquad z = u - v.$$

Show that the inverse transformation is

$$u = x + y + z, \qquad v = x + y, \qquad w = \frac{y}{x}.$$

Let R be the region in the first octant of xyz-space under the plane $x + y + z = 2$ and directly above the trapezoid in the xy-plane bounded by the lines $x + y = 1$, $x + y = 2$, $y = 0$, $y = x$. Show that the corresponding region R' in uvw-space is a prism with two faces perpendicular to the w-axis and three faces parallel to that axis. Then evaluate

$$\iiint_R \frac{xy + y^2}{x^3}\, dx\, dy\, dz$$

by transforming to an integral over R'.

6. Use the transformation

$$x = u(1 - v), \qquad y = uv(1 - w), \qquad z = uvw$$

to calculate

$$\iiint_R x\, dx\, dy\, dz \quad \text{and} \quad \iiint_R \frac{dx\, dy\, dz}{y + z}$$

where R is the tetrahedron cut from the first octant by the plane $x + y + z = 1$.

7. If curvilinear co-ordinates u, v, w are introduced into the first octant by the transformation

$$x = v \cos w, \qquad y = v \sin w, \qquad z = \sqrt{u - v^2},$$

describe the surfaces on which u, v, w respectively are constant. Describe the region R' in uvw-space which corresponds to the region R described as follows:

$$9 \leqq x^2 + y^2 + z^2 \leqq 16,\ 1 \leqq x^2 + y^2 \leqq 4,\ x \geqq 0,\ y \geqq 0,\ z \geqq 0.$$

Calculate $\iiint_R z\, dx\, dy\, dz$ with the aid of the transformation.

8. Let curvilinear co-ordinates be introduced into the first octant by the transformation

$$x = v, \qquad y = w, \qquad z = (u - v^2 - w^2)^{1/2};$$

show that, for small values of Δu, Δv, Δw, the volume in the first octant bounded by the planes $x = v$, $x = v + \Delta v$, $y = w$, $y = w + \Delta w$, and the spheres $x^2 + y^2 + z^2 = u$, $x^2 + y^2 + z^2 = u + \Delta u$ is approximately

$$\frac{\Delta u\, \Delta v\, \Delta w}{2\sqrt{u - v^2 - w^2}}.$$

9. If R is the region defined by $x^2+y^2+z^2 \leqq 1$, and $p^2=a^2+b^2+c^2>0$, show that

$$\iiint_R \cos(ax+by+cz)\,dx\,dy\,dz = \frac{4\pi}{p^3}(\sin p - p\cos p).$$

HINT: Make a rotation of the co-ordinate axes so that the plane $ax+by+cz=0$ becomes the plane $z'=0$. Then use cylindrical co-ordinates.

10. Suppose each point (x, y, z) moves according to the law

$$x=\xi(1+t^2), \qquad y=\eta e^t, \qquad z=\zeta e^{2t},$$

so that the point is at (ξ, η, ζ) when $t=0$. If the components of velocity are computed in terms of x, y, z we find

$$\frac{\partial x}{\partial t}=2t\xi=\frac{2tx}{1+t^2}, \qquad \frac{\partial y}{\partial t}=\eta e^t=y, \qquad \frac{\partial z}{\partial t}=2\zeta e^{2t}=2z,$$

so that the velocity is the vector

$$\mathbf{F}=\frac{2tx}{1+t^2}\mathbf{i}+y\mathbf{j}+2z\mathbf{k}.$$

Now let R_0 be any fixed region in $\xi\eta\zeta$-space, and let R be the region in xyz-space into which R_0 has been carried at time t. If V is the volume of R, show that

$$\frac{dV}{dt}=\iiint_R (\nabla\cdot\mathbf{F})\,dx\,dy\,dz.$$

Actually, the validity of this equation does not depend on the particular equations of motion.

11. Show the last formula in Exercise 10 holds for the case in which the equations of motion are

$$x=\frac{\xi}{1+\xi t}, \qquad y=\eta+t, \qquad z=\zeta e^{-t}.$$

In this case the velocity vector is $\mathbf{F}=-x^2\mathbf{i}+\mathbf{j}-z\mathbf{k}$, and it is to be assumed that $\xi>0$ in R_0.

15.7 / STOKES'S THEOREM

Stokes's theorem is the formula

$$\iint_S \left[\left(\frac{\partial R}{\partial y}-\frac{\partial Q}{\partial z}\right)\cos\alpha+\left(\frac{\partial P}{\partial z}-\frac{\partial R}{\partial x}\right)\cos\beta + \left(\frac{\partial Q}{\partial x}-\frac{\partial P}{\partial y}\right)\cos\gamma\right] dA = \int_C P\,dx+Q\,dy+R\,dz. \quad (15.7\text{–}1)$$

This formula has an important vector interpretation, given in (15.7–6). Stokes's theorem plays a role in three dimensions much like that of Green's theorem in the plane (§15.3). Both the divergence theorem and Stokes's theorem are valuable tools in mathematical physics. We use Stokes's theorem in §15.8.

In formula (15.7–1) S denotes a surface bounded by a curve C; P, Q, and R denote functions of x, y, z. The normal to S is constructed at each point, and a certain direction is chosen along each normal so that all these directions point away from S on the same side of S. This is called the positive side of S. The angles α, β, γ refer to the direction of the normal at (x, y, z) on S. The curve C is oriented in such a way that, if a person walks along C in the positive direction, standing on the positive side of S, the surface S is always on his left. If S is a simple surface element, this means that the orientation of C appears as counterclockwise to a person standing on the positive side of the surface (see Fig. 159). A more exact description of the assumptions which we make about the surface S and the functions P, Q, R will be given later on.

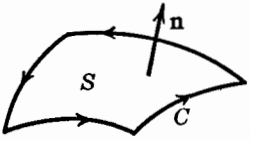


Fig. 159.

For the particular case in which S is a region in the xy-plane, Stokes's theorem is identical with Green's theorem in the plane; for in that case, if we choose the positive side of S as that toward the positive z direction, we see that $\cos\alpha = 0$, $\cos\beta = 0$, $\cos\gamma = 1$. Therefore, since $dz = 0$ along C in this case, (15.7–1) becomes

$$\iint_S \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA = \int_C P\,dx + Q\,dy,$$

and this is Green's theorem in the plane.

We shall now prove formula (15.7–1) for the case in which S is a simple surface element. If we pick out of (15.7–1) just those terms which involve P, we have the formula

$$\iint_S \left(\frac{\partial P}{\partial z}\cos\beta - \frac{\partial P}{\partial y}\cos\gamma\right) dA = \int_C P\,dx. \tag{15.7–2}$$

There are two other similar formulas, one involving Q and one involving R. We shall prove (15.7–2). The other two formulas will then be established by cyclic permutation of the letters in the sets (P, Q, R), (x, y, z), (α, β, γ). Combination of the three formulas will then give us (15.7–1).

We shall assume that the surface element S is represented parametrically by a one-to-one mapping

$$x = f(u, v), \qquad y = g(u, v), \qquad z = h(u, v),$$

the parameters (u, v) ranging over a region G in the uv-plane. We assume that G is a regular region bounded by a simple closed curve Γ which is sectionally smooth. We also assume that the functions f, g, h have continuous first and second derivatives in G and that the Jacobians

$$j_1 = \frac{\partial(y, z)}{\partial(u, v)}, \qquad j_2 = \frac{\partial(z, x)}{\partial(u, v)}, \qquad j_3 = \frac{\partial(x, y)}{\partial(u, v)}$$

are never all zero at once. Let Γ be oriented counterclockwise in the uv-plane, and let C be oriented so that as (u, v) goes along Γ in the positive sense, (x, y, z) goes along C in the positive sense. Let

$$D = (j_1^2 + j_2^2 + j_3^2)^{1/2}.$$

Then for the direction of the normal toward the positive side of S we have

$$\cos\alpha = \pm\frac{j_1}{D}, \quad \cos\beta = \pm\frac{j_2}{D}, \quad \cos\gamma = \pm\frac{j_3}{D}, \tag{15.7–3}$$

where the same choice of sign is to be made in each equation. We shall show that the + sign must be chosen. Select any point of S not on C. At this point j_1, j_2, and j_3 are not all zero, and we shall suppose for definiteness that $j_3 \neq 0$. Then near this point j_3 always has the same sign. Also, the normal near the point is never perpendicular to the z-axis and so $\cos\gamma$ is always of the same sign. The problem is to show that the sign of $\cos\gamma$ is the same as that of j_3. Let C_0 be a small closed curve on the portion of S we are considering. Let Γ_0 be the corresponding curve in the uv-plane, and let C_0' be the projection of C_0 in the xy-plane (see Fig. 160). Let the orientations of C_0 and Γ_0 agree with those of C and Γ respectively, and let the orientation of C_0' be that which is naturally induced by the orientation of C_0. Now the equations $x = f(u, v)$, $y = g(u, v)$ may be regarded as defining a mapping from the uv-plane to the xy-plane. Under this mapping Γ_0 is mapped into C_0'. The Jacobian of the mapping is j_3, so, by the discussion in §15.32, the orientation of C_0' will be counterclockwise (as viewed from the positive z-axis) if $j_3 > 0$, and clockwise if $j_3 < 0$. A little consideration of the relation between C_0' and C_0 then shows that $\cos\gamma > 0$ if $j_3 > 0$ and $\cos\gamma < 0$ if $j_3 < 0$. The case $j_3 > 0$ is illustrated in Fig. 160. This completes the demonstration that the + signs are to be chosen in (15.7–3).

We are now ready to proceed with the proof of (15.7–2). In terms of the parameters u, v the surface integral becomes

$$\iint_G \left(\frac{\partial P}{\partial z}\cos\beta - \frac{\partial P}{\partial y}\cos\gamma\right) D\,du\,dv = \iint_G \left(\frac{\partial P}{\partial z}j_2 - \frac{\partial P}{\partial y}j_3\right) du\,dv.$$

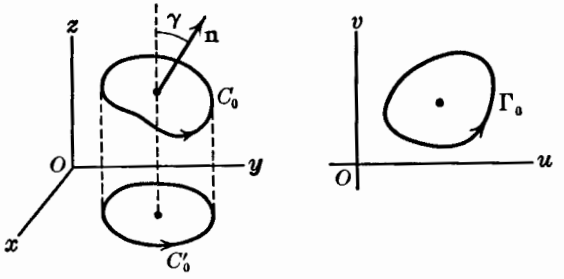


Fig. 160.

Here we have used (15.51–4) and (14.6–7). Next we show that

$$\frac{\partial P}{\partial z}\,j_2 - \frac{\partial P}{\partial y}\,j_3 = \frac{\partial P}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial P}{\partial v}\frac{\partial x}{\partial u}. \tag{15.7–4}$$

In fact,

$$\frac{\partial P}{\partial u} = \frac{\partial P}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial P}{\partial y}\frac{\partial y}{\partial u} + \frac{\partial P}{\partial z}\frac{\partial z}{\partial u},$$

with a similar formula for $\dfrac{\partial P}{\partial v}$. Therefore,

$$\frac{\partial P}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial P}{\partial v}\frac{\partial x}{\partial u} = \frac{\partial P}{\partial y}\left(\frac{\partial y}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial y}{\partial v}\frac{\partial x}{\partial u}\right) + \frac{\partial P}{\partial z}\left(\frac{\partial z}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial z}{\partial v}\frac{\partial x}{\partial u}\right),$$

and this is equivalent to (15.7–4).

The surface integral in (15.7–2) has now been reduced to the form

$$\iint_G \left(\frac{\partial P}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial P}{\partial v}\frac{\partial x}{\partial u}\right) du\,dv.$$

An easy calculation shows that this is the same as

$$\iint_G \left[\frac{\partial}{\partial u}\left(P\frac{\partial x}{\partial v}\right) - \frac{\partial}{\partial v}\left(P\frac{\partial x}{\partial u}\right)\right] du\,dv. \tag{15.7–5}$$

To this integral we now apply Green's theorem in the uv-plane. As a result, (15.7–5) is equal to the line integral

$$\int_\Gamma P\frac{\partial x}{\partial u}\,du + P\frac{\partial x}{\partial v}\,dv.$$

But this is just

$$\int_C P\,dx,$$

and so we have completed the proof of (15.7–2) under the assumptions on S as stated earlier. We have assumed that P, Q, and R have continuous partial derivatives in some region containing S.

Stoke's theorem may be extended to more general surfaces by a process entirely similar to that employed in the proof of Green's theorem in the plane. The process is suggested by Fig. 161. It consists in dividing S into a finite number of simple surface elements by the construction of one or more "cuts," or interior dividing lines. We assume that each element and its boundary takes its orientation from the overall orientation of S and C, and that each "cut" occurs as part of the boundary of just two surface elements, with opposite orientations in the two cases. If now we add the formulas

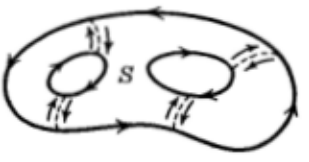


Fig. 161.

of Stokes's theorem for the several elements, the contributions from the cuts to the line integrals cancel each other in pairs, and we obtain (15.7–1) as the final result.

Notice that it is taken for granted that there is an orientation of the boundary C of S, and a corresponding orientation of S itself, that is, a designation of a positive side of S. The assumption that such orientation of S is possible is a restriction placed on S, for there are surfaces which cannot be oriented, as we shall presently show.

The concept of orientation of a surface may be developed in the following way: We limit our discussion to surfaces which may be thought of as formed out of a finite number of simple surface elements in such a way that if two elements have an arc of common boundary, then no other element has any sub-arc of this arc as part of its boundary. Moreover, we require that if two elements have any points in common, the common points shall form a finite number of arcs on the boundary of each element. A simple surface element is oriented by assigning a positive direction to the simple closed curve forming its boundary. If S is a surface formed out of elements in the manner described, and if each element is given an orientation, we shall say that the elements are *coherently* oriented if the following two conditions are fulfilled:

(a) If Γ is an arc of common boundary between two elements S_1, S_2, then the orientation of Γ in S_1 is opposite to its orientation in S_2;

(*b*) If Γ is a simple closed curve forming part of the boundary of the whole surface S, then the orientations given to various arcs of Γ by the several elements of S shall all be consistent and given an orientation to Γ as a whole.

If the elements forming S can be coherently oriented, we shall say that S is

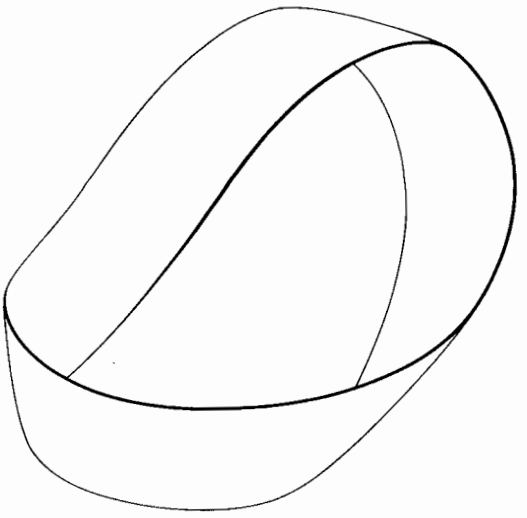


Fig. 162.

orientable. Otherwise, S is said to be *nonorientable.* These definitions apply to both closed and nonclosed surfaces. When S is orientable we can designate a positive side of S by designating a positive side to each element in accord with the orientation of the boundary of the element, as explained near the beginning of this section. But if S is nonorientable we cannot designate a positive side of S, for there will be two adjacent elements of S for which the designations of the positive sides will be in conflict as we pass from one element to the other. A nonorientable surface is in fact *one-sided,* whereas an orientable surface is two-sided. The simplest nonorientable surface is the Moebius band, represented in Fig. 162. The student can easily make a model of this surface by cutting a long narrow strip of paper, and gluing the two ends together after giving one end a half twist. The Moebius band has a single simple closed curve C as boundary.

Stokes's theorem may be given a vector form, as follows:

THEOREM XI. *Let S be an orientable surface as described above, formed from smooth surface elements with sectionally smooth boundaries. Let $\mathbf{F}$ be a vector field which is continuously differentiable in some open set containing S. Let $\mathbf{n}$ be the unit normal vector on the positive side of S. Then if S is not closed, and if C is the boundary of S,*

$$\iint_S (\nabla \times \mathbf{F}) \cdot \mathbf{n}\, dA = \int_C \mathbf{F} \cdot \mathbf{T}\, ds, \tag{15.7–6}$$

where $\mathbf{T}$ is the unit vector tangent to C in the positive sense. If S is closed, the surface integral in (15.7–6) *has the value zero.*

The form (15.7–6) is equivalent to (15.7–1) with $\mathbf{F} = P\mathbf{i} + Q\mathbf{j} + R\mathbf{k}$, since, by (14.3–1),

$$\mathbf{T} = \frac{dx}{ds}\mathbf{i} + \frac{dy}{ds}\mathbf{j} + \frac{dz}{ds}\mathbf{k}.$$

If S is closed, the line integrals around the boundaries of the elements forming S all cancel out, because S has no boundary in this case. Hence, the surface integral over S must vanish.

EXERCISES

1. Let C be the curve of intersection of $x + y = 2b$ and $x^2 + y^2 + z^2 = 2b(x + y)$, oriented in the clockwise sense as viewed from the origin. Use Stokes's theorem to find the value of $\int_C y\, dx + z\, dy + x\, dz$.

2. Let S be the part of the surface $z = x^2 - y^2$ inside the cylinder $x^2 + y^2 = 1$, and let the positive side of S be such that γ is acute. Show that, if C is the boundary of S,

$$\iint_S (x \cos \alpha - y \cos \beta)\, dA = \int_C xy\, dz$$

is a special case of Stokes's theorem. Calculate the values of the line integral and the surface integral independently, without use of Stokes's theorem, and verify that they are equal. Observe that C may be parametrized by setting $x = \cos\theta$, $y = \sin\theta$.

3. Calculate the value of $\int_C y\,dx + z\,dy + x\,dz$ where C is the curve of intersection of $bz = xy$ and $x^2 + y^2 = a^2$, oriented counterclockwise around the cylinder as viewed from a point high upon the positive z-axis. Use Stokes's theorem and then convert to a double integral over a circle in the xy-plane.

4. Let C be the curve in which the cylinder $x^2 + y^2 = a^2$ is intersected by a plane parallel to the x-axis, and let C be oriented counterclockwise around the cylinder as viewed from a point high on the positive z-axis.

(a) Show that $\int_C z(x^2 - 1)\,dy + y(x + 1)\,dz = 0$.

(b) Show that $\int_C y(z - 1)\,dx + x(z + 1)\,dy = 2\pi a^2$.

(c) If the plane is $y + z = a$, show that $\int_C z\,dx - x\,dz = 2\pi a^2$.

(d) What is the value of the integral in part (c) if the plane is $z = 2(a + y)$?

5. Calculate $\int_C y^2\,dx + xy\,dy + zx\,dz$, where C is the curve of intersection of $x^2 + y^2 = 2ay$ and $y = z$.

6. Let S be the first octant portion of the surface of the sphere $x^2 + y^2 + z^2 = 4a^2$ which is inside the cylinder $x^2 + y^2 = 2ax$. Let the outer side of the sphere be the positive side of S. Calculate the integrals

$$\textbf{(a)}\ \int_C z\,dx - x\,dz, \quad \textbf{(b)}\ \int_C x\,dy - y\,dx, \quad \textbf{(c)}\ \int_C y\,dz - z\,dy,$$

where C is the boundary of S, oriented as in Stokes's theorem. Observe that, by suitable use of projection in evaluating the surface integrals which result from using Stokes's theorem, each line integral is equal to twice the area of an appropriate plane region.

7. If S is a simple surface element whose equation is $z = f(x, y)$, with (x, y) varying over a plane region R, let C be the boundary of S and Γ the boundary of R. Prove Stokes's theorem for S by converting the line integral over C to a line integral over Γ and the integral over S to an integral over R. Then use Green's theorem in the plane.

15.8 / EXACT DIFFERENTIALS IN THREE VARIABLES

An expression such as

$$P\,dx + Q\,dy + R\,dz \tag{15.8–1}$$

is called a first-order differential form in three variables. Here P, Q, R denote functions of x, y, z. In this section we consider exact differentials and line integrals independent of the path.

Definition. *The differential form* (15.8–1) *is said to be* exact at the point (a, b, c) *if there is some single-valued differentiable function* $u = f(x, y, z)$ *defined in some neighborhood of* (a, b, c) *such that*

$$du = P\,dx + Q\,dy + R\,dz$$

at all points of the neighborhood.

The concept of an exact differential can be expressed in vector language, using the notion of the gradient of a scalar function. Let **F** be a vector field. It is called a *gradient field at a particular point* in case there is a scalar differentiable function u defined in some neighborhood of that point, with the property that **F** is the gradient of u in that neighborhood. In terms of the rectangular xyz-coordinate system, suppose that

$$\mathbf{F} = P\mathbf{i} + Q\mathbf{j} + R\mathbf{k}. \tag{15.8–2}$$

Then $\nabla u = \mathbf{F}$ means that

$$\frac{\partial u}{\partial x} = P, \quad \frac{\partial u}{\partial y} = Q, \quad \frac{\partial u}{\partial z} = R,$$

or, equivalently, that

$$du = P\,dx + Q\,dy + R\,dz. \tag{15.8–3}$$

Therefore, the differential form (15.8–1) *is exact at a point if and only if* **F** *is a gradient field at that point.*

We now state a theorem about line integrals independent of the path.

THEOREM XII. *Suppose that P, Q, R are continuous in some region, and that in this region there is defined a single-valued differentiable function $u = f(x, y, z)$ such that* (15.8–3) *holds at all points of the region. Then, if C is any sectionally smooth curve in the region, with initial point (x_0, y_0, z_0) and terminal point (x_1, y_1, z_1), the equation*

$$\int_C P\,dx + Q\,dy + R\,dz = f(x_1, y_1, z_1) - f(x_0, y_0, z_0) \tag{15.8–4}$$

holds. Therefore the line integral is independent of the path from (x_0, y_0, z_0) to (x_1, y_1, z_1).

This theorem corresponds to Theorem V, §15.4, and is proved in the same way.

Our next theorem corresponds to Theorem IV, §15.4.

THEOREM XIII. *Let D be an open region in the shape of a rectangular box, each face of which is parallel to a co-ordinate plane. Suppose P, Q, R are defined and have continuous first partial derivatives in D. Then, in order that there shall be a single-valued differentiable function $u = f(x, y, z)$ defined in D such that $du = P\,dx + Q\,dy + R\,dz$ at each point, it is necessary and sufficient that*

$$\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z} = 0, \quad \frac{\partial P}{\partial z} - \frac{\partial R}{\partial x} = 0, \quad \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} = 0 \tag{15.8–5}$$

at each point. Such a function can then be found by calculating the line integral of the differential form from an arbitrary fixed point (x_0, y_0, z_0) to the variable point (x, y, z) along any path lying in D.

Proof. The conditions (15.8–5) are necessary, for if the function u exists we have

$$\frac{\partial u}{\partial z} = R, \quad \frac{\partial u}{\partial y} = Q, \quad \frac{\partial^2 u}{\partial y\, \partial z} = \frac{\partial^2 u}{\partial z\, \partial y},$$

and so $\dfrac{\partial R}{\partial y} = \dfrac{\partial Q}{\partial z}$; similar arguments show that all the equations (15.8–5) must be satisfied.

Let us now start with the assumption that equations (15.8–5) hold in D. Choose any fixed point (x_0, y_0, z_0) in D, and let (x_1, y_1, z_1) be any other point in D. Let C be the straight line segment from (x_0, y_0, z_0) to (x_1, y_1, z_1) and define

$$f(x_1, y_1, z_1) = \int_C P\,dx + Q\,dy + R\,dz. \tag{15.8–6}$$

This gives us a definition for the value of the function at each point of D. It remains only to show that

$$\frac{\partial f}{\partial x} = P, \quad \frac{\partial f}{\partial y} = Q, \quad \frac{\partial f}{\partial z} = R. \tag{15.8–7}$$

Now suppose that (a, b, c) is any third point of D, and let C_1, C_2 be the segments as shown in Fig. 163. We shall show that

$$f(x_1, y_1, z_1) = f(a, b, c) + \int_{C_2} P\,dx + Q\,dy + R\,dz. \tag{15.8–8}$$

This is equivalent to showing that

$$\int_C = \int_{C_1} + \int_{C_2}$$

(with $P\,dx + Q\,dy + R\,dz$ understood as appearing under each integral sign). This in turn is equivalent to showing that

$$\int_\Gamma P\,dx + Q\,dy + R\,dz = 0 \tag{15.8–9}$$

where Γ is the oriented path all the way around the triangle in one direction. If S is the plane triangular region bounded by Γ, we see that (15.8–9) holds by virtue of (15.8–5) and Stokes's theorem (15.7–1).

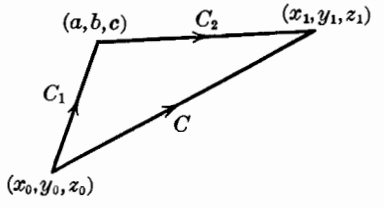


Fig. 163.

Let us now consider x_1 as a variable, and y_1, z_1 as fixed, and let us take C_2 parallel to the x-axis, so that $b = y_1$, $c = z_1$. Then (15.8–8) becomes

$$f(x_1, y_1, z_1) = f(a, y_1, z_1) + \int_a^{x_1} P(x, y_1, z_1)\, dx.$$

From this equation it is clear that

$$\frac{\partial f(x_1, y_1, z_1)}{\partial x_1} = P(x_1, y_1, z_1).$$

We have thus proved the first formula in (15.8–7). The other two are proved in the same way. This completes the proof of Theorem XIII. It will be observed that, for the proof we have given, the open region D need not be a rectangular box, provided it has the property that any two points of D can be joined by a line segment lying wholly in D. Such a region is called *convex*.

The substance of Theorems XII and XIII is capable of succinct statement in vector notation. Theorem XII says that

$$\int_C \mathbf{F} \cdot \mathbf{T}\, ds = f(x_1, y_1, z_1) - f(x_0, y_0, z_0)$$

provided $\mathbf{F} = \nabla f$ in the region where C lies. Theorem XIII says that $\mathbf{F}$ is the gradient of some function f in the region D if and only if $\nabla \times \mathbf{F} = \mathbf{0}$ in D. For, with $\mathbf{F}$ defined by (15.8–2), equations (15.8–5) mean precisely that $\nabla \times \mathbf{F} = \mathbf{0}$.

Example. Of what function is

$$(y^2 + 2z^2x - 1)\, dx + 2yx\, dy + 2zx^2\, dz$$

the differential?

This differential form satisfies the conditions (15.8–5) at all points. We shall compute a function $f(x, y, z)$ by integrating from $(0, 0, 0)$ to (x, y, z) along a broken line as shown in Fig. 164. We use (r, s, t) as co-ordinates of a variable point along the path. Then

$$f(x, y, z) = \int_C (s^2 + 2t^2r - 1)\, dr + 2sr\, ds + 2tr^2\, dt$$

$$= \int_0^x -dr + \int_0^y 2sx\, ds + \int_0^z 2tx^2\, dt,$$

$$f(x, y, z) = -x + y^2x + z^2x^2.$$

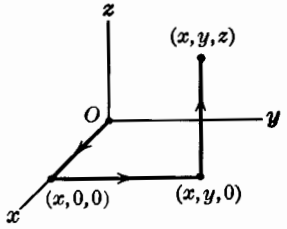


Fig. 164.

The differential of this function is the originally given differential form.

Theorem XIII is not true for *all* open regions D. Consider, for example, the differential form

$$\frac{-y\, dx + x\, dy}{x^2 + y^2} + 2z\, dz,$$

where D is all of space except the z-axis. In this region the equations (15.8–5) are satisfied, and yet there is no single-valued function defined in D whose differential is equal to the foregoing differential form. The explanation lies in Example 1, §15.41. The essential trouble is that, if a closed curve C in D encircles the z-axis, it is impossible to find a surface S in D, with C as a boundary, to which we can apply Stokes's theorem.

To get away from difficulties of this kind we need to develop a notion which for space of three dimensions is analogous to the notion of a simply connected region in two dimensions, as explained in §15.41. We shall not attempt any systematic developments of this kind, however.

EXERCISES

1. In each of the following cases determine, by inspection or otherwise, a function whose differential is the differential form under the integral sign. Then use Theorem XII to evaluate the line integral over the specified path.

(a) $\int_C yz\,dx + zx\,dy + xy\,dz$, from $(1, 2, 3)$ to $(4, 5, 6)$;

(b) $\int_C 2xy^2z\,dx + 2x^2yz\,dy + x^2y^2\,dz$, from $(0, 0, 0)$ to (a, b, c);

(c) $\int_C 2xyz^3\,dx + x^2z^3\,dy + 3x^2yz^2\,dz$, from $(1, 1, 1)$ to (p, q, r);

(d) $\int_C yz\cos(xz)\,dx + \sin(xz)\,dy + xy\cos(xz)\,dz$, from $(\pi^2/4, 1, 2/\pi)$ to $(\pi/2, -2, 3)$.

2. Let $u = x/r^2$ where $r^2 = x^2 + y^2 + z^2$, and let $\mathbf{F} = \nabla u$. Compute $\int_C \mathbf{F} \cdot \mathbf{T}\,ds$, where C is any curve from $(1, 0, 0)$ to (a, b, c) not passing through the origin.

3. Find a function whose differential is $2xy\,dx + (x^2 + \log z)\,dy + (y/z)\,dz$, by integrating along a path from $(0, 0, 1)$ to (x, y, z). Solve the problem in several ways: **(a)** Use a broken-line path from $(0, 0, 1)$ to $(x, 0, 1)$ to $(x, y, 1)$ to (x, y, z); **(b)** use a broken-line path from $(0, 0, 1)$ to $(0, 0, z)$ to $(0, y, z)$ to (x, y, z); **(c)** use the straight-line path from $(0, 0, 1)$ to (x, y, z).

4. Find a function whose differential is

$$\frac{2x}{z}\,dx + \frac{2y}{z}\,dy + \left(1 - \frac{x^2 + y^2}{z^2}\right)dz,$$

by integrating over some path from $(1, 1, 1)$ to (x, y, z).

5. Find a function whose differential is

$$e^{-xy}[(y - xy^2 + yz)\,dx + (x - x^2y + xz)\,dy - dz],$$

by integrating over some path from $(0, 0, 0)$ to (x, y, z).

6. Let C be the circle $x^2 + y^2 = a^2$, $z = 0$. If $P(x, y, z)$ is any point, let P' be the point $(x, y, 0)$, and let Q be the point in which the ray OP' (produced if necessary) intersects C. Let ϕ be the angle between the directions OQ and QP (see Fig. 165), with the convention that $0 < \phi < \pi$ if $z > 0$, $-\pi < \phi < 0$ if $z < 0$, and, if $z = 0$, that $\phi = 0$ if $x^2 + y^2 > a^2$ and $\phi = \pi$ if $x^2 + y^2 < a^2$. This leaves ϕ undefined at points of the circle C, but it is defined everywhere else.

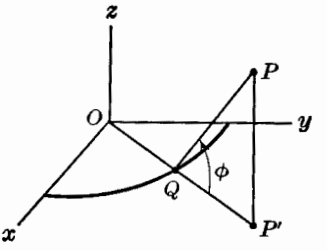


Fig. 165.

(a) Show that, in general,

$$d\phi = \frac{-xz\,dx - yz\,dy}{\sqrt{x^2+y^2}[(\sqrt{x^2+y^2}-a)^2+z^2]} + \frac{\sqrt{x^2+y^2}-a}{(\sqrt{x^2+y^2}-a)^2+z^2}\,dz.$$

(b) At what points is ϕ discontinuous?

(c) At what points is ϕ continuous but not differentiable?

(d) If D is the region consisting of all of space except the z-axis and the circle C, show that the differential form in (a) satisfies conditions (15.8–5) in D.

(e) Describe a closed curve in D which is not the boundary of any surface lying in D and to which Stokes's theorem may be applied.

MISCELLANEOUS EXERCISES

1. If S is defined by $z = f(x, y)$, with (x, y) ranging over R, and if the positive side of S is chosen so that $\cos\gamma > 0$, show that

$$\iint_S (x\cos\alpha - y\cos\beta)\,dA = \iint_R \left(y\frac{\partial f}{\partial y} - x\frac{\partial f}{\partial x}\right)dx\,dy.$$

2. Find a function whose differential is

$$\left(e^x\cos y - \frac{3}{\sqrt{1-x^2}}\right)dx - (e^x\sin y - 7\sec^2 y)\,dy.$$

3. Suppose $a > 0$, $b > 0$. Let P be a point of intersection of $y^2 = -4a^2(x - a^2)$ and $y^2 = 4b^2(x + b^2)$, and let R be the region bounded by the first parabola, the x-axis, and the line OP. Show that $\iint_R \frac{dx\,dy}{\sqrt{x^2+y^2}} = 2ab$ by using the transformation $x = u^2 - v^2$, $y = 2uv$.

4. Use (15.4–6) with $a = b = 0$ to find a function u such that

$$du = \frac{4xy\,dx + 2(1 + y^2 - x^2)\,dy}{(x^2+y^2-1)^2+4y^2}.$$

The function so found has certain discontinuities. Where are they?

5. Find a region in the $r\theta$-plane which is mapped, by $x = r\cos\theta$, $y = r\sin\theta$, into the region R between $x^2+y^2 = 1$, $x^2+y^2 = 4$, and inside $x^2+y^2 = 2x$. Hence calculate $\iint_R \frac{x}{\sqrt{x^2+y^2}}\,dx\,dy$.

6. Find a first-quadrant region in the uv-plane which maps into the region R defined by $1 \leqq x^2+y^2 \leqq 4$, $y \geqq 0$, if the mapping is $x = u^2 - v^2$, $y = 2uv$. Calculate $\iint_R \frac{dx\,dy}{\sqrt{x^2+y^2}}$ by transforming to the uv-plane, and check your result by calculating the given integral in terms of polar co-ordinates.

7. Calculate $\iint_R \cos\left[\frac{\pi(2x-y)}{2(x-2y)}\right]dx\,dy$, where R is the triangle bounded by $x = y$, $x + y = 0$, and $x - 2y = 2$. Use the transformation $u = 2x - y$, $v = x - 2y$.

8. Let S be the surface of the elliptic cone $(x^2/a^2)+(y^2/b^2)=(z^2/c^2)$ between $z=0$ and $z=c$. Show that the centroid is at $(0,0,\frac{2}{3}c)$. Show that both the area and the first moment of S can be expressed in terms of the same complete elliptic integral of the second kind. For a hint as to a useful procedure, see Exercise 5, §15.32.

9. Show that the potential at O due to a uniform charge on the surface of the torus $x=(a+b\cos\phi)\cos\theta$, $y=(a+b\cos\phi)\sin\theta$, $z=b\sin\phi$ is expressible in terms of the standard elliptic integrals

$$\int_0^{\pi/2}\sqrt{1-k^2\sin^2\phi}\,d\phi, \quad \int_0^{\pi/2}\frac{d\phi}{\sqrt{1-k^2\sin^2\phi}},$$

where

$$k^2=\frac{4ab}{(a+b)^2}.$$

10. Find the area of the part of the surface $(x^2+y^2+z^2)^2=x^2-y^2$ in the first octant.

16 / POINT-SET THEORY

16 / PRELIMINARY REMARKS

This chapter is a continuation of the study of limits and convergence. Chapter 2 may be regarded as part of the theory of point sets on a straight line. In Chapter 5 we began some study of point sets in a plane. Point-set theory is fundamental in the study of functions. Our primary aim in this chapter is to develop enough theory to deal with our later study of continuous functions, integration, and the theory of convergence of sequences and series. The principal theorems are the Bolzano-Weierstrass theorem, Cauchy's convergence condition, and the Heine-Borel theorem.

16.1 / FINITE AND INFINITE SETS

A *set* of points is defined by some condition which distinguishes the points which belong to the set from those which do not belong to it. The points which belong to a set are called *elements* of the set. In the study of limit concepts we are often obliged to deal with sets having an infinite number of elements. A set is called a *finite* set if there is some positive integer n such that the set has exactly n elements.

Sometimes we may define a set of points by a condition which is such that no points satisfy the condition. In this case we call the set empty, or *void.* The empty set is considered to be a finite set, the number of its elements being 0.

A set which is not finite is called *infinite.*

Example 1. Let S be the set of all points (x, y) in the xy-plane which are such that x and y are integers and $x^2 + y^2 < 100$. This is a finite set. Without determining exactly how many elements there are in S, we can easily see that the number does not exceed $(19)^2 = 361$.

Example 2. Let S be the set of all points (x, y) inside the circle $(x-2)^2 + y^2 = 4$ and such that $y > x + 1$. This set is void, for the circle lies entirely below the line $y = x + 1$.

Perhaps the simplest infinite set is the set whose elements are the positive integer points 1, 2, 3, ... on the real number scale. Let us, for the time being, denote this set by the letter I.

Definition. *A point set S is called denumerable, or denumerably infinite, if there is a one-to-one correspondence between the elements of S and the elements of the set I of positive integers.*

The existence of a one-to-one correspondence between the elements of S

and those of I means that each point of S can be labeled with a unique positive integer n, with no two distinct points having the same label n, and that every positive integer is used as the label for some point. The elements of S can then be symbolized by $P_1, P_2, P_3, \ldots$.

The words *enumerable* and *countable* are synonyms for *denumerable.*

Infinite sets may be classified into those which are denumerable and those which are not. It can be shown that the set consisting of all the real numbers is nondenumerable, whereas the set consisting merely of the rational numbers is denumerable. The demonstrations of these facts may be based on Examples 3 and 4, below.

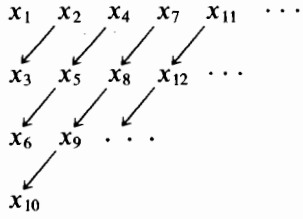
If a set is infinite, we can select out of it a subset which is denumerable. This simple but important fact is used in many arguments.

Example 3. The set S of all rational numbers x such that $0 < x < 1$ is denumerable.

To demonstrate the denumerability of S, let us make a display as follows:

$$\begin{array}{ccccccc} \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \cdots \\ \frac{2}{3} & \frac{2}{5} & \frac{2}{7} & \frac{2}{9} & \frac{2}{11} & \cdots \\ \frac{3}{4} & \frac{3}{5} & \frac{3}{7} & \frac{3}{8} & \frac{3}{10} & \cdots \\ \cdots \end{array}$$

The numerators in the successive rows of the display are $1, 2, 3, \ldots$. The denominators in each row are increasing, but such that each fraction is proper and in its lowest terms. A little reflection shows us that each member of S appears just once in this array, and that each member of the array is an element of S. By labeling the elements $x_1, x_2, x_3, x_4, \ldots$ as shown in the following array, we see that S is denumerable.



Example 4. Let S be the set of all nonterminating decimal fractions of the form $0.a_1a_2a_3\ldots$, where $a_1, a_2, \ldots$ represent digits $0, 1, \ldots, 9$, and the case in which all the a's from some point onward are 0 is ruled out. This set is nondenumerable.

If S were denumerable, we could label *all* its members by a double subscript

scheme as follows:

$$\text{1st element: } 0.a_{11}a_{21}a_{31}\cdots$$
$$\text{2d element: } 0.a_{12}a_{22}a_{32}\cdots$$
$$\text{3d element: } 0.a_{13}a_{23}a_{33}\cdots$$
$$\cdots$$

This leads to a contradiction, however, for there are elements of S not included in the foregoing list. In fact, consider the decimal $0.c_1c_2c_3\ldots$, where the c's are chosen from $0, 1, \ldots, 9$ according to the rules: $c_1 \neq a_{11}$, $c_2 \neq a_{22}$, $c_3 \neq a_{33}, \ldots$, and each c_i is different from 0. This decimal belongs to S, but is different from each decimal in the proposed enumeration of the elements of S. Therefore, S is nondenumerable.

16.2 / POINT SETS ON A LINE

In dealing with point sets on a line we shall assume that a number scale has been established on the line. We shall usually think of it as the x-axis. Then each point is identified by its x-co-ordinate, and we shall refer to a point by referring to the corresponding real number.

Definition. *By a* neighborhood *of a point x_0 we mean the set of all points x such that $x_0 - h < x < x_0 + h$, where h is some positive number. The neighborhood thus consists of all points between $x_0 - h$ and $x_0 + h$, not including these two points. For each choice of h we get a neighborhood of x_0.*

This definition should be compared with the definition of a neighborhood in the discussion of point sets in the xy-plane; see §5.1.

Definition. *A point set S is called* open *if for each point x_0 of S there is some neighborhood of x_0 which belongs entirely to S.*

Example 1. If $a < b$, the set S of all x such that $a < x < b$ is an open set. It is called an *open interval.*

Example 2. The set of all rational numbers r such that $0 < r < 1$ is not open. This is because any neighborhood of any points contains both rational and irrational numbers.

Example 3. Let S be the set of all points on the x-axis except the points $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$. This set is not open, for 0 is in the set, but no neighborhood of 0 lies wholly in the set.

Definition. *In speaking of point sets on a line, if S is a set, the* complement *of S is defined as the set of all the points on the line which are not in S. The complement of S is denoted by $C(S)$.*

For instance, if S is the set of Example 1, $C(S)$ consists of all x such that $x \leqq a$, together with all x such that $x \geqq b$.

Definition. *A set is called* closed *if its complement is open.*

Example 4. If $a < b$, the set of all x such that $a \leqq x \leqq b$ is a closed set. It is called a closed interval.

It is very important to notice that, in our use of "open" and "closed" as adjectives describing sets, "*closed*" *is not the opposite of* "*open.*" If a set is not open, that does not mean that it is closed.

Example 5. The set S of all x such that $0 < x \leqq 1$ is neither open nor closed. For, no neighborhood of 1 lies in S, which shows that S is not open. And no neighborhood of 0 lies in $C(S)$, which shows that $C(S)$ is not open, and hence that S is not closed.

The entire x-axis is an open set, by the definition of openness. The complement of this set is the void set. Also, the complement of the void set is the entire x-axis. It is convenient to agree, by convention, that the void set is open. This makes the entire x-axis a closed set. Thus, the entire x-axis is both open and closed, and the same is true of the void set. There are no other sets on the line which are both open and closed, however.

One of the most important concepts in this chapter is that of an *accumulation point* of a set.

Definition. *Let S be a point set, and let y be a point which is not necessarily in S. We call y an accumulation point of S if in each neighborhood of y there is at least one point x which is in S and distinct from y.*

In some books the term *limit point* is used instead of *accumulation point.*

A finite set can have no accumulation point, for if S is finite and y is any point, it is easily seen that if a small enough neighborhood of y is selected, the condition of the definition cannot be fulfilled. An infinite set may have no accumulation points, but it may also have one or several or infinitely many. There is no requirement that an accumulation point of S shall belong to S, but it may happen to belong to S in a particular case.

Example 6. Let S be the set of all rational numbers. Every point on the number scale is an accumulation point of S.

This follows from the fact, demonstrated in §2.5, that there is a rational number between any two distinct numbers. As a consequence, if x_0 is any number, every neighborhood of x_0 contains many rational numbers, so that x_0 is an accumulation point of S.

In the definition of y as a point of accumulation of S, the requirement of the definition actually makes it necessary for each neighborhood of y to contain infinitely many points of S. This is seen as follows: Let I_1 be some neighborhood

of y, and let x_1 be a point of S in I_1, with $x_1 \neq y$. Now choose a neighborhood I_2 of y, small enough so that x_1 is not in I_2. There must be a point x_2 of S in I_2, with $x_2 \neq y$. We then repeat the argument, getting smaller and smaller neighborhoods $I_3, I_4, \ldots$, and points $x_3, x_4, \ldots$ all in S such that, for each k, x_k is in I_k but not in I_{k+1}, and $x_k \neq y$. The infinite sequence of distinct points belongs to the original neighborhood I_1.

When the full import of the foregoing paragraph has been realized by the student, he or she will not find it difficult to recognize points of accumulation in the situations which come to his or her attention.

There is an important relation between the concept of a closed set and the notion of accumulation points, as we see in the following theorem:

THEOREM I. *A set is closed provided that each accumulation point of S is a member of S. Conversely, if S is closed, it contains each of its points of accumulation.*

Proof. Suppose every accumulation point of S belongs to S. If S were not closed, its complement would not be open. But, if $C(S)$ were not open, this would mean that $C(S)$ contains a point y which has no neighborhood lying entirely in $C(S)$. But if a neighborhood of y fails to lie wholly in $C(S)$, it must contain a point of S, and since y is in $C(S)$, the point of S cannot be the point y. Thus, if $C(S)$ were not open, it would contain a point y whose every neighborhood contains a point of S, and thus y would be a point of accumulation of S. This would contradict the supposition that every accumulation point of S belongs to S. Consequently, $C(S)$ must be open, and S closed. This proves the first assertion in Theorem I.

For the converse we suppose S is closed. If y is any accumulation point of S, we have to show that y is in S. If it were not, it would be in the complement $C(S)$, which is open. Then some neighborhood of y would also lie in $C(S)$. Such a neighborhood could not contain any points of S, and so y could not be an accumulation point of S. Thus we arrive at a contradiction unless we conclude that y is in S. This completes the proof.

EXERCISES

1. For each of the following sets answer the question: Is the set open, closed, or neither?

(a) All x such that $x < 1$.

(b) All x such that $x \geqq 0$.

(c) All x such that either $x < 0$ or $x \geqq 1$.

(d) All rational numbers.

(e) All irrational numbers.

2. If S is the set of all x such that $0 \leqq x \leqq 1$, what points, if any, are points of accumulation of both S and $C(S)$?

3. **(a)** If S is the set of all x such that $0 < x < 1$, are there any points of S which are points of accumulation of $C(S)$? **(b)** Are there any points of $C(S)$ which are points of accumulation of S?

4. Prove that any finite set is closed.

5. Prove that, if S is open, each of its points is a point of accumulation of S.

6. For each of the following sets find the points of accumulation and specify whether or not the set is closed.

(a) All numbers of the form $(n-1)/n$, $n = 1, 2, 3, \ldots$.

(b) All numbers of the form $n/(n^2+1)$, $n = 0, 1, 2, 3, \ldots$.

(c) All numbers of the form $(-1)^n n/(n+1)$, $n = 0, 1, 2, 3, \ldots$.

16.3 / THE BOLZANO-WEIERSTRASS THEOREM

A point set S on the x-axis is called *bounded* if there is some finite interval which contains all of S. In other words, S is bounded if there exist numbers a, b (with $a < b$) such that $a \leqq x \leqq b$ for every x in S.

THEOREM II. *Suppose that S is a bounded, infinite set. Then there is at least one point of accumulation of S.*

This theorem is generally known as the Bolzano-Weierstrass theorem. Bernard Bolzano (1781–1848) of Prague was a pioneer in the rigorous study of point sets and other matters fundamental in analysis. Karl Weierstrass (1815–1897) was one of the great mathematicians of the nineteenth century

Proof. For the proof we appeal to the theorem about nested intervals (Theorem VI, §2.8). We assume that S lies in the closed interval $[a, b]$, and we denote this interval by I_1. Now divide I_1 into two equal parts by the midpoint $(a+b)/2$, and consider the two closed intervals $[a, (a+b)/2]$, $[(a+b)/2, b]$. Since every point of S lies in one or the other of these intervals, at least one of them must contain an infinite number of points of S. Let I_2 denote such a one of the two. We then bisect I_2 and obtain a new closed interval I_3. By repetition of the process we generate a nest of closed intervals $\{I_n\}$. By the manner of construction I_n contains infinitely many points of S, and the length of I_n is $(b-a)/2^{n-1}$. By the theorem referred to earlier in this paragraph, there is exactly one point common to all the intervals of the nest. This point, call it z, is an accumulation point of S, for any given neighborhood of z will contain I_n if n is sufficiently large. This is true because z is in I_n and the length of I_n approaches 0 as n increases. The neighborhood must therefore contain an infinite number of points of S.

We shall use the Bolzano-Weierstrass theorem as a basis for much of the later work in this chapter.

EXERCISES

1. Suppose S is a set having the number M as its least upper bound. If M is not a member of S, show that it is a point of accumulation of S. Give an example showing that, if M does belong to S, it need not be a point of accumulation of S. For the definition of least upper bounds see §2.7.

2. For an alternative proof of Theorem II, let two classes L, R be defined as follows:

A number x belongs to L if there are infinitely many elements of S to the right of x on the scale; otherwise x belongs to R. Show that L and R form a cut (see §2.4), and that the cut number is a point of accumulation of S.

16.31 / CONVERGENT SEQUENCES ON A LINE

The notion of limit of a sequence has been discussed briefly in §1.62. The student should review this earlier section at this time. Our present purpose is to develop some relations between the concept of a convergent sequence and the notion of a point of accumulation of a set.

If $\{x_n\}$ is a sequence of real numbers, we can consider the set consisting of all the distinct points on the number scale corresponding to the successive numbers $x_1, x_2, x_3, \ldots$. *It is to be emphasized that the set thus defined is not the same thing as the sequence.* A point set is just a collection of points; a sequence $\{x_n\}$ is a function defined on the positive integers, with x_n the value of the function corresponding to the particular positive integer n.

If $\{x_n\}$ is a sequence, the set of values $x_1, x_2, x_3, \ldots$ may be a finite set. This is true in the example

$$x_n = 1 - (-1)^n,$$

where the set of values consists of the two points 0, 2. If the set of values $x_1, x_2, x_3, \ldots$ is a finite set, the sequence will be convergent if and only if there is some integer N such that all the values x_n are the same when $n \geqq N$. This is easily seen by direct consideration of the definition of convergence (§1.62). The case in which the set of values is infinite is taken up in the theorem which follows:

THEOREM III. *Suppose $\{x_n\}$ is a convergent sequence of real numbers, and suppose that S, the set of distinct points among the values $x_1, x_2, x_3, \ldots$, is an infinite set. Then S has just one point of accumulation, and this point is the limit of the sequence.*

Proof. Denote the limit by x. Consider any neighborhood of x, say the open interval from $x - \epsilon$ to $x + \epsilon$, where $\epsilon > 0$. By the definition of convergence there is some integer N such that x_n is in the above neighborhood if $n \geqq N$. Since S is infinite, there must be infinitely many distinct points of S represented among the values $x_N, x_{N+1}, x_{N+2}, \ldots$. These are all in the specified neighborhood of x, and so x must be an accumulation point of S. There can be no other accumulation point. For, if $y \neq x$, choose neighborhoods I_x and I_y of x and y, respectively, which are small enough so that no point belongs to both I_x and I_y. Then all but a finite number of points of S are in I_x, whereas if y were an accumulation point, an infinite number of points of S would have to be in I_y. This completes the proof of Theorem III.

The following theorem is useful:

THEOREM IV. *Let S be any point set having an accumulation point y. Then*

there is a sequence $\{x_n\}$ *such that* $\lim_{n\to\infty} x_n = y$, *and such that each* x_n *is an element of* S *and the values* $x_1, x_2, \ldots$ *are all distinct.*

Most of the argument needed to prove this theorem has already been given, in the second paragraph following Example 6, §16.2. All that is necessary is to specify that the intervals $I_1, I_2, I_3, \ldots$ be chosen so that the length of I_n approaches 0 as $n \to \infty$. This will insure that $x_n \to y$.

We conclude this section with some considerations of *subsequences*. If $\{x_n\}$ is a sequence, a subsequence is another sequence $\{y_k\}$ formed by dropping out certain of the x_n's and retaining the rest in the order originally given. If the indices retained are $n_1, n_2, n_3, \ldots$, where $n_1 < n_2 < n_3 < \cdots$, then $y_1 = x_{n_1}$, $y_2 = x_{n_2}$, and so on. As particular examples of subsequences we cite:

$$x_2, x_4, x_6, x_8, \ldots \quad (n_k = 2k),$$
$$x_1, x_2, x_6, x_{24}, \ldots \quad (n_k = k!).$$

If $\{x_n\}$ is a sequence, convergent or not, we may consider a subsequence and inquire as to its convergence. A sequence may be divergent and yet contain a convergent subsequence. Indeed, it may contain many convergent subsequences, each with a different limit.

If a sequence is convergent, every subsequence formed from it is also convergent, and they all have the same limit, namely the limit of the original sequence.

There are many uses of the following theorem:

THEOREM V. *Let* $\{x_n\}$ *be a bounded sequence. Then* $\{x_n\}$ *contains a convergent subsequence.*

Proof. Let S be the set of distinct points represented by the values $x_1, x_2, \ldots$. There are two cases to consider, according as S is finite or infinite. If S is finite, it has no accumulation point. But some point of S must be repeated an infinite number of times in the sequence, so that for some subsequence we must have $x_{n_1} = x_{n_2} = x_{n_3} = \cdots$. This subsequence is certainly convergent, the limit being the repeated value. If S is infinite, it must have at least one accumulation point, say y, by the Bolzano-Weierstrass theorem. But then, by Theorem IV, y is the limit of a sequence of distinct terms chosen from S. This sequence might not be a subsequence of $\{x_n\}$. For example, it might be $x_{10}, x_5, x_{20}, x_{30}, x_{25}, \ldots$, where the indices are not arranged in increasing order. But, by dropping out the terms whose indices are not in the natural order we get a subsequence ($x_{10}, x_{20}, x_{30}, \ldots$ in the foregoing example), and this subsequence will converge to y. This completes the proof.

EXERCISES

1. Suppose that $\{x_n\}$ is a sequence which is bounded and such that all the values $x_1, x_2, x_3, \ldots$ are distinct. Assume that the set of these values has just one point of accumulation, denoted by x. Prove that the sequence is convergent and that the limit is x.

2. Consider the sequence with terms $2, \frac{1}{2}, \frac{4}{3}, \frac{1}{4}, \frac{6}{5}, \frac{1}{6}, \ldots$ where $x_n = \frac{1}{2}[1-(-1)^n]+(1/n)$. Find two convergent subsequences with different limits.

3. Give an example of a sequence for which the set of values $x_1, x_2, \ldots$ is finite and there are three different convergent subsequences with distinct limits.

4. If all the terms of a sequence are distinct, and the set of values has just one point of accumulation, can the sequence fail to be convergent? Consider the case where $x_n = n[1+(-1)^n]+(1/n)$.

16.4 / POINT SETS IN HIGHER DIMENSIONS

In §5.2 we defined neighborhoods for points in the xy-plane. Then we went on to define open sets, closed sets, and some other concepts. It was indicated that similar definitions can be made for point sets in three-dimensional space, starting from the definition of spherical neighborhoods.

If S is a point set in the plane, we define an accumulation point of S in exactly the same way as we did for point sets on the line: A point Q is called an accumulation point of S if each neighborhood of Q contains at least one point P which is in the set S and distinct from Q. This same definition is also used for point sets in three-dimensional space.

When we analyze and compare the fundamentals of point-set theory for point sets in 1, 2, or 3 dimensions, we observe the following things: (1) The verbal definition of an open set is the same in all three cases, but the word "neighborhood" has to be given the interpretation appropriate to the dimensionality. (2) The verbal definition of a closed set, namely, that S is closed if $C(S)$ is open, is the same in all three cases. (3) Theorem I of §16.2, which states that a set is closed if and only if it contains all its points of accumulation, is valid for point sets in the plane or in space as well as for point sets on a line. The proof as given in §16.2 does not depend upon the dimensionality.

It must be recognized that a point set on the x-axis may be open when it is considered as a point set on the line, but that it will not be open when considered as a point set in the xy-plane. This is because of the different meanings of the word "neighborhood" in the theories for one and two dimensions, respectively. Likewise, for example, the set of all points in the xy-plane for which $x^2+y^2<1$ is open in the theory for two dimensions, but is not open when we regard it as a point set in the three-dimensional space of points (x, y, z).

We now turn to the Bolzano-Weierstrass theorem for higher dimensions. The statement is exactly the same as that already given in §16.3 (Theorem II). Bounded sets in the plane or in space were defined in §5.3. To prove the theorem for sets in the xy-plane we shall first discuss a generalization of Theorem VI, §2.8.

Let $R_1, R_2, \ldots, R_n, \ldots$ be a sequence of closed rectangular regions, with sides parallel to the co-ordinate axes and such that R_2 is contained in R_1, R_3 is contained in R_2, and so on. Furthermore, let the dimensions of R_n approach 0 as $n \to \infty$. We shall call such a sequence of rectangles a *nest*. By an extension of the argument used in proving Theorem VI, §2.8, it is easy to see that *if R_n is any nest*

of closed rectangles, there is one and only one point which is common to all the rectangles.

Now let S be a bounded infinite point set in the xy-plane. Since S is bounded, it is contained in some sufficiently large closed rectangle with sides parallel to the axes, say R_1, defined by $a_1 \leqq x \leqq b_1$, $c_1 \leqq y \leqq d_1$. Now divide R_1 into four equal smaller rectangles, by the lines $x = (a_1 + b_1)/2$, $y = (c_1 + d_1)/2$. In at least one of the four closed smaller rectangles there must be an infinite number of points of S. Call such a one of the smaller rectangles R_2. The same procedure is now applied to R_2. By continuation we obtain a sequence R_1, $R_2, \ldots, R_n, \ldots$ forming a nest of closed rectangles, and, for each n, there are an infinite number of points of S in R_n. Let P be the unique point belonging to all the rectangles. This point P is easily seen to be an accumulation point of S.

This method of proof may be suitably extended to apply to point sets in three dimensions.

16.41 / CONVERGENT SEQUENCES IN HIGHER DIMENSIONS

The concept of a convergent sequence may be extended to the case of sequences whose terms are points in space of higher dimensions.

Definition. *Let P_n be a sequence of points. We say that the sequence is convergent to the limit P, and write $\lim_{n\to\infty} P_n = P$, if each neighborhood of P contains all but a finite number of the points $P_1, P_2, P_3, \ldots$.*

For some purposes it is convenient to rephrase this definition in terms of distances between points. Let $d(P, Q)$ denote the distance between P and Q. This distance can be expressed in terms of the co-ordinates of P and Q, but just now it is more convenient to deal directly with the points and not with the co-ordinates. The meaning of $\lim_{n\to\infty} P_n = P$ is the following: If $\epsilon > 0$, there is some integer N such that $d(P_n, P) < \epsilon$ if $N \leqq n$.

From this definition it can easily be shown that, if $\{P_n\}$ is a sequence in the xy-plane, with co-ordinates (x_n, y_n), and if P is the point (x, y), then P_n converges to P if and only if x_n converges to x and y_n converges to y. A like situation obtains in three dimensions.

The relations between convergent sequences and points of accumulation are essentially the same in higher dimensions as in the case of points restricted to the x-axis. The Theorems III, IV, V of §16.31 all remain true in higher dimensions. The proofs as given in §16.31 remain valid with only a few minor modifications in notation to suit the higher dimensional cases.

EXERCISES

1. Let $\{P_n\}$ be the sequence $([1-(-1)^n]/n, [1+(-1)^n]/n)$. Is it convergent?

2. Let S be the set of points on the curve $y = \sin(1/x)$ $(x \neq 0)$ in the xy-plane. **(a)** Find a sequence of points P_n on the x-axis and in S such that P_n converges to $(0, 0)$. **(b)** Find a

sequence of points P_n on the line $y = 1$ and in S such that P_n converges to $(0, 1)$. (c) What points must be adjoined to S in order to get a closed set?

3. Let S be the set of points (x, y) such that x is rational and $0 \leqq x \leqq 1$ while y is irrational and $0 < y < 1$. Is S open, closed, or neither? Describe the totality of points of accumulation of S.

16.5 / CAUCHY'S CONVERGENCE CONDITION

The definition of convergence of a sequence is stated in a way which involves the limit of the sequence. Therefore one cannot use the definition to prove that a sequence is convergent unless one already knows what the limit is. This is often inconvenient, for it frequently happens that we want to prove that a sequence is convergent, even if we do not know precisely what the limit is. The following theorem provides a way out of the difficulty.

THEOREM VI. *Let $\{P_n\}$ be a sequence of points. Let $d(P_m, P_n)$ denote the distance between P_m and P_n. A necessary and sufficient condition that the sequence be convergent is that $d(P_m, P_n)$ approach 0 as m and n become infinite. In other words, for each $\epsilon > 0$ there must be some N such that $d(P_m, P_n) < \epsilon$ whenever $N \leqq m$ and $N \leqq n$.*

We note that in the case of a sequence $\{x_n\}$ of points on the x-axis, the distance is

$$d(x_m, x_n) = |x_m - x_n|,$$

so that the condition becomes

$$|x_m - x_n| < \epsilon \qquad \text{if} \qquad N \leqq m \qquad \text{and} \qquad N \leqq n.$$

This condition on the sequence is called *Cauchy's condition* (see the reference to Cauchy in §1.61).

Proof that the condition is necessary. If $P_n \to P$ as $n \to \infty$, and $\epsilon > 0$, choose N so large that $d(P_n, P) < \epsilon/2$ if $N \leqq n$. This is possible, by the definition of convergence. If now $m \geqq N$ and $n \geqq N$, we have $d(P_m, P_n) \leqq d(P_m, P) + d(P, P_n) < (\epsilon/2) + (\epsilon/2) = \epsilon$, for the three points P_m, P_n, P lie either on a straight line or at the vertices of a triangle, and in either case the distance between one pair is not greater than the sum of the distances between the other two pairs. For points on a line we could write

$$|x_m - x_n| \leqq |x_m - x| + |x - x_n|,$$

which is an application of the inequality (2.2–9) with $a = x_m - x$, $b = x - x_n$.

Proof that the condition is sufficient. We now assume that Cauchy's condition is satisfied by the sequence $\{P_n\}$. First we shall prove that the sequence is bounded. Taking $\epsilon = 1$, Cauchy's condition assures us that there is some N

such that $d(P_m, P_n) < 1$ if m and n are not less than N. In particular, $d(P_N, P_n) < 1$ if $n \geqq N$. This means that all points of the sequence with the possible exception of $P_1, \ldots, P_{N-1}$ are less than unit distance from P_N. Certainly, then, there is some neighborhood with center at P_N large enough to contain *all* the points of the sequence, and the sequence must be bounded.

We now apply Theorem V, §16.31 (the extension of this theorem to higher dimensions was discussed in §16.41). The sequence $\{P_n\}$ contains a convergent subsequence $P_{n_1}, P_{n_2}, \ldots$. Let the limit of this subsequence be P. We shall complete the proof by showing that the original sequence is convergent to the limit P. Suppose $\epsilon > 0$. Choose N so that $d(P_m, P_n) < \epsilon/2$ if $m \geqq N$ and $n \geqq N$ (Cauchy's condition). Since the subsequence converges to P, we can choose one of the indices n_i so large that $d(P_{n_i}, P) < \epsilon/2$. We can also, at the same time, choose it larger than N. Then, if $n \geqq N$, we have

$$d(P_n, P) \leqq d(P_n, P_{n_i}) + d(P_{n_i}, P) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

by the way things have been arranged. Here we have used the same kind of inequality involving the distances between three pairs of points that we encountered in the necessity argument. The proof of Theorem VI is now complete.

Theorem VI is most often used as a tool for carrying on theoretical developments.

EXERCISE

Let a sequence be defined on the x-axis as follows: $x_1 = 1$, $x_2 = 2$, $x_3 = \frac{1}{2}(x_1 + x_2)$, and in general $x_{n+1} = \frac{1}{2}(x_{n-1} + x_n)$, $n = 2, 3, \ldots$. Show that $|x_m - x_n| \leqq 1/2^{N-1}$ if $N \leqq m$ and $N \leqq n$, so that Cauchy's condition is fulfilled. SUGGESTION: Note that each term is midway between the two preceding terms.

16.6 / THE HEINE-BOREL THEOREM

In this section we are concerned with an important theorem of point-set theory which is widely known by the names of the two mathematicians Eduard Heine and Émile Borel. The central idea of this theorem probably originated in connection with the concept of *uniform continuity*. Heine defined this concept, and proved the very important theorem that if a function of one variable x is continuous at each point of a closed interval $a \leqq x \leqq b$, then it is uniformly continuous on that interval. We shall discuss the concept and the theorem in Chapter 17, at which time we shall need to appeal to the Heine-Borel theorem. You will doubtless find difficulty at first in appreciating the motivation of the Heine-Borel theorem; it may be best that you take up the study of the theorem when you need it in reading parts of Chapter 17. But since the theorem itself is a theorem of point-set theory, we put the exposition of it here in Chapter 16.

Before stating the theorem we make a definition and consider some illustrations to motivate the theorem.

Definition. *Let S be a point set, and suppose we have a collection of a certain number of open sets such that each point of S belongs to at least one of the open sets. Then we say that S is* covered *by the collection of open sets.*

The number of open sets in the collection may be either finite or infinite.

Example 1. Let S be the set of all points on the x-axis such that $0 < x \leqq 1$. Let the collection of open sets be the sequence of open intervals

$$(\tfrac{1}{2}, \tfrac{3}{2}), (\tfrac{1}{4}, 1), (\tfrac{1}{8}, \tfrac{5}{8}), (\tfrac{1}{16}, \tfrac{3}{8}), \ldots,$$

the nth interval being $I_n = (1/2^n, (n+2)/2^n)$. These intervals are shown schematically in Fig. 166. The set S is covered by this collection, for every x such that $0 < x \leqq 1$ lies in I_n for some value of n.

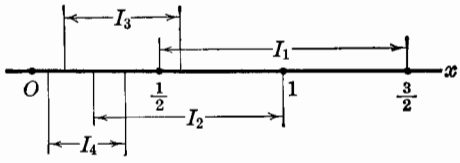


Fig. 166.

Example 2. Let S be the set of points $0, \frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \ldots, 1/(2n), \ldots$ on the x-axis. Let the collection of open sets be the open intervals

$$(-\tfrac{1}{10}, \tfrac{1}{10}), (\tfrac{1}{3}, 1), (\tfrac{1}{5}, \tfrac{1}{3}), (\tfrac{1}{7}, \tfrac{1}{5}), \ldots, \left(\frac{1}{2n+1}, \frac{1}{2n-1}\right), \ldots.$$

Note that S is covered by the collection of intervals.

Example 3. Let S be the set of all points on the x-axis such that $0 < x \leqq 2$. Consider the function $f(x) = 1/x$, and define a collection of open intervals as follows: Suppose $0 < \epsilon < \frac{1}{2}$. If x_0 is in S, denote by $I(x_0)$ the set of all x such that $x > 0$ and $|f(x) - f(x_0)| < \epsilon$, that is, the set of all $x > 0$ such that

$$\frac{1}{x_0} - \epsilon < \frac{1}{x} < \frac{1}{x_0} + \epsilon.$$

It is clear from Fig. 167 that $I(x_0)$ is the open interval from x_0' to x_0'' where x_0' and x_0'' are found by solving the equations

$$\frac{1}{x_0'} = \frac{1}{x_0} + \epsilon, \quad \frac{1}{x_0''} = \frac{1}{x_0} - \epsilon.$$

One easily finds

$$x_0' = \frac{x_0}{1 + \epsilon x_0}, \quad x_0'' = \frac{x_0}{1 - \epsilon x_0}.$$

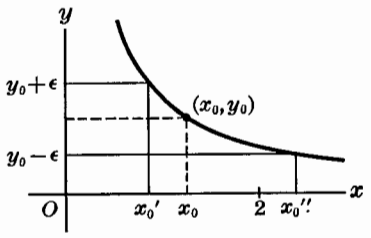


Fig. 167.

Clearly $I(x_0)$ contains x_0. Consequently, as x_0 varies over S we get a collection of intervals $I(x_0)$ which covers S.

The Heine-Borel theorem is concerned with this question: If S is covered by a collection of open sets, under what conditions on S is it possible to choose a *finite* number of open sets from the collection in such a way that the new finite collection still covers S? Let us examine the preceding examples with this question in mind.

In Example 1 no finite number of the open intervals will suffice to cover S, for the first n intervals leave uncovered all the points x of S for which $0 < x \leqq 1/2^n$. Example 2 is different. The point O is covered by the special open interval $(-\frac{1}{10}, \frac{1}{10})$. This same interval also covers all but a finite number of the remaining points of S, namely, all points $1/(2n)$ with $1/(2n) < \frac{1}{10}$, or $n = 6, 7, 8, \ldots$. The remaining points are $\frac{1}{2}, \frac{1}{4}, \ldots, \frac{1}{10}$, and these are covered by the five intervals $(\frac{1}{5}, 1)$, $(\frac{1}{5}, \frac{1}{3}), \ldots, (\frac{1}{11}, \frac{1}{9})$. Hence six intervals suffice to cover S in Example 2. In Example 3 no finite selection of the intervals $I(x_0)$ will suffice to cover S, for, of any finite number of the $I(x_0)$, there will be one for which x_0' is furthest to the left, and this will leave all the x such that $0 < x \leqq x_0'$ uncovered.

We now proceed directly to a formal statement and proof of the Heine-Borel theorem.

THEOREM VII. *Let S be a bounded and closed point set, and let S be covered by a collection of open sets. Then a finite number of open sets may be chosen from the collection in such a way that S is covered by the new finite collection.*

All our illustrative examples were given for point sets on a line. However, the theorem holds equally for point sets on a line, in the plane, or in space. It is merely necessary to bear in mind the proper definitions of open and closed sets for each dimensionality.

Proof of the theorem. We shall give the proof for point sets on a line. But the method of proof may be immediately applied in higher dimensions.

The set S will lie on some closed interval $a \leqq x \leqq b$, since it is bounded. If S is a finite set there is no problem; we have only to choose one open set to cover each point of S, and this will be a finite collection. Hence let us assume that S is an infinite point set. We make the assumption that no finite number of the open sets of the given collection will suffice to cover S. From this assumption we shall deduce a contradiction, and thus the theorem will be proved.

We bisect the interval $[a, b]$ and consider the parts of S lying in each of the two closed subintervals. If each of these parts could be covered by a finite subcollection of the open sets, so could S itself. Therefore, for at least one of the subintervals the corresponding part of S cannot be covered by a finite number of the given open sets. We bisect this interval and proceed as before. Each time we bisect, an infinite number of points of S must belong in the

subinterval which is retained, for otherwise this part of S could surely be covered by a finite number of the open sets. The repeated bisection process gives us a nest of closed intervals, and there is a unique point x_0 common to all the intervals of the nest (Theorem VI, §2.8). This point x_0 is clearly an accumulation point of S. It must therefore belong to S (Theorem I, §16.2), because S is closed. Since x_0 is in S, there is some open set of the collection which contains x_0. This open set must therefore contain all except a finite number of the closed intervals of the nest which is shrinking down on x_0. This brings us to a contradiction, for each interval of the nest has the property that the part of S in it cannot be covered by a finite number of the given open sets. As explained earlier, our arrival at this contradiction completes the proof.

EXERCISE

Let S be the set $1 \leqq x \leqq 2$, and define the open intervals $I(x_0)$ as in Example 3. Show that S can be covered with a finite number N of these intervals, where N is the smallest integer exceeding $1/\epsilon$. SUGGESTION: Consider the intervals $I(x_0)$ for $x_0 = 1,\ 1+\epsilon,\ 1+2\epsilon, \ldots$.

17 / FUNDAMENTAL THEOREMS ON CONTINUOUS FUNCTIONS

17 / PURPOSE OF THE CHAPTER

In Chapter 3, we proved several important theorems about continuous functions, confining ourselves always to the case of a single real variable. Analogues of some of these theorems for functions of two variables were considered in §5.3, but no proofs were given. With the aid of the concepts and theorems developed in Chapter 16 we are now in a position to make a deeper study of continuity.

One important new concept will be introduced in this chapter: the concept of uniform continuity. This is needed for the theory of integration, in Chapter 18.

17.1 / CONTINUITY AND SEQUENTIAL LIMITS

For the definition of continuity of a function of one variable, we refer the student to §3. There is an alternative way of expressing continuity, using the notion of a convergent sequence. We express this as a theorem.

THEOREM I. *Suppose f is a function of x, defined on some interval I. Let x_0 be a point of I. Then f is continuous at x_0 if and only if $\lim_{n\to\infty} f(x_n) = f(x_0)$ for every sequence $\{x_n\}$ which has terms that belong to I and which is such that $\lim_{n\to\infty} x_n = x_0$.*

We shall not prove this theorem, because we are presently going to prove a more general theorem of which Theorem I is a special case.

In dealing with continuity it is quite worth while to arrange matters so that what we say applies just as well to functions of two or three variables as to functions of one variable. The terminology of point-set theory enables us to do this. In particular, the concept of continuity can be defined in terms of the concept of neighborhood, without explicit mention of ϵ and δ or use of inequalities.

Definition. *Let S be a point set, and let f be a function defined at each point of S. If P_0 is a certain point of S, the function is said to be continuous at P_0 provided that to each neighborhood V of the value $f(P_0)$ there corresponds a neighborhood*

U of the point P_0 such that the value $f(P)$ lies in the neighborhood V whenever the point P lies in the set S and also in the neighborhood U.

This definition is equivalent to the previously given definitions (in §3 and §5.3), though it is somewhat different in form from these earlier definitions. For example, if f is a function of one real variable, we may take the set S to be on the x-axis, and denote the point by x instead of P. The neighborhood V of $f(x_0)$ will be an interval defined by $|y - f(x_0)| < \epsilon$, or $f(x_0) - \epsilon < y < f(x_0) + \epsilon$, where ϵ is some positive number. The neighborhood U will be an interval defined by $|x - x_0| < \delta$, or $x_0 - \delta < x < x_0 + \delta$, where δ is some positive number. The statement that $f(x)$ is in V if x is in U is then the same as the statement that $|f(x) - f(x_0)| < \epsilon$ if $|x - x_0| < \delta$.

One thing which deserves to be emphasized is that the point set S on which f is defined is not subject to any restrictions. Thus, if we are talking about a function of two variables, and P has co-ordinates (x, y), the point set S can be any kind of point set in the plane. It does not need to be a region (as defined in §5.1). For example, S could be the set of points on the parabola $y = x^2$, and $f(x, y)$ might be the radius of curvature at the point (x, y). Or, in three dimensions, S might be the surface of an ellipsoid, and $f(x, y, z)$ might be the distance from (x, y, z) to the origin.

We now come to the generalization of Theorem I.

THEOREM II. *Suppose f is a function defined at each point of the set S. If P_0 is a particular point of S, the function is continuous at P_0 if and only if $\lim_{n\to\infty} f(P_n) = f(P_0)$ whenever $\{P_n\}$ is a sequence of points belonging to S such that $\lim_{n\to\infty} P_n = P_0$.*

Proof. Suppose f is continuous at P_0, and suppose P_n belongs to S and P_n converges to P_0. If V is any neighborhood of $f(P_0)$, we have to show that $f(P_n)$ is in V when n is sufficiently large. Now, since f is continuous at P_0, there is a neighborhood U of P_0 such that $f(P)$ is in V if P is in S and U. The fact that P_n converges to P_0 means that P_n is in U if n is sufficiently large. But then $f(P_n)$ must be in V, and one part of the proof is complete.

We now assume that $f(P_n)$ converges to $f(P_0)$ whenever $\{P_n\}$ is a sequence in S converging to P_0. We shall make the required proof by supposing that f is not continuous at P_0, and deducing a contradiction. The denial of continuity at P_0 may be phrased in this way: There is *some* neighborhood V of $f(P_0)$ such that, no matter what neighborhood U of P_0 is selected, some point P in U has the property that P is in S but $f(P)$ is not in V. Accordingly, let us choose a succession of neighborhoods $U_1, U_2, \ldots, U_n, \ldots$ closing down on P_0 (e.g., U_n consisting of all points at distance less than $1/n$ from P_0), and let P_n be a point in U_n with properties as described above. Then P_n must converge to P_0, but $f(P_n)$ does not converge to $f(P_0)$, because $f(P_n)$ is always outside the neighborhood V. This contradicts our initial assumption. Therefore the proof is complete.

17.2 / THE BOUNDEDNESS THEOREM

We give a generalization of Theorem II, §3.1, and Theorem I, §5.3.

THEOREM III. *Let S be a bounded and closed point set, and let f be a function defined on S which is continuous at each point of S. Then the values of f are bounded, or, as we customarily say, f is bounded on S.*

The proof is rather simple. Suppose the values of f were not bounded. Then no finite interval of the real number scale contains all the values, and for each positive integer n there must be at least one point P_n in S such that $|f(P_n)| > n$. The sequence $\{P_n\}$ is bounded, and must therefore contain a convergent subsequence (see Theorem V, §16.31, and the remarks about it in §16.41). Denote the subsequence by $\{P_{n_i}\}$ and its limit by Q. Since S is closed, Q belongs to S. Then, since f is continuous at Q, $f(P_{n_i})$ converges to $f(Q)$, by Theorem II, §17.1. This is in contradiction to $|f(P_{n_i})| > n_i$, which has the consequence that $|f(P_{n_i})| \to \infty$ as $i \to \infty$. To avoid the contradiction we are forced to conclude that f is bounded.

17.3 / THE EXTREME-VALUE THEOREM

We give a generalization of Theorem III, §3.2, and Theorem II, §5.3.

THEOREM IV. *Let S be a non-empty bounded and closed set, and suppose f is a function defined on S and continuous at each point of S. Let m and M be the greatest lower bound and least upper bound, respectively, of the values of f on S. Then there is some point of S at which f has the value M, and there is also a point at which f has the value m.*

The existence of m and M is guaranteed by the theorem in §17.2 and Theorem II, §2.7.

We begin the proof by observing that since M is the least upper bound of the values $f(P)$, there must exist a sequence $\{P_n\}$ in S such that $f(P_n)$ converges to M. The sequence $\{P_n\}$ is bounded, and just as in the proof of Theorem III we obtain a subsequence $\{P_{n_i}\}$ converging to a point Q, with $f(P_{n_i})$ converging to $f(Q)$. But $f(P_{n_i})$ also converges to M, so $M = f(Q)$. This proves the theorem as regards M. The proof for m is essentially the same.

17.4 / UNIFORM CONTINUITY

When we say that a function is continuous at a certain point, this describes a certain relation between the values of the function and the values of the independent variable near a particular value of the independent variable. The new concept of *uniform continuity*, which we are concerned with in this section, has to do with continuity of a function at many different points. The word

"uniform" is used because, when we say that a function is uniformly continuous on a certain point set, this means that there is a certain quality about the continuity which is the same at all the points of the set. Before giving a formal definition of uniform continuity we shall have to see in what sense there can be recognizable differences in the continuity of a function at different points.

The type of difference we have in mind has to do with the amount by which the independent variable may be allowed to change if the value of the function is not permitted to change more than a specified amount.

Example 1. Consider the function $f(x) = x^2$. Suppose we start with some particular value x_0, and ask: How much may x differ from x_0 if $f(x)$ is required to differ from $f(x_0)$ by less than 2 units? The answer depends on the value of x_0. If $x_0 = 0$, the requirement on x is that $x^2 < 2$, or $|x| < \sqrt{2}$, so that x must differ from x_0 by less than $\sqrt{2}$. But if $x_0 = 8$ the requirement is that $|x^2 - 64| < 2$, or $62 < x^2 < 66$. Now $\sqrt{66} = 8.124\ldots$, $\sqrt{62} = 7.874\ldots$, so that the permissible difference between x and 8 is less than $\sqrt{66} - 8 = 0.124\ldots$. The student will see at once from a graph of $y = x^2$ that the larger we make x_0, the smaller is the amount by which x may differ from x_0 if x^2 is to differ from x_0^2 by less than 2 units.

If instead of 2 units we specify ϵ units, where $\epsilon > 0$, the situation is essentially the same. If we want to find a number δ so that $|x^2 - x_0^2| < \epsilon$ whenever $|x - x_0| < \delta$ the value of δ will depend on x_0 as well as on ϵ, and with ϵ fixed, δ must be made smaller and smaller as x_0 gets larger. There is no single positive value of δ which is small enough to serve simultaneously for *all* values of x_0. This illustrates the *opposite* of uniform continuity.

Definition. *Let S be a point set on the x-axis, forming part or all of the set of values of x for which $f(x)$ is defined. Then f is said to be* uniformly continuous *on S provided that to each $\epsilon > 0$ there corresponds a $\delta > 0$ such that $|f(x) - f(x_0)| < \epsilon$ whenever x and x_0 are any points of S such that $|x - x_0| < \delta$.*

Let us carry the discussion of Example 1 further in the light of this definition. The function $f(x) = x^2$ is *not* uniformly continuous on the set S of all x. In fact if S is any set which is not bounded, the function is not uniformly continuous on S, for if there is no limit to how large x_0 can be, no single δ can be found which meets the requirements of the definition. On the other hand, if S is a bounded set, the function *is* uniformly continuous on S. For instance, if S is the closed interval $0 \leqq x \leqq 8$, we can take $\delta = \epsilon/16$ in the definition of uniform continuity. To see this, observe that if x and x_0 belong to S,

$$|x^2 - x_0^2| = |(x - x_0)(x + x_0)| \leqq 16|x - x_0|$$

and so

$$|x^2 - x_0^2| < \epsilon \quad \text{if} \quad |x - x_0| < \frac{\epsilon}{16}.$$

Example 2. The function $f(x) = 1/x$ is not uniformly continuous on the set S

defined by $0 < x \leqq 1$, but it *is* uniformly continuous on the set S defined by $x \geqq 1$. The first assertion follows from the fact that if $x_0 > 0$ and δ is chosen so that

$$\left|\frac{1}{x} - \frac{1}{x_0}\right| < \epsilon \quad \text{if} \quad |x - x_0| < \delta,$$

the value of δ must approach 0 as $x_0 \to 0$. On the other hand, for the set S defined by $x \geqq 1$ we can take $\delta = \epsilon$, because if x and x_0 belong to S and $|x - x_0| < \epsilon$ we have

$$\left|\frac{1}{x} - \frac{1}{x_0}\right| = \frac{|x_0 - x|}{xx_0} \leqq |x - x_0| < \epsilon.$$

The essential theorem about uniform continuity will now be given.

THEOREM V. *Suppose S is a closed and bounded point set, and suppose the function f is defined and continuous at each point of S. Then f is uniformly continuous on S.*

Proof. We make use of the Heine-Borel theorem. Suppose $\epsilon > 0$. If x' is any point of S, the definition of continuity assures us that there is some positive number h such that $|f(x) - f(x')| < \epsilon/2$ if x and x' belong to S and $|x - x'| < h$. The size of h will usually vary as x' is varied. Now consider the open interval $x' - (h/2) < x < x' + (h/2)$. When x' varies over S, the collection of all these open intervals covers the set S. By the Heine-Borel theorem (§16.6) a finite number of these intervals suffice to cover S. Let the centers of these intervals be denoted by $x_1, \ldots, x_n$ and let the corresponding values of h be $h_1, \ldots, h_n$. Choose δ as the smallest of the numbers $h_1/2, \ldots, h_n/2$. We shall show that this δ will serve as required in the definition of uniform continuity. Suppose x and x_0 belong to S and $|x - x_0| < \delta$. Then x_0 belongs to one of the finite set of open intervals, say the one with end points $x_i \pm (h_i/2)$, so that $|x_0 - x_i| < h_i/2$. Now

$$|x - x_i| \leqq |x - x_0| + |x_0 - x_i| < \delta + \frac{h_i}{2}.$$

But $\delta \leqq \dfrac{h_i}{2}$, and so $|x - x_i| < h_i$. The inequalities satisfied by $|x_0 - x_i|$ and $|x - x_i|$ guarantee that

$$|f(x_0) - f(x_i)| < \frac{\epsilon}{2} \quad \text{and} \quad |f(x) - f(x_i)| < \frac{\epsilon}{2}.$$

Therefore $|f(x) - f(x_0)| \leqq |f(x) - f(x_i)| + |f(x_i) - f(x_0)| < \epsilon$.
This completes the proof.

The definition of uniform continuity can be worded so as to apply to functions of more than one variable. It is merely necessary to write the condition involving ϵ and δ in the form "$|f(P) - f(P_0)| < \epsilon$ whenever P and P_0 are points of S such that $d(P, P_0) < \delta$." Here $d(P, P_0)$ is the distance between P and P_0.

Theorem V remains valid for functions of several variables, and the proof as given above can be adapted easily to the new situation by using distances in place of absolute values.

17.5 / CONTINUITY OF SUMS, PRODUCTS, AND QUOTIENTS

The process of addition may be considered as defining a function of two variables:

$$s(x, y) = x + y.$$

The same may be said of multiplication and division:

$$p(x, y) = xy, \quad q(x, y) = \frac{x}{y}.$$

We use the letters s, p, q for these three functions because of the names sum, product, quotient. The following theorem states a fundamental fact about these functions:

THEOREM VI. *The functions s and p are continuous at all points (x, y). The function q is continuous at all points where it is defined (i.e., at all points except those for which $y = 0$).*

The proof of this theorem is very similar to the proof of Theorem XIV, §1.64, and on that account we omit the formal proof. Some suggestions for the student who wishes to work out the details are given in the following exercises.

EXERCISES

1. Show that $|s(x, y) - s(x_0, y_0)| < \epsilon$ if (x, y) is in the square neighborhood of side ϵ with center at (x_0, y_0), and so certainly if (x, y) is in the circular neighborhood of radius $\epsilon/2$ with center at (x_0, y_0).

2. Let M be a number larger than $|x_0|$ and $|y_0|$. Let δ be the smaller of the numbers $M - |y_0|$, $\epsilon/(2M)$. Show that $|p(x, y) - p(x_0, y_0)| < \epsilon$ if $|x - x_0| < \delta$ and $|y - y_0| < \delta$. Start from the fact that $xy - x_0y_0 = (x - x_0)y + x_0(y - y_0)$ and so $|xy - x_0y_0| \leqq |x - x_0|\,|y| + |x_0|\,|y - y_0|$.

3. Study the proof of (1.64–4) and so show that $|q(x, y) - q(x_0, y_0)| < \epsilon$ if $|x - x_0| < \delta$ and $|y - y_0| < \delta$, where δ is the smaller of the numbers $\frac{1}{2}|y_0|$, $\dfrac{|y_0|^2\epsilon}{2(|x_0| + |y_0|)}$. Assume that $y_0 \neq 0$.

17.6 / PERSISTENCE OF SIGN

We shall now give a generalization of Theorem V, §3.3.

THEOREM VII. *Suppose f is continuous at a point P_0 of the set S on which it is defined, and suppose $f(P_0) \neq 0$. Then there is some neighborhood of P_0 such*

that at all points of S in this neighborhood the sign of $f(P)$ is the same as the sign of $f(P_0)$.

The truth of this theorem depends merely on the fact that $f(P)$ is near $f(P_0)$ if P is near P_0. If $f(P_0) \neq 0$, all values of $f(P)$ sufficiently near $f(P_0)$ will have the same sign as $f(P_0)$.

17.7 / THE INTERMEDIATE-VALUE THEOREM

An intermediate-value theorem was stated as Theorem IV, §3.3. We shall now consider how this theorem is to be generalized so as to apply to functions of more than one variable. The problem is of this nature: Suppose a function is defined on some point set, and is continuous at each point. If the function takes on the value 2 at one point and the value -3 at another point, does it also take on all values between -3 and 2? An example will show that such is not always true. For instance, let $f(x, y) = 1/(xy)$. Then $f(2, \frac{1}{4}) = 2$, $f(-\frac{1}{6}, 2) = -3$. But $f(x, y)$ never takes on the value 0, and 0 is between -3 and 2, in spite of the fact that f is continuous at each point of the set where it is defined. The explanation of the difficulty lies in the fact that the points $(2, \frac{1}{4})$ and $(-\frac{1}{6}, 2)$ are separated by the line $x = 0$, on which f is not defined.

If f is defined and continuous on a set S which is not separated into several disconnected parts, it can be proved that f has the property of taking on all values between every pair of distinct values. In order to make this assertion in an exact form, and prove it, we must first make some definitions.

Definition. *Two point sets S_1 and S_2 are called* separated *if the following three conditions hold:*

1. *Neither set is empty.*
2. *No point belongs to both sets.*
3. *Neither set has a point of accumulation belonging to the other set.*

Example 1. The intervals $-1 < x < 0$ and $0 < x < 1$ are separated sets. The interior and exterior of the circle $x^2 + y^2 = 1$ are separated sets in the plane. The points on opposite sides of the plane $z = 0$ form two separated sets in space of three dimensions.

Definition. *A set S is called* connected *if it cannot be divided into two parts which are separated.*

Example 2. An interval on the x-axis, with or without either end point, is a connected set. The first quadrant in the xy-plane ($x > 0$, $y > 0$) is a connected set. The surface of a sphere is a connected set.

If S is an open set which is connected, it may be shown that any two points of S can be joined by an arc which lies entirely in S. In fact, the arc may be

taken to consist of a finite number of line segments joined end to end. In §7.4 we *defined* an open set to be connected if it had this property. For open sets the two definitions are equivalent, but for sets which are not open the present definition is the proper one to use.

We now come to the general intermediate-value theorem.

THEOREM VIII. *Suppose S is a connected set and that f is a function which is continuous at each point of S. Suppose f takes on two different values C_1 and C_2 at points P_1 and P_2 in S. Then, for every number k between C_1 and C_2, there is some point of S at which f takes on the value k.*

Proof. Let us suppose the notation is such that $C_1 < k < C_2$. We shall suppose that for some particular k there is no point P such that $f(P) = k$. We shall show that with this assumption we can divide S into two separated parts. This contradiction will complete the proof of the theorem.

Let S_1 be the set of those points of S for which $f(P) < k$, and S_2 the set for which $k < f(P)$. Observe that P_1 is in S_1, P_2 is in S_2, and each point of S belongs either to S_1 or to S_2. The sets S_1 and S_2 are separated. In fact, since conditions (1) and (2) in the definition of separated sets are clearly satisfied, we have only to verify the third condition. Now, it is impossible for a point of accumulation of S_2 to belong to S_1. For suppose Q were such a point. Then we could select a sequence $\{Q_n\}$ of points in S_2 such that Q_n converges to Q. Then $f(Q_n)$ converges to $f(Q)$, by continuity. But $f(Q_n) > k$, since Q_n is in S_2. Therefore the limit $f(Q)$ cannot be less than k. But $f(Q) < k$ since Q was assumed to be in S_1. This shows that no such point as Q can exist. The same argument shows that no point of accumulation of S_1 can belong to S_2. This proves that S_1 and S_2 are separated, and the proof is complete.

18 / THE THEORY OF INTEGRATION

18 / THE NATURE OF THE CHAPTER

Heretofore, in considering integrals defined as limits of sums, we have taken it for granted that the limits do exist. The applications of integrals in calculating various geometrical and physical magnitudes are commonly of such a kind that we consider integrals of continuous functions, and it is plausible, from an intuitive standpoint, that the process of defining the integral of a continuous function does actually lead to a definite limit.

For certain purposes it becomes desirable and necessary to consider the integration of functions which may be discontinuous. Consequently it is necessary to lay down the fundamental definitions about integrals in such a way that the theory can be developed without depending on hypotheses about continuity. In this chapter we shall lay the groundwork of a general theory of integration. The form of this theory goes back to Bernhard Riemann (1826–1866), a German mathematician. The first purpose of this chapter is to set forth clearly the concept of the Riemann integral. A function which can be integrated according to Riemann's definition is called *integrable.* We show that continuous functions and certain kinds of discontinuous functions are integrable.

In certain ways the theory of multiple integrals is much more complicated than the theory of integrals of functions of one variable. Our discussion of multiple integrals is designed mainly to lead up to an analytical treatment of the relation between a multiple integral and its evaluation by iterated integrals. This is a topic on which we have not aimed to give the greatest possible generality, but a reasonably simple and comprehensible treatment of the kind of multiple integrals which ordinarily arise in calculus.

The chapter concludes with a very brief, conceptual, nontheoretical discussion of Stieltjes integrals.

18.1 / THE DEFINITION OF INTEGRABILITY

Suppose that a, b are any real numbers such that $a < b$. We shall consider functions defined on the closed interval $[a, b]$. We shall not require the functions to be continuous, but we shall assume that each function is bounded.

Now let us subdivide the interval into any number of subintervals. This is done by inserting points between a and b. Thus, suppose

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b.$$

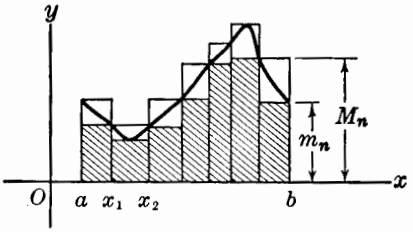


Fig. 168.

Then we have n subintervals

$$[x_0, x_1], [x_1, x_2], \ldots, [x_{n-1}, x_n].$$

We shall refer to such a subdivision as a *partition* of $[a, b]$. The partition determined by these particular points will be called the partition $(x_0, x_1, \ldots, x_n)$.

In the closed subinterval $[x_{i-1}, x_i]$ let m_i and M_i denote respectively the greatest lower bound and least upper bound of the values $f(x)$. Then we form the two sums

$$s = m_1(x_1 - x_0) + m_2(x_2 - x_1) + \cdots + m_n(x_n - x_{n-1}), \tag{18.1-1}$$

$$S = M_1(x_1 - x_0) + M_2(x_2 - x_1) + \cdots + M_n(x_n - x_{n-1}). \tag{18.1-2}$$

We call s a *lower sum* and S an *upper sum*. Observe that these sums are dependent upon the particular function f and the particular partition.

Let us denote the greatest lower bound and least upper bound of $f(x)$ on all of $[a, b]$ by m, M respectively. Then $m \leqq m_i$ and $M_i \leq M$ for $i = 1, 2, \ldots, n$. Of course it is also true that $m_i \leqq M_i$. We observe that

$$(x_1 - x_0) + (x_2 - x_1) + \cdots + (x_n - x_{n-1}) = b - a.$$

Therefore, it is readily seen that

$$m(b - a) \leqq s \leqq S \leqq M(b - a). \tag{18.1-3}$$

The next step in our procedure is motivated by geometry. Suppose for a moment that the function f is continuous and that all its values are positive. The lower sum s then represents the area of a sum of rectangles all of which lie between the x-axis and the curve $y = f(x)$ (see Fig. 168). It then seems plausible to suppose that if we consider all possible partitions and the corresponding lower sums, the least upper bound of all these lower sums s will be exactly the area between the x-axis and the curve and between $x = a$ and $x = b$. It likewise seems plausible to suppose that this area is the greatest lower bound of all the upper sums S when all possible partitions are considered (again, see Fig. 168).

Definition. *As a result of the foregoing considerations, let us define*

$$I = \textit{least upper bound of all lower sums } s,$$

$$J = \textit{greatest lower bound of all upper sums } S.$$

Then we shall say that the function f is integrable *on* $[a, b]$ *if* $I = J$, *and in that case we shall call the common value of* I *and* J *the* definite integral *of* f *from* $x = a$ *to* $x = b$, *and denote it by*

$$\int_a^b f(x)\,dx.$$

This definition applies to each function which is defined and bounded on $[a, b]$. For each such function we get unique values for I and J, but it may happen that $I \neq J$. It is only when $I = J$ that the function is integrable.

Example 1. Consider the function defined on [0, 1] by setting $f(x) = 0$ if x is rational and $f(x) = 1$ if x is irrational. For this function and this interval $I = 0$, $J = 1$, so that the function is not integrable. To see the truth of this assertion, note that any subinterval contains both rational and irrational values of x, and therefore that $m_i = 0$, $M_i = 1$. Consequently, from the definitions of s and S we see that $s = 0$, $S = 1$, no matter how the partition is chosen. Accordingly, $I = 0$, $J = 1$.

The student will observe that the foregoing definition of $\int_a^b f(x)\,dx$ is not the same as the definition given in §1.5. The definition which we are now considering makes no reference to the concept of a limit of approximating sums. Later on in this chapter we shall see that the two different forms of definition are equivalent. Our first concern, however, is to develop enough theory to furnish a practical means of telling when a function is integrable.

LEMMA 1. *Suppose we start with a certain partition, and then obtain a new partition by inserting some additional points. Let* s, S *refer to the lower and upper sums for the original partition, while* s', S' *are the sums for the new partition. Then* $s \leqq s'$ *and* $S' \leqq S$.

For simplicity we shall prove this on the supposition that just one new point is inserted. Suppose for definiteness that the new point ξ is between x_0 and x_1. Let

$$M_1' = \text{least upper bound of } f(x) \text{ for } x_0 \leqq x \leqq \xi,$$

$$M_1'' = \text{least upper bound of } f(x) \text{ for } \xi \leqq x \leqq x_1.$$

Then certainly $M_1' \leqq M_1$ and $M_1'' \leqq M_1$. Therefore

$$M_1'(\xi - x_0) + M_1''(x_1 - \xi) \leqq M_1(\xi - x_0) + M_1(x_1 - \xi) = M_1(x_1 - x_0).$$

From this it follows that $S' \leqq S$, because S' differs from S only in having $M_1'(\xi - x_0) + M_1''(x_1 - \xi)$ in place of $M_1(x_1 - x_0)$. The proof that $s \leqq s'$ is similar. If more than one point is inserted we need only apply the argument several times.

LEMMA II. *For a given function and any two partitions, the lower sum for one partition is algebraically less than or equal to the upper sum for the other partition.*

Proof. Denote the sums corresponding to the two partitions by s_1, S_1 and s_2, S_2. Now consider the third partition which is obtained by considering simultaneously all the points of subdivision of the first two partitions; denote the corresponding sums by s_3, S_3. We know by (18.1–3) that $s_3 \leqq S_3$. Also, we know by Lemma I that $s_1 \leqq s_3$ and $S_3 \leqq S_2$. Therefore, combining the inequalities, we see that $s_1 \leqq S_2$. This is what Lemma II asserts.

LEMMA III. *It is always true that* $I \leqq J$.

This is an immediate consequence of Lemma II and the definitions of I and J. We can now state an important criterion for integrability.

THEOREM I. *Suppose* f *is bounded on* $[a, b]$, *and suppose that corresponding to each positive* ϵ *there is a partition of* $[a, b]$ *such that the corresponding upper and lower sums satisfy the inequality* $S - s < \epsilon$. *Then* f *is integrable.*

Proof. If the conditions as stated in the theorem are fulfilled, we have $S < s + \epsilon$. But $J \leqq S$ and $s \leqq I$, by the definitions of I and J. Therefore, combining inequalities, we see that $J \leqq S < s + \epsilon \leqq I + \epsilon$, or $J < I + \epsilon$. Since this conclusion is valid for every positive ϵ, we infer that $J \leqq I$. But we also know that $I \leqq J$ (Lemma III). Therefore $I = J$. This means that f is integrable, by definition.

Example 2. Suppose f is defined on $[0, 2]$ as follows:

$$f(x) = 1 \quad \text{if} \quad 0 \leqq x < 1, \qquad f(x) = 2 \quad \text{if} \quad 1 \leqq x \leqq 2.$$

We can show by Theorem I that f is integrable.

For this purpose suppose a positive ϵ is assigned. Let h be a positive number smaller than 1 and also smaller than $\epsilon/2$. Consider the partition defined by

$$x_0 = 0, \quad x_1 = 1 - h, \quad x_2 = 1 + h, \quad x_3 = 2.$$

It is readily seen from Fig. 169 that

$$m_1 = M_1 = 1, \quad m_2 = 1, \quad M_2 = 2, \quad m_3 = M_3 = 2,$$

and therefore that

$$s = 1 \cdot (1 - h) + 1 \cdot 2h + 2(1 - h) = 3 - h,$$

$$S = 1 \cdot (1 - h) + 2 \cdot 2h + 2(1 - h) = 3 + h.$$

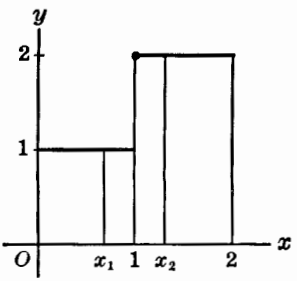


Fig. 169.

Consequently $S - s = 2h < \epsilon$. Therefore f is integrable, by Theorem I. In giving this argument we did not need to find the value of the integral, but it is not hard to see that

$$\int_0^2 f(x)\, dx = 3. \tag{18.1–4}$$

Note that the function is discontinuous at $x = 1$.

The following theorem is the converse of Theorem I.

THEOREM II. *If f is integrable on $[a, b]$, and if $\epsilon > 0$, there is a partition with upper and lower sums such that $S - s < \epsilon$.*

The proof is left as an exercise.

EXERCISES

1. Prove (18.1–4) in Example 2 by the following steps: First, $I \geqq 3$; next, $J \leqq 3$; and finally, $I = J = 3$. Explain each step fully.

2. Suppose f is defined as follows: $f(x) = 2$ if $0 \leqq x < 1$, $f(1) = 0$, $f(x) = -1$ if $1 < x < 2$, $f(2) = 3$, $f(x) = 0$ if $2 < x < 3$, $f(3) = 1$. **(a)** Prove that f is integrable, using an argument something like that in Example 2, but with six subintervals. **(b)** Find the value of $\int_0^3 f(x)\,dx$, using an argument like that of Exercise 1.

3. Suppose f is defined by the requirement that $f(x) = 2$ if x is a rational number of the form $p/2^q$, where p can take on all the values $0, \pm 1, \pm 2, \ldots$ and q can take on all the values $1, 2, \ldots$, and $f(x) = 1$ for all other values of x. Calculate I and J for this function on the interval $[0, 2]$, and thus prove that f is not integrable.

4. Prove Theorem II.

5. If f is integrable, so is the absolute-value function $|f(x)|$. Prove this by showing that if s, S refer to f, and s', S' refer to $|f|$, then $S' - s' \leqq S - s$. Then use Theorem I.

6. If f is integrable over the inverval $[a, b]$, it is also integrable over any closed interval of $[a, b]$. Prove this, using Lemma I and Theorem I.

7. Suppose $a < b < c$ and that f is integrable over $[a, b]$ and also over $[b, c]$. Prove by Theorems I and II that f is integrable over $[a, c]$.

18.11 / THE INTEGRABILITY OF CONTINUOUS FUNCTIONS

Every continuous function is integrable. We state this in a formal theorem.

THEOREM III. *If a function f is continuous at each point of $[a, b]$, it is integrable on that interval.*

Proof. The argument hinges on Theorem I and on the uniform continuity of the function. Suppose $\epsilon > 0$. Choose δ so that

$$|f(x') - f(x'')| < \frac{\epsilon}{2(b - a)} \tag{18.11–1}$$

if x' and x'' are points of $[a, b]$ such that $|x' - x''| < \delta$. This may be done, since f is uniformly continuous (Theorem V, §17.4). Now consider any partition $(x_0, x_1, \ldots, x_n)$ such that all the subintervals have length less than δ, and let s, S be the corresponding lower and upper sums. The boundedness of f is guaranteed by Theorem II, §3.1. Now, in the interval $[x_{i-1}, x_i]$ we can choose a point x' so that $f(x')$ is as close as we like to M_i, and a point x'' so that $f(x'')$ is as close as we

like to m_i. Since (18.11–1) holds, we conclude that

$$M_i - m_i \leqq \frac{\epsilon}{2(b-a)}. \tag{18.11–2}$$

But then, by (18.1–1) and (18.1–2),

$$S - s = (M_1 - m_1)(x_1 - x_0) + \cdots + (M_n - m_n)(x_n - x_{n-1}),$$

and so

$$|S - s| \leqq \frac{\epsilon}{2(b-a)}[(x_1 - x_0) + \cdots + (x_n - x_{n-1})] = \frac{\epsilon}{2}.$$

Thus f is integrable, by Theorem I.

EXERCISE

If f is integrable on $[a, b]$, if $f(x) \geqq 0$ for each x, and if there is at least one point ξ where f is continuous and $f(\xi) > 0$, prove that $\int_a^b f(x)\,dx > 0$.

18.12 / INTEGRABLE FUNCTIONS WITH DISCONTINUITIES

A function such that $f(x)$ never decreases as x increases can have points of discontinuity. Nevertheless, such a function is integrable over any closed interval on which it is defined.

THEOREM IV. *Suppose $f(x)$ is defined when $a \leqq x \leqq b$, and that $f(x') \leqq f(x'')$ if $x' < x''$. Then f is integrable on $[a, b]$.*

Proof. There are two cases to consider. If $f(a) = f(b)$, the hypothesis guarantees that $f(x)$ is constant on the interval, and in this case certainly the upper and lower sums are both equal to $f(a)(b-a)$, no matter what partition is chosen, and so the function is integrable. The other case is that in which $f(a) < f(b)$. If $\epsilon > 0$ we can choose a partition in which all the subintervals are so short that the conditions

$$x_i - x_{i-1} < \frac{\epsilon}{f(b) - f(a)}, \quad i = 1, \ldots, n$$

are satisfied. Now, the special hypothesis on f assures that $m_i = f(x_{i-1})$ and $M_i = f(x_i)$. That is, as x goes from x_{i-1} to x_i, $f(x)$ increases from its smallest to its largest value in the subinterval. Consequently

$$S - s = [f(x_1) - f(x_0)](x_1 - x_0) + \cdots + [f(x_n) - f(x_{n-1})](x_n - x_{n-1}),$$

$$S - s < \{[f(x_1) - f(x_0)] + \cdots + [f(x_n) - f(x_{n-1})]\}\frac{\epsilon}{f(b) - f(a)}.$$

Because of cancellation of terms and the facts that $x_0 = a$, $x_n = b$, this last inequality becomes simply $S - s < \epsilon$. We then conclude that f is integrable, by Theorem I.

It is of course also true that f is integrable if $f(x)$ never increases as x increases. In this case $m_i = f(x_i)$ and $M_i = f(x_{i-1})$ on any subinterval $[x_{i-1}, x_i]$.

It is beyond the scope of this book to make an intensive study of the question as to exactly what kinds of discontinuous functions are integrable. The function of Example 1, § 18.1, is not integrable, but it is discontinuous at every point of the interval [0, 1]. For many practical purposes it is sufficient to know that if f is bounded and has only a finite number of points of discontinuity on $[a, b]$, then it is integrable. The discussion of the following example will illustrate a method by which the foregoing assertion may be proved.

Example 1. Suppose f is defined on [0, 2] by

$$f(x) = x \quad \text{if} \quad 0 \leqq x < 1, \qquad f(x) = x - 1 \quad \text{if} \quad 1 \leqq x \leqq 2.$$

This function is bounded on [0, 2], and continuous except at $x = 1$. The graph is shown in Fig. 170. We shall use Theorem I to show that f is integrable. Suppose $\epsilon > 0$. We form a partition of [0, 2] by taking $1 - (\epsilon/4)$ and $1 + (\epsilon/4)$ as two consecutive points in the partition. The remaining points are taken between 0 and $1 - (\epsilon/4)$ and $1 + (\epsilon/4)$ and 2, in a manner to be specified presently. Observe that f is continuous on the intervals $[0, 1 - (\epsilon/4)]$, $[1 + (\epsilon/4), 2]$; the point of discontinuity has been enclosed in the subinterval $[1 - (\epsilon/4), 1 + (\epsilon/4)]$, whose length is $\epsilon/2$. Now consider the lower and upper sums s, S. Let s_1 represent the contributions coming to s from the subintervals of $[0, 1 - (\epsilon/4)]$, and let s_3 represent the contributions from the subintervals of $[1 + (\epsilon/4), 2]$. Let s_2 represent the contribution from the single subinterval $[1 - (\epsilon/4), 1 + (\epsilon/4)]$, so that $s = s_1 + s_2 + s_3$. With similar notations for upper sums we have $S = S_1 + S_2 + S_3$. Now, the least upper bound of $f(x)$ when $1 - (\epsilon/4) \leqq x \leqq 1 + (\epsilon/4)$ is 1, and the least value is 0. Therefore

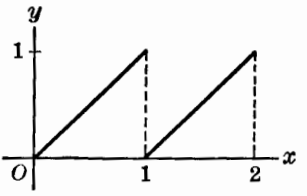


Fig. 170.

$$s_2 = 0\left(\frac{\epsilon}{2}\right) = 0, \quad S_2 = 1\left(\frac{\epsilon}{2}\right) = \frac{\epsilon}{2}.$$

Since f is continuous on $[0, 1 - (\epsilon/4)]$ and $[1 + (\epsilon/4), 2]$, we can choose the part of the partition in these intervals so that

$$S_1 - s_1 < \frac{\epsilon}{4} \quad \text{and} \quad S_3 - s_3 < \frac{\epsilon}{4}.$$

This is by Theorems III and II. Then

$$S - s = (S_1 - s_1) + (S_2 - s_2) + (S_3 - s_3) < \frac{\epsilon}{4} + \frac{\epsilon}{2} + \frac{\epsilon}{4} = \epsilon.$$

We then conclude by Theorem I that the function is integrable on [0, 2].

The kind of argument employed in the foregoing discussion has a much more general application, and by means of it we can prove the following theorem.

THEOREM V. *If f is bounded on $[a, b]$, and if the points of $[a, b]$ at which f is discontinuous can be enclosed in a finite number of subintervals the sum of whose lengths can be made as small as we please, then f is integrable. In particular, f is integrable if it is bounded and has only a finite number of points of discontinuity.*

We omit the details of proof because of the similarity to the argument used in connection with Example 1.

Example 2. Suppose f is defined and bounded on $[0, 1]$, and is continuous except at the points $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots$. Then it is integrable; for, no matter how small a positive ϵ we choose, we can enclose all the points of discontinuity in a finite number of subintervals of total length less than ϵ. To see that this is so, suppose $0 < \epsilon < 1$. As a first subinterval choose $[0, \epsilon/4]$. This interval will enclose all but a finite number of the points $\frac{1}{2}, \frac{1}{3}, \ldots$. If there are N points *not* so enclosed, we can enclose each of these remaining points in a subinterval of length $\epsilon/(2N)$. This gives us $N + 1$ subintervals in all, the sum of whose lengths is $3\epsilon/4$, and they enclose all the points at which f is discontinuous. Thus f is integrable, by Theorem V.

The condition for integrability in Theorem V is *sufficient*, but not *necessary*. The following theorem is true, but we shall not attempt to prove it.

THEOREM VI. *If f is bounded on $[a, b]$, it is integrable if and only if the points at which f is discontinuous can be enclosed in a finite or denumerably infinite set of subintervals of total length as small as we please.*

18.2 / THE INTEGRAL AS A LIMIT OF SUMS

In this section we shall show the equivalence between the definition of the integral as given in §18.1 and the definition which is used in elementary calculus (the one given in § 1.5).

It is desirable to introduce some special terminology and notation in connection with partitions. If $[a, b]$ is a fixed closed interval, let us use symbols such as $P, P_0, P', \ldots$ for various partitions of the interval. If P is the partition determined by points $(x_0, x_1, \ldots, x_n)$, where $a = x_0$, $b = x_n$, the length of the longest subinterval in the partition is called the *mesh fineness* of P, and is denoted in symbols by $|P|$. Thus, by definition, $|P|$ is the maximum of the differences

$$x_1 - x_0, x_2 - x_1, \ldots, x_n - x_{n-1}.$$

The notion of the integral as a limit of sum is based on consideration of sums of the type

$$f(x_1')(x_1 - x_0) + f(x_2')(x_2 - x_1) + \cdots + f(x_n')(x_n - x_{n-1}). \tag{18.2-1}$$

Here $x_0, x_1, \ldots, x_n$ are the points of a partition P, and $x_1', \ldots, x_n'$ are any points

chosen so that $x_{i-1} \leqq x_i' \leqq x_i$, $i = 1, \ldots, n$. The following theorem is fundamental in the theory of integration:

THEOREM VII. *Suppose f is bounded on the interval $[a, b]$. Then it is integrable if and only if the sums* (18.2–1) *approach a limit as the mesh fineness $|P|$ approaches 0. This limit is then the same as the integral defined in* §18.1.

In order to prove Theorem VII it is best to begin by proving the following theorem, usually named after the French mathematician J. G. Darboux (1842–1917).

THEOREM VIII. *Suppose f is bounded on $[a, b]$, and let s, S be the lower and upper sums corresponding to a partition P. Then s approaches I and S approaches J as $|P| \to 0$. This means that for any $\epsilon > 0$ there is some $\delta > 0$ such that*

$$|s - I| < \epsilon \quad \text{and} \quad |S - J| < \epsilon \quad \text{if} \quad |P| < \delta.$$

If we grant the truth of Darboux's theorem, it is rather easy to prove Theorem VII. Let us suppose that f is integrable. Now, if $x_{i-1} \leqq x_i' \leqq x_i$, we certainly have $m_i \leqq f(x_i') \leqq M_i$ and therefore

$$s \leqq \sum_{i=1}^{n} f(x_i')(x_i - x_{i-1}) \leqq S. \tag{18.2–2}$$

As $|P| \to 0$, Darboux's theorem asserts that s and S approach I and J respectively. But $I = J = \int_a^b f(x)\,dx$, and so we see by (18.2–2) that the sums (18.2–1) must approach $\int_a^b f(x)\,dx$ as $|P| \to 0$. On the other hand, if we assume that the sums (18.2–1) approach some limit A, this means that all such sums lie between $A - \epsilon$ and $A + \epsilon$ if $|P|$ is sufficiently small. But, if we choose such a partition and keep it fixed, then by varying the choice of $x_1', \ldots, x_n'$ we can bring the sum (18.2–1) as close as we please to either s or S. Consequently we must have

$$A - \epsilon \leqq s \quad \text{and} \quad S \leqq A + \epsilon.$$

But then $S - s \leqq 2\epsilon$. Since ϵ can be chosen as small as we please, we know by Theorem I that f is integrable. This concludes the proof of Theorem VII.

We still have to prove Theorem VIII. This proof is a bit intricate in detail. Let us first establish the following fact: If S is the upper sum corresponding to a partition P, and if S' is the new upper sum corresponding to a partition P' obtained from P by inserting a single additional point, then

$$S - S' \leqq 2C|P|, \tag{18.2–3}$$

where C is the least upper bound of $|f(x)|$ on $[a, b]$. To see this let us suppose for definiteness that the new point ξ is between x_0 and x_1, and use the notation as in the proof of Lemma I, §18.1. Then

$$S - S' = M_1(x_1 - x_0) - M_1'(\xi - x_0) - M_1''(x_1 - \xi).$$

But since each of the M's is in absolute value less than or equal to C, we certainly have the inequality

$$S - S' \leqq C(x_1 - x_0) + C(\xi - x_0) + C(x_1 - \xi) = 2C(x_1 - x_0),$$

and so (18.2–3) is true, for $|x_1 - x_0| \leqq |P|$. By an extension of this argument we see that

$$S - S' \leqq 2NC|P| \tag{18.2–4}$$

if P' is obtained from P by inserting N additional points, not more than one in any one of the original subintervals of P.

Now we are ready to prove Darboux's theorem. Suppose $\epsilon > 0$. Choose a fixed partition P_0 such that $S_0 < J + (\epsilon/2)$. This can be done, since J is the greatest lower bound of all possible upper sums. Suppose that h is the length of the *shortest* subinterval in P_0, and suppose the points forming P_0 are $\xi_0, \xi_1, \ldots, \xi_{N+1}$ (with $\xi_0 = a$, $\xi_{N+1} = b$). Choose δ so small that $\delta < h$ and $2NC\delta < \epsilon/2$. Consider any partition P such that $|P| < \delta$. Let P' be the partition formed by inserting the points $\xi_1, \ldots, \xi_N$ along with the points forming P. No more than one of the ξ's can fall into the same subinterval of P, so (18.2–4) will hold. Also, $S' \leqq S_0$, by Lemma I. Thus

$$\begin{aligned} S - J &= (S - S') + (S' - J) \leqq (S - S') + (S_0 - J) \\ &< 2NC|P| + \frac{\epsilon}{2} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \end{aligned}$$

by the way things have been arranged. We know, moreover, that $0 \leqq S - J$. Therefore $|S - J| < \epsilon$ if $|P| < \delta$. This proves the part of Darboux's theorem referring to S and J. A similar argument can be given for s and I, but there is a device for deducing the case of s and I from the result already proved for S and J; see Exercise 3.

EXERCISES

1. Suppose f and g are integrable on $[a, b]$. Prove that the functions cf and $f + g$ are integrable on $[a, b]$, and that

$$\int_a^b cf(x)\,dx = c\int_a^b f(x)\,dx,$$

$$\int_a^b [f(x) + g(x)]\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx.$$

Use Theorem VII and the basic theorems about limits, as stated in §1.6.

2. If f is integrable on $[a, c]$ and $a < b < c$, prove that

$$\int_a^c f(x)\,dx = \int_a^b f(x)\,dx + \int_b^c f(x)\,dx.$$

Use Theorem VII and the result of Exercise 6, §18.1.

3. Let us write $s(f)$, $I(f)$ to denote the dependence of s and I on the function f, and use similar notations for S, J. Show that $s(f) = -S(-f)$ and $I(f) = -J(-f)$. Now deduce the part of Theorem VIII dealing with s and I from the part dealing with S and J.

18.21 / DUHAMEL'S PRINCIPLE

It sometimes happens that we need to consider sums which resemble those occurring in the expression of an integrable as a limit of sums but which do not have quite the right form to enable us to apply Theorem VII. It is useful to have a theorem which enables us to express the limit of certain sums as an integral, even though the situation is not one to which Theorem VII is applicable. We shall state such a theorem and discuss a few of its typical applications.

First let us introduce the concept of what we shall call a law of *partition weighting*. By this we mean a rule whereby to each partition P corresponds a set of numbers, one number being assigned as a "weight" to each subinterval of the partition. Thus, if P is the partition $(x_0, x_1, \ldots, x_n)$ of an interval $[a, b]$, the law of weighting will assign weights $\phi_1, \ldots, \phi_n$ to the respective subintervals $[x_0, x_1], \ldots, [x_{n-1}, x_n]$. As an example of a law of partition weighting, suppose f is a function defined on $[a, b]$, and let the weights $\phi_1, \ldots, \phi_n$ be the values $f(x_1'), \ldots, f(x_n')$, where x_i' is chosen in $[x_{i-1}, x_i]$ according to some kind of rule. In connection with a law of partition weighting we are going to consider sums of the form

$$\phi_1(x_1 - x_0) + \phi_2(x_2 - x_1) + \cdots + \phi_n(x_n - x_{n-1}).$$

We shall call these *weighted partition sums*. Our general purpose is to describe certain conditions under which these partition sums will approach an integral as a limit when the mesh fineness $|P|$ approaches 0. In the applications of this sort of thing, if we have a certain law of partition weighting, this law is likely to be of such a sort that it appears to be very nearly the same as a law in which the weights are of the type $f(x_1'), \ldots, f(x_n')$ derived from some function f. This will suggest that the limit of the partition sums is the integral $\int_a^b f(x)\,dx$. The problem is to justify this surmise by sound reasoning. The key lies in considering the size of the differences $\phi_1 - f(x_1'), \ldots, \phi_n - f(x_n')$.

THEOREM IX. *Suppose we have a certain law of partition weighting, $\phi_1, \ldots, \phi_n$ corresponding to the subintervals of the partition P: $(x_0, x_1, \ldots, x_n)$. Suppose also that f is a function integrable on $[a, b]$, and suppose that there is some choice of points $x_1', \ldots, x_n'$ in each partition such that the maximum of the absolute values*

$$|\phi_1 - f(x_1')|, \ldots, |\phi_n - f(x_n')| \tag{18.21–1}$$

approaches 0 as $|P| \to 0$. Then

$$\lim_{|P|\to 0} \sum_{i=1}^{n} \phi_i(x_i - x_{i-1}) = \int_a^b f(x)\,dx. \tag{18.21–2}$$

Proof. Suppose $\epsilon > 0$ and choose δ so small that the maximum of the absolute values in (18.21–1) is less than ϵ if $|P| < \delta$. Our hypothesis means that

such a choice of δ is possible for each positive ϵ. Then $|P| < \delta$ implies

$$\left|\sum_{i=1}^{n} [\phi_i - f(x_i')](x_i - x_{i-1})\right| \leqq \sum_{i=1}^{n} |\phi_i - f(x_i')|(x_i - x_{i-1})$$
$$< \epsilon \sum_{i=1}^{n} (x_i - x_{i-1}) = \epsilon(b - a);$$

in other words,

$$\lim_{|P|\to 0} \sum_{i=1}^{n} [\phi_i - f(x_i')](x_i - x_{i-1}) = 0.$$

But this is the same as the assertion

$$\lim_{|P|\to 0} \sum_{i=1}^{n} \phi_i(x_i - x_{i-1}) = \lim_{|P|\to 0} \sum_{i=1}^{n} f(x_i')(x_i - x_{i-1}),$$

and this is equivalent to (18.21–2), by Theorem VII.

We shall refer to Theorem IX as Duhamel's principle. Duhamel dealt with the problem of finding limits of sums of the same general type as those we have called partition sums associated with a law of partition weighting. In Duhamel's time and for long afterward in calculus textbooks the treatment of such matters was regularly couched in the phraseology of "infinitesimals." The treatment given here, avoiding the term "infinitesimal," is patterned after a formulation given by W. F. Osgood.

Example 1. Suppose that f and g are continuous functions defined on $[a, b]$. Let $F(x) = f(x)g(x)$. Consider a law of partition weighting which arises by choosing *two* points x_i', x_i'' in each subinterval of an arbitrary partition $(x_0, x_i, \ldots, x_n)$, and letting $\phi_i = f(x_i')g(x_i'')$. Then the limit of the corresponding partition sums is the integral of the function F. That is,

$$\lim_{|P|\to 0} \sum_{i=1}^{n} f(x_i')g(x_i'')(x_i - x_{i-1}) = \int_a^b f(x)g(x)\,dx. \tag{18.21–4}$$

This is an application of Duhamel's principle. To see that it is a valid application we have to verify that the maximum of the quantities

$$|f(x_i')g(x_i'') - F(x_i')| \quad (i = 1, \ldots, n) \tag{18.21–5}$$

approaches 0 as $|P| \to 0$. The truth of (18.21–4) will then follow, for F is continuous, and therefore integrable. We make use of the fact that f is bounded and g is *uniformly* continuous. Let A be the maximum of $|f(x)|$. The uniform continuity of g assures us that the maximum of the differences

$$|g(x_i'') - g(x_i')| \quad (i = 1, \ldots, n)$$

approaches 0 as $|P| \to 0$, for $|x_i' - x_i''| \leqq |P|$. But, for the expression in (18.21–5) we can write

$$|f(x_i')[g(x_i'') - g(x_i')]| \leqq A|g(x_i'') - g(x_i')|.$$

Since A is fixed, the assertion about (18.21–5) is seen to be true.

Formula (18.21–4) holds true under less restrictive assumptions on f and g; see Exercise 3.

Formula (18.21–4) can be extended in an obvious way to the case of products of three or more functions.

Example 2. Consider the derivation of the integral formula for arc length of a curve, as discussed in §14.2. Let C be a smooth arc defined by

$$x = f(t), \quad y = g(t), \quad z = h(t), \quad a \leqq t \leqq b,$$

where f, g, h have continuous first derivatives. If $(t_0, t_1, \ldots, t_n)$ is a partition of $[a, b]$, the discussion in §14.2 shows that the arc length of C is the limit

$$\lim_{|P|\to 0} \sum_{i=1}^{n} \{[f'(\alpha_i)]^2 + [g'(\beta_i)]^2 + [h'(\gamma_i)]^2\}^{1/2} \Delta t_i, \tag{18.21–6}$$

where $\Delta t_i = t_i - t_{i-1}$ and α_i, β_i, γ_i are certain points of the interval $[t_{i-1}, t_i]$. Now if we set

$$\phi_i = \{[f'(\alpha_i)]^2 + [g'(\beta_i)]^2 + [h'(\gamma_i)]^2\}^{1/2},$$

we have a law of partition weighting. If it were true that $\alpha_i = \beta_i = \gamma_i$, the limit of the partition sums would be the integral

$$\int_a^b \{[f'(t)]^2 + [g'(t)]^2 + [h'(t)]^2\}^{1/2}\, dt. \tag{18.21–7}$$

To show that the limit in (18.21–6) is equal to the integral in (18.21–7) we compare ϕ_i with the corresponding expression in which β_i and γ_i are replaced by α_i, and show that Theorem IX is applicable (with the necessary changes in notation). To do this we use the fact that the function

$$\{[f'(t)]^2 + [g'(u)]^2 + [h'(v)]^2\}^{1/2}$$

is uniformly continuous in the cubical region

$$a \leqq t \leqq b, \quad a \leqq u \leqq b, \quad a \leqq v \leqq b.$$

We omit the details.

EXERCISES

1. Suppose F is continuous on $[a, b]$ and that f is defined and has a continuous derivative on $[a, b]$. Using the standard notation relating to partitions, find the limit

$$\lim_{|P|\to 0} \sum_{i=1}^{n} F(x_i')[f(x_i) - f(x_{i-1})]$$

where x_i' is any point on the ith subinterval.

2. Suppose f and g are continuous on $[a, b]$, and $F(u, v)$ is a continuous function of u and v for all values of u and v obtained by setting $u = f(x)$, $v = g(y)$ and letting x, y vary over $[a, b]$. Use Theorem IX to show that

$$\lim_{|P|\to 0} \sum_{i=1}^{n} F[f(x_i'), g(x_i'')](x_i - x_{i-1}) = \int_a^b F[f(x), g(x)]\, dx.$$

Explain carefully where uniform continuity comes into the argument.

3. Show that (18.21–4) is true if it is assumed that f is bounded and that $f(x)g(x)$ and $g(x)$ are integrable. For the proof it is enough to show that

$$\lim_{|P|\to 0} \sum_{i=1}^{n} f(x_i')[g(x_i'') - g(x_i')](x_i - x_{i-1}) = 0.$$

But, if $|f(x)| \leqq A$, and if S and s are upper and lower sums for g, show that the expression following the limit symbol above is not larger than $A(S - s)$. Then use a certain theorem employing the fact that g is integrable.

18.3 / FURTHER DISCUSSION OF INTEGRALS

In our theory of the Riemann integral we have been assuming that $a < b$. The cases of equal upper and lower limits and of lower limit larger than upper limit are handled exactly as in (1.52–2), (1.52–3). The way in which integrals over adjacent intervals are combined is indicated in Exercise 7, §18.1, and Exercise 2, §18.2. All these things are familiar from elementary calculus.

Inequalities for integrals are used a great deal. The basic fact is that if $f(x) \leqq g(x)$ on $[a, b]$, then (assuming $a < b$)

$$\int_a^b f(x)\,dx \leqq \int_a^b g(x)\,dx. \tag{18.3–1}$$

In particular, since $f(x) \leqq |f(x)|$ and $-|f(x)| \leqq f(x)$, we have

$$\left|\int_a^b f(x)\,dx\right| \leqq \int_a^b |f(x)|\,dx.$$

The inequality (18.3–1) is easily proved by analytical reasoning on either upper or lower sums. For the fact that $|f|$ is integrable if f is integrable, see Exercise 5, §18.1.

Finally, in this miscellany of remarks, it should be stated that the product of two integrable functions is integrable. This result can be proved in various ways. It is an immediate consequence of the rather advanced Theorem VI. It can be proved by more elementary means, however, though the arguments are rather ingenious. We shall not give the details.

18.4 / THE INTEGRAL AS A FUNCTION OF THE UPPER LIMIT

In this section we consider

$$F(x) = \int_a^x f(t)\,dt, \tag{18.4–1}$$

where $a \leqq x \leqq b$ and f is assumed to be integrable on $[a, b]$.

THEOREM X. *The function F is continuous.*

This is very easily proved. Consider any two points x', x'', and suppose

$x' < x''$. Then

$$F(x'') - F(x') = \int_a^{x''} f(t)\,dt - \int_a^{x'} f(t)\,dt = \int_{x'}^{x''} f(t)\,dt.$$

Let C be the least upper bound of $|f(t)|$ on $[a, b]$. Then

$$|F(x'') - F(x')| \leqq \int_{x'}^{x''} |f(t)|\,dt \leqq C|x'' - x'|.$$

The continuity of F is a direct consequence of this inequality.

In §1.52 we considered the same formula (18.4–1), but on the assumption that f was continuous at each point of $[a, b]$. It was then proved (Theorem VII of §1.52) that $F'(x) = f(x)$ at each point. In the present case we are not assuming the continuity of f, but merely its integrability. With this less restrictive assumption we cannot conclude quite as much as before.

THEOREM XI. *The formula $F'(x) = f(x)$ holds at each point where f is continuous.*

Proof. Let x_0 be a point at which f is continuous. If $\epsilon > 0$, choose δ so that

$$|f(x) - f(x_0)| < \epsilon \quad \text{if} \quad |x - x_0| < \delta. \tag{18.4–2}$$

We wish to show that

$$\lim_{h \to 0} \left[\frac{F(x_0 + h) - F(x_0)}{h} - f(x_0) \right] = 0. \tag{18.4–3}$$

Now

$$F(x_0 + h) - F(x_0) = \int_a^{x_0+h} f(t)\,dt - \int_a^{x_0} f(t)\,dt = \int_{x_0}^{x_0+h} f(t)\,dt,$$

and

$$f(x_0) = \frac{1}{h} \int_{x_0}^{x_0+h} f(x_0)\,dt.$$

Therefore

$$\left| \frac{F(x_0 + h) - F(x_0)}{h} - f(x_0) \right| = \left| \frac{1}{h} \int_{x_0}^{x_0+h} [f(t) - f(x_0)]\,dt \right|. \tag{18.4–4}$$

We now suppose that $0 < |h| < \delta$. Then the expression under the integral sign on the right side of (18.4–4) is in absolute value less than ϵ, by (18.4–2), since $|t - x_0| \leqq |h| < \delta$. Consequently the whole right side of (18.4–4) is less than ϵ when $0 < |h| < \delta$. This proves (18.4–3) and thus proves the theorem.

EXERCISES

1. Use Theorem X to prove that the value of $\int_a^b f(x)\,dx$ is not changed if the function f is altered merely by changing the value of $f(b)$.

2. If f is integrable and $G(x) = \int_x^b f(t)\,dt$, show that G is continuous by finding the relation between F and G. (Assume F defined by (18.4–1).)

3. Prove the result corresponding to that in Exercise 1 if f is altered merely by changing the value of $f(a)$.

4. Suppose f is defined on $[0, 1]$ as follows: $f(x) = (-1)^n$ if $1/2^{n+1} < x \leqq 1/2^n$, $n = 0, 1, 2, \ldots$, and $f(0) = 0$. Find the value of $\int_0^1 f(x)\,dx$ by using the result of Exercise 2.

5. If $F(x) = \int_0^x \sqrt{(1-t^2)(2-t^2)}\,dt$, sketch the graph of $y = F(x)$ when $-1 \leqq x \leqq 1$, using information derived from $F'(x)$ and $F''(x)$. The integral is an elliptic integral, and cannot be expressed in terms of elementary functions.

6. Show that $\dfrac{d}{dx}\int_x^b f(t)\,dt = -f(x)$ if f is continuous at x.

7. Let $F(x) = \int_{-1}^x |t|\,dt$. **(a)** Find $F'(x)$ by Theorem XI. **(b)** Find explicit formulas for $F(x)$ for the separate cases $x < 0$, $x > 0$. **(c)** Draw the graph of $y = F(x)$.

8. Let $f(x) = 1$ if $0 \leqq x \leqq 1$, $f(x) = -1$ if $-1 \leqq x < 0$. Define $F(x)$ by (18.4–1) with $a = -1$. **(a)** Find separate explicit formulas for $F(x)$ if $-1 \leqq x \leqq 0$ and $0 \leqq x \leqq 1$. **(b)** Draw the graph of $y = F(x)$ and check the validity of Theorems X, XI. What about $F'(0)$?

18.41 / THE INTEGRAL OF A DERIVATIVE

In this section we shall prove a generalization of Theorem VIII, §1.53.

THEOREM XII. *Suppose f is integrable on $[a, b]$, and suppose there is a function F which is continuous on $[a, b]$ and such that, for each x on the open interval $a < x < b$, F has a derivative given by $F'(x) = f(x)$. Then*

$$\int_a^b f(x)\,dx = F(b) - F(a). \tag{18.41–1}$$

Proof. Consider any partition $(x_0, x_1, \ldots, x_n)$ of $[a, b]$. We apply the law of the mean (Theorem IV, §1.2) to $F(x)$ on each of the subintervals $[x_{-1}, x_i]$, $i = 1, \ldots, n$. There is some point x_i' such that $x_{i-1} < x_i' < x_i$ and

$$F(x_i) - F(x_{i-1}) = (x_i - x_{i-1})F'(x_i').$$

Thus, noting that $F'(x_i') = f(x_i')$, we have

$$\begin{aligned} F(x_1) - F(x_0) &= (x_1 - x_0)f(x_1') \\ F(x_2) - F(x_1) &= (x_2 - x_1)f(x_2') \\ &\vdots \\ F(x_n) - F(x_{n-1}) &= (x_n - x_{n-1})f(x_n'). \end{aligned}$$

Adding, and recalling that $x_0 = a$, $x_n = b$, we see that

$$F(b) - F(a) = \sum_{i=1}^{n} f(x_i')(x_i - x_{i-1}). \tag{18.41–2}$$

This kind of relation holds for each partition. Hence by Theorem VII we see that (18.41–1) is true.

The student will observe, on comparing the present theorem with Theorem

VIII, §1.53, that the conclusions in the two theorems are identical. In the earlier theorem we assumed that f was continuous, whereas here we have assumed merely that f is integrable.

Example. Find the value of

$$\int_0^{2/\pi} \left(2x \sin \frac{1}{x} - \cos \frac{1}{x}\right) dx.$$

Here

$$f(x) = 2x \sin \frac{1}{x} - \cos \frac{1}{x}$$

when $x \neq 0$. We may define $f(0)$ in any manner we please, but f will be discontinuous at $x = 0$ in any case. However, f is continuous for other values of x, and bounded; therefore, it is integrable. To evaluate the integral we define

$$F(x) = x^2 \sin \frac{1}{x} \quad \text{if} \quad x \neq 0, \quad F(0) = 0.$$

The definition $F(0) = 0$ makes F continuous at $x = 0$. For $x \neq 0$ we readily verify that $F'(x) = f(x)$. We observe, incidentally, that $F'(0) = 0$; this may or may not be $f(0)$, depending on how we define $f(0)$. The conditions of Theorem XII are satisfied, and so

$$\int_0^{2/\pi} \left(2x \sin \frac{1}{x} - \cos \frac{1}{x}\right) dx = F\left(\frac{2}{\pi}\right) - F(0) = \frac{4}{\pi^2}.$$

18.5 / INTEGRALS DEPENDING ON A PARAMETER

Our concern in this section is with integrals in which a parameter occurs under the integral sign or in the limits of integration, or in both places. The main objective is to learn about differentiation with respect to the parameter in such cases. We begin with the situation in which the parameter occurs solely in the integrand. Consider

$$F(y) = \int_a^b f(x, y)\, dx. \tag{18.5–1}$$

We shall suppose that f is defined when (x, y) is a point of the rectangle in the xy-plane defined by $a \leqq x \leqq b$, $c \leqq y \leqq d$. We denote this rectangle by R.

It is important for us to know conditions on the function f that will guarantee the continuity of F. We do not attempt to give the most general (i.e., the least restrictive) conditions. For usefulness in practice the following theorem is convenient.

THEOREM XIII. *If f is continuous at each point of R, then F is continuous for each y on the interval $[c, d]$.*

The proof depends upon the fact that f is uniformly continuous in R: see Exercise 9.

The next theorem deals with finding the derivative of $F(y)$.

THEOREM XIV. *Suppose that $f(x, y)$ is an integrable function of x for each value of y, and that the partial derivative $\dfrac{\partial f(x, y)}{\partial y}$ exists and is a continuous function of x and y in the rectangle R. Then $F(y)$ has a derivative given by*

$$F'(y) = \int_a^b \frac{\partial f(x, y)}{\partial y}\, dx. \tag{18.5–2}$$

The formula (18.5–2) is often called Leibniz's rule.

Proof. To give the proof let us use the notation $f_2(x, y)$ for the partial derivative with respect to y. Formula (18.5–2) is equivalent to the statement

$$\lim_{h\to 0}\left\{\frac{F(y+h)-F(y)}{h} - \int_a^b f_2(x, y)\, dx\right\} = 0. \tag{18.5–3}$$

Now

$$F(y+h)-F(y) = \int_a^b [f(x, y+h)-f(x, y)]\, dx.$$

We apply the law of the mean to $f(x, y)$ as a function of y:

$$f(x, y+h)-f(x, y) = hf_2(x, y+\theta h).$$

Here θ is a number depending on x, y, h, and such that $0<\theta<1$. From this last formula we see that

$$\frac{F(y+h)-F(y)}{h} - \int_a^b f_2(x, y)\, dx = \int_a^b [f_2(x, y+\theta h)-f_2(x, y)]\, dx. \tag{18.5–4}$$

At this stage we make use of the fact that f_2 is uniformly continuous in R (by Theorem V, §17.4). Suppose $\epsilon>0$. Choose δ so that the values of f_2 at two different points of R differ by less than ϵ if the distance between the points is less than δ. Then certainly

$$|f_2(x, y+\theta h)-f_2(x, y)|<\epsilon$$

if $|h|<\delta$. Consequently, we see by (18.5–4) that

$$\left|\frac{F(y+h)-F(y)}{h} - \int_a^b f_2(x, y)\, dx\right| < \epsilon(b-a)$$

if $0<|h|<\delta$. since ϵ can be as small as we please, this proves (18.5–3).

Example 1. Find $F'(y)$ if

$$F'(y) = \int_0^1 \log(x^2+y^2)\, dx.$$

We can apply Theorem XIV with $0\leqq x\leqq 1$ and y on any closed interval not

containing $y = 0$. The result is

$$F'(y) = \int_0^1 \frac{2y}{x^2 + y^2}\, dx.$$

The integration is readily performed, and we find

$$F'(y) = 2 \tan^{-1}\left(\frac{1}{y}\right).$$

If the parameter occurs merely in the limits of integration we can use Theorem XI and the standard chain rule for differentiating composite functions. We illustrate by an example.

Example 2. Suppose

$$F(y) = \int_{\sin y}^{e^y} \sqrt{1 + x^3}\, dx.$$

How is $F'(y)$ found in this case? We set

$$G(u, v) = \int_u^v \sqrt{1 + x^3}\, dx.$$

Then $G(u, v)$ becomes $F(y)$ if we put $u = \sin y$, $v = e^y$. Therefore

$$\frac{dF}{dy} = \frac{\partial G}{\partial u}\frac{du}{dy} + \frac{\partial G}{\partial v}\frac{dv}{dy}.$$

But, by Theorem XI,

$$\frac{\partial G}{\partial u} = -\sqrt{1 + u^3}, \qquad \frac{\partial G}{\partial v} = \sqrt{1 + v^3}.$$

(How does the minus sign get into the first formula?) Therefore

$$F'(y) = -\sqrt{1 + \sin^3 y}\cos y + \sqrt{1 + e^{3y}}\, e^y.$$

If the parameter occurs both under the integral sign and in the limits of integration, we use both Theorem XI and Theorem XIV and the chain rule.

Example 3. Suppose

$$F(y) = \int_0^{y^2} x^5(y - x)^7\, dx.$$

Here we can define

$$G(u, v) = \int_0^u x^5(v - x)^7\, dx$$

and put $u = y^2$, $v = y$. Then

$$\frac{\partial G}{\partial u} = u^5(v - u)^7, \qquad \frac{\partial G}{\partial v} = \int_0^u 7x^5(v - x)^6\, dx.$$

Since $\frac{du}{dy} = 2y$, $\frac{dv}{dy} = 1$, the chain rule gives

$$F'(y) = y^{10}(y - y^2)^7 \cdot 2y + \int_0^{y^2} 7x^5(y - x)^6\, dx.$$

With a little practice the student will be able to handle problems of this kind without explicit introduction of auxiliary variables.

EXERCISES

1. If $F(u) = \int_0^4 \log(1 - u^2x^2)\, dx$, find $F'(u)$ by Leibniz's rule. What is the value of $F'(0)$? Show without performing any integrations that $F(u)$ has a maximum value at $u = 0$ and that the graph of $v = F(u)$ is concave downward in the uv-plane.

2. If $F(a) = \int_0^1 \frac{a\, dx}{\sqrt{1 - a^2x^2}}$, find $F'(a)$ by two different methods.

3. If $F(x) = \int_0^x e^{-x^2t^2}\, dt$, find $F'(x)$.

4. If $F(x) = \int_{x^2}^{\sin x} (x^2 - t^2)^n\, dt$, find $F'(x)$.

5. Suppose $\phi(y) = \int_0^y x^n(y - x)^m\, dx$, where m and n are positive integers. Calculate $\phi'(y)$, $\phi''(y)$, ... without integration, and show that

$$\phi^{(m)}(y) = \frac{m!}{n+1}\, y^{n+1}.$$

Then, observing that $\phi(0) = \phi'(0) = \cdots = \phi^{(m-1)}(0) = 0$, integrate $\phi^{(m)}(y)$ successively m times and so arrive at the formula

$$\phi(y) = \frac{m!n!}{(m+n+1)!}\, y^{m+n+1}.$$

6. If $u = \int_{x-ct}^{x+ct} \phi(s)\, ds$, where ϕ has a continuous derivative, show that $\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$.

7. If $u = \int_{1/x}^{y} dt \int_{1/t}^{x} f(s, t)\, ds$, show that $\frac{\partial^2 u}{\partial x\, \partial y} = f(x, y)$, assuming that f is continuous. Find the other second derivatives of u, assuming that f has continuous first partial derivatives.

8. Suppose $M(x, y)$ and $N(x, y)$ are continuous in a rectangle with center at (a, b), and that the partial derivatives $\frac{\partial M}{\partial y}$ and $\frac{\partial N}{\partial x}$ are continuous and equal in the rectangle. Show by Theorems XI, XII, and XIV that $\frac{\partial f}{\partial x} = M$ and $\frac{\partial f}{\partial y} = N$ if $f(x, y) = \int_a^x M(s, b)\, ds + \int_b^y N(x, t)\, dt$.

9. Prove Theorem XIII, using the uniform continuity of f in much the same way that the uniform continuity of f_2 was used in proving Theorem XIV.

18.6 / RIEMANN DOUBLE INTEGRALS

In this section we shall discuss the theory of Riemann double integrals

$$\iint_R f(x, y)\, dA \tag{18.6–1}$$

in a manner paralleling the theory for single integrals, as developed in the first part of this chapter. We shall be much briefer in the theory of double integrals, and many proofs will be omitted.

It is simplest to begin with the case in which R is a rectangle with sides parallel to the co-ordinate axes. Suppose these sides are $x = a$, $x = b$, $y = c$, $y = d$, where $a < b$ and $c < d$. If we form a partition of $[a, b]$ by points $(x_0, x_1, \ldots, x_m)$, and a partition of $[c, d]$ by points $(y_0, y_1, \ldots, y_n)$, the lines $x = x_i$ and $y = y_j$ form a rectangular partition of the rectangle R into rectangular cells. The numbers m, n need not be equal; the total number of cells in the partition of R is $N = mn$.

Now suppose that f is a function which is defined and bounded in R. If the cells are numbered in any order, let $m_1, \ldots, m_N$ be the greatest lower bounds of $f(x, y)$ in the various cells, and let $M_1, \ldots, M_N$ denote the corresponding least upper bounds of $f(x, y)$. If the areas of the cells are $\Delta A_1, \ldots, \Delta A_N$, we form the *lower sum*

$$s = m_1 \Delta A_1 + \cdots + m_N \Delta A_N$$

and the *upper sum*

$$S = M_1 \Delta A_1 + \cdots + M_N \Delta A_N.$$

Then we denote the least upper bound of all possible lower sums by I, and the greatest lower bound of all possible upper sums by J.

Definition. *If $I = J$, we say that f is* integrable *over R, and we define the double integral* (18.6–1) *as the common value of I and J.*

Starting from this definition we can obtain analogues of the lemmas and theorems in §18.1. It can be proved that if f is continuous in the closed rectangle R, it is integrable. The proof, using uniform continuity, is much like the proof of Theorem III, § 18.11.

A function may have certain points of discontinuity and yet be integrable. For example, if f is continuous in R except at certain points which lie on a finite number of smooth curves, it can be shown that f is integrable. An instance is furnished by the function defined by $f(x, y) = 1$ if $x^2 + y^2 \leqq 1$, $f(x, y) = 0$ if $x^2 + y^2 > 1$ and $-2 \leqq x \leqq 2$, $-2 \leqq y \leqq 2$. This function is discontinuous at the points of the circle $x^2 + y^2 = 1$, but it is integrable over the square in which it is defined.

It will be very useful in some later work to be able to deal with integrable functions having some points of discontinuity. On that account we shall discuss such matters a little further here. As an aid in this discussion we introduce the concept of the *outer content* of a set of points in the xy-plane.

Definition. *Let T be a point set in the rectangle R. Consider any rectangular partition of R. Select all those cells of the partition which contain points of T, and let A denote the sum of the areas of all these cells. If we consider all possible partitions, the greatest lower bound of the values of A is defined as the* outer content *of T.*

It will be seen that if we define a function f such that $f(x, u) = 1$ if (x, y) is a point of T, and $f(x, y) = 0$ elsewhere in R, then A is the upper sum S for this function, and consequently the number J is the outer content of T.

It is also quite evident that the outer content of T depends on T alone, not on the choice of the rectangle R containing T.

THEOREM XV. *Let f be any function which is bounded in R, and let T be the set of points in R at which f is discontinuous. Then f is integrable provided the outer content of T is 0.*

This theorem is analogous to Theorem V, §18.12. We omit the proof. The condition of zero outer content is sufficient, but not necessary, for f to be integrable.

It is easy to see that a straight line segment has zero outer content. Also, any smooth arc has zero outer content (see Exercise 5). The boundaries of regions of familiar shapes, such as squares, circles, ellipses, all have zero outer content.

Next we turn to the definition of double integrals over regions which are not rectangles with sides parallel to the axes. We confine our attention to bounded regions with boundaries of outer content zero. For convenience we shall refer to regions of this type as *Riemann regions.*

Suppose G is a Riemann region. Let R be a rectangle containing G and having its sides parallel to the axes. Suppose f is a function which is defined in G. We shall define a new function throughout R by setting $g(x, y) = f(x, y)$ at points of G and $g(x, y) = 0$ at points of R not in G. We then say that f is integrable over G if g is integrable over R, and in that case we define

$$\iint_G f(x, y)\, dA = \iint_R g(x, y)\, dA. \tag{18.6–2}$$

This procedure makes everything depend on the theory of integrals over rectangles. Alternative procedures are possible. We could define upper and lower sums directly for f and the region G. In that case it becomes necessary to decide what to do with cells of a partition in case the cells contain points in G as well as points not in G. The fact that the boundary of G has zero outer content makes it immaterial whether such cells are ignored or not.

Darboux's theorem (Theorem VIII, §18.2) can be generalized to the case of double integrals, and as a result we get the important fact that a double integral can be expressed as a limit of sums (the analogue of Theorem VII). Thus we make connection with the earlier definition of double integrals in §13.2.

EXERCISES

1. Suppose $f(x, y) = \tan^{-1}\left(\dfrac{1}{x - y}\right)$ if $x \neq y$, and define $f(x, y) = 0$ when $x = y$. Is f integrable over the square $0 \leqq x \leqq 1$, $0 \leqq y \leqq 1$?

2. Suppose $f(x, y) = \sin(1/xy)$ if $xy \neq 0$, and define $f(x, y) = 1$ if x or $y = 0$. Is f integrable over the square $-1 \leqq x \leqq 1$, $-1 \leqq y \leqq 1$?

3. Suppose $f(x, y) = 0$ if $x + y$ is rational, and $f(x, y) = 1$ if $x + y$ is irrational. Is f integrable over the square $0 \leqq x \leqq 1$, $0 \leqq y \leqq 1$?

4. Suppose $f(x)$ is a continuous function defined when $a \leqq x \leqq b$. Let T be the set of points on the graph of $y = f(x)$ in the xy-plane. Is it possible for T to have positive outer content? Justify your answer.

5. Let C be a smooth arc with parametric equations $x = f(t)$, $y = g(t)$, $a \leqq t \leqq b$. Divide $[a, b]$ into n equal parts by points $t_0, \ldots, t_n$: let $\delta = (b - a)/n$ and let (x_i, y_i) be the point of C corresponding to t_i. Choose M so that $|f'(t)|$ and $|g'(t)|$ do not exceed M on $[a, b]$. Use the law of the mean to show that the points of C for $t_{i-1} \leqq t \leqq t_i$ satisfy the inequalities $|x - x_i| \leqq M\delta$, $|y - y_i| \leqq M\delta$, and hence that C is covered by n squares of total area $4M^2(b - a)\delta$. What do you conclude about the outer content of C?

18.61 / DOUBLE INTEGRALS AND ITERATED INTEGRALS

In §13.3 we gave an account of the procedure for expressing a double integral as an iterated integral. The procedure is summed up in the two formulas (13.3–5), (13.3–6), and a formal statement about the first of these formulas was made in Theorem III, §13.3. However, we did not actually prove the theorem; we merely made an argument for its plausibility, by interpreting both the double integral and the iterated integral as expressions for a certain volume. The student will do well to read §13.3 as far as Example 1 before proceeding further with the present section.

We now wish to give a strictly analytical proof of the relation between double integrals and iterated integrals. This proof is necessary if we are to have a firm logical justification for the procedures used in evaluating double integrals. The proof is, moreover, very instructive for the prospective student of more advanced mathematics, for it is fairly representative of a type of proof which is encountered in a variety of different forms in higher analysis. The basic principle is that of showing that certain limit processes of rather complicated nature can be replaced by two successive limit processes of simpler nature.

We begin with the case of a double integral over a rectangular region.

THEOREM XVI. *Let R be the rectangle $a \leqq y \leqq b$, $c \leqq x \leqq d$, where $a < b$, $c < d$. Suppose that $f(x, y)$ is defined and integrable over R. For each x suppose that $f(x, y)$ is integrable with respect to y over $[a, b]$, and let the function*

$$\phi(x) = \int_a^b f(x, y)\, dy \tag{18.61–1}$$

be integrable over $[c, d]$. Then

$$\iint_R f(x, y)\, dA = \int_c^d dx \int_a^b f(x, y)\, dy. \tag{18.61–2}$$

Proof. Let $(x_0, \ldots, x_m)$ be a partition of $[c, d]$ into m equal parts and $(y_0, \ldots, y_n)$ a partition of $[a, b]$ into n equal parts. We then obtain a partition of

R into mn cells, each of the same area. Let $\Delta x_i = x_i - x_{i-1}$, $\Delta y_j = y_j - y_{j-1}$. Now, if we add up all the mn terms $f(x_i, y_j)\,\Delta x_i\,\Delta y_j$ where $1 \leqq i \leqq m$ and $1 \leqq j \leqq n$, we get an approximation to the value of the double integral of $f(x, y)$ over R, and we can make this approximation as close as we please by taking m and n sufficiently large. That is, if $\epsilon > 0$, there is some integer q depending on ϵ such that

$$\left| \sum_{i=1}^{m} \sum_{j=1}^{n} f(x_i, y_j)\,\Delta x_i\,\Delta y_j - \iint_R f(x, y)\,dA \right| < \epsilon \tag{18.61–3}$$

if $q \leqq m$ and $q \leqq n$.

Next, since $f(x_i, y)$ is an integrable function of y, we know by Theorem VII, §18.2, that for each i

$$\lim_{n\to\infty} \sum_{j=1}^{n} f(x_i, y_j)\,\Delta y_j = \int_a^b f(x_i, y)\,dy.$$

It then follows from (18.61–3) that

$$\left| \sum_{i=1}^{m} \left(\int_a^b f(x_i, y)\,dy \right) \Delta x_i - \iint_R f(x, y)\,dA \right| \leqq \epsilon \tag{18.61–4}$$

if $q \leqq m$. Using the definition of $\phi(x)$ [see [18.61–1)], we can rewrite (18.61–4) as

$$\left| \sum_{i=1}^{m} \phi(x_i)\,\Delta x_i - \iint_R f(x, y)\,dA \right| \leqq \epsilon.$$

This means that

$$\lim_{m\to\infty} \sum_{i=1}^{m} \phi(x_i)\,\Delta x_i = \iint_R f(x, y)\,dA.$$

But since ϕ is integrable we know that this last limit is $\int_c^d \phi(x)\,dx$. Therefore

$$\int_c^d \phi(x)\,dx = \iint_R f(x, y)\,dA.$$

This is the same as (18.61–2).

There is, naturally, a corresponding theorem in which the roles of x and y are reversed.

If $f(x, y)$ is continuous in the rectangle, all the assumptions in Theorem XVI are satisfied, as we see by using Theorem III, §18.11, and Theorem XIII, §18.5.

Let us now consider Theorem III, §13.3, in either of the forms (13.3–5), (13.3–6). It is assumed that f is continuous in the region under consideration (see Fig. 96 and Fig. 97). The boundaries of these regions have zero outer content, for they are made up of graphs of continuous functions, and it is easy to show that the graph of a continuous function, e.g., $x = X_1(y)$, $a \leqq y \leqq b$, has zero outer

content (the proof depends on uniform continuity). Now let the region be enclosed in the smallest possible rectangle with sides parallel to the axes, and extend the definition of f by setting $f(x, y) = 0$ in the part of the rectangle outside the region R. This makes f integrable over the rectangle, and the conditions for applying Theorem XVI are satisfied. Thus the double integral over the rectangle is equal to the iterated integral in each of the two possible orders. But, since $f = 0$ outside of the region R, the integrations need not be carried beyond the boundary of R. Thus, for example, the integral with respect to x across the width of the rectangle becomes simply

$$\int_{x_1}^{x_2} f(x, y)\, dx.$$

In this way we see that (13.3–5) and (13.3–6) are both true.

18.7 / TRIPLE INTEGRALS

The theory of Riemann triple integrals may be developed in a manner closely analogous to the theory of double integrals. The ideas are not fundamentally different from those in § 18.6. In dealing with integrals over regions in three dimensions, it is assumed that the regions have boundaries of outer content zero. Regions of familiar shapes, e.g., spheres, cones, and cylinders, are of the required type. Likewise, the reduction of triple integrals to iterated integrals of lower order is handled in much the same way as the corresponding problem for double integrals in § 18.61. We omit the details.

18.8 / IMPROPER INTEGRALS

In the Riemann theory of integration the functions are assumed to be bounded, and the intervals or regions of integration are assumed to be bounded. Under these conditions, if the function is integrable, the integral is said to be a *proper* integral. There are some extensions of the definitions of integrals. When the Riemann theory is taken as basic, as it is in this book, any integral whose definition does not come within the framework of the Riemann theory, but which is defined by a limiting process depending on the Riemann theory, will be called an *improper* integral. An integral may be improper because it is an integral over an unbounded interval or region; or it may be improper because it is the integral of an unbounded function.

Example 1. The integral $\int_0^1 x^{-1/2}\, dx$ is improper because the function $f(x) = x^{-1/2}$ is not bounded on $[0, 1]$. However, the integral $\int_c^1 x^{-1/2}\, dx$ is proper for each c if $0 < c < 1$, and we define

$$\int_0^1 x^{-1/2}\, dx = \lim_{c \to 0^+} \int_c^1 x^{-1/2}\, dx.$$

The limit exists and has the value 2, as is easily verified.

Example 2. $\int_1^\infty \frac{dx}{x^\alpha}$ is improper, because the interval of integration is not finite. We define

$$\int_1^\infty \frac{dx}{x^\alpha} = \lim_{b\to\infty} \int_1^b \frac{dx}{x^\alpha}$$

if the limit exists, which is if and only if $\alpha > 1$, as may be verified by the student. The value in that case is $(\alpha - 1)^{-1}$.

These two examples are instances of the definition of improper integrals as limits of proper integrals. Similar procedures can be used to define improper double and triple integrals. For a more thorough discussion of improper integrals see Chapter 22.

18.9 / STIELTJES INTEGRALS

The purpose of this section is to give a brief elementary introduction to the subject of Stieltjes integrals. These integrals are named after a Dutch mathematician, T. J. Stieltjes (1856–1894). They have long been used by mathematicians as a tool in theoretical investigations. The current tendency is in the direction of a wider usage, and the student of pure or applied mathematics will sometimes encounter the Stieltjes integral in his reading before he has had a chance to learn much about the integral. The common practice is to discuss the theory of the Stieltjes integral in courses in the theory of functions of a real variable, along with (or after) the study of functions of bounded variation. Our intent here is to break the ice much earlier, particularly for the student who may never take the courses just mentioned. Our discussion will necessarily deal only with the rudiments, and will stress definitions, basic properties, and illustrations. Proofs are omitted. In large part they are very similar to proofs in the theory of Riemann integrals.

The Stieltjes integral involves two functions f and g, each defined on a closed interval $[a, b]$. It is denoted by

$$\int_a^b f(x)\, dg(x). \tag{18.9–1}$$

In the special case in which g is the simple function $g(x) = x$, the Stieltjes integral (18.9–1) becomes the Riemann integral

$$\int_a^b f(x)\, dx.$$

To define the Stieltjes integral (18.9–1) we start with a partition P of $[a, b]$: $(x_0, x_1, \ldots, x_n)$ and a set of points $x_1', \ldots, x_n'$, one in each subinterval of the partition. We then define

$$\int_a^b f(x)\, dg(x) = \lim_{|P|\to 0} \sum_{i=1}^n f(x_i')[g(x_i) - g(x_{i-1})], \tag{18.9–2}$$

provided the sums converge to a unique limit as $|P| \to 0$. Here $|P|$ denotes the mesh fineness of the partition P, as defined in §18.2.

Example 1. Find the value of the integral if $f(x) = 1$ and $g(x) = x^2$. The sum in this case is

$$(x_1^2 - x_0^2) + (x_2^2 - x_1^2) + \cdots + (x_n^2 - x_{n-1}^2) = x_n^2 - x_0^2 = b^2 - a^2.$$

Therefore

$$\int_a^b f(x)\, dg(x) = b^2 - a^2$$

in this case.

This argument works just as well for any $g(x)$ when $f(x) = 1$ for all x. Thus

$$\int_a^b 1 \cdot dg(x) = g(b) - g(a), \tag{18.9–3}$$

no matter what kind of a function g is. In particular, g need not be continuous in (18.9–3).

If g is constant on $[a, b]$, all the differences $g(x_i) - g(x_{i-1})$ are zero, and so we see that

$$\int_a^b f(x)\, dg(x) = 0 \tag{18.9–4}$$

when g is constant on $[a, b]$. This is true no matter what kind of function f is.

One of the important practical uses of Stieltjes integrals involves the case in which g is a discontinuous function which has a finite number of discontinuities at which it jumps suddenly in value, but remains constant in value in the open intervals between the points of discontinuity. The simplest case is that in which the only discontinuities are at the endpoints a and b.

Example 2. Suppose $g(x) = c$ if $a < x < b$, and let $g(a)$ and $g(b)$ have any values whatever (see Fig. 171). In this case we can show that

$$\int_a^b f(x)\, dg(x) = f(a)[c - g(a)] + f(b)[g(b) - c], \tag{18.9–5}$$

where f is any continuous function.

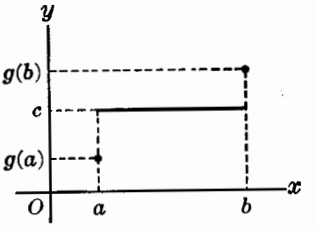


Fig. 171.

To see this, let $(x_0, x_1, \ldots, x_n)$ be any partition of $[a, b]$. Then the differences $g(x_i) - g(x_{i-1})$ will all be zero as long as both points x_{i-1}, x_i are in the open interval, since $g(x) = c$ there. Hence in (18.9–2) only the terms for $i = 1$, $i = n$ remain, and

$$\int_a^b f(x)\, dg(x) = \lim_{|P| \to 0} \{f(x_1')[g(x_1) - g(x_0)] + f(x_n')[g(x_n) - g(x_{n-1})]\}.$$

But $x_0 = a$, $x_n = b$, $g(x_1) = g(x_{n-1}) = c$. Also, $x_1' \to a$ and $x_n' \to b$ when $|P| \to 0$.

Therefore, because of the continuity of f, $f(x_1') \to f(a)$, $f(x_n') \to f(b)$. When these results are assembled, we see that (18.9–5) holds.

Observe that the differences $c - g(a)$, $g(b) - c$ are the amounts by which the value $g(x)$ jumps at the points of discontinuity as x moves to the right.

If there are more points of discontinuity it is easy to extend the formula (18.9–5).

Example 3. Suppose $[a, b]$ is divided into N parts by the partition $(a_0, a_1, \ldots, a_N)$, and suppose $g(x)$ has the value c_i in the interior of the ith subinterval (see Fig. 172). The values of $g(x)$ at the points $a_0, \ldots, a_N$ can be

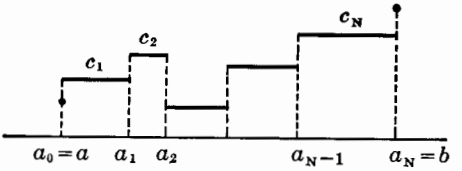


Fig. 172.

quite independent of the values $c_1, \ldots, c_N$. If f is continuous on $[a, b]$ we can show that

$$\int_a^b f(x)\, dg(x) = f(a)[c_1 - g(a)] + f(a_1)(c_2 - c_1) + \cdots + f(a_{N-1})(c_N - c_{N-1}) + f(b)[g(b) - c_N]. \tag{18.9–6}$$

Observe that the values of g at $a_1, \ldots, a_{N-1}$ (the *interior* points of discontinuity) do not enter into the formula.

The proof of (18.9–6) can be given with the aid of the general formula

$$\int_a^b f(x)\, dg(x) = \int_a^{a_1} f(x)\, dg(x) + \int_{a_1}^{a_2} f(x)\, dg(x) + \cdots + \int_{a_{N-1}}^b f(x)\, dg(x). \tag{18.9–7}$$

This last formula is valid for any functions f, g such that all the integrals exist.

As a concrete illustration of the kind of thing arising in Example 3, consider the moment of inertia of a number of mass particles distributed along the x-axis. Let the masses $m_1, \ldots, m_n$ be located as shown in Fig. 173, with $x_1 < x_2 < \cdots <$

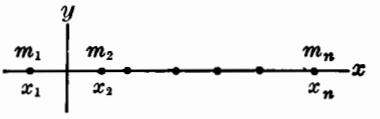


Fig. 173.

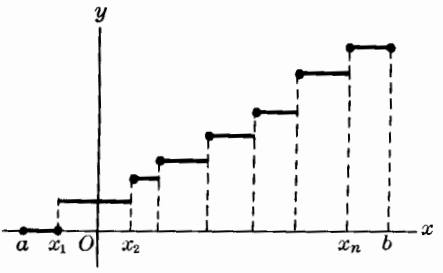


Fig. 174.

x_n. The moment of inertia of this mass system about the y-axis is

$$I = m_1x_1^2 + \cdots + m_nx_n^2. \tag{18.9–8}$$

We shall show how (18.9–8) can be written as a Stieltjes integral. Take any closed interval $[a, b]$ containing all the points $x_1, \ldots, x_n$. For any x in $[a, b]$ define $g(x) = 0$ if $a \leqq x \leqq x_1$, and, if $x_1 < x$, define $g(x)$ as the sum of all the masses m_i for which $a \leqq x_i \leqq x$. The graph of $g(x)$ appears in Fig. 174.
The exact definition of $g(x)$ is

$$\begin{aligned}
&g(x) = 0 && \text{if } a \leqq x \leqq x_1\\
&g(x) = m_1 && \text{if } x_1 < x < x_2\\
&g(x) = m_1 + m_2 && \text{if } x_2 \leqq x < x_3\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&g(x) = m_1 + m_2 + \cdots + m_n && \text{if } x_n \leqq x \leqq b.
\end{aligned}$$

In this case there is no jump at b, and there is a jump at a if and only if $a = x_1$. Formula (18.9–6) now becomes

$$\int_a^b f(x)\, dg(x) = f(x_1)m_1 + \cdots + f(x_n)m_n.$$

In particular, if $f(x) = x^2$ we see that

$$\int_a^b x^2\, dg(x) = m_1x_1^2 + \cdots + m_nx_n^2 = I.$$

This is the desired expression of (18.9–8) as a Stieltjes integral.

It is important to know some of the general conditions under which it is certain that the limit (18.9–2) defining the Stieltjes integral will exist. As we have seen, there are interesting and important cases where the integral exists when g has certain discontinuities. The most important conditions sufficient to guarantee the existence of the integral are that f be continuous and that g be either (1) a nondecreasing function, (2) a nonincreasing function, or (3) the sum of a nondecreasing function and a nonincreasing function. A function g is called

nondecreasing if its values never decrease as x increases; it is called *nonincreasing* if $g(x)$ never increases as x increases. A function g of the third type in the foregoing classification is said to be of *bounded variation* on $[a, b]$. There is an alternative way of defining such functions, based on the notion of computing the sum of the total increase and the total decrease of the values of the function as x goes from a to b. We do not have enough space to discuss in detail the interesting and important properties of functions of bounded variation.

It can be shown that if $\int_a^b f(x)\,dg(x)$ exists, then $\int_a^b g(x)\,df(x)$ also exists, and

$$\int_a^b f(x)\,dg(x) = f(b)g(b) - f(a)g(a) - \int_a^b g(x)\,df(x). \tag{18.9–9}$$

This is called the formula of integration by parts.

Still another important fact about Stieltjes integrals is that if g has a continuous derivative and if f is integrable in the Riemann sense, then

$$\int_a^b f(x)\,dg(x) = \int_a^b f(x)g'(x)\,dx, \tag{18.9–10}$$

where the integral on the right is a Riemann integral.

Example 4. Consider a linear mass distribution on the interval $a \leqq x \leqq b$. Suppose the density ρ at the point x is a continuous function of x, denoted by $\rho(x)$. Then the mass on the interval $[a, x]$ is

$$m(x) = \int_a^x \rho(t)\,dt.$$

We know that $m'(x) = \rho(x)$; therefore (18.9–10) applies with $m(x)$ in place of $g(x)$. In particular, the total mass is

$$M = \int_a^b 1 \cdot dm(x) = \int_a^b \rho(x)\,dx$$

and the abscissa $\bar{x}$ of the center of mass is given by

$$M\bar{x} = \int_a^b x\,dm(x) = \int_a^b x\rho(x)\,dx.$$

The function $m(x)$ in Example 4 is called a *cumulative mass distribution function.* This same name applies to the function $g(x)$ which is discussed in connection with Fig. 174, but in that case there is no continuous density function. One of the great uses of the Stieltjes integrals is in the unification of ideas and formulas about mass distributions, whether discrete or continuous.

The notion of a distribution function occurs also in the theory of probability, with applications in statistics, and Stieltjes integrals play an important part in the general formulation of such concepts as *first moment, variance, and mathematical expectation.*

EXERCISES

1. Suppose $g(x) = 1$ if $0 \leqq x < 1$ and $g(x) = 4$ if $1 < x \leqq 2$. Calculate the values of $\int_0^2 x^k \, dg(x)$ for $k = 0, 1, 2$.

2. Suppose $g(x) = x$ if $0 < x < 1$, $g(x) = 2 - x$ if $1 < x < 2$, $g(0) = 1$, $g(1) = 0$, $g(2) = 2$. Calculate $\int_0^1 x \, dg(x)$, $\int_1^2 x \, dg(x)$, $\int_0^2 x \, dg(x)$, and check by (18.9–7).

3. Suppose $g(x) = 5 - n$ if $n - 1 < x < n$, for $n = -1, 0, 1, 2, 3$, and $g(n) = n$. Compute $\int_{-2}^3 x^k \, dg(x)$ for $k = 0, 1, 3$.

4. If $g(x) = e^{-x}$, find **(a)** $\int_{0^-}^1 x \, dg(x)$, **(b)** $\int_0^1 g(x) \, dg(x)$.

5. If $g(x) = \tan^{-1} x$ find **(a)** $\int_{\sqrt{3}}^1 x \, dg(x)$ and **(b)** $\int_{-1}^{\sqrt{3}} g(x) \, dg(x)$.

6. Let $g(x) = [x]$ (the greatest integer n such that $n \leqq x$).

(a) Find $\int_0^a x \, dg(x)$ for the successive values $a = \frac{1}{2}, 1, \frac{3}{2}, 2$.

(b) Find $\int_0^3 g(x) \, d(\sqrt{1 + x^3})$.

7. A uniform rod 10 feet long weighs 5 pounds and carries three concentrated weights of 1 pound each fastened at each end and at the point $x = 5$ (the rod extending from $x = 0$ to $x = 10$). Define $w(x) =$ total weight on $[0, x]$ if $x > 0$, $w(0) = 0$. **(a)** Draw the graph of $w(x)$. **(b)** Calculate $\int_0^{10} x \, dw(x)$ and $\int_0^{10} (x - 5)^2 \, dw(x)$.

8. Let $\phi(x)$ be continuous and positive on $[a, b]$. Let $A(x)$ be the area between $y = \phi(x)$ and $y = 0$ from a to x.

(a) Express the total area under the curve, and the first and second moments of this area with respect to the y-axis, as Stieltjes integrals.

(b) Express the y-co-ordinate of the centroid of the area by a formula involving a Stieltjes integral with $dA(x)$.

9. If the area in Exercise 8 is revolved around the x-axis, let $V(x)$ be the volume generated by $A(x)$. Express the co-ordinate $\bar{x}$ for the centroid of the total volume by a formula involving a Stieltjes integral.

10. Use (18.9–5) and (18.9–7) to prove (18.9–6).

19 / INFINITE SERIES

19 / DEFINITIONS AND NOTATION

It is likely that a person studying this book has already learned some things about infinite series at a more elementary level of study. However, to meet the needs of students with varying degrees of experience, the basic ideas of the subject are presented here without depending on what the student may already know about infinite series. The main reason for studying infinite series is that they are widely used to define or represent functions. One must know some of the tests, methods, and techniques explained in this chapter in order to follow the reasoning and understand the use of infinite series in the theory of functions. Also, one must have some facility in using infinite series in the study of differential equations and in applied mathematics. Chapter 19 is mainly concerned with questions of convergence and divergence of infinite series. Chapters 20 and 21 are concerned with the study of functions defined by infinite series or infinite sequences.

Perhaps the simplest of all infinite series is the geometric series

$$1+x+x^2+x^3+\cdots+x^n+\cdots. \tag{19–1}$$

Let us briefly summarize the main facts about this series. We start from the algebraic identity

$$1-x^{n+1}=(1-x)(1+x+x^2+\cdots+x^n), \quad n\geqq 1$$

and rewrite it in the form

$$\frac{1}{1-x}=1+x+\cdots+x^n+\frac{x^{n+1}}{1-x}, \quad x\neq 1. \tag{19–2}$$

If $-1<x<1$ we see that $\lim_{n\to\infty}x^{n+1}=0$; therefore

$$\lim_{n\to\infty}(1+x+\cdots+x^n)=\frac{1}{1-x}.$$

We write this last result in the form

$$\frac{1}{1-x}=1+x+x^2+\cdots+x^n+\cdots, \quad -1<x<1. \tag{19–3}$$

Here we have an infinite-series representation of the function $(1-x)^{-1}$, valid subject to certain limitations on x.

Let us now make some formal definitions. Suppose that u_0, u_1,

$u_2, \ldots, u_n, \ldots$ is an infinite sequence of numbers. The expression

$$u_0 + u_1 + u_2 + \cdots + u_n + \cdots \tag{19–4}$$

is called an *infinite series*, and the numbers $u_0, u_1, \ldots$ are called the *terms* of the series. The series is in a formal sense just a certain collection of mathematical symbols arranged in a certain way. Now we consider the sequence of numbers $s_0, s_1, s_2, \ldots, s_n, \ldots$ formed as follows:

$$\begin{aligned} s_0 &= u_0 \\ s_1 &= u_0 + u_1 \\ s_2 &= u_0 + u_1 + u_2 \\ &\vdots \\ s_n &= u_0 + u_1 + \cdots + u_n. \end{aligned}$$

If the sequence s_n has a limit as $n \to \infty$ we say that the series (19–4) is *convergent*; if the limit of s_n is s we say that s is the *sum* (or value) of the series, and we write

$$s = u_0 + u_1 + u_2 + \cdots + u_n + \cdots.$$

If the series is not convergent we say that it is *divergent*, and we do not assign it any sum. The numbers $s_0, s_1, \ldots, s_n, \ldots$ are called the *partial sums* of the series.

We have made the definitions in the case where the terms are numerical constants. If the u's are functions of a variable x, they assume definite numerical values when x is given a fixed value, and then all the foregoing definitions apply. In such a case the partial sums s_n will also be functions of x.

Let us return to the series (19–1). We have seen that it is convergent if $-1 < x < 1$, the sum then being given by (19–3). For example,

$$\frac{3}{2} = 1 + \frac{1}{3} + \frac{1}{3^2} + \frac{1}{3^3} + \cdots + \frac{1}{3^n} + \cdots.$$

But if $x \leqq -1$ or $1 \leqq x$ the series is divergent. If $x \geqq 1$ this is clear at once, for then $s_n \geqq n + 1$, and so $s_n \to +\infty$ as $n \to \infty$. If $x = -1$ the series becomes

$$1 - 1 + 1 - 1 + \cdots,$$

with $s_0 = 1$, $s_1 = 0$, $s_2 = 1$, $s_3 = 0, \ldots$. This sequence $\{s_n\}$ is not convergent, and so the series is divergent. How do the partial sums behave when $x < -1$?

From a given series (19–4) we may form new series by omitting a certain finite number of terms at the beginning, e.g.,

$$u_2 + u_3 + u_4 + \cdots,$$
$$u_{100} + u_{101} + \cdots.$$

All such series will be convergent if the original series is convergent, and divergent if the original series is divergent (why?). Likewise we may form a new series by multiplying each term of the original series by the same constant c:

$$cu_0 + cu_1 + \cdots + cu_n + \cdots.$$

If $c \neq 0$ this series will be convergent if and only if the original series is convergent (why?). If the original series had sum s, the new series has sum cs.

We frequently use the summation symbol Σ in dealing with series:

$$\sum_{n=1}^{N} u_n = u_1 + u_2 + \cdots + u_N, \tag{19–5}$$

$$\sum_{i=m}^{n} u_i = u_m + u_{m+1} + \cdots + u_n. \tag{19–6}$$

Observe that the index n in (19–5) and the index i in (19–6) are *dummy indices*; this means that the expression is not altered if the index letter is changed:

$$\sum_{n=1}^{N} u_n = \sum_{k=1}^{N} u_k.$$

A sum such as (19–6) is in certain respects analogous to an integral $\int_a^b f(x)\,dx$, in which s is a dummy variable. The u_i of (19–6) corresponds to $f(x)$; the limits of summation $i = m$ and $i = n$ correspond to the limits $x = a$ and $x = b$ on the integral.

The infinite series (19–4) is now denoted by

$$\sum_{n=0}^{\infty} u_n, \tag{19–7}$$

whether it is convergent or not. Often, for convenience in printing, we write Σu_n in place of (19–7), suppressing the limits 0 and ∞.

Finally, we note that u_n is not necessarily the nth term of the series. The index notation merely expresses the fact that u_n is a function of n. *It is very important to observe that if the series* (19–7) *is convergent, then* $u_n \to 0$ *as* $n \to \infty$. For $u_n = s_n - s_{n-1}$. If the series is convergent, s_n and $s_{n-1} \to s$ as $n \to \infty$, and so $\lim u_n = \lim s_n - \lim s_{n-1} = 0$. But although $u_n \to 0$ is a *necessary* condition for convergence, it is by no means *sufficient*. A series may diverge, even though $u_n \to 0$ (see Example 1, §19.2).

EXERCISES

1. For what values of x is each of the following series convergent? Express the sum of the series as a simple function of x in each case.

(a) $a + ax + ax^2 + \cdots + ax^n + \cdots$, $a \neq 0$;

(b) $8x + 8x^3 + 8x^5 + \cdots + 8x^{2n+1} \cdots$;

(c) $cx^2 + cx^4 + cx^6 + \cdots + cx^{2n} + \cdots$;

(d) $\dfrac{10}{x} + \dfrac{10}{x^2} + \dfrac{10}{x^3} + \cdots$;

(e) $1 + \dfrac{1}{1+x} + \dfrac{1}{(1+x)^2} + \cdots + \dfrac{1}{(1+x)^n} + \cdots$;

(f) $2 + 2\left(\dfrac{1-x}{1+x}\right) + 2\left(\dfrac{1-x}{1+x}\right)^2 + \cdots$;

(g) $e^x + e^{2x} + e^{3x} + \cdots + e^{nx} + \cdots$;

(h) $\log x + (\log x)^2 + \cdots + (\log x)^n + \cdots$.

2. Explain why the series

$$\frac{1+1}{1+2}+\frac{1+2}{1+4}+\cdots+\frac{1+n}{1+2n}+\cdots$$

is divergent.

3. Prove that the series

$$\frac{2}{\sqrt{1000+4}}+\frac{3}{\sqrt{1000+9}}+\cdots+\frac{n}{\sqrt{1000+n^2}}+\cdots$$

is divergent.

4. Discuss the behavior of the partial sums s_n in the case of each of the following series:

(a) $1-2+3-4+5-\cdots$;

(b) $1+\frac{1}{2}-1+\frac{1}{4}-1+\frac{1}{8}-1+\cdots$;

(c) $(1-\frac{1}{2})+(\frac{1}{2}-\frac{1}{3})+(\frac{1}{3}-\frac{1}{4})+\cdots$;

(d) $(\sqrt{2}-\sqrt{1})+(\sqrt{3}-\sqrt{2})+(\sqrt{4}-\sqrt{3})+\cdots$.

Which, if any, of the series are convergent? One can tell that two of the series are divergent without investigating the partial sums s_n. Which are these series, and how does one know by quick inspection that they are divergent?

5. Is the series

$$\frac{1^1}{(101)!}+\frac{2^2}{(102)!}+\cdots+\frac{n^n}{(100+n)!}+\cdots$$

convergent or divergent? Justify your answer.

19.1 / TAYLOR'S SERIES

In order to become familiar with a number of interesting examples of infinite series, let us examine how we are led to study the representation of functions by infinite series. The starting point is Taylor's formula with remainder (Chapter 4, especially §4.3). For instance, in Examples 3 and 4 of §4.3 we saw that if $x>-1$,

$$\log(1+x)=x-\tfrac{1}{2}x^2+\tfrac{1}{3}x^3-\cdots+(-1)^{n-1}\frac{1}{n}x^n+R_{n+1},$$

where

$$|R_{n+1}|\leqq\frac{|x|^{n+1}}{n+1}\quad\text{if}\quad 0\leqq x\leqq 1$$

and

$$|R_{n+1}|\leqq\frac{|x|^{n+1}}{1+x}\quad\text{if}\quad -1<x\leqq 0.$$

These inequalities show that $R_{n+1}\to 0$ as $n\to\infty$ when x is limited as indicated. Therefore

$$\log(1+x)=\lim_{n\to\infty}\left[x-\tfrac{1}{2}x^2+\tfrac{1}{3}x^3-\cdots+(-1)^{n-1}\frac{1}{n}x^n\right]$$

if $-1<x\leqq 1$. According to the definitions in §19 this result may be written in the

form

$$\log(1+x) = x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 - \cdots + (-1)^{n-1}\frac{1}{n}x^n + \cdots. \tag{19.1-1}$$

As a special instance, let $x = 1$; then

$$\log 2 = 1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} + \cdots + (-1)^{n-1}\frac{1}{n} + \cdots. \tag{19.1-2}$$

Another series representation of $\log 2$ may be obtained by putting $x = -\frac{1}{2}$ in (19.1–1). We find

$$\log \tfrac{1}{2} = -\log 2 = -\tfrac{1}{2} - \tfrac{1}{2}(-\tfrac{1}{2})^2 + \tfrac{1}{3}(-\tfrac{1}{2})^3 - \cdots,$$

$$\log 2 = \frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2^2} + \frac{1}{3}\cdot\frac{1}{2^3} + \cdots + \frac{1}{n}\cdot\frac{1}{2^n} + \cdots. \tag{19.1-3}$$

Now suppose that f is a function which has derivatives of all orders on an open interval containing the point $x = a$. According to Taylor's formula (4.3–8) we have, for any x of this interval,

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_{n+1},$$

where the remainder R_{n+1} can be expressed in a variety of different forms. If we can show, for a particular x, that $\lim R_{n+1} = 0$, then it follows that $f(x)$ can be represented by an infinite series

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \tag{19.1-4}$$

(with the usual convention that $n! = 1$ if $n = 0$). This is called *Taylor's series expansion of $f(x)$ about the point $x = a$* (also sometimes called the expansion of $f(x)$ in powers of $x - a$). Observe that we are *not* asserting that (19.1–4) is always true. It *will* be true if $\lim_{n\to\infty} R_{n+1} = 0$, but this is a matter to be investigated for each particular function and each particular x.

Example 1. The function e^x can be represented by Taylor's series for all values of x, no matter how the point $x = a$ is chosen.

In order to prove what has just been stated, let us first choose $a = 0$. Then, since $f^{(n)}(x) = e^x$ for all orders n, Taylor's formula with remainder is

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + R_{n+1}.$$

Lagrange's form of the remainder is

$$R_{n+1} = \frac{e^X}{(n+1)!}x^{n+1},$$

with X between 0 and x. Thus certainly $-|x| < X < |x|$, and $0 < e^X < e^{|x|}$; there-

fore

$$|R_{n+1}| \leqq e^{|x|}\frac{|x|^{n+1}}{(n+1)!}. \tag{19.1–5}$$

We shall prove that $\lim_{n\to\infty} R_{n+1} = 0$. From (19.1–5) we see that it will be sufficient to prove that $\lim_{n\to\infty} \frac{|x|^n}{n!} = 0$. Now choose an integer N such that $N \geqq 2|x|$. Then, if $n > N$,

$$\frac{|x|^n}{n!} = \frac{|x|^N}{N!}\frac{|x|^{n-N}}{(N+1)(N+2)\cdots n}$$

$$\leqq \frac{|x|^N}{N!}\left(\frac{N}{N+1}\right)\left(\frac{N}{N+2}\right)\cdots\left(\frac{N}{n}\right)\frac{1}{2^{n-N}},$$

$$\frac{|x|^n}{n!} \leqq \frac{|x|^N}{N!}\left(\frac{1}{2}\right)^{n-N}.$$

Keeping N fixed and letting $n \to \infty$, we see that $(\frac{1}{2})^{n-N} \to 0$, and so the desired result is attained. Therefore the series representation

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots \tag{19.1–6}$$

is valid for all values of x.

Now consider any point $x = a$. Taylor's formula for e^x now takes the form

$$e^x = e^a + e^a(x-a) + \cdots + e^a\frac{(x-a)^n}{n!} + e^X\frac{(x-a)^{n+1}}{(n+1)!},$$

with X between a and x. A very slight modification of the foregoing arguments shows that the remainder approaches zero as $n \to \infty$, so that the representation

$$e^x = e^a + e^a(x-a) + \cdots + e^a\frac{(x-a)^n}{n!} + \cdots \tag{19.1–7}$$

is valid for all values of x.

Example 2. The functions $\sin x$ and $\cos x$ can be represented by Taylor's series for all values of x, no matter how the point $x = a$ is chosen. The series take particularly simple forms when $a = 0$, namely

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \tag{19.1–8}$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots. \tag{19.1–9}$$

The details are left to the student (see Exercise 1).

In general the problem of finding a Taylor's series representation for a given function, and proving that the representation is valid, can be a very difficult task, especially if we approach the problem directly as in the foregoing examples. It

may prove to be a very complicated matter to calculate the derivatives $f^{(n)}(x)$ for higher values of n, and one cannot always expect to find a manageable general formula for the coefficient $\frac{f^{(n)}(a)}{n!}$. Moreover, if one cannot get a reasonably simple formula for $f^{(n+1)}(x)$, there is very little chance of proving that $R_{n+1} \to 0$ as $n \to \infty$ by means of the standard formulas for the remainder (Lagrange, Cauchy, or the integral form (4.2–8)). There are, however, more advanced methods for attacking the problem of finding out whether a function can be represented by Taylor's series. The most important of these methods belongs to the theory of functions of a complex variable, and is beyond the scope of this book.

In certain particular cases which are interesting and important there are special devices which enable us to find series representations without the necessity of calculating a general formula for $f^{(n)}(x)$. One such device is considered in §19.11. Other devices are considered in Chapter 21.

EXERCISES

1. Prove the validity of the series expansions (19.1–8) and (19.1–9) for $\sin x$ and $\cos x$.
2. Prove the validity of the expansion

$$(1-x)^{-1/2} = 1 + \tfrac{1}{2}x + \frac{1 \cdot 3}{2 \cdot 4}x^2 + \cdots + \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots 2n}x^n + \cdots$$

when $-1 < x < 1$. SUGGESTION: See Exercise 7, §4.3. For the case $0 < x < 1$ it is convenient to use the fact that $na^n \to 0$ as $n \to \infty$ if $|a| < 1$. Prove this fact separately.

19.11 / A SERIES FOR THE INVERSE TANGENT

In the formula (19–2) let us put $x = -t^2$. Then

$$\frac{1}{1+t^2} = 1 - t^2 + t^4 - \cdots + (-1)^n t^{2n} + (-1)^{n+1}\frac{t^{2n+2}}{1+t^2}. \qquad (19.11\text{–}1)$$

Integrating both sides of this algebraic identity, we have, for any x,

$$\int_0^x \frac{dt}{1+t^2} = \tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + (-1)^n \frac{x^{2n+1}}{2n+1} + (-1)^{n+1}\int_0^x \frac{t^{2n+2}}{1+t^2}dt.$$

We apply Theorem V of §4.4 to the integral on the right:

$$\int_0^x \frac{t^{2n+2}}{1+t^2}\,dt = \frac{1}{1+X^2}\int_0^x t^{2n+2}\,dt = \frac{1}{1+X^2}\frac{x^{2n+3}}{2n+3},$$

where X is between 0 and x. Now if $|x| \leqq 1$,

$$\lim_{n\to\infty} \frac{x^{2n+3}}{2n+3} = 0,$$

and therefore we conclude that

$$\tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + (-1)^n \frac{x^{2n+1}}{2n+1} + \cdots. \qquad (19.11\text{–}2)$$

This series representation is valid if $-1 \leqq x \leqq 1$. In particular, if $x = 1$, we find the interesting result

$$\frac{\pi}{4} = 1 - \tfrac{1}{3} + \tfrac{1}{5} - \tfrac{1}{7} + \tfrac{1}{9} - \cdots. \tag{19.11-3}$$

The series (19.11–2) is actually the Taylor's series expansion of $\tan^{-1} x$ in powers of x, even though we did not find the terms of the series by calculating the successive derivatives of $\tan^{-1} x$. This latter procedure would lead to great complications, as the student may find if he attempts it. But it can be shown in the general theory of power series, that if a function can be represented by a series of powers of $x - a$, then the series is actually the Taylor's series expansion of the function about the point $x = a$.

EXERCISES

1. Show that $\pi = 16 \tan^{-1} \frac{1}{5} - 4 \tan^{-1} \frac{1}{239}$. Start by setting $\phi = \tan^{-1} \frac{1}{5}$ and computing $\tan 2\phi$ and $\tan 4\phi$ by repeated use of the formula for the tangent of twice an angle. Then compute $\tan\left(4\phi - \frac{\pi}{4}\right)$. If the inverse tangents in this formula for π are computed by the series (19.11–2), a value of π accurate to a number of decimal places may be found without excessive labor. As a sample, let the student obtain the approximation $\pi = 3.141593$, of which the first five decimal places are accurate and the sixth is correct as a rounded-off figure.

2. Show that $2 \tan^{-1} \frac{1}{10} - \tan^{-1} \frac{1}{5} = \tan^{-1} \frac{1}{515}$. Combine this with the result of Exercise 1 to obtain the formula

$$\frac{\pi}{4} = 8 \tan^{-1} \frac{1}{10} - 4 \tan^{-1} \frac{1}{515} - \tan^{-1} \frac{1}{239}.$$

Try this formula along with (19.11–2) in the computation of π.

19.2 / SERIES OF NONNEGATIVE TERMS

For a deeper study of the general problem of representation of functions by infinite series it is essential to have a certain fund of general knowledge about infinite series as such. We turn then to a different problem. Suppose we have before us a series, no matter how obtained. What can we do to determine whether or not the series is convergent? We begin by considering the special case in which all the terms of the series are nonnegative. Special though this case is, what we learn about it will have important applications to the general problem.

THEOREM I. *Suppose that $u_n \geqq 0$ for every n. Then the series Σu_n is convergent if and only if the sequence $\{s_n\}$ of partial sums is bounded.*

Proof. By definition $s_n = u_0 + u_1 + \cdots + u_n$. Since $u_n \geqq 0$, it is clear that $0 \leqq s_n \leqq s_{n+1}$. If the partial sums are bounded, the sequence $\{s_n\}$ has a limit, by Theorem III, §2.7. On the other hand, if the sequence is not bounded, then $s_n \to +\infty$ as $n \to \infty$. This completes the proof of the theorem.

Example 1. The series

$$1+\tfrac{1}{2}+\tfrac{1}{3}+\cdots+\frac{1}{n}+\cdots \tag{19.2-1}$$

is divergent. It is called the *harmonic series.*

We prove the divergence by showing that the partial sums are not bounded. Let

$$s_n = 1+\tfrac{1}{2}+\cdots+\frac{1}{n}.$$

Then

$$s_{2n} = s_n + \frac{1}{n+1}+\frac{1}{n+2}+\cdots+\frac{1}{2n} > s_n + \frac{1}{2},$$

since

$$\frac{1}{n+1}+\frac{1}{n+2}+\cdots+\frac{1}{2n} > n\cdot\frac{1}{2n} = \frac{1}{2}.$$

With $s_{2n} > s_n + \frac{1}{2}$ for every n, it is plainly impossible for $\{s_n\}$ to be bounded. We have

$$s_1 = 1,\ s_2 = \tfrac{3}{2},\ s_4 > s_2 + \tfrac{1}{2} = 2,\ s_8 > s_4 + \tfrac{1}{2} > \tfrac{5}{2},$$

and in general

$$s_{2^n} > \frac{n+2}{2}.$$

For a series with negative as well as positive terms, convergence of the series is *not* guaranteed by boundedness of the partial sums. For instance, the series

$$1-1+1-1+1-\cdots$$

is not convergent, yet for its partial sums we have $s_n = 1$ or $s_n = 0$, depending on the oddness or evenness of n. Thus these partial sums are bounded.

THEOREM II. *Let Σa_n and Σb_n be two series of nonnegative terms, and suppose that, for all values of n after some fixed index N, it is true that $a_n \leqq b_n$. Then if the series Σb_n is convergent, so is Σa_n, and if the series Σa_n is divergent, so is Σb_n.*

Proof. In discussing convergence or divergence we may drop the terms with index less than N. Then, for any $n > N$,

$$a_N + a_{N+1}+\cdots+a_n \leqq b_N + b_{N+1}+\cdots+b_n.$$

The proof of the theorem is an immediate consequence of this inequality and Theorem I.

Example 2. The series

$$1+\frac{1}{2^2}+\frac{1}{3^3}+\cdots+\frac{1}{n^n}+\cdots \tag{19.2–2}$$

is convergent. For $\frac{1}{n^n}<\frac{1}{n!}$ if $n>1$, and the series

$$1+\frac{1}{2!}+\frac{1}{3!}+\cdots+\frac{1}{n!}+\cdots \tag{19.2–3}$$

is known to be convergent (see the series (19.1–6) for e^x with $x=1$; see also Example 5, §1.62).

THEOREM III. *Let Σa_n and Σb_n be two series of positive terms, and suppose that a_n/b_n approaches a nonzero limit as $n\to\infty$. Then either both series are convergent, or both are divergent.*

Proof. Suppose that $a_n/b_n\to c$, and $c\neq 0$. Then, for all sufficiently large values of n we shall have $\frac{1}{2}c<a_nb_n<\frac{3}{2}c$, whence

$$b_n<\left(\frac{2}{c}\right)a_n \quad \text{and} \quad a_n<\left(\frac{3}{2}c\right)b_n.$$

Since the convergence or divergence of a series is not affected by multiplying each of its terms by the same nonzero constant, the foregoing inequalities together with Theorem II are sufficient to prove Theorem III.

Example 3. The series

$$\frac{2^2}{3!}+\frac{3^2}{4!}+\cdots+\frac{(n+1)^2}{(n+2)!}+\cdots$$

is convergent. We prove this by using the convergent series (19.2–3) and applying Theorem III:

$$\frac{(n+1)^2}{(n+2)!}\Big/\frac{1}{n!}=\frac{(n+1)^2}{(n+1)(n+2)}=\frac{n+1}{n+2}\to 1.$$

Example 4. The series

$$1+\frac{1}{3}+\frac{1}{5}+\cdots+\frac{1}{2n-1}+\cdots$$

is divergent. We prove this by using the series (19.2–1) and applying Theorem III.

$$\frac{1}{2n-1}\Big/\frac{1}{n}=\frac{n}{2n-1}\to\frac{1}{2}.$$

EXERCISES

1. Test the following series for convergence or divergence by Theorem II.

(a) $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n}}$.

(b) $\sum_{n=2}^{\infty} \frac{1}{\log n}$.

(c) $1+\frac{1}{1\cdot 3}+\frac{1}{1\cdot\cdot 3\cdot\cdot 5}+\cdots+\frac{1}{1\cdot 3\cdots(2n-1)}+\cdots$.

(d) $\frac{1}{2\cdot 2}+\frac{2}{3\cdot 2^2}+\cdots+\frac{n}{(n+1)2^n}+\cdots$.

2. Prove that $(n+1)/n! < 8/2^n$ if $n \geqq 1$. What do you conclude about the following series?

$$\frac{2}{1!}+\frac{3}{2!}+\cdots+\frac{n+1}{n!}+\cdots.$$

3. Compare the series

$$\frac{4\cdot 2}{3!}+\frac{5\cdot 3}{4!}+\cdots+\frac{(n+3)(n+1)}{(n+2)!}+\cdots$$

with a suitable multiple of the series (19.2–3), and use Theorem II to establish convergence. Also prove the convergence of the series by using the series (19.2–3) and Theorem III.

4. Compare the series

$$\frac{3}{5}+\frac{3\cdot 5}{5\cdot 10}+\cdots+\frac{3\cdot 5\cdot 7\cdots(2n+1)}{5\cdot 10\cdot 15\cdots 5n}+\cdots$$

with a suitable multiple of the series

$$1+\frac{1}{2}+\cdots+\frac{1}{2^{n-1}}+\cdots,$$

and use Theorem II to establish convergence.

5. Test the following series for convergence or divergence by Theorem III.

(a) $\sum_{n=1}^{\infty} \frac{n+1}{(n+2)n!}$.

(b) $\sum_{n=1}^{\infty} \frac{2n+1}{n^2+n}$.

(c) $\sum_{n=1}^{\infty} \frac{n^2}{n^3+1}$.

(d) $\sum_{n=1}^{\infty} \frac{\sqrt{n+1}}{n^{n+1/2}}$.

(e) $\sum_{n=1}^{\infty} \frac{(n+1)(n+2)}{n^2\cdot 2^n}$.

(f) $\sum_{n=1}^{\infty} n^{(1-n)/n}$ (see Exercise 17, §1.62).

(g) $\sum_{n=1}^{\infty} \frac{(n+1)^n}{n^{2n}}$ (see (1.62–5)).

6. Prove that, if Σa_n and Σb_n are series of positive terms with Σb_n convergent and $a_n/b_n \to 0$, then Σa_n is convergent. State and prove a comparison theorem in which part of the hypothesis is that $a_n/b_n \to \infty$.

7. Let Σa_n be a convergent series of positive terms, and let Σb_k be a series in which $b_k = a_{n_k}$, where $n_1, n_2, \ldots, n_k, \ldots$ is a sequence of positive integers such that $n_i \neq n_j$ if $i \neq j$. Prove that the series Σb_k is convergent.

8. Show that the series

$$\frac{1}{1\cdot 2}+\frac{1}{2\cdot 3}+\cdots+\frac{1}{n(n+1)}\cdots$$

is convergent by noting that

$$\frac{1}{n(n+1)}=\frac{1}{n}-\frac{1}{n+1}$$

and finding a simple expression for the sum of the first n terms. What can you infer about the series

$$\frac{1}{1^2}+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2}+\cdots$$

and

$$\frac{1}{1\cdot 2}+\frac{1}{3\cdot 4}+\frac{1}{5\cdot 6}+\cdots?$$

9. Prove that the series

$$\frac{1}{1}\log\frac{2^2}{1\cdot 3}+\frac{1}{2}\log\frac{3^2}{2\cdot 4}+\cdots+\frac{1}{n}\log\frac{(n+1)^2}{n(n+2)}+\cdots$$

is convergent by showing directly that the sum of the first n terms is less than $\log\frac{2(n+1)}{n+2}<\log 2$.

19.21 / THE INTEGRAL TEST

For this section the student will need to be familiar with a few of the most elementary things about improper integrals of the type

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

The integral is called *convergent* if the limit exists, and *divergent* if the limit does not exist. We refer the student to §18.8 and §22.

The idea of the integral test is to relate the convergence or divergence of a series

$$u_1+u_2+u_3+\cdots+u_n+\cdots$$

to the convergence or divergence of a certain improper integral $\int_a^\infty f(x)\,dx$. A relationship of this sort can be established for certain kinds of series. For the series

$$\frac{1}{1^2}+\frac{1}{2^2}+\frac{1}{3^2}+\cdots+\frac{1}{n^2}+\cdots$$

the appropriate integral to consider is $\int_1^\infty \frac{dx}{x^2}$. In general one wants a function $f(x)$ such that $f(n)=u_n$.

THEOREM IV. *Let $f(x)$ be a function which is positive, continuous, and nonincreasing as x increases for all values of $x \geqq N$, where N is some fixed positive integer. Let the terms of an infinite series be given by $u_n = f(n)$ when $n \geqq N$. Then the series Σu_n converges or diverges according as the improper integral $\int_N^\infty f(x)\,dx$ is convergent or divergent.*

Proof. If $m \leqq x \leqq m+1$, we have $f(m+1) \leqq f(x) \leqq f(m)$, and therefore

$$u_{m+1} = f(m+1) \leqq \int_m^{m+1} f(x)\,dx \leqq f(m) = u_m.$$

We set m successively equal to $N, N+1, \ldots, n$ and add. The result is

$$u_{N+1} + \cdots + u_{n+1} \leqq \int_N^{n+1} f(x)\,dx \leqq u_N + \cdots + u_n. \tag{19.21-1}$$

Suppose now that the integral $\int_N^\infty f(x)\,dx$ is convergent. Then from the first inequality in (19.21–1) it follows that

$$u_{N+1} + \cdots + u_{n+1} \leqq \int_N^\infty f(x)\,dx.$$

This shows that the partial sums of the series $\Sigma_{n=N+1}^\infty u_n$ are bounded, and hence that this series is convergent (by Theorem I, §19.2). The series $\Sigma_{n=1}^\infty u_n$ is then convergent also. On the other hand, suppose the integral $\int_N^\infty f(x)\,dx$ is divergent Since $f(x) > 0$ this can only happen if $\int_N^{n+1} f(x)\,dx \to +\infty$ as $n \to \infty$. It then follows from the second inequality in (19.21–1) that $u_N + \cdots + u_n \to +\infty$, and hence that the series is divergent. This completes the proof.

Example. The series

$$\frac{1}{1^p} + \frac{1}{2^p} + \frac{1}{3^p} + \cdots + \frac{1}{n^p} + \cdots \tag{19.21-2}$$

is convergent if $p > 1$ and divergent if $p \leqq 1$.

These results are established by considering the integral $\int_1^\infty \frac{dx}{x^p}$ and applying Theorem IV. If $p > 1$ we have

$$\int_1^b \frac{dx}{x^p} = \frac{x^{-p+1}}{-p+1}\bigg|_1^b = \frac{1}{1-p}\left(\frac{1}{b^{p-1}} - 1\right),$$

$$\lim_{b\to\infty} \int_1^b \frac{dx}{x^p} = \int_1^\infty \frac{dx}{x^p} = \frac{1}{p-1}.$$

We leave it for the student to verify that the integral diverges if $p \leqq 1$. If $p = 1$ this gives us a new proof that the harmonic series is divergent (see Example 1, §19.2).

The proof of Theorem IV gives us some estimates of the sum of the series if

it is convergent. From (19.21–1) we have, on letting $n \to \infty$,

$$\sum_{n=N+1}^{\infty} u_n \leqq \int_N^{\infty} f(x)\,dx \leqq \sum_{n=N}^{\infty} u_n. \tag{19.21–3}$$

Actually neither inequality can be an equality. That is, (19.21–3) remains true if the sign $\leqq$ is replaced by $<$ at both places where it occurs. We leave it for the student to supply the argument in justification of this assertion.

EXERCISES

1. Test the following series for convergence or divergence by Theorem IV. For the convergent series give an upper bound for the sum of the series with the aid of (19.21–3).

(a) $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n+1}-1}$.

(b) $\sum_{n=2}^{\infty} \frac{2n}{n^4-4}$.

(c) $\sum_{n=1}^{\infty} \frac{1}{(2n-1)(2n)}$.

(d) $\sum_{n=2}^{\infty} \frac{1}{n \log n}$.

(e) $\sum_{n=2}^{\infty} \frac{1}{n(\log n)^2}$.

(f) $\sum_{n=4}^{\infty} \frac{1}{n \log n[\log(\log n)]^2}$.

(g) $\sum_{n=2}^{\infty} \frac{1}{(\log n)^{10}}$.

(h) $\sum_{n=2}^{\infty} \frac{(\log n)^2}{n^3}$.

2. Show that the series

$$\frac{1}{2(\log 2)^p} + \frac{1}{3(\log 3)^p} + \cdots + \frac{1}{n(\log n)^p} + \cdots$$

is convergent if $p > 1$, divergent if $p \leqq 1$.

3. For what values of p does the series $\sum_{n=4}^{\infty} \frac{1}{n \log n[\log(\log n)]^p}$ converge?

4. Show that the series $\sum_{n=2}^{\infty} \frac{1}{(\log n)^p}$ diverges for all values of p.

5. For what values of p and q is the series $\sum_{n=2}^{\infty} \frac{(\log n)^q}{n^p}$ convergent?

6. Let $C_n = 1 + \frac{1}{2} + \cdots + (1/n) - \log n$. Put $N = 1$, $f(x) = 1/x$ in (19.21–1), with $n - 1$ in place of n, and show that $0 < 1/n \leqq C_n$. Also show that $C_n - C_{n+1} > 0$. The sequence C_n is therefore convergent, since it is decreasing and bounded below. The number $C = \lim_{n\to\infty} C_n$ is known as Euler's constant. Its value is approximately 0.577.

Show by a similar argument that

$$\lim_{n\to\infty}\left[1 + \frac{1}{\sqrt{2}} + \cdots + \frac{1}{\sqrt{n}} - 2(\sqrt{n}-1)\right] \text{exists.}$$

19.22 / RATIO TESTS

In this section we deal with series all of whose terms are positive. In many cases it proves to be useful to consider the ratio u_{n+1}/u_n of two successive terms of the series Σu_n.

THEOREM V. *Let Σu_n and Σv_n be two series of positive terms, and suppose that the inequality*

$$\frac{v_{n+1}}{v_n} \leqq \frac{u_{n+1}}{u_n} \tag{19.22–1}$$

holds for all values of n. Then if Σu_n is convergent, so is Σv_n.

The same conclusion can of course be drawn if (19.22–1) *holds merely for all sufficiently large values of n.*

Proof. We have

$$\frac{v_2}{v_1} \leqq \frac{u_2}{u_1}, \qquad \frac{v_3}{v_2} \leqq \frac{u_3}{u_2}, \ldots, \frac{v_n}{v_{n-1}} \leqq \frac{u_n}{u_{n-1}}.$$

Hence

$$v_n = v_1 \frac{v_2}{v_1} \cdot \frac{v_3}{v_2} \cdots \frac{v_n}{v_{n-1}} \leqq v_1 \frac{u_2}{u_1} \cdot \frac{u_3}{u_2} \cdots \frac{u_n}{u_{n-1}} = \frac{v_1}{u_1} u_n,$$

so that $v_n \leqq C u_n$, where C is the positive constant v_1/u_1. The convergence of Σv_n now follows by an application of Theorem II (§19.2).

One of the most useful applications of Theorem V is obtained by choosing for Σu_n the geometric series $1 + r + r^2 + \cdots + r^n + \cdots$, which for positive r is convergent if $r < 1$. For this choice of u_n we have $u_{n+1}/u_n = r$. The following theorem is an immediate corollary of Theorem V.

THEOREM VI. *The series Σv_n of positive terms is convergent if there is a positive number $r < 1$ such that $v_{n+1}/v_n \leqq r$ for all sufficiently large values of n.*

The ratio v_{n+1}/v_n need not approach a limit as $n \to \infty$, but it does so in many cases arising in practice.

THEOREM VII. *Let Σv_n be a series of positive terms, and suppose that the ratio v_{n+1}/v_n approaches a limit t as $n \to \infty$. Then the series is convergent if $0 \leqq t < 1$ and divergent if $t > 1$. If $v_{n+1}/v_n \to +\infty$ we write $t = +\infty$; in this case the series diverges.*

Proof. If $0 \leqq t < 1$, choose r so that $t < r < 1$. Then v_{n+1}/v_n is close to t when n is large, and so $v_{n+1}/v_n < r$ for all sufficiently large values of n. The convergence of the series is then assured by Theorem VI. On the other hand, if $t > 1$, then $v_{n+1} > v_n$ if n is large enough, and in this case the terms of the series cannot approach zero as $n \to \infty$. Hence the series cannot be convergent (see the final paragraph of §19).

No conclusion can be drawn if $t = 1$, for this case can occur both with convergent and with divergent series. For instance, in the case of the series (19.21–2) it turns out that $t = 1$ for all values of p.

Example. The series $\sum_{n=1}^{\infty} \frac{n^2}{2^n}$ is convergent.

Here we have $v_n = \frac{n^2}{2^n}$, $\frac{v_{n+1}}{v_n} = \frac{(n+1)^2 2^n}{2^{n+1} n^2} = \frac{1}{2}\left(\frac{n+1}{n}\right)^2$,

and

$$\lim_{n\to\infty} \frac{v_{n+1}}{v_n} = \frac{1}{2}.$$

Further uses of the theorems of this section will be developed in §19.4.

EXERCISES

1. Test the following series for convergence or divergence by Theorem VII.

(a) $\sum_{n=1}^{\infty} \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 5 \cdots (3n-1)}$.

(b) $\sum_{n=1}^{\infty} \frac{(n!)^2 2^n}{(2n+2)!}$.

(c) $\sum_{n=1}^{\infty} \frac{n!}{2^n}$.

(d) $\sum_{n=1}^{\infty} n(\frac{3}{4})^n$.

(e) $\sum_{n=1}^{\infty} \frac{n^n}{n!}$.

(f) $\sum_{n=1}^{\infty} \frac{2 \cdot 4 \cdots 2n}{4 \cdot 7 \cdots (3n+1)}$.

(g) $\sum_{n=2}^{\infty} \frac{n^{n-1}}{3^{n-1}(n-1)!}$.

2. If $0 < r < 1$ and p is any positive integer, show that the series

$$\sum_{n=p+1}^{\infty} n(n-1)\cdots(n-p)r^n$$

is convergent.

3. Suppose Σu_n and Σv_n are series of positive terms satisfying (19.22–1). Prove that, if Σv_n is divergent, so is Σu_n. Is this theorem equivalent to Theorem V?

4. Suppose $u_n > 0$ and $\frac{u_{n+1}}{u_n} \leqq 1 - \frac{2}{n} + \frac{1}{n^2}$ if $n \geq 2$. Show that Σu_n is convergent.

5. Suppose $u_n > 0$ and $\frac{u_{n+1}}{u_n} \geqq 1 - \frac{1}{n}$. Show that Σu_n is divergent.

19.3 / ABSOLUTE AND CONDITIONAL CONVERGENCE

Theorems I–VII all deal with series whose terms are nonnegative. Thus far we have developed no general methods for discussing the convergence of a given series if its terms are unrestricted as to sign. On the other hand, in §19.1 and §19.11 we have incidentally met some examples of convergent series with both positive and negative terms, e.g.,

$$\log 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots + (-1)^{n-1}\frac{1}{n} + \cdots, \tag{19.3–1}$$

$$\frac{1}{e} = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!} + \cdots. \tag{19.3–2}$$

Now there is an important difference between the two series (19.3–1) and (19.3–2). Each of the series is convergent as it stands. But let us consider the series which are obtained if we make the alteration of changing the signs of all the negative terms. We then have

$$1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}+\cdots \tag{19.3–3}$$

$$1+1+\frac{1}{2!}+\cdots+\frac{1}{n!}+\cdots. \tag{19.3–4}$$

The change from (19.3–1) to (19.3–3) has given us a divergent series (the harmonic series (19.2–1)), whereas in the change from (19.3–2) to (19.3–4) we still have a convergent series (its value is e; see (19.1–6)).

If a series of positive terms is to be convergent, the terms with large index n must be so small that even the sum of arbitrarily many of them is small. But with a series of terms in which infinitely many are positive and infinitely many are negative, the series may be convergent because of a partial cancelling out effect, e.g. a negative term offsetting the cumulative effect of one or more positive terms. This kind of process may operate to produce a convergent series even though the series would not converge if all the terms were replaced by their absolute values. This situation is illustrated by the convergent series (19.3–1) and its divergent counterpart (19.3–3). With this background of explanation we now make a definition.

Definition. *Let Σu_n be a given series. Consider the series $\Sigma|u_n|$ obtained by putting $|u_n|$ in place of u_n. We say that Σu_n is* absolutely convergent *provided that the series $\Sigma|u_n|$ is convergent.*

The careful student will observe at once that the definition of *absolute convergence* of Σu_n does not in itself make any statement about the mere *convergence* of Σu_n. It is in fact true, however, that if $\Sigma|u_n|$ is convergent, then so is Σu_n. Before proving this we must take up a general criterion for convergence.

THEOREM VIII. *A series Σu_n is convergent if and only if to each $\epsilon>0$ there corresponds some integer N (depending on ϵ and the particular series) such that*

$$|u_{m+1}+u_{m+2}+\cdots+u_n|<\epsilon \tag{19.3–5}$$

for all integers m, n such that $N \leqq m < n$.

Proof. This is a direct consequence of Cauchy's convergence condition (Theorem VI, §16.5). We write

$$s_n = u_1+u_2+\cdots+u_n.$$

Then

$$s_n - s_m = u_{m+1} + \cdots + u_n.$$

The condition (19.3–5) is now seen to be equivalent to $|s_n - s_m| < \epsilon$ and the theorem is merely the statement of Cauchy's condition for the sequence $\{s_n\}$.

Theorem VIII is not very useful for making direct tests to find whether a series is convergent. But it is of fundamental importance in the general theory of infinite series, and we shall use it in proving various things about series. The first such application deals with absolute convergence.

THEOREM IX. *If the series Σu_n is absolutely convergent, it is convergent.*

Proof. We are assuming that $\Sigma |u_n|$ is convergent. By Theorem VIII, therefore, we know that to any $\epsilon > 0$ there corresponds some N such that

$$|u_{m+1}| + |u_{m+2}| + \cdots + |u_n| < \epsilon$$

whenever $N \leqq m < n$. Now

$$|u_{m+1} + u_{m+2} + \cdots + u_n| \leqq |u_{m+1}| + \cdots + |u_n|.$$

Hence $|u_{m+1} + \cdots + u_n| < \epsilon$ whenever $N \leqq m < n$. This is precisely condition (19.3–5), and so we conclude that the series Σu_n is convergent.

Definition. *A series is called* conditionally *convergent if it is convergent, but not absolutely convergent.*

The series (19.3–1) is conditionally convergent.

We shall consider some of the differences between absolutely convergent and conditionally convergent series. In the series Σu_n let us denote by $a_1, a_2, a_3, \ldots$ the positive terms, taken in the order of their occurrence; let the negative terms be denoted by $-b_1, -b_2, -b_3, \ldots$. Thus, in the series

$$1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} + \cdots$$

we have

$$a_1 = 1,\ a_2 = \tfrac{1}{3},\ a_3 = \tfrac{1}{5}, \ldots,$$
$$b_1 = \tfrac{1}{2},\ b_2 = \tfrac{1}{4},\ b_3 = \tfrac{1}{6}, \ldots.$$

We now consider the two series Σa_n and Σb_n, each of which consists entirely of positive terms.

THEOREM X. *If the series Σu_n is absolutely convergent, then each of the series Σa_n, Σb_n is convergent, and $\Sigma u_n = \Sigma a_n - \Sigma b_n$. But if the series Σu_n is conditionally convergent, then each of the series Σa_n, Σb_n is divergent.*

The theorem is illustrated for the absolutely convergent case by the series

(19.3–2), which is the difference of the two convergent series

$$\left(1+\frac{1}{2!}+\frac{1}{4!}+\cdots\right)-\left(1+\frac{1}{3!}+\frac{1}{5!}+\cdots\right).$$

In the case of the series (19.3–1) each of the constituent series

$$1+\tfrac{1}{3}+\tfrac{1}{5}+\cdots, \quad \tfrac{1}{2}+\tfrac{1}{4}+\tfrac{1}{6}+\cdots$$

is divergent.

Proof of the theorem. Suppose that Σu_n is absolutely convergent, with $M = \Sigma|u_n|$. Then

$$|u_1|+|u_2|+\cdots+|u_n| \leqq M,$$

no matter how large n is. Now consider a partial sum of the series Σa_n, say $a_1+\cdots+a_m$. Since each a_i is a positive term somewhere in the series Σu_n, the terms $a_1, \ldots, a_m$ all occur in the sum $|u_1|+\cdots+|u_n|$ if n is sufficiently large. But then we see that

$$a_1+a_2+\cdots+a_m \leqq M.$$

It follows by Theorem I (§19.2) that the series Σa_n is convergent. In the same way we see that $b_1+\cdots+b_m \leqq M$, since each b_i is a term somewhere in the series $\Sigma|u_n|$. Thus the series Σb_n is convergent.

Now suppose that, in the sum $u_1+\cdots+u_n$, the number of positive terms is p_n and the number of negative terms is q_n. Then

$$u_1+\cdots+u_n = (a_1+\cdots+a_{p_n})-(b_1+\cdots+b_{q_n}). \tag{19.3–6}$$

In the case of absolute convergence we let $n \to \infty$ and obtain the result

$$\sum_{n=1}^{\infty} u_n = \sum_{n=1}^{\infty} a_n - \sum_{n=1}^{\infty} b_n.$$

It may happen that there are only a finite number of a's or a finite number of b's, or possibly none of one kind or the other. In these cases the series is of course absolutely convergent if it is convergent at all, since its terms from some point onward are all of one sign. Let us then consider the case in which there are infinitely many terms of each sign, so that p_n and $q_n \to \infty$ as $n \to \infty$. Let

$$s_n = u_1+\cdots+u_n, \; A_{p_n} = a_1+\cdots+a_{p_n}, \; B_{q_n} = b_1+\cdots+b_{q_n},$$

so that (19.3–6) becomes $s_n = A_{p_n} - B_{q_n}$. We also have

$$\begin{aligned} |u_1|+\cdots+|u_n| &= (a_1+\cdots+a_{p_n})+(b_1+\cdots+b_{q_n}) \\ &= A_{p_n}+B_{q_n}. \end{aligned} \tag{19.3–7}$$

Now suppose that the series Σu_n is convergent, and consider the series Σa_n, Σb_n. If either of these latter series is convergent, so is the other, by virtue of the relation $s_n = A_{p_n} - B_{q_n}$. For instance, if Σa_n is convergent, B_{q_n} approaches a limit, since s_n and A_{p_n} each approach limits, and $B_{q_n} = A_{p_n} - s_n$. But to say that

$\lim_{n\to\infty} B_{q_n}$ exists is equivalent to saying that Σb_n is convergent, since $\{B_{q_n}\}$ is a nondecreasing sequence, and is therefore convergent if and only if the partial sums of the series Σb_n are bounded. But if both the series Σa_n, Σb_n are convergent, the series $\Sigma |u_n|$ is convergent, by (19.3–7). Thus we see that, if Σu_n is conditionally convergent, both of the series Σa_n, Σb_n must be divergent. This completes the proof of Theorem *X*.

EXERCISES

1. If Σc_n is a convergent series of positive terms, the series $\Sigma c_n x^n$ is absolutely convergent when $|x| \leqq 1$. Prove this.

2. Show that, if $0 \leqq r < 1$, the series $\Sigma r^n \sin n\theta$ and $\Sigma r^n \cos n\theta$ are absolutely convergent for all values of θ.

3. Which of the following series are absolutely convergent?

(a) $1 - \frac{1}{3^2} + \frac{1}{5^2} - \frac{1}{7^2} + \cdots$.

(b) $1 - \frac{1}{2} + \frac{1}{3!} - \frac{1}{4} + \frac{1}{5!} - \cdots$.

(c) $\frac{1}{\log 2} - \frac{1}{2\log 3} + \frac{1}{3\log 4} - \cdots + (-1)^n \frac{1}{(n-1)\log n} + \cdots$.

(d) $\sum_{n=1}^{\infty} \frac{\sin n\theta}{n^2}$.

4. Let $\{a_n\}$ and $\{b_n\}$ be sequences of real numbers with the following properties:

(i) $b_n > 1$

(ii) to each $\epsilon > 0$ corresponds some positive integer N such that $C_{n,p} \log(b_n b_{n+1} \ldots b_{n+p}) < \epsilon$ if p is any positive integer, where $C_{n,p}$ is the maximum of $|a_k|$ for $n \leq k \leq n+p$, and $N \leq n$.

Then prove that

(a) $\Sigma a_n \log b_n$ is absolutely convergent, and that

(b) if there are infinitely many nonzero a_n's, the series $\Sigma \log b_n$ is also convergent.

19.31 / REARRANGEMENT OF TERMS

Suppose that we have two convergent series with values s and t respectively:

$$s = u_1 + u_2 + \cdots + u_n + \cdots,$$
$$t = v_1 + v_2 + \cdots + v_n + \cdots.$$

We may combine these series by adding corresponding terms, and the result will be a new convergent series whose value is $s + t$:

$$s + t = w_1 + w_2 + \cdots + w_n + \cdots,$$

where $w_1 = u_1 + v_1$, $w_2 = u_2 + v_2, \ldots, w_n = u_n + v_n, \ldots$. This assertion is easily proved as a direct application of the rule that the limit of a sum is the sum of the limits, for

$$(w_1 + \cdots + w_n) = (u_1 + \cdots + u_n) + (v_1 + \cdots + v_n).$$

Now let us consider the question: Suppose the series Σu_n is convergent; what can we say about a series which is obtained by using the same terms u_n, *but in a different order*? It may surprise the student to learn that (*a*) if the new series is convergent, it does not necessarily have the same sum as the original series, and (*b*) the new series may be divergent.

Example 1. By rearranging the terms in the series (19.3–1), whose sum is $\log 2$, we can obtain the following result:

$$\tfrac{3}{2}\log 2 = 1+\tfrac{1}{3}-\tfrac{1}{2}+\tfrac{1}{5}+\tfrac{1}{7}-\tfrac{1}{4}+\tfrac{1}{9}+\tfrac{1}{11}-\tfrac{1}{6}+\cdots. \tag{19.31–1}$$

The scheme in the rearrangement is to take two positive terms and one negative term, then two more positive terms and another negative term, and so on. To prove (19.31–1) we note that

$$\log 2 = 1-\tfrac{1}{2}+\tfrac{1}{3}-\tfrac{1}{4}+\tfrac{1}{5}-\tfrac{1}{6}+\tfrac{1}{7}-\cdots, \tag{19.31–2}$$
$$\tfrac{1}{2}\log 2 = \tfrac{1}{2}-\tfrac{1}{4}+\tfrac{1}{6}-\tfrac{1}{8}+\tfrac{1}{10}-\tfrac{1}{12}+\tfrac{1}{14}-\cdots.$$

In this last series we may insert zero terms without affecting the value. Thus

$$\tfrac{1}{2}\log 2 = 0+\tfrac{1}{2}+0-\tfrac{1}{4}+0+\tfrac{1}{6}+0-\cdots. \tag{19.31–3}$$

On adding (19.31–2) and (19.31–3) term by term we get

$$\tfrac{3}{2}\log 2 = 1+0+\tfrac{1}{3}-\tfrac{1}{2}+\tfrac{1}{5}+0+\tfrac{1}{7}-\tfrac{1}{4}+\cdots.$$

This is the same as (19.31–1) when we drop out the zero terms.

Example 2. In this example we shall not be so explicit. The series (19.31–2) is conditionally convergent, so that the series composed of its positive terms is divergent. This means that by taking k large enough we can make

$$1+\frac{1}{3}+\frac{1}{5}+\cdots+\frac{1}{2k-1}$$

as large as we please. We now rearrange the series (19.31–2) according to the following plan: First take just enough of the positive terms to obtain a sum greater than 2. Then take the first negative term, $-\frac{1}{2}$. This leaves us with more than $\frac{3}{2}$. Now take just enough more positive terms to increase the total sum beyond 4, and then take one more negative term, $-\frac{1}{4}$. Continuing in this way, we build up partial sums such that when the term $-1/(2n)$ is taken, the total exceeds $2n-(1/n)$. The series so formed must diverge.

What we have just seen in Examples 1 and 2 is typical of conditionally convergent series. In fact, it may be proved that the terms in a conditionally convergent series may be rearranged so as to produce a series which has any desired sum, or such that the partial sums tend to $+\infty$ or to $-\infty$. By contrast, no such thing can happen with absolutely convergent series.

THEOREM XI. *Let the series Σu_n be absolutely convergent, with sum s. Let Σv_n be any series obtained by a rearrangement of the terms of Σu_n (i.e., every v_i is some u_j and every u_k is some v_i). Then Σv_n is convergent, with sum s.*

Proof. First let us prove the assertion on the assumption that all the u's (and hence all the v's) are nonnegative. Since $s = \Sigma u_n$ and since each v_i is some u_j, it is clear that the partial sums of the series Σv_n cannot exceed s. Thus the series Σv_n must be convergent and its sum s' must satisfy the inequality $s' \leqq s$. Reversing the role of Σu_n and Σv_n, we see that $s \leqq s'$. Therefore $s' = s$. This completes the proof for the case of series of nonnegative terms. Now, in the general case of an absolutely convergent series, we have

$$\sum u_n = \sum a_n - \sum b_n$$

in the notation of Theorem X (§19.3). In the rearranged series Σv_n, the separation into positive and negative terms yields

$$\sum v_n = \sum a'_n - \sum b'_n,$$

where $\Sigma a'_n$ is a rearrangement of Σa_n and $\Sigma b'_n$ is a rearrangement of Σb_n. By what has just been proved for series of positive terms, we have

$$\sum a'_n = \sum a_n, \sum b'_n = \sum b_n.$$

Hence the series Σv_n is convergent, with the same sum as Σu_n.

EXERCISES

1. Show that $\log 2 = 1 + \frac{1}{3} + \frac{1}{5} - \frac{1}{2} - \frac{1}{4} - \frac{1}{6} + + + - - - \cdots$.

2. Show that $\log 2 - 1 = \frac{1}{3} - \frac{1}{2} + \frac{1}{5} - \frac{1}{4} + \frac{1}{7} - \frac{1}{6} + \cdots$.

Here the positive terms have odd denominators, and the negative terms have even integers as denominators. The terms alternate in sign.

3. Show that $\frac{1}{2}\log 2 = 1 - \frac{1}{2} - \frac{1}{4} + \frac{1}{3} - \frac{1}{6} - \frac{1}{8} + \frac{1}{5} - - + \cdots$.

4. Let Σu_n be a convergent series, and let Σv_n be a rearrangement of it. In the rearrangement, suppose that no term of the original series is moved more than N places from its original position, where N is a fixed number. Show that the new series is convergent and has the same value as the old one.

19.32 / ALTERNATING SERIES

The simplest type of series having both positive and negative terms is the type in which the successive terms alternate in sign. Many commonly occurring series are of this type, and have the additional property that the magnitude of the terms steadily decreases toward zero as n increases. Concerning such series we have a theorem of practical importance as a test for convergence.

THEOREM XII. *Suppose the terms of the series Σu_n are alternately positive and negative, that $|u_{n+1}| \leqq |u_n|$ for all n, and that $u_n \to 0$ as $n \to \infty$. Then the series is convergent.*

Proof. For convenience let us write the series in the form

$$c_1 - c_2 + c_3 - c_4 + \cdots + c_{2n-1} - c_{2n} + \cdots,$$

where $c_n > 0$. There are two kinds of partial sums, depending on whether the sum ends with a positive or a negative term:

$$s_{2n-1} = c_1 - c_2 + \cdots + c_{2n-1}$$
$$s_{2n} = c_1 - c_2 + \cdots - c_{2n}.$$

If we consider the sequences $\{s_{2n-1}\}$ and $\{s_{2n}\}$ separately, we observe that

$$s_1 \geqq s_3 \geqq s_5 \geqq \cdots$$
$$s_2 \leqq s_4 \leqq s_6 \leqq \cdots.$$

These inequalities depend on the fact that $c_{n+1} \leqq c_n$. For example,

$$s_7 = s_5 - c_6 + c_7 = s_5 - (c_6 - c_7) \leqq s_5$$

since $c_6 - c_7 \geqq 0$, and

$$s_8 = s_6 + c_7 - c_8 \geqq s_6$$

since $c_7 - c_8 \geqq 0$. Moreover, $c_1 - c_2 \leqq s_n \leqq c_1$ for all values of n (see Fig. 175). Therefore each of the sequences $\{s_{2n-1}\}$, $\{s_{2n}\}$ is convergent, being monotonic and

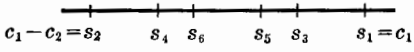


Fig. 175.

bounded. But $s_{2n-1} - s_{2n} = c_{2n}$, and since $c_{2n} \to 0$, it follows that both sequences have the same limit. But then the sequence $\{s_n\}$, where n runs through *all* positive integers, is convergent, its limit being the common limit of s_1, s_3, $s_5, \ldots$ and s_2, s_4, $s_6, \ldots$. The theorem is therefore proved.

It is worth while noting that in a series of the kind just described, the partial sum $u_1 + \cdots + u_n$ differs from the sum of the series by not more than $|u_{n+1}|$, i.e., by not more than the first term not included in the partial sum.

Example 1. The series

$$\frac{1}{\sqrt{1}} - \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} - \cdots + (-1)^{n-1}\frac{1}{\sqrt{n}} + \cdots$$

is convergent, since the conditions of Theorem XII are satisfied.

Example 2. Compute e^{-1} with an error not exceeding 0.005 by using (19.3–2):

$$e^{-1} = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots.$$

It will be sufficient to stop with the term involving $\frac{1}{n!}$ if n is chosen so that

$\frac{1}{(n+1)!} < 0.005$. Thus we use the approximation

$$e^{-1} = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!},$$

the next term in the series being

$$\frac{1}{6!} = \frac{1}{720} = 0.0014 \text{ (approximately)}.$$

This gives

$$e^{-1} = \frac{44}{120} = \frac{11}{30} = 0.3666\cdots,$$

with an error less than 0.0014. Our work actually shows that

$$0.3666 < e^{-1} < 0.3681.$$

EXERCISES

1. Show by Theorem XII that each of the following series is convergent. Prove carefully that each condition of the theorem is fulfilled.

(a) $\frac{1}{2^2} - \frac{2}{3^2} + \frac{3}{4^2} + \cdots + (-1)^{n+1}\frac{n}{(n+1)^2} + \cdots$.

(b) $\log\frac{3}{2} - \log\frac{4}{3} + \log\frac{5}{4} - \cdots$.

(c) $\frac{1}{1+1} - \frac{\sqrt{2}}{1+2} + \frac{\sqrt{3}}{1+3} - \cdots$.

(d) $\frac{1}{2\cdot 4} - \frac{1\cdot 3}{2\cdot 4\cdot 6} + \cdots + (-1)^{n+1}\frac{1\cdot 3\cdots(2n-1)}{2\cdot 4\cdots 2n\cdot(2n+2)} + \cdots$.

(e) $\sum_{n=1}^{\infty}(-1)^{n+1}\frac{3\cdot 6\cdots 3n}{1\cdot 4\cdots(3n-2)}\cdot\frac{1}{n^2}$.

2. Test the following series for convergence or divergence.

(a) $\frac{\log 2}{\sqrt{2}} - \frac{\log 3}{\sqrt{3}} + \frac{\log 4}{\sqrt{4}} - \cdots$.

(b) $\frac{1}{2} - \left(\frac{2}{3}\right)^2 + \left(\frac{3}{4}\right)^3 - \cdots + (-1)^{n+1}\left(\frac{n}{n+1}\right)^n + \cdots$.

(c) $(1 - \log 2) - (1 - 2\log\frac{3}{2}) + (1 - 3\log\frac{4}{3}) - \cdots$.

(d) $\frac{1}{3} - \frac{2}{3} + \frac{1}{6} - \frac{2}{6} + \frac{1}{9} - \frac{2}{9} + \cdots$.

3. Show that the conditions of Theorem XII are fulfilled by the series $\sum_{n=N}^{\infty}(-1)^n\frac{(\log n)^p}{n^q}$ for any fixed positive q and any p, provided the initial value $n = N$ is large enough (how large will depend on p and q).

4. The series

$$1 - \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} - \frac{1}{\sqrt{4}} + \frac{1}{\sqrt{5}} - \cdots$$

is convergent, by Theorem XII. Why is it conditionally convergent? Show that the

rearrangement

$$1+\frac{1}{\sqrt{3}}-\frac{1}{\sqrt{2}}+\frac{1}{\sqrt{5}}+\frac{1}{\sqrt{7}}-\frac{1}{\sqrt{4}}++-\cdots$$

diverges to $+\infty$. SUGGESTION. Compare the sum S_{3n} of the first $3n$ terms of the rearranged series with the sum s_{2n} of the first $2n$ terms of the original series, and show that

$$S_{3n}-s_{2n}>\frac{n}{\sqrt{4n-1}}.$$

Since s_{2n} approaches a limit as $n\to\infty$, it follows that $S_{3n}\to+\infty$. How does one show that $S_n\to+\infty$?

19.4 / TESTS FOR ABSOLUTE CONVERGENCE

In a very large number of important cases the most convenient test for absolute convergence is the following one, which is known as d'Alembert's ratio test (after J. le R. d'Alembert, 1717?–1783).

THEOREM XIII. *Let Σu_n be a series with all its terms different from zero. Then the series is absolutely convergent if there is a positive number $r<1$ such that $\left|\frac{u_{n+1}}{u_n}\right|<r$ for sufficiently large values of n. In particular, this condition is satisfied if the limit*

$$t=\lim_{n\to\infty}\left|\frac{u_{n+1}}{u_n}\right|$$

exists and $t<1$. But if $t>1$ the series is divergent.

Proof. The assertions regarding absolute convergence are direct applications of Theorems VI and VII (§19.22) to the series $\Sigma|u_n|$. If $t>1$ the general term u_n cannot approach zero, and the series Σu_n cannot be convergent.

The theorem makes no assertion about what happens if $t=1$. The series may then converge absolutely, or conditionally, or it may diverge.

Theorem XIII is very useful in dealing with power series, as the following example shows.

Example 1. The series

$$1-\tfrac{1}{2}x+\frac{1\cdot3}{2\cdot4}x^2-\frac{1\cdot3\cdot5}{2\cdot4\cdot6}x^3+\cdots+(-1)^n\frac{1\cdot3\cdots(2n-1)}{2\cdot4\cdots2n}x^n+\cdots \qquad (19.4\text{–}1)$$

is absolutely convergent if $|x|<1$, and divergent if $|x|>1$.

Here we have (calling the first term u_0)

$$u_n=(-1)^n\frac{1\cdot3\cdots(2n-1)}{2\cdot4\cdots2n}x^n,$$

$$u_{n+1}=(-1)^{n+1}\frac{1\cdot3\cdots(2n+1)}{2\cdot4\cdots(2n+2)}x^{n+1},$$

$$\left|\frac{u_{n+1}}{u_n}\right| = \frac{2n+1}{2n+2}|x|, \qquad \lim_{n\to\infty}\left|\frac{u_{n+1}}{u_n}\right| = |x|.$$

The assertions about (19.4–1) are now seen to follow by application of Theorem XIII (with $t = |x|$). For the present we avoid a discussion of the behavior of the series (19.4–1) when $|x| = 1$ (but see Example 3).

Example 2. If m is any number except 0 or a positive integer, the series

$$1 + \frac{m}{1!}x + \frac{m(m-1)}{2!}x^2 + \cdots + \frac{m(m-1)\cdots(m-n+1)}{n!}x^n + \cdots \quad (19.4\text{–}2)$$

is absolutely convergent if $|x| < 1$ and divergent if $|x| > 1$.

The ratio $\frac{u_{n+1}}{u_n}$ is readily found to be

$$\frac{m(m-1)\cdots(m-n)}{m(m-1)\cdots(m-n+1)}\frac{n!}{(n+1)!}\frac{x^{n+1}}{x^n} = \frac{m-n}{n+1}x,$$

so that

$$\lim_{n\to\infty}\left|\frac{u_{n+1}}{u_n}\right| = \lim_{n\to\infty}\left|\frac{m-n}{n+1}\right||x| = |x|.$$

Application of Theorem XIII gives the result as asserted.

For cases in which $\lim_{n\to\infty}\left|\frac{u_{n+1}}{u_n}\right| = 1$ there is a convenient test which is sometimes effective. It was first established by L. J. Raabe in 1832, and is known as Raabe's test. The essential idea of the test is to use Theorem V (§19.22), taking one of the series in that theorem to be the series

$$\frac{1}{1^p} + \frac{1}{2^p} + \cdots + \frac{1}{n^p} + \cdots,$$

which is known to converge if $p > 1$ and diverge if $p \leqq 1$. Let $a_n = n^{-p}$. Then

$$\frac{a_{n+1}}{a_n} = \frac{n^p}{(n+1)^p} = \left(\frac{n+1}{n}\right)^{-p} = \left(1 + \frac{1}{n}\right)^{-p}.$$

Hence, by Theorem V, we have the following criterion. *If $p > 1$ and if*

$$\left|\frac{u_{n+1}}{u_n}\right| \leqq \left(1 + \frac{1}{n}\right)^{-p} \quad (19.4\text{–}3)$$

for all sufficiently large values of n, the series Σu_n is absolutely convergent. This criterion in itself is not very useful, however, for it is not easy in practice to tell whether the inequality (19.4–3) is satisfied. To improve matters let us proceed as follows: We set $f(x) = (1+x)^{-p}$ and expand $f(x)$ in powers of x by Taylor's formula with remainder. We use (4.3–7) with $a = 0$, $h = x$, $n = 1$. The result is

$$(1+x)^{-p} = 1 - px + \frac{x^2}{2}\frac{p(p+1)}{(1+\theta x)^{p+2}}, \quad 0 < \theta < 1.$$

This is valid if $x > -1$. We put $x = 1/n$. Thus

$$\left(1+\frac{1}{n}\right)^{-p} = 1-\frac{p}{n}+\frac{A_n}{n^2}, \quad A_n = \frac{p(p+1)}{2\left(1+\frac{\theta}{n}\right)^{p+2}}.$$

The exact form of A_n is unimportant. The only essential thing is that A_n is bounded as $n \to \infty$. Now (19.4–3) can be put in the form

$$\left|\frac{u_{n+1}}{u_n}\right| \leqq 1-\frac{p}{n}+\frac{A_n}{n^2};$$

we rewrite this as

$$p-\frac{A_n}{n} \leqq n\left\{1-\left|\frac{u_{n+1}}{u_n}\right|\right\}. \tag{19.4–4}$$

Now when $p > 1$ and n is very large, the expression $p - (A_n/n)$ is greater than 1. The form of the inequality now suggests that we consider the limit

$$t = \lim_{n\to\infty} n\left\{1-\left|\frac{u_{n+1}}{u_n}\right|\right\}, \tag{19.4–5}$$

provided this limit exists.

THEOREM XIV. (RAABE'S TEST.) *Suppose that the limit* (19.4–5) *exists (either as a finite limit or as* $+\infty$ *or* $-\infty$*). Then the series* Σu_n *is absolutely convergent if* $t > 1$, *but not if* $t < 1$.

Proof. Suppose $t > 1$, and choose p so that $1 < p < t$ (here t may be $+\infty$). Then certainly we shall have

$$p < n\left\{1-\left|\frac{u_{n+1}}{u_n}\right|\right\}$$

for all sufficiently large values of n. Consequently the inequality (19.4–4) must hold, since $A_n > 0$. This is equivalent to (19.4–3), and therefore the series Σu_n is absolutely convergent. To show that the series does not converge absolutely if $t < 1$ it suffices to show that if we take p such that $t < p < 1$, then

$$\left(1+\frac{1}{n}\right)^{-p} \leqq \left|\frac{u_{n+1}}{u_n}\right|$$

for all sufficiently large values of n. For, by Theorem V, the convergence of $\Sigma|u_n|$ would imply the convergence of Σn^{-p}, and the latter series is divergent. We omit the details of the demonstration, which are very similar to the details of the first part of the proof, with all the inequalities reversed.

Raabe's test is indecisive if $t = 1$ (see Exercise 9).

Example 3. Consider the series of Example 1 when $x = \pm 1$. In that case

$$\left|\frac{u_{n+1}}{u_n}\right| = \frac{2n+1}{2n+2}, \qquad n\left\{1-\left|\frac{u_{n+1}}{u_n}\right|\right\} = \frac{n}{2n+2},$$

so that the limit t in (19.4–5) is $\frac{1}{2}$. Therefore, by Theorem XIV, the series (19.4–1) is not absolutely convergent when $|x| = 1$. This shows that it actually diverges when $x = -1$, for the terms are all positive in that case. When $x = 1$, however, the terms of the series are alternating in sign, and we can show that the series is conditionally convergent. The series is

$$1 - \frac{1}{2} + \frac{1 \cdot 3}{2 \cdot 4} - \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6} + \cdots + (-1)^n \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots 2n} + \cdots .$$

It is clear that the terms steadily decrease in magnitude as $n \to \infty$, so that by Theorem XII all we need to show is that

$$\lim_{n \to \infty} \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots 2n} = 0 \tag{19.4–6}$$

Let

$$c_n = \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots 2n}.$$

Then

$$c_n < \frac{2 \cdot 4 \cdots 2n}{3 \cdot 5 \cdots (2n+1)} = \frac{1}{c_n} \frac{1}{2n+1},$$

and therefore

$$c_n^2 < \frac{1}{2n+1}, \qquad c_n < \frac{1}{\sqrt{2n+1}};$$

the truth of (19.4–6) is now evident.

There is a somewhat more powerful test than Raabe's test. It is essentially a test for series of positive terms, and therefore may be used to test for absolute convergence by applying the test to $\Sigma |u_n|$. The test reads as follows:

THEOREM XV. *Suppose that* $\left|\frac{u_{n+1}}{u_n}\right|$ *can be expressed in the form*

$$\left|\frac{u_{n+1}}{u_n}\right| = 1 - \frac{p}{n} + \frac{A_n}{n^q}, \tag{19.4–7}$$

where $q > 1$ *and the sequence* $\{A_n\}$ *is bounded. Then the series* Σu_n *is absolutely convergent if* $p > 1$, *and not absolutely convergent (either divergent or conditionally convergent) if* $p \leqq 1$.

The test in Theorem XV is due to Karl Friedrich Gauss (1777–1855), one of the greatest mathematicians in history. We leave the proof to the student, with certain guiding suggestions; see Exercise 10.

Example 4. Consider the series (19.4–2) when $|x| = 1$. As before, assume that

m is not zero or a positive integer. Then

$$\left|\frac{u_{n+1}}{u_n}\right| = \left|\frac{m-n}{n+1}\right|.$$

For sufficiently large values of n we shall have $|m-n| = n-m$, and hence

$$\left|\frac{u_{n+1}}{u_n}\right| = \frac{n-m}{n+1} = \frac{1-\frac{m}{n}}{1+\frac{1}{n}} = 1 - \frac{m+1}{n} + \frac{m+1}{n^2\left(1+\frac{1}{n}\right)},$$

so that (19.4–7) holds with $p = m+1$, $q = 2$, $A_n = \dfrac{m+1}{1+(1/n)}$. It follows that the series (19.4–2) is absolutely convergent when $|x| = 1$ if and only if $m+1 > 1$, i.e., if and only if $m > 0$.

There is a test due to Cauchy which, although not as easy to apply in many cases as d'Alembert's ratio test (Theorem XIII), is more powerful (i.e., has wider theoretical applicability). It has important applications in the theory of power series.

THEOREM XVI. (CAUCHY'S ROOT TEST.) *The series Σu_n is absolutely convergent if there is a positive number $r < 1$ such that $|u_n|^{1/n} \leqq r$ for all sufficiently large values of n. This condition will be satisfied if $\lim_{n\to\infty} |u_n|^{1/n}$ exists and is less than unity. But, if $|u_n|^{1/n} \geqq 1$ for an infinite number of values of n, the series Σu_n is divergent.*

Proof. In the first case we have $|u_n| \leqq r^n$, and the absolute convergence follows by comparison with the convergent geometric series Σr^n. In the second case u_n cannot approach zero as $n \to \infty$ (since $|u_n| \geqq 1$ for infinitely many values of n). Hence the series is divergent.

Example 5. The series

$$\sum_{n=0}^{\infty}\left(1+\sin\frac{n\pi}{2}\right)^n \frac{x^n}{2^n}$$

is absolutely convergent if $|x| < 1$, and divergent if $|x| \geqq 1$. For,

$$|u_n|^{1/n} = \left(1+\sin\frac{n\pi}{2}\right)\frac{|x|}{2} \leqq |x|$$

for all n, and

$$|u_n|^{1/n} = |x|$$

if $n = 1, 5, 9, 13, \ldots$. *The statements about the series now follow immediately from Theorem XVI.*

Cauchy's test is indecisive if $|u_n|^{1/n} \to 1$, as we see by applying it to the series Σn^{-p} with $p = 1$ and 2, successively.

EXERCISES

1. Apply d'Alembert's ratio test to each series and state explicitly the conclusions which you draw from use of the test. In parts **(a)**–**(e)** make a separate investigation and decide whether the series converges or diverges for each of the values of x for which the ratio test is indecisive.

(a) $\sum_{n=1}^{\infty} \frac{x^n}{n^2}$.

(b) $\sum_{n=1}^{\infty} \frac{n+1}{n^2(n+2)}x^n$.

(c) $\sum_{n=1}^{\infty} \frac{n!}{x^n}$.

(d) $\sum_{n=1}^{\infty} \frac{3 \cdot 5 \cdots (2n+1)}{5 \cdot 10 \cdots 5n}x^n$.

(e) $\sum_{n=1}^{\infty} \frac{n!}{n^n}x^n$.

(f) $\sum_{n=1}^{\infty} \frac{n!}{3 \cdot 5 \cdots (2n+1)}x^n$.

(g) $\sum_{n=1}^{\infty} \frac{n^p}{2^{n-1}}x^n$.

(h) $\sum_{n=1}^{\infty} \frac{(2n)^n}{(n+1)^{n+1}}x^n$.

2. Apply Cauchy's root test to each series and state explicitly the conclusions which you draw from use of the test.

(a) $\sum_{n=1}^{\infty} \left(\frac{x}{n}\right)^n$.

(b) $\sum_{n=1}^{\infty} n^p x^n$, $p > 0$.

(c) $\sum_{n=2}^{\infty} \frac{1}{(\log n)^n}$.

(d) $\sum_{n=2}^{\infty} \frac{1}{(\log n)^{\log n}}$.

(e) $\frac{1}{3} + \frac{1}{2} + (\frac{1}{3})^2 + (\frac{1}{2})^2 + (\frac{1}{3})^3 + (\frac{1}{2})^3 + \cdots$.

(f) $\frac{1}{2}x + \frac{1}{2} \cdot \frac{4}{3}x^2 + (\frac{1}{2})^2 \cdot \frac{4}{3}x^3 + (\frac{1}{2})^2(\frac{4}{3})^2x^4 + (\frac{1}{2})^3(\frac{4}{3})^2x^5 + (\frac{1}{2})^3(\frac{4}{3})^3x^6 + \cdots$.

(g) $\sum_{n=1}^{\infty} \frac{[1+(1/n)]^{2n}}{e^n}$.

(h) $\sum_{n=0}^{\infty} \left(3 + 2\cos\frac{n\pi}{2}\right)^n \left(\frac{x}{2}\right)^n$.

3. Apply Raabe's test to each series and state explicitly the conclusions which you draw from use of the test. In **(e)** and **(f)** use the formula for $(1+x)^{-p}$ developed for the proof of Raabe's test.

(a) $\sum_{n=1}^{\infty} \frac{2 \cdot 4 \cdots 2n}{3 \cdot 5 \cdots (2n+1)}$.

(b) $\sum_{n=1}^{\infty} \frac{2 \cdot 4 \cdots 2n}{5 \cdot 7 \cdots (2n+3)}$.

(c) $\sum_{n=1}^{\infty} \frac{5 \cdot 6 \cdots (4+n)}{8 \cdot 9 \cdots (7+n)}(n+1)$.

(d) $\sum_{n=1}^{\infty} \frac{4 \cdot 5 \cdots (3+n)}{8 \cdot 9 \cdots (7+n)} \cdot \frac{5 \cdot 6 \cdots (4+n)}{n!}$.

(e) $\sum_{n=1}^{\infty} \left[\frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots 2n}\right]^p$.

(f) $\sum_{n=1}^{\infty} \left[\frac{2 \cdot 4 \cdots 2n}{5 \cdot 7 \cdots (2n+3)}\right]^p$.

(g) $\sum_{n=1}^{\infty} \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots (2n)} \cdot \frac{3 \cdot 7 \cdots (3n-1)}{5 \cdot 9 \cdots (4n+1)}$.

4. Test each series in Exercise 3 by Gauss's test. Observe that Gauss's test is more effective than Raabe's test in parts **(e)**, **(f)**, **(g)**.

5. Let $0 < a < b < 1$. Show by Cauchy's root test that the series

$$a + b + a^2 + b^2 + a^3 + b^3 + \cdots$$

is convergent. Attempt to test the series by d'Alembert's ratio test and state clearly the outcome of your attempt.

6. Consider the series

$$a + ab + a^2b + a^2b^2 + a^3b^2 + \cdots,$$

in which each term is derived from its predecessor by multiplying alternately by b and by

a. Show that, if $0<a<1$ and $0<b<1$, the series may be proved convergent by d'Alembert's ratio test. Assuming a and b to be positive, prove that the series is convergent if $ab<1$ and divergent if $ab \geqq 1$.

7. Assume that none of the numbers a, b, c is a negative integer or zero. Prove that the series

$$\frac{ab}{c}+\frac{a(a+1)b(b+1)}{2!c(c+1)}+\frac{a(a+1)(a+2)b(b+1)(b+2)}{3!c(c+1)(c+2)}+\cdots$$

is absolutely convergent if $c>a+b$ and divergent if $c \leqq a+b$. Why is conditional convergence not possible?

8. (**a**) Suppose $a_n>0$ and $t=\lim_{n\to\infty}\dfrac{\log(1/a_n)}{\log n}$. Show that the series Σa_n is convergent if $t>1$ and divergent if $t<1$. SUGGESTION: Show that $t>1$ implies $a_n<n^{-p}$ if $1<p<t$ and n is sufficiently large.

(**b**) Apply the test of part (**a**) to each of the series

$$\sum_{n=1}^{\infty}\left(1-\frac{1}{\sqrt{n}}\right)^n, \quad \sum_{n=2}^{\infty}\frac{1}{(\log n)^{\log n}}, \quad \sum_{n=2}^{\infty}\frac{1}{(\log n)^{\log(\log n)}}.$$

9. For each of the series

$$\sum_{n=1}^{\infty}\left[\frac{1\cdot 3\cdots(2n-1)}{2\cdot 4\cdots 2n}\right]^2, \quad \sum_{n=2}^{\infty}\frac{1}{n(\log n)^2}$$

show that $t=1$ in Raabe's test. Show by other methods that the first series is divergent and the second is convergent. This shows that Raabe's test is indecisive when $t=1$.

10. Prove Theorem XV with the aid of the following suggestions. If $p \neq 1$, the conclusions are easily drawn with the aid of Raabe's test. Write out the argument explicitly. For the case $p=1$, let

$$v_n=\frac{1}{(n-1)\log(n-1)};$$

use Taylor's formula with remainder to show that

$$\log\left(1-\frac{1}{n}\right)=-\frac{1}{n}-\frac{A_n}{n^2},$$

where A_n is bounded as $n\to\infty$. Then show that

$$\frac{v_{n+1}}{v_n}=1-\frac{1}{n}-\frac{1}{n\log n}-\frac{B_n}{n^2},$$

where B_n is bounded as $n\to\infty$. From this show that

$$\left|\frac{u_{n+1}}{u_n}\right|-\frac{v_{n+1}}{v_n}$$

is positive when n is sufficiently large. Now complete the proof of Theorem XV.

19.5 / THE BINOMINAL SERIES

Let us consider the Taylor's series expansion in powers of x of the function $f(x) = (1+x)^m$, where m is any real number. We have

$$\begin{aligned} f(x) &= (1+x)^m \\ f'(x) &= m(1+x)^{m-1} \\ f''(x) &= m(m-1)(1+x)^{m-2} \\ &\vdots \\ f^{(k)}(x) &= m(m-1)\cdots(m-k+1)(1+x)^{m-k} \\ &\vdots \end{aligned}$$

Thus, $f(0) = 1$, and for $k = 1, 2, \cdots$

$$f^{(k)}(0) = m(m-1)(m-2)\cdots(m-k+1).$$

Taylor's formula with remainder is

$$(1+x)^m = 1 + mx + \frac{m(m-1)}{2!}x^2 + \cdots + \frac{m(m-1)\cdots(m-n+1)}{n!}x^n + R_{n+1}.$$

We shall use Cauchy's formula for R_{n+1} (Theorem IV, §4.3). From (4.3–14) with $a = 0$ and $h = x$ we have

$$R_{n+1} = \frac{f^{(n+1)}(\theta x)}{n!}x^{n+1}(1-\theta)^n, \quad 0 < \theta < 1.$$

If m happens to be a positive integer or zero we see that $f^{(k)}(x) \equiv 0$ when $k > m$. Hence in this case $R_{m+1} \equiv 0$ and we have

$$(1+x)^m = 1 + mx + \frac{m(m-1)}{2!}x^2 + \frac{m(m-1)(m-2)}{3!}x^3 + \cdots + x^m.$$

This is just the ordinary binomial-expansion formula.

In the rest of this section we shall assume that m is not a positive integer or zero. The formula for R_{n+1} becomes

$$R_{n+1} = \frac{m(m-1)\cdots(m-n)(1+\theta x)^{m-n-1}}{n!}x^{n+1}(1-\theta)^n,$$

$$R_{n+1} = \frac{m(m-1)\cdots(m-n)}{n!}\left(\frac{1-\theta}{1+\theta x}\right)^n(1+\theta x)^{m-1}x^{n+1}. \tag{19.5–2}$$

Now suppose that $-1 < x < 1$. Then

$$0 < \frac{1-\theta}{1+\theta x} < 1.$$

If $m > 1$ we have $0 < (1+\theta x)^{m-1} < (1+|x|)^{m-1}$, while if $m < 1$ we have

$$(1+\theta x)^{m-1} = \frac{1}{(1+\theta x)^{1-m}} < \frac{1}{(1-|x|)^{1-m}},$$

so that $0<(1+\theta x)^{m-1}<(1-|x|)^{m-1}$. Thus, if $-1<x<1$,

$$|R_{n+1}| \leqq \frac{m(m-1)\cdots(m-n)}{n!}(1\pm|x|)^{m-1}|x|^{n+1}, \tag{19.5–3}$$

the choice of the double sign in $(1\pm|x|)^{m-1}$ depending on the sign of $m-1$. From (19.5–3) we shall show that $R_{n+1}\to 0$ as $n\to\infty$ if $|x|<1$. Since $(1\pm|x|)^{m-1}$ is independent of n, it is sufficient to show that

$$\lim_{n\to\infty}\frac{m(m-1)\cdots(m-n)}{n!}x^{n+1}=0. \tag{19.5–4}$$

Now let

$$u_n=\frac{m(m-1)\cdots(m-n)}{n!}x^{n+1}.$$

and consider the series Σu_n. This series is convergent if $|x|<1$, as we see by applying the ratio test of Theorem XIII, for

$$\left|\frac{u_{n+1}}{u_n}\right|=\left|\frac{m-n+1}{n+1}\right||x|\to|x| \text{ as } n\to\infty.$$

Since the series converges, it follows that $u_n\to 0$ as $n\to\infty$; thus (19.5–4) is established.

We have thus proved the validity of the series expansion

$$(1+x)^m=1+mx+\frac{m(m-1)}{2!}x^2+\cdots+\frac{m(m-1)\cdots(m-n+1)}{n!}x^n+\cdots \tag{19.5–5}$$

when $|x|<1$. This is called the *binomial series*. Except when m is a positive integer of zero it is a nonterminating series, none of the coefficients vanishing. For a discussion of the validity of (19.5–5) when $x=\pm 1$, see the Exercises. The student will note that the series (19.5–5) is the same as the series (19.4–2), which was discussed in Examples 2 and 4 of §19.4. Observe, however, that there is a logical difference between proving merely that the series converges and proving that its sum is equal to $(1+x)^m$.

EXERCISES

The purpose of this set of exercises is to guide the student in completing the discussion of the binomial series when $x=\pm 1$. Certain results of a general nature are needed, and these are taken up first.

1. Suppose $u_n>0$ and

$$\frac{u_{n+1}}{u_n}=1-\frac{p}{n}+\frac{A_n}{n^q},$$

where $p>0$, $q>1$, and A_n is bounded as $n\to\infty$. Show that, if $0<r<p$, u_n satisfies an inequality of the form $u_n\leqq Cn^{-r}$, where C is a constant, for all sufficiently large values of n. As a consequence, $u_n\to 0$ as $n\to\infty$. SUGGESTION: Let $v_n=n^{-r}$ and use the formula

(developed in §19.4)

$$\frac{v_{n+1}}{v_n} = 1 - \frac{r}{n} + \frac{B_n}{n^2},$$

where B_n is bounded as $n \to \infty$, to show that

$$\frac{v_{n+1}}{v_n} - \frac{u_{n+1}}{u_n}$$

has the same sign as $p - r$ when n is sufficiently large. Then, by an argument like that in the proof of Theorem V, draw the desired conclusion.

2. Let

$$u_n = \frac{a(a+1)\cdots(a+n-1)b(b+1)\cdots(b+n-1)}{n!\,c(c+1)\cdots(c+n-1)}$$

where none of the numbers a, b, c is zero or a negative integer. Suppose that $1 + c - (a + b) > 0$. Show that $u_n \to 0$ as $n \to \infty$. Use the result of Exercise 1.

3. Consider the binomial series for $x = -1$. Assume throughout that m is not zero or a positive integer. **(a)** Explain why the terms of the series are all of the same sign for sufficiently large values of n, and show that the series converges if $m > 0$, but diverges if $m < 0$. Note Example 4, §19.4. Actually this is a special case of Exercise 7, §19.4. **(b)** Follow the suggestions given and so prove that the binomial expansion (19.5–5) is valid if $m > 0$ and $x = -1$. Start with Taylor's formula with integral remainder (Theorem II, §4.2), taking $f(x) = (1 + x)^m$, $a = 0$, $x = -1$. In this way one finds

$$(1-1)^m = 1 - m + \frac{m(m-1)}{2!} + \cdots + \frac{m(m-1)\cdots(m-n+1)}{n!}(-1)^n + R_{n+1},$$

$$R_{n+1} = (-1)^{n+1} \cdot \frac{(m-1)(m-2)\cdots(m-n)}{n!}.$$

Then

$$R_{n+1} = -\frac{(-m+1)(-m+2)\cdots(-m+n)}{n!}.$$

Now use the result of Exercise 2 to show that $\lim_{n\to\infty} R_{n+1} = 0$. This proves that if $x = -1$ the binomial series is convergent, with sum 0, when $m > 0$.

4. Consider the binomial series for $x = 1$.

(a) Show that, for sufficiently large values of n, the terms of the series alternate in sign. Show also that the series is divergent if $m \leqq -1$ and convergent if $m > -1$. For this last part use the result of Exercise 2. The general term is

$$u_n = \frac{m(m-1)\cdots(m-n+1)}{n!}.$$

(b) Show that the binomial expansion (19.5–5) is valid if $x = 1$ and $m > -1$. Start with Taylor's formula as in Exercise **3(b)**, this time putting $x = 1$, and getting

$$(1+1)^m = u_0 + u_1 + \cdots + u_n + R_{n+1},$$

$$R_{n+1} = \frac{m(m-1)\cdots(m-n)}{n!}\int_0^1 (1-t)^n (1+t)^{m-n-1}\,dt.$$

If $n > m - 1$ the value of the integral is less than $1/(n+1)$. Explain why this is so.

Thus

$$|R_{n+1}| \leqq \left|\frac{m(m-1)\cdots(m-n)}{(n+1)!}\right| = |u_{n+1}|.$$

What is the rest of the argument?

19.6 / MULTIPLICATION OF SERIES

We have already noted, at the beginning of §19.31, that term-by-term addition (or subtraction) of convergent series is a legitimate operation. This is convenient for obtaining new series expansions from expansions already established.

Example 1. Show that, if $|x| < 1$,

$$\tfrac{1}{2}\log\frac{1+x}{1-x} = x + \tfrac{1}{3}x^3 + \tfrac{1}{5}x^5 + \cdots. \tag{19.6–1}$$

We get this series expansion by observing that

$$\log\frac{1+x}{1-x} = \log(1+x) - \log(1-x)$$

and using (19.1–1). From this latter formula we have

$$\log(1+x) = x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 - \tfrac{1}{4}x^4 + \tfrac{1}{5}x^5 - \cdots$$

if $-1 < x \leqq 1$. Consequently, replacing x by $-x$, we have

$$\log(1-x) = -x - \tfrac{1}{2}x^2 - \tfrac{1}{3}x^3 - \tfrac{1}{4}x^4 - \tfrac{1}{5}x^5 - \cdots$$

if $-1 \leqq x < 1$. Combining by subtraction, which is valid if $-1 < x < 1$, since both series are then convergent, we have

$$\log(1+x) - \log(1-x) = 2x + \tfrac{2}{3}x^3 + \tfrac{2}{5}x^5 + \cdots.$$

This is equivalent to (19.6–1).

It is likewise convenient to be able to find new series expansions by multiplying known series expansions. For instance, suppose we wish to expand

$$\left(\frac{1+x}{1+x^2}\right)^{1/2} = (1+x)^{1/2}(1+x^2)^{-1/2}$$

in powers of x. We might proceed as follows: The expansion of $(1+x)^{1/2}$ is a particular case of (19.5–5); to a few terms it is

$$(1+x)^{1/2} = 1 + \tfrac{1}{2}x - \tfrac{1}{8}x^2 + \tfrac{1}{16}x^3 - \tfrac{5}{128}x^4 + \cdots. \tag{19.6–2}$$

To get the series for $(1+x^2)^{-1/2}$ we put x^2 in place of x in (19.5–5) and set $m = -\frac{1}{2}$. This is legitimate if $|x| < 1$, since then $x^2 < 1$ also. To a few terms the series is

$$(1+x^2)^{-1/2} = 1 - \tfrac{1}{2}x^2 + \tfrac{3}{8}x^4 - \tfrac{5}{16}x^6 + \cdots. \tag{19.6–3}$$

Thus

$$\left(\frac{1+x}{1+x^2}\right)^{1/2} = (1 + \tfrac{1}{2}x - \tfrac{1}{8}x^2 + \tfrac{1}{16}x^3 - \cdots)(1 - \tfrac{1}{2}x^2 + \tfrac{3}{8}x^4 - \cdots).$$

The next question is: How do we multiply the two series together to get a new series? Proceeding just as though the series were finite sums, we might write down the following scheme, which arises by multiplying the second series successively by each term of the first series:

$$\begin{array}{ccccc} 1 & -\frac{1}{2}x^2 & +\frac{3}{8}x^4 & -\frac{5}{16}x^6+\cdots \\ \frac{1}{2}x & -\frac{1}{4}x^3 & +\frac{3}{16}x^5 & -\cdots \\ -\frac{1}{8}x^2 & +\frac{1}{16}x^4 & -\frac{3}{64}x^6+\cdots \\ & \frac{1}{16}x^3 & -\frac{1}{32}x^5 & +\cdots \\ & \cdots & \cdots & \end{array}$$

There will be an infinite number of rows, each row being an infinite series. But we observe that there are only a finite number of terms of each degree, so that if we collect together terms of like degree, we obtain for the first few terms

$$1+\tfrac{1}{2}x-\tfrac{5}{8}x^2-\tfrac{3}{16}x^3+\cdots. \tag{19.6–4}$$

It is clear that there is here a systematic process, but it remains to prove that the process gives a series which has as its sum the product of the sums of the two original series. There is a general theorem which justifies the process.

THEOREM XVII. *Suppose that each of the series* Σu_n, Σv_n *is absolutely convergent, with sums* U *and* V *respectively:*

$$U = u_0+u_1+u_2+\cdots, \tag{19.6–5}$$

$$V = v_0+v_1+v_2+\cdots. \tag{19.6–6}$$

Let $w_0 = u_0v_0$, $w_1 = u_0v_1+u_1v_0$, *and in general*

$$w_n = u_0v_n+u_1v_{n-1}+\cdots+u_nv_0. \tag{19.6–7}$$

Then the series Σw_n *is absolutely convergent, and its sum is* UV*:*

$$UV = w_0+w_1+w_2+\cdots. \tag{19.6–8}$$

Moreover, any infinite series which has as its terms the products u_iv_j *(*i *and* $j \geqq 0$*) arranged in any order, each product occurring once and only once, is absolutely convergent, with sum* UV*.*

Proof. Let us consider the array

$$\begin{array}{cccccc} u_0v_0 & u_0v_1 & \cdots & u_0v_n & \cdots \\ u_1v_0 & u_1v_1 & \cdots & u_1v_n & \cdots \\ \cdots & & & \cdots & \\ \cdots & & & \cdots & \\ u_nv_0 & u_nv_1 & \cdots & u_nv_n & \cdots \\ \cdots & & & & \\ \cdots & & & & \end{array} \tag{I}$$

and the similar array in which u_iv_j is replaced by $|u_i|\,|v_j|$. Let us denote the second array by (II). Finally, let

$$A = |u_0| + |u_1| + \cdots + |u_n| + \cdots$$
$$B = |v_0| + |v_1| + \cdots + |v_n| + \cdots.$$

These last series are convergent because of the assumption that (19.6–5) and (19.6–6) are absolutely convergent. Now consider any series formed from the terms of the array (II), taken in some definite order. Such a series is convergent, since each of its partial sums is less than or equal to AB. For, any such partial sum is less than

$$(|u_0| + |u_1| + \cdots + |u_n|)(|v_0| + |v_1| + \cdots + |v_n|)\,,$$

if n is taken sufficiently large, and this last product is certainly no larger than AB, since the first factor does not exceed A and the second factor does not exceed B. It follows that any infinite series formed by taking the terms of the array (I) in some definite order is absolutely convergent. Since, of any two such series, one is merely a rearrangement of the other, Theorem XI (§19.31) assures us that all such series have the same sum.

Now one possible arrangement is that in which we take u_0v_0 first, then u_0v_1, u_1v_1, and u_1v_0; then u_0v_2, u_1v_2, u_2v_2, u_2v_1, and u_2v_0; and at the nth stage all terms u_iv_j, for which i and j do not exceed n and at least one of them is equal to n. At the nth stage our partial sum is exactly

$$(u_0 + u_1 + \cdots + u_n)(v_0 + v_1 + \cdots + v_n),$$

which is the sum of all the terms in an upper left square portion of the array (I). Hence, by (19.6–5) and (19.6–6) the limit of the partial sums is UV, and this must be the value of the series (by the fact that the limit of a product is the product of the limits).

Another arrangement is that in which we take first u_0v_0, then u_0v_1 and u_1v_0, then u_0v_2, u_1v_1, and u_2v_0, and at the $n$the stage all terms u_iv_j for which $i + j = n$. This arrangement gives the same sum UV, of course. But if we group the terms according to the stage, we get

$$u_0v_0 + (u_0v_1 + u_1v_0) + (u_0v_2 + u_1v_1 + u_2v_0) + \cdots,$$

which is exactly the series (19.6–8). The insertion of the parentheses technically changes the series, of course, since the last series has the sum of all products in one parenthesis as a single term, whereas previously each product u_iv_j was a single term. But the insertion of parentheses into a convergent series always leads to a convergent series whose sum is the same as that of the original series (see Exercise 7). Therefore (19.6–8) holds, and the proof of the theorem is complete.

As an application of the theorem let us return to (19.6–4). This series was obtained by multiplication of the series (19.6–2) and (19.6–3) by the rule of Theorem XVII. Since each of the two latter series is absolutely convergent if

$|x| < 1$, the series (19.6–4) is absolutely convergent if $|x| < 1$, and its sum is the product of the sums of the other two series, namely $(1+x)^{1/2}(1+x^2)^{-1/2}$.

Observe that, if the series

$$a_0 + a_1x + a_2x^2 + \cdots + a_ix^i + \cdots,$$
$$b_0 + b_1x + b_2x^2 + \cdots + b_jx^j + \cdots$$

are multiplied according to Theorem XVII, the resulting series is

$$a_0b_0 + (a_0b_1 + a_1b_0)x + (a_0b_2 + a_1b_1 + a_2b_0)x^2 + \cdots,$$

the coefficient of x^n in the last series being

$$a_0b_n + a_1b_{n-1} + \cdots + a_nb_0.$$

To justify this application of Theorem XVII in any given case we need merely to check on the absolute convergence of the first two series.

Example 2. Find the series expansion of $\dfrac{1}{1-x}\log\dfrac{1}{1-x}$ in powers of x.

We know from (19–3) that

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots = \sum_{n=0}^{\infty} x^n \tag{19.6–9}$$

if $|x| < 1$; also, from (19.1–1),

$$\log(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots = -\sum_{n=1}^{\infty} \frac{x^n}{n}$$

if $-1 \leqq x < 1$, so that

$$\log\frac{1}{1-x} = \sum_{n=1}^{\infty} \frac{x^n}{n}. \tag{19.6–10}$$

Each of these series is absolutely convergent if $|x| < 1$, as may be verified by Theorem XIII. The coefficients in (19.6–9) are $a_n = 1$, $n = 0, 1, 2, \ldots$; those in (19.6–10) are $b_0 = 0$, $b_n = \dfrac{1}{n}$, $n = 1, 2, \ldots$. Therefore $a_0b_0 = 0$, and if $n \geqq 1$,

$$a_0b_n + a_1b_{n-1} + \cdots + a_nb_0 = \frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{2} + 1.$$

Thus

$$\frac{1}{1-x}\log\frac{1}{1-x} = \sum_{n=1}^{\infty} \left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right) x^n$$
$$= x + (1+\tfrac{1}{2})x^2 + (1+\tfrac{1}{2}+\tfrac{1}{3})x^3 + \cdots.$$

This series is absolutely convergent if $|x| < 1$.

EXERCISES

1. Multiply the series for $(1-x)^{-1}$ (series (19–3), valid when $|x|<1$) by itself and so obtain the expansion

$$(1-x)^{-2}=\sum_{n=0}^{\infty}(n+1)x^n, \quad |x|<1.$$

2. Derive the expansion

$$(1-x)^{-3}=\sum_{n=0}^{\infty}\frac{(n+1)(n+2)}{2}x^n \quad (|x|<1)$$

by multiplication of series. Use the result of Exercise 1.

3. Obtain the expansion, valid when $0\leq x<1$,

$$\frac{1}{1-x}\frac{\tan^{-1}\sqrt{x}}{\sqrt{x}}=1+(1-\tfrac{1}{3})x+(1-\tfrac{1}{3}+\tfrac{1}{5})x^2+(1-\tfrac{1}{3}+\tfrac{1}{5}-\tfrac{1}{7})x^3+\cdots.$$

4. Prove by multiplication of series that if

$$f(x)=\sum_{n=0}^{\infty}\frac{x^n}{n!}, \text{ then } f(x)f(y)=f(x+y)$$

for arbitrary x and y.

5. Prove that

$$2\sum_{m=0}^{\infty}(-1)^m\frac{x^{2m+1}}{(2m+1)!}\cdot\sum_{n=0}^{\infty}(-1)^n\frac{x^{2n}}{(2n)!}=\sum_{p=0}^{\infty}(-1)^p\frac{(2x)^{2p+1}}{(2p+1)!}.$$

6. Using no other properties of $\sin x$ and $\cos x$ than the fact that the series representations (19.1–8) and (19.1–9) are valid, prove that $\sin^2 x+\cos^2 x=1$.

7. Let Σu_n be a convergent series with sum s. Let sets of parentheses be introduced into the series, e.g.:

$$(u_1+u_2)+(u_3+u_4+u_5)+(u_6)+(u_7+u_8)+\cdots.$$

Let the sum of the terms in the nth parenthesis be v_n. Prove that the series Σv_n is convergent and that its sum is s.

8. The rule for multiplying two series, as given in Theorem XVII, may give a wrong result if the series are not absolutely convergent. Verify this by taking $u_n=v_n=(-1)^n/\sqrt{n+1}$ in the theorem and showing that w_n does not approach zero as $n\to\infty$. SUGGESTION: Show that, for $k=0, 1, \ldots, n$, $(k+1)(n-k+1)\leqq\frac{1}{4}(n+2)^2$; then show that $|w_n|\geqq 2(n+1)/(n+2)$.

19.7 / DIRICHLET'S TEST

Thus far all our tests, with one exception, have been tests for series of positive terms, or tests for absolute convergence. The one exception is the alternating series test of §19.32. We shall now discuss a more general test which is useful on occasion for proving that a series is convergent, though not necessarily absolutely convergent. Most tests of the character of this one depend upon an algebraic device known as *summation by parts*. The device is analogous to integration by parts.

Let $a_0, a_1, \ldots$ and $b_0, b_1, \ldots$ be two arbitrary sequences of numbers. Let

$$s_n = a_0 + a_1 + \cdots + a_n.$$

Then

$$a_0b_0 + a_1b_1 + \cdots + a_nb_n = s_0(b_0 - b_1) + s_1(b_1 - b_2) + \cdots + s_{n-1}(b_{n-1} - b_n) + s_nb_n. \tag{19.7-1}$$

This is the identity of summation by parts. It is known as *Abel's summation identity.* Niels H. Abel was a Norwegian mathematician (1802–1829). The proof of the formula is very simple. We observe that $a_n = s_n - s_{n-1}$ if $n \geqq 1$, and $a_0 = s_0$. Therefore

$$\begin{aligned} a_0b_0 &= s_0b_0 \\ a_1b_1 &= s_1b_1 - s_0b_1 \\ a_2b_2 &= s_2b_2 - s_1b_2 \\ &\vdots \quad\quad \vdots \quad\quad \vdots \\ a_{n-1}b_{n-1} &= s_{n-1}b_{n-1} - s_{n-2}b_{n-1} \\ a_nb_n &= s_nb_n - s_{n-1}b_n. \end{aligned}$$

Adding these results and grouping the terms on the right appropriately, we obtain (19.7–1).

We can now state a test known by the name of P. G. Lejeune Dirichlet, a German mathematician of the first half of the nineteenth century.

THEOREM XVIII. (DIRICHLET'S TEST.) *Consider a series of the form*

$$a_0b_0 + a_1b_1 + a_2b_2 + \cdots + a_nb_n + \cdots \tag{19.7-2}$$

which satisfies the following conditions:
(a) the terms b_n are positive, $b_{n+1} \leqq b_n$, and $b_n \to 0$ as $n \to \infty$;
(b) there is some constant M independent of n such that

$$|a_0 + a_1 + \cdots + a_n| \leqq M \text{ for all values of } n.$$

Then the series (19.7–2) *is convergent.*

Proof. Let

$$S_n = a_0b_0 + \cdots + a_nb_n.$$

We have to show that $\lim_{n\to\infty} S_n$ exists. Now by (19.7–1) we have

$$S_n = T_n + s_nb_n \tag{19.7-3}$$

where

$$T_n = s_0(b_0 - b_1) + \cdots + s_{n-1}(b_{n-1} - b_n).$$

Now $|s_nb_n| \leqq Mb_n$, by condition (b), and so $s_nb_n \to 0$, since $b_n \to 0$. Next we show

that $\lim_{n\to\infty} T_n$ exists. The desired conclusion will then follow from (19.7–3). Now T_n is the partial sum of the series

$$s_0(b_0 - b_1) + s_1(b_1 - b_2) + \cdots, \tag{19.7–4}$$

and certainly T_n will approach a limit if we show that the series (19.7–4) is absolutely convergent. Now

$$\begin{aligned} &|s_0(b_0 - b_1)| + |s_1(b_1 - b_2)| + \cdots + |s_{n-1}(b_{n-1} - b_n)| \\ &\leqq M(b_0 - b_1) + M(b_1 - b_2) + \cdots + M(b_{n-1} - b_n) = M(b_0 - b_n), \end{aligned}$$

or

$$\sum_{k=0}^{n-1} |s_k(b_k - b_{k+1})| \leqq Mb_0.$$

Here we have used the conditions (a) and (b). Since these sums are bounded, we conclude that the series (19.7–4) is absolutely convergent. This is all that was needed to complete the proof.

Theorem XVIII is useful in connection with trigonometric series.

Example. The series

$$\frac{\cos x}{1} + \frac{\cos 3x}{3} + \frac{\cos 5x}{5} + \cdots \tag{19.7–5}$$

is convergent if x is not one of the values $0, \pm\pi, \pm 2\pi, \ldots,$

To prove this we apply Theorem XVIII, taking $a_0 = 0$, $a_1 = \cos x$, $a_2 = \cos 3x, \ldots, a_n = \cos(2n-1)x$, $n = 1, 2, \ldots,$ and $b_0 = 1$, $b_n = (2n-1)^{-1}$, $n = 1, 2, \ldots$. Condition (a) is clearly satisfied, so it remains only to show that condition (b) is satisfied. This requires a bit of ingenuity. We use the trigonometric identity

$$2\cos A \sin B = \sin(A + B) - \sin(A - B).$$

Taking $B = x$ and A successively equal to x, $3x$, $5x, \ldots,$ we have

$$\begin{aligned} 2\cos x \sin x &= \sin 2x - 0, \\ 2\cos 3x \sin x &= \sin 4x - \sin 2x, \\ 2\cos 5x \sin x &= \sin 6x - \sin 4x, \\ &\vdots \quad\quad \vdots \quad\quad \vdots \\ 2\cos(2n-1)x \sin x &= \sin 2nx - \sin 2(n-1)x. \end{aligned}$$

Adding these results, we find

$$2\sin x(\cos x + \cos 3x + \cdots + \cos(2n-1)x) = \sin 2nx,$$

or

$$\cos x + \cos 3x + \cdots + \cos(2n-1)x = \frac{\sin 2nx}{2\sin x}. \tag{19.7–6}$$

If now x is not one of the values $0, \pm\pi, \pm 2\pi, \ldots$, then $\sin x \neq 0$, and

$$|\cos x + \cos 3x + \cdots + \cos(2n-1)x| \leqq \frac{1}{2|\sin x|}.$$

Hence condition (b) of Theorem XVIII is satisfied with $M = \dfrac{1}{2|\sin x|}$. Therefore the series (19.7–5) is convergent.

EXERCISES

1. Prove that the series

$$\frac{\sin x}{1} + \frac{\sin 3x}{\sqrt{3}} + \frac{\sin 5x}{\sqrt{5}} + \cdots$$

is convergent for all values of x. Use the identity

$$2 \sin A \sin B = \cos(A-B) - \cos(A+B).$$

2. Prove the identities

$$2 \sin \frac{x}{2}(\sin x + \sin 2x + \cdots + \sin nx) = \cos \frac{x}{2} - \cos \frac{2n+1}{2} x,$$

$$2 \sin \frac{x}{2}(\cos x + \cos 2x + \cdots + \cos nx) = \sin \frac{2n+1}{2} x - \sin \frac{x}{2}.$$

3. Suppose that $a_n > 0$, $a_{n+1} \leqq a_n$, $a_n \to 0$ as $n \to \infty$. Prove that the series $\Sigma_{n=1}^{\infty} a_n \sin nx$ is convergent for all values of x, and that $\Sigma_{n=1}^{\infty} a_n \cos nx$ is convergent with the possible exception of the cases in which $x = 2\pi m$, $m = 0, \pm 1, \pm 2, \ldots$. Use the identities in Exercise 2.

4. Suppose that $\Sigma_{n=1}^{\infty} a_n/n^p$ is convergent. Show that $\Sigma_{n=1}^{\infty} a_n/n^q$ is convergent if $p < q$.

5. Deduce Theorem XII as a special case of Dirichlet's test.

6. The following theorem is known as Abel's test: Suppose Σa_n is convergent, and that $b_n > 0$, $b_{n+1} \leqq b_n$. Then $\Sigma a_n b_n$ is convergent. Prove this.

7. Show that the series

$$\frac{1}{2 \log 2} + \frac{1}{3 \log 3} - \frac{1}{4 \log 4} - \frac{1}{5 \log 5} - \frac{1}{6 \log 6} - \frac{1}{7 \log 7} + \cdots$$

is convergent, the rule of signs being that successive terms with the same sign come in groups of 2, 4, 8, 16, ... Begin by considering the series

$$\tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} - \tfrac{1}{5} - \tfrac{1}{6} - \tfrac{1}{7} + \tfrac{1}{8} + \cdots.$$

8. Show that the series

$$1 - \tfrac{1}{2} - \tfrac{1}{3} + \tfrac{1}{4} + \tfrac{1}{5} - \tfrac{1}{6} - \tfrac{1}{7} + + - - \cdots$$

is convergent. Successive terms of like sign come in groups of two.

MISCELLANEOUS EXERCISES

1. For what values of x is $\Sigma n^x x^n$ convergent?

2. Find the sum of the first n terms of the series $\Sigma_{n=1}^{\infty} \log[1 + (1/n)]$. Is the series convergent or divergent?

3. Show that $\Sigma(1/n)\log[1+(1/n)]$ is convergent.

4. Show by Theorem III, or otherwise, that $\Sigma_{n=2}^{\infty} 1/(\log n)^c$ is divergent for all values of c.

5. Express $(\log n)^{\log n}$ as a power of n, and use the result to show that $\Sigma \frac{1}{(\log n)^{\log n}}$ is convergent.

6. Examine each of the following series for convergence or divergence.

(a) $\Sigma(-1)^n \frac{n^n}{(n+1)^{n+1}}$, (c) $\Sigma(-1)^n\left(\frac{n}{n+1}\right)^n$.

(b) $\Sigma \frac{(n+1)^n}{n^{n+1}}$. (d) $\Sigma \frac{[1+(1/n)]^{n^2}}{e^n}$.

7. Show that

(a) $\frac{2\cdot 4\cdots 2n}{3\cdot 5\cdots(2n+1)} < \frac{1}{\sqrt{n+1}}$.

(b) Classify the values of x according to whether the series

$$\Sigma \frac{n!}{3\cdot 5\cdots(2n+1)}x^n$$

is convergent or divergent.

8. Find all values of x for which the series $\Sigma \frac{1\cdot 3\cdots(2n-1)}{2\cdot 4\cdots 2n}\frac{1}{n^x}$ is convergent.

9. (a) Is $\Sigma \sin \pi[n+(1/n)]$ absolutely convergent? Is it convergent?

(b) Show that $\Sigma \sin^2 \pi[n+(1/n)]$ is convergent.

(c) For what values of θ is $\Sigma(-1)^n(1/n)\cos(\theta/n)$ convergent?

(d) Is $\Sigma[1-\cos(\pi/n)]$ convergent or divergent?

10. Discuss the convergence of each series, classifying the values of x into those for which the series converges and those for which it diverges.

(a) $\sum_{n=1}^{\infty}(-1)^n \frac{1\cdot 3\cdots(2n-1)}{2\cdot 4\cdots 2n}\frac{2}{4n+1}\left(\frac{x}{2}\right)^{4n+1}$.

(b) $\sum_{n=2}^{\infty} 2^{2n-2}\frac{((n-1)!)^2}{(2n)!}x^{2n}$.

(c) $\sum_{n=1}^{\infty}\frac{(n+1)^n}{n!}x^n$.

11. Show that $\Sigma \frac{\log[1+(1/n)]}{a^{\log(\log n)}}$ is convergent if $a > e$ and divergent if $0 < a \leqq e$.

12. (a) If $x_n = 1-\frac{1}{2}+\frac{1}{3}-\cdots-1/(2n)$, show that $x_n \to \log 2$ as $n \to \infty$, by using the definition of Euler's constant (see Exercise 6, §19.21). HINT: Show that $x_n = C_{2n} - C_n + \log 2$ in the notation of the exercise just mentioned.

(b) Prove that the partial sums of the series

$$1+\tfrac{1}{2}-\tfrac{1}{3}+\tfrac{1}{4}+\tfrac{1}{5}-\tfrac{1}{6}++-\cdots$$

approach $+\infty$ as $n \to \infty$. Make use of Euler's constant.

13. Prove that, if $u_n > 0$ and Σu_n is convergent, so is Σu_n^2.

14. Show that the series $\Sigma \log(n \sin(1/n))$ is convergent.

15. Consider the series

$$1-\tfrac{1}{2}-\tfrac{1}{3}+\tfrac{1}{4}+\tfrac{1}{5}+\tfrac{1}{6}+\tfrac{1}{7}-\tfrac{1}{8}-\cdots,$$

the rule of signs being that successive terms with the same sign come in groups of 1, 2, 4, 8, 16, ... Show that this series does not satisfy the condition of Theorem VIII, and is therefore divergent. Show also that, nevertheless, the sum s_n of the first n terms of the series satisfies the condition $0 < s_n \leqq 1$.

16. Consider the series

$$1-\tfrac{1}{2}-\tfrac{1}{3}+\tfrac{1}{4}+\tfrac{1}{5}+\tfrac{1}{6}-\tfrac{1}{7}-\cdots,$$

in which successive terms of like sign come in groups of 1, 2, 3, 4, ... Show that this series is convergent.

17. Prove that the series $\sum_{n=2}^{\infty}(1/n)\cos(a\log n)$ is divergent by showing that it does not satisfy the condition of Theorem VIII. SUGGESTION: Suppose $0 < \delta < \pi/2$. Let n_1 denote the greatest integer $\leqq \exp[(2\pi m+\delta)/a]$ and let n_2 denote the greatest integer $\leqq \exp[(2\pi m+\delta)/a]$, where m is a fixed positive integer, and $\exp u = e^u$. Show that $\cos(a\log n) \geqq \cos\delta > 0$ if $n_1 < n \leqq n_2$. Then show that

$$\frac{\cos\delta}{n_1}+\sum_{n=n_1+1}^{n_2}\frac{\cos(a\log n)}{n} > \int_{n_1}^{n_2+1}\frac{\cos\delta}{x}\,dx > \frac{2\delta\cos\delta}{a}.$$

Now note that n_1 and $n_2 \to \infty$, if $m \to \infty$, and so finish the proof. (Assume $a > 0$.)

18. Show by an argument similar to that of Exercise 17 that the series $\sum_{n=4}^{\infty}\dfrac{\cos(\log(\log n))}{\log n}$ is divergent.

20 / UNIFORM CONVERGENCE

20 / FUNCTIONS DEFINED BY CONVERGENT SEQUENCES

In the more advanced parts of analysis we often deal with functions which are defined by means of infinite series. Suppose that

$$f(x) = \sum_{n=1}^{\infty} u_n(x), \tag{20–1}$$

where it is assumed that the terms of the series are each defined on the same interval $a \leqq x \leqq b$, and that the series is convergent for each x of the interval. The sum of the series is then also a function of x, and we denote it by $f(x)$. The main purpose of this chapter is to deal with questions A–D below.

Assuming that we know a good deal about the functions $u_1(x)$, $u_2(x), \ldots,$ but nothing about the function $f(x)$ except in so far as we can draw certain conclusions from the series (20–1), we raise four questions:

A. What conditions will assure us that f is continuous?

B. What conditions will assure us that f is differentiable?

C. What conditions will justify us in writing

$$\int_a^b f(x)\,dx = \int_a^b u_1(x)\,dx + \int_a^b u_2(x)\,dx + \cdots ? \tag{20–2}$$

D. What conditions will justify us in writing

$$f'(x) = u_1'(x) + u_2'(x) + \cdots ? \tag{20–3}$$

That is, when can we integrate and differentiate an infinite series just as though it were a *finite* sum of functions?

The answers which we shall give to all these equations are dependent on the concept of *uniform convergence.* Before proceeding further, however, let us look at the problems raised by these questions in another way. Consider the partial sums

$$s_n(x) = u_1(x) + u_2(x) + \cdots + u_n(x) \tag{20–4}$$

of the series (20–1). Then

$$f(x) = \lim_{n\to\infty} s_n(x); \tag{20–5}$$

that is, the function $f(x)$ is the limit of a *sequence* of functions. Now we may consider the notion of the limit of a sequence of functions quite apart from the

notion of a function defined by an infinite series, by presenting the sequence directly, rather than as a sum formed from the terms of a series.

Example 1. Consider the sequence $f_n(x) = x^n$, and define

$$f(x) = \lim_{n\to\infty} f_n(x), \quad 0 \leqq x \leqq 1. \tag{20–6}$$

The sequence is convergent if $0 \leqq x \leqq 1$, and we see that (20–6) is equivalent to the definition

$$\begin{cases} f(x) = 0 & \text{if} \quad 0 \leqq x < 1, \\ f(1) = 1. \end{cases} \tag{20–7}$$

The graphs of several of the functions $f_n(x)$ are shown in Fig. 176.

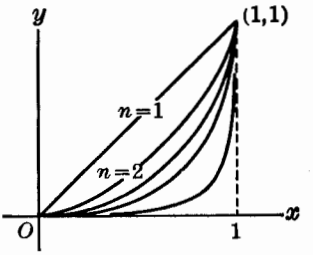


Fig. 176.

We shall find that, in framing our answers to the questions A–D, the important consideration is that of functions defined as the limits of sequences. For this reason the subsequent discussion in this chapter is often phrased in terms of sequences rather than in terms of series. Let us rephrase the questions as follows:

Assuming that a function $f(x)$ is known to us entirely by

$$f(x) = \lim_{n\to\infty} f_n(x), \quad a \leqq x \leqq b, \tag{20–8}$$

we ask:

A′. What conditions on the functions f_n will assure us that f is continuous?

B′. What conditions will assure us that f is differentiable?

C′. Under what conditions will it be true that

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \int_a^b f_n(x)\,dx? \tag{20–9}$$

D′. Under what conditions will it be true that

$$f'(x) = \lim_{n\to\infty} f_n'(x)? \tag{20–10}$$

In the particular case when $f_n(x) = s_n(x)$ and $s_n(x)$ is defined by (20–4), the questions A′–D′ are the same as the questions A–D.

We do not aim to get answers to these questions in the form of *necessary and sufficient* conditions; the conditions which we shall impose will be *sufficient* but not *necessary*.

Let us examine the situation in Example 1 with respect to question A′. Here we observe that each of the functions $f_n(x) = x^n$ is continuous on the closed interval [0, 1]; nevertheless, the limit function $f(x)$ is not continuous on the whole interval, but is discontinuous at the point $x = 1$. From this example we can con-

clude that in question A′ the assumption that each of the functions f_n is continuous on $[a, b]$ is *not* sufficient to assure us that the limit function f is continuous on $[a, b]$. Something more is needed, not about the functions f_n themselves, but about the way in which $f_n(x)$ converges to $f(x)$.

Next we consider an example which will be instructive with respect to questions B′ and D′.

Example 2. Let $f_n(x) = x^n/n$, $n = 1, 2, \ldots$.
On the interval $0 \leqq x \leqq 1$, $0 \leqq f_n(x) \leqq 1/n$, and hence

$$\lim_{n\to\infty} f_n(x) = 0, \quad 0 \leqq x \leqq 1.$$

The limit function $f(x)$ has the value zero for each x in the interval [0, 1]. Thus f is differentiable. But consider the sequence of derivatives $f_n'(x) = x^{n-1}$. The successive members of this sequence are $1, x, x^2, x^3, \ldots$, and as we saw in Example 1,

$$\lim_{n\to\infty} f_n'(x) = 0 \quad \text{if} \quad 0 \leqq x < 1$$

$$\lim_{n\to\infty} f_n'(1) = 1.$$

Thus

$$\lim_{n\to\infty} f_n'(1) \neq f'(1).$$

This example shows us that in question D′ the relation (20–10) may fail to be true (at least for *some* values of x in the interval) even when the functions f_n and f are differentiable and the sequence of derivatives $f_n'(x)$ is convergent for *every* x in the interval.

Finally we give an example bearing on question C′.

Example 3. Let us define a sequence $f_n(x)$ so that its graph is as indicated in Fig. 177. That is, the graph of $y = f_n(x)$ for $0 \leqq x \leqq 1$ consists of three line segments: the line $y = 4n^2x$ from $x = 0$ to $x = 1/(2n)$, the line $y = -4n^2x + 4n$ from $x = 1/(2n)$ to $x = 1/n$, and the line $y = 0$ from $x = 1/n$ to $x = 1$. The high point of the graph is at $x = 1/(2n)$, the height being $2n$. From the definition we see that, for each x in [0, 1],

$$\lim_{n\to\infty} f_n(x) = 0.$$

For, $f_n(0) = 0$ for all n, and if $0 < x \leqq 1$, $f_n(x) = 0$ as soon as n is large enough to make $1/n < x$. Thus the limit function $f(x)$ has the value zero for each x, and

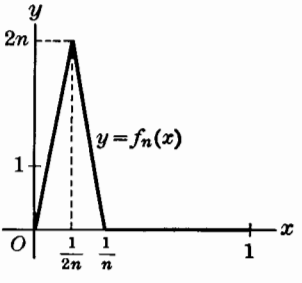


Fig. 177.

accordingly

$$\int_0^1 f(x)\,dx = 0.$$

But

$$\int_0^1 f_n(x)\,dx = 1$$

for each value of n, because the integral is just equal to the area of a triangle of base $1/n$ and height $2n$. Therefore

$$\int_0^1 f(x)\,dx \neq \lim_{n\to\infty} \int_0^1 f_n(x)\,dx.$$

This shows that for the truth of (20–9) in question C′ we must have something more than the mere convergence of $f_n(x)$ to $f(x)$, even when the functions f_n and f are all continuous.

In concluding this section let us observe that any function which is defined by a convergent *sequence* of functions may also be defined by a convergent *series* of functions. For if $f(x) = \lim_{n\to\infty} f_n(x)$, let us form the series

$$f_1(x) + [f_2(x) - f_1(x)] + [f_3(x) - f_2(x)] + \cdots, \tag{20–11}$$

that is, define the terms $u_n(x)$ by the formulas $u_1(x) = f_1(x)$, and $u_n(x) = [f_n(x) - f_{n-1}(x)]$ if $n > 1$. Then

$$s_n(x) = f_1(x) + [f_2(x) - f_1(x)] + \cdots + [f_n(x) - f_{n-1}(x)],$$

and when we simplify the right side of this equation we find that $s_n(x) = f_n(x)$. Thus the series (20–11) has the sum $f(x)$.

For instance, the limit function $f(x)$ in Example 1 can be expressed as the infinite series

$$f(x) = x + (x^2 - x) + (x^3 - x^2) + (x^4 - x^3) + \cdots.$$

20.1 / THE CONCEPT OF UNIFORM CONVERGENCE

Let us begin by recalling the definition of the limit of a sequence, as it applies to a sequence of functions. Suppose that

$$f(x) = \lim_{n\to\infty} f_n(x) \tag{20.1–1}$$

for each x of the interval $a \leqq x \leqq b$. According to the definition in §1.62 this means that if ϵ is any positive number and if x is any point of the interval, there is some integer N, the size of which will usually depend on ϵ and x, such that

$$|f_n(x) - f(x)| < \epsilon \tag{20.1–2}$$

if $N \leqq n$. The inequality (20.1–2) is equivalent to the two inequalities

$$f(x) - \epsilon < f_n(x) < f(x) + \epsilon. \tag{20.1–3}$$

Let us examine in a particular case the way in which the choice of N may depend on ϵ and x. Consider, for instance, $f_n(x) = x^n$. We saw in Example 1 of §20 that $f_n(x) \to f(x)$ if $0 \leqq x \leqq 1$, where $f(x) = 0$ if $0 \leqq x < 1$ and $f(1) = 1$. Suppose that $0 < \epsilon < 1$ and $0 < x < 1$. Then (20.1–2) is equivalent to each of the following inequalities:

$$\begin{aligned} x^n &< \epsilon, \\ n \log x &< \log \epsilon, \\ \log(1/\epsilon) &< n \log(1/x), \\ \frac{\log(1/\epsilon)}{\log(1/x)} &< n. \end{aligned} \tag{20.1–4}$$

(Observe that $\log x < 0$ and $\log(1/x) > 0$ if $0 < x < 1$.) Now, to have (20.1–4) true for all $n \geqq N$ it is necessary to choose the integer N large enough so that

$$\frac{\log(1/\epsilon)}{\log(1/x)} < N. \tag{20.1–5}$$

The dependence on ϵ and x shows up clearly here. As $\epsilon \to 0$, $\log(1/\epsilon) \to +\infty$, and hence $N \to \infty$. Also, for a fixed ϵ, as $x \to 1-$, $\log(1/x) \to 0+$ and we see from (20.1–5) that $N \to \infty$. *There is no value of N such that* (20.1–5) *holds simultaneously for all values of x in the range $0 < x < 1$.*

The foregoing situation illustrates what we shall call *nonuniform* convergence. We shall now define what we mean by *uniform* convergence.

Definition. *Suppose that the sequence of functions $\{f_n(x)\}$ and the function $f(x)$ are defined on the interval $[a, b]$ and satisfy the following condition: To each $\epsilon > 0$ corresponds some integer N such that, for every x in the interval,*

$$|f_n(x) - f(x)| < \epsilon$$

provided that $N \leqq n$. Then we say that $f_n(x)$ converges uniformly *to $f(x)$ on the interval $[a, b]$. The essential thing is that N is to be independent of x; it will usually depend on ϵ, however.*

Example 1. If $f_n(x) = x^n/n$, then $\lim_{n\to\infty} f_n(x) = 0$, the convergence being uniform, on the interval $0 \leqq x \leqq 1$. For, if x is on the stated interval,

$$\left|\frac{x^n}{n}\right| \leqq \frac{1}{n} < \epsilon$$

if $1/\epsilon < n$. Hence we may take N as the smallest integer greater than $1/\epsilon$. This choice is independent of x.

Example 2. The convergence in Example 3 of §20 is *not* uniform on

$0 \leqq x \leqq 1$. For, since $f(x) = 0$ for each value of x involved, uniform convergence would mean that $|f_n(x)| < \epsilon$ when $n \geqq N$, where N depends only on ϵ, not on x. In particular, then, we should have $\left|f_n\left(\frac{1}{2n}\right)\right| < \epsilon$ if n is sufficiently large. But $\left|f_n\left(\frac{1}{2n}\right)\right| = 2n$, and certainly the inequality $2n < \epsilon$ is false if n is sufficiently large. This is a case where the choice of N depends on x, and as $x \to 0$, $N \to \infty$.

Uniform convergence may be portrayed graphically by interpreting the inequalities (20.1–3). Let the graph of $y = f(x)$ be displaced upward by the addition of ϵ. This gives us the graph of $y = f(x) + \epsilon$. Similarly, a downward displacement of amount ϵ gives us the graph of $y = f(x) - \epsilon$. Let us visualize these two displaced graphs, on the assumption that the function f is continuous. Considering the portion of the xy-plane between these two curves, for $a \leqq x \leqq b$, we have a ribbon-like region of width 2ϵ in the y-direction (see Fig. 178). The

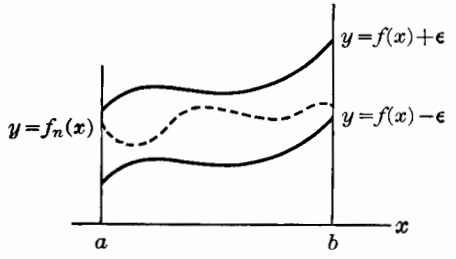


Fig. 178.

inequalities (20.1–3) state that the graph of $y = f_n(x)$ lies within the ribbon throughout its length. For each $\epsilon > 0$ there is a ribbon; to have uniform convergence the graph of $y = f_n(x)$ must lie entirely in the ribbon when n is large enough ($n \geqq N$), and this must be true for each ϵ.

The concept of uniform convergence relates to a whole set of values of the variable x; we have stated the definition for the case in which the set of values is a closed interval. The definition is essentially the same for any interval, whether open, closed, or neither. The crux of the matter is to have N the same for all values of x in the given interval or set.

Example 3. Let $f_n(x) = \dfrac{nx}{1 + n^2x^2}$. Here we have

$$\lim_{n \to \infty} f_n(x) = 0$$

for all values of x. The convergence is uniform in any closed interval which does not include $x = 0$, but it is not uniform in any interval having $x = 0$ in its interior or at one end. We shall investigate these statements about uniform convergence.

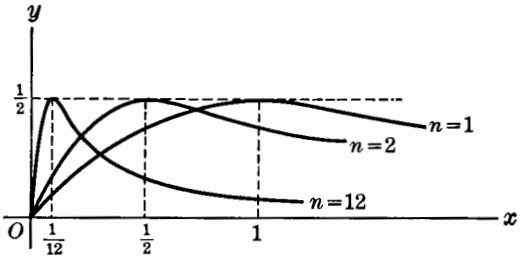


Fig. 179.

The function f_n is odd, i.e., $f_n(-x) = -f_n(x)$, so that the graph of $y = f_n(x)$ is symmetric with respect to the origin. Hence we confine our attention to values of x for which $x \geqq 0$. The graph is easily constructed with the aid of information obtained from the derivative. An easy calculation shows that

$$f_n'(x) = \frac{n(1 - n^2x^2)}{(1 + n^2x^2)}.$$

The graph of $y = f_n(x)$, for $x > 0$, rises to a maximum at $x = 1/n$, and then diminishes toward zero as $x \to +\infty$. The maximum value is $f_n(1/n) = \frac{1}{2}$. The graphs of $n = 1, 2, 12$ are shown in Fig. 179. It is clear from this figure that the convergence is nonuniform in any interval containing the point $x = 0$, for such an interval will contain the point $x = 1/n$ if n is large enough, and so we can never have $|f_n(x)| < \epsilon$ for *all* values of x in this interval if $\epsilon < \frac{1}{2}$, no matter how large we take n. However, the convergence is uniform for all values of x such that $x \geqq \delta$, where δ is any fixed positive number. For, $\epsilon > 0$ being given, if we choose N so large that

$$\frac{1}{N} < \delta \quad \text{and} \quad \frac{n\delta}{1 + n^2\delta^2} < \epsilon$$

if $N \leqq n$, then it will also be true that

$$\frac{nx}{1 + n^2x^2} < \epsilon$$

if $N \leqq n$ and $\delta \leqq x$, so that the convergence is uniform for these latter values of x. The situation is illustrated in Fig. 180.

When we are dealing with infinite series, we say that a series of functions, each defined on a certain fixed interval, is uniformly convergent on that interval provided that the sequence of partial sums $s_n(x)$ is uniformly convergent on the interval.

Sometimes it is convenient to express the condition for uniform convergence without reference to the limit function $f(x)$, but entirely in terms of the sequence $\{f_n(x)\}$.

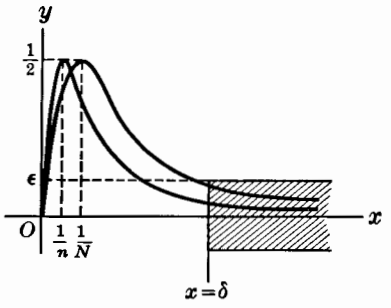


Fig. 180.

THEOREM I. *Let the functions of the sequence* $\{f_n(x)\}$ *be defined on the interval* $[a, b]$. *In order that the sequence be uniformly convergent to a limit function* $f(x)$ *on the interval, it is necessary and sufficient that to each* $\epsilon > 0$ *there correspond some integer* N *independent of* x *such that*

$$|f_n(x) - f_m(x)| < \epsilon \tag{20.1–6}$$

whenever $N \leqq m$ *and* $N \leqq n$.

The proof is a direct consequence of Cauchy's convergence criterion (Theorem VI, §16.5). We shall prove only the sufficiency of condition (20.1–6), and leave the proof of the necessity to the student. It follows from (20.1–6) by Cauchy's criterion that the sequence $\{f_n(x)\}$ is convergent for each x of the interval, and so defines a limit function $f(x)$. We may then let $m \to \infty$ in (20.1–6) and obtain the result that $|f_n(x) - f(x)| \leqq \epsilon$ if $N \leqq n$. Since N is independent of x, this shows that the convergence is uniform.

Later on this chapter we shall see that answers to questions A–D or A′–D′ may be stated in terms of the concept of uniform convergence.

EXERCISES

1. Discuss the functions $f(x) = \lim_{n\to\infty} f_n(x)$ defined by the following sequences for the values of x indicated. In each case sketch several typical graphs $y = f_n(x)$, showing the character of the approximation of $f(x)$ by $f_n(x)$ for rather large values of n. Indicate the discontinuities, if any, of $f(x)$, and answer the particular questions in each case.

(a) $f_n(x) = (\sin x)^n$, $0 \leqq x \leqq \pi$. Is the convergence uniform? Explain.

(b) $f_n(x) = (\sin x)^{1/n}$, $0 \leqq x \leqq \pi$. For what values of a and b is the convergence uniform on $a \leqq x \leqq b$?

(c) $f_n(x) = (1/n)e^{-n^2x^2}$, all x. Is the convergence uniform? Prove your answer.

(d) $f_n(x) = nxe^{-nx}$, $x \geqq 0$. Prove that the convergence is uniform for $x \geqq \delta$, where δ is any positive number. Why is the convergence not uniform on the interval $0 \leqq x \leqq \delta$?

(e) $f_n(x) = \tan^{-1} nx$, all x. What conditions must a and b satisfy if the convergence is to be uniform on $a \leqq x \leqq b$?

(f) $f_n(x) = (x/n)e^{-x/n}$, $x \geqq 0$. For given positive ϵ and A, find N such that $|f_n(x) - f(x)| < \epsilon$

if $0 \leqq x \leqq A$ and $N \leqq n$. What is the corresponding verbal statement about uniform convergence? Is the convergence uniform for *all* x such that $x \geqq 0$?

2. Follow the general directions of Exercise 1 in the case of $f_n(x) = x^n/(1+x^{2n})$, $x \geqq 0$. Get what help you can from examination of $f_n'(x)$. Show that $|f_n(x)| \leqq x^n$ if $x \geqq 0$, and $|f_n(x)| \leqq x^{-n}$ if $x > 0$. Use these inequalities to prove that the sequence converges uniformly if $0 \leqq x \leqq a < 1$, and also if $1 < b \leqq x$, where a and b are constants. Is the convergence uniform in the interval $\frac{1}{2} \leqq x \leqq \frac{3}{2}$?

3. Let $f_n(x) = x^{a_n}$, $a_n = \frac{1}{2} + \frac{1}{4} + \cdots + 1/2^n$.

(a) Find $f(x) = \lim_{n\to\infty} f_n(x)$.

(b) Show that $|f_n(x) - f(x)| \leqq 1 - a_n$ if $0 \leqq x \leqq 1$. What do you infer about uniform convergence? SUGGESTION: Consider where $f_n(x) - f(x)$ is a maximum on the interval $0 \leqq x \leqq 1$.

4. **(a)** Find a simple expression for the function $f(x) = \sum_{n=0}^{\infty} x^2/(1+x^2)^n$ when $x \neq 0$. What is $f(0)$? Does $f(x)$ have any discontinuities?

(b) If $s_n(x)$ is the sum of the first n terms of the series, show that $0 < f(x) - s_n(x) \leqq 1/(1+\delta^2)^{n-1}$ if $x^2 \geqq \delta^2 > 0$. What do you conclude about uniform convergence?

(c) Make a sketch showing the appearance of the curve $y = s_n(x)$ for a very large value of n. Is the series uniformly convergent when $|x| \leqq \delta$?

5. Consider the function

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

(a) Find the sum $s_n(x)$ of the first n terms of the series and hence find the value of $f(x)$ for each x. **(b)** Plot the curves $y = s_n(x)$ for several values of n, indicating how they look for large values of n. Is the convergence uniform when $0 \leqq x \leqq 1$? when $-1 \leqq x \leqq 0$? when $1 \leqq x \leqq 2$? **(c)** Show that $|s_n(x) - f(x)| \leqq 1/(n\delta - 1)$ if $|x| \geqq \delta > 0$ and $n\delta > 1$. What can you conclude about uniform convergence from this?

6. Give a proof of the necessity of condition (20.1–6) in Theorem I.

7. Suppose $f_0(x)$ is continuous when $0 \leqq x \leqq a$, and $f_n(x) = \int_0^x f_{n-1}(t)\,dt$. Prove that $f_n(x)$ converges uniformly to 0 when $0 \leqq x \leqq a$.

8. Prove that the series $\sum_{n=1}^{\infty} (-1)^{n+1}/(n+x^2)$ is uniformly convergent for all values of x. Also prove that it is not absolutely convergent for any value of x.

20.2 / A COMPARISON TEST FOR UNIFORM CONVERGENCE

We shall now give a theorem which is useful for proving that a series is uniformly convergent. While a series may be uniformly convergent without meeting the requirements of the theorem, it is nevertheless true that in a large number of important cases the theorem provides the simplest practical method for establishing the fact of uniform convergence.

THEOREM II. *Let the terms of the series*

$$u_1(x) + u_2(x) + \cdots + u_n(x) + \cdots \tag{20.2–1}$$

be defined on an interval. Let

$$M_1 + M_2 + \cdots + M_n + \cdots \tag{20.2–2}$$

be a convergent series of positive constants, and let the inequality $|u_n(x)| \leqq M_n$ *be satisfied for all values of* n *and for all* x *on the interval. Then the series* (20.2–1) *is uniformly convergent on the interval. The word "interval" may be replaced by "point set" and the theorem is still true.*

Example 1. The geometric series

$$1 + x + x^2 + x^3 + \cdots + x^n + \cdots \tag{20.2–3}$$

is uniformly convergent on the interval $-r \leqq x \leqq r$ if $0 < r < 1$. For, let

$$u_n(x) = x^n, \quad M_n = r^n.$$

Then, on the stated interval, $|x| \leqq r$ and so $|u_n(x)| \leqq M_n$. Since the series $\Sigma\, r^n$ is convergent, the uniform convergence of (20.2–3) follows by virtue of Theorem II.

We may note that although the series (20.2–3) converges at each point x of the *open* interval $-1 < x < 1$, it does not converge uniformly on that interval. For the partial sums of (20.2–3) may be written

$$1 + x + x^2 + \cdots + x^{n-1} = \frac{1 - x^n}{1 - x},$$

and the sequence $(1 - x^n)/(1 - x)$ is not uniformly convergent in the interval $0 \leqq x < 1$, as may be seen by referring to the discussion in the second paragraph of §20.1.

Proof of theorem II. Let

$$s_n(x) = u_1(x) + \cdots + u_n(x).$$

Then, if $m < n$, $s_n(x) - s_m(x) = u_{m+1}(x) + \cdots + u_n(x)$, and

$$|s_n(x) - s_m(x)| \leqq M_{m+1} + \cdots + M_n,$$

since $|u_k(x)| \leqq M_k$. Now, if $\epsilon > 0$, there is some integer N such that

$$M_{m+1} + \cdots + M_n < \epsilon$$

if $N \leqq m < n$ (see Theorem VIII, §19.3). Since the M's are constants, the choice of N is completely independent of x. But then we have

$$|s_n(x) - s_m(x)| < \epsilon$$

if $N \leqq m < n$; by Theorem I this establishes the uniform convergence of the series (20.2–1).

Theorem II is often called the M-test, or the Weierstrass M-test.

Example 2. The series

$$\frac{\sin x}{1^2} + \frac{\sin 2x}{2^2} + \frac{\sin 3x}{3^2} + \cdots + \frac{\sin nx}{n^2} + \cdots$$

is uniformly convergent on the entire x-axis. Here the comparison series may be

taken to be

$$\frac{1}{1^2}+\frac{1}{2^2}+\cdots+\frac{1}{n^2}+\cdots,$$

since $|\sin nx| \leqq 1$ for all n and all x. The series of constants is known to be convergent.

EXERCISES

1. Prove what you can about the uniform convergence of each of the following series by the method of Theorem II. State specifically the range of values of x about which you make an assertion.

(a) $\sum_{n=1}^{\infty} \frac{n}{n+1} x^n$.

(b) $\frac{\cos x}{1\cdot 2}+\frac{\cos 3x}{3\cdot 4}+\frac{\cos 5x}{5\cdot 6}+\cdots$.

(c) $\sum_{n=1}^{\infty} \frac{1}{n^2+x^2}$.

(d) $\sum_{n=0}^{\infty} e^{-nx}$.

(e) $\sum_{n=1}^{\infty} \left(\frac{1}{n}-\frac{1}{x^2+n}\right)$.

2. Prove that the series $\sum_{n=1}^{\infty} \frac{2x}{n^2-x^2}$ is uniformly convergent on any finite closed interval not containing any of the points $\pm 1, \pm 2, \ldots$.

3. Show that the series $\Sigma_{n=1}^{\infty} 1/n^x$ converges uniformly for values of x such that $x \geqq p$, if $p > 1$.

4. Suppose that $a_n \geqq 0$ and that the series $\Sigma_{n=1}^{\infty} a_n n^{-x}$ is convergent when $x = x_0$. Show that it converges for values of x such that $x \geqq x_0$.

5. Consider the series $\sum_{n=1}^{\infty} \frac{x^n}{n^2(1+x^n)}$.

(a) Show that it converges uniformly when $x \geqq 0$.

(b) Show that it converges uniformly if $|x| \leqq a$, where $0 < a < 1$.

(c) Show that it converges uniformly if $b \leqq x$, where $-1 < b$.

(d) Show that it converges uniformly if $x \leqq -c$, where $c > 1$.

20.3 / CONTINUITY OF THE LIMIT FUNCTION

We now return to the questions A and A′ of §20.

THEOREM III. *Let the functions $\{f_n(x)\}$ be defined on the interval $a \leqq x \leqq b$, and let them converge uniformly on this interval to a limit function $f(x)$. Then, if each of the functions f_n is continuous at a point x_0, the limit function f is also continuous at x_0. In particular, if each f_n is continuous on the whole interval, so is f.*

Proof. We write

$$f(x)-f(x_0) = (f(x)-f_n(x))+(f_n(x)-f_n(x_0))+(f_n(x_0)-f(x_0)),$$

$$|f(x)-f(x_0)| \leqq |f(x)-f_n(x)|+|f_n(x)-f_n(x_0)|+|f_n(x_0)-f(x_0)|. \qquad (20.3\text{–}1)$$

Now let $\epsilon > 0$ be given; x_0 is fixed, and we are to show that $|f(x) - f(x_0)| < \epsilon$ provided x is sufficiently close to x_0, for this is what is meant by saying that f is continuous at x_0. Now, by the uniform convergence, we can choose n, independent of x, so large that

$$|f_n(x) - f(x)| < \frac{\epsilon}{3}$$

for every x in the interval. If this is done, (20.3–1) shows that

$$|f(x) - f(x_0)| < |f_n(x) - f_n(x_0)| + \tfrac{2}{3}\epsilon. \tag{20.3–2}$$

Now n has been fixed; since f_n is continuous at x_0, we shall have $|f_n(x) - f_n(x_0)| < \epsilon/3$ if x is sufficiently close to x_0. But then by (20.3–2) we shall have $|f(x) - f(x_0)| < \epsilon$. This completes the proof.

As a corollary of Theorem III we have the following assertion about series: *If the terms of an infinite series are continuous on an interval $a \leqq x \leqq b$, and if the series is uniformly convergent on the interval, the function defined as the sum of the series is continuous on the interval.*

EXERCISES

1. (**a**) Find $f(x) = \lim_{n\to\infty} x^{1/(2n-1)}$ for each value of x. From the nature of $f(x)$ what do you conclude about uniform convergence of the sequence in intervals containing the origin?

(**b**) Proceed as in (**a**) with $f(x) = \lim_{n\to\infty} x^{2/(2n-1)}$.

2. From an examination of the function $f(x) = \lim_{n\to\infty} 1/(1 + x^{2n})$ and use of Theorem III, state what you conclude about lack of uniform convergence of the sequence. Prove what you can of an affirmative nature about uniform convergence of the sequence.

3. If $f(x) = 1 + \sum_{n=1}^{\infty} x^n/n$, find $\lim_{x\to 0} f(x)$ and justify your answer.

4. Is the function $f(x) = \sum_{n=0}^{\infty} xe^{-nx^2}$ continuous at $x = 0$? Explain your answer.

5. Explain why the function

$$f(x) = \sum_{n=1}^{\infty} \frac{1}{2x}\left(\frac{1}{n-x} - \frac{1}{n+x}\right)$$

is continuous except at the points $x = 0, \pm 1, \pm 2, \ldots$. What is $\lim_{x\to 0} f(x)$?

20.4 / INTEGRATION OF SEQUENCES AND SERIES

The questions C and C′ of §20 may be answered, in one way at least, by the following theorem.

THEOREM IV. *Let the functions $f_n(x)$ be continuous on the closed interval $a \leqq x \leqq b$, and let them converge uniformly on this interval to the limit function $f(x)$. Then*

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \int_a^b f_n(x)\,dx. \tag{20.4–1}$$

For infinite series this is equivalent to the statement that if the series

$$f(x) = \sum_{n=1}^{\infty} u_n(x)$$

is uniformly convergent on the interval $a \leqq x \leqq b$, *and if the terms* $u_n(x)$ *are continuous on the interval, then*

$$\int_a^b f(x)\,dx = \sum_{n=1}^{\infty} \int_a^b u_n(x)\,dx. \tag{20.4–2}$$

Proof. The proof is quite easy. Consider (20.4–1).

$$\int_a^b f_n(x)\,dx - \int_a^b f(x)\,dx = \int_a^b [f_n(x) - f(x)]\,dx,$$

$$\left|\int_a^b f_n(x)\,dx - \int_a^b f(x\,dx\right| \leqq \int_a^b \left|f_n(x) - f(x)\right| dx. \tag{20.4–3}$$

Now suppose $\epsilon > 0$ is given. Choose N, independent of x, so large that if $N \leqq n$ we have

$$|f_n(x) - f(x)| < \frac{\epsilon}{b-a} \tag{20.4–4}$$

if $a \leqq x \leqq b$. This is possible because of the assumed uniform convergence. Now from (20.4–3) and (20.4–4) we see that

$$\left|\int_a^b f_n(x)\,dx - \int_a^b f(x)\,dx\right| < \epsilon \tag{20.4–5}$$

if $N \leqq n$, for

$$\int_a^b |f_n(x) - f(x)|\,dx < \int_a^b \frac{\epsilon}{b-a}\,dx = \epsilon.$$

But (20.4–5) is just the condition which means that, by definition (20.4–1) is true.

For the case of the infinite series, (20.4–1) becomes, with $f_n(x)$ as the partial sums,

$$\begin{aligned}\int_a^b f(x)\,dx &= \lim_{n\to\infty} \int_a^b [u_1(x) + \cdots + u_n(x)]\,dx \\ &= \lim_{n\to\infty}\left[\int_a^b u_1(x)\,dx + \cdots + \int_a^b u_n(x)\,dx\right].\end{aligned}$$

This last relation, in turn, is just another way of expressing (20.4–2).

Example. We saw in Example 1, §20.2, that the geometric series

$$\frac{1}{1-x} = 1 + x + \cdots + x^n + \cdots$$

is uniformly convergent if $-r \leqq x \leqq r$, where $0 < r < 1$. Let us then apply

(20.4–2), integrating from $-r$ to r. The result is

$$\int_{-r}^{r} \frac{dx}{1-x} = \int_{-r}^{r} dx + \int_{-r}^{r} x\,dx + \cdots + \int_{-r}^{r} x^n\,dx + \cdots. \tag{20.4–6}$$

Now

$$\int_{-r}^{r} \frac{dx}{1-x} = -\log(1-x)\Big|_{-r}^{r} = \log\frac{1+r}{1-r},$$

$$\int_{-r}^{r} x^n\,dx = \frac{r^{n+1}}{n+1} - \frac{(-r)^{n+1}}{n+1}.$$

The latter expression is equal to 0 if $n+1$ is even, and equal to $2r^{n+1}/(n+1)$ if $n+1$ is odd. Therefore (20.4–6) becomes

$$\log\frac{1+r}{1-r} = 2\left(r + \frac{r^3}{3} + \frac{r^5}{5} + \cdots\right).$$

This result was obtained by a different method in Example 1, §19.6.

As was suggested by this example, Theorem IV has important applications in deriving certain series expansions from other series expansions by integration.

The conclusions (20.4–1) and (20.4–2) of Theorem IV *may* be false if the convergence is not uniform. This is illustrated by Example 3, §20, in which the convergence is not uniform. To have (20.4–1) and (20.4–2) it is *sufficient* to have uniform convergence; but uniform convergence is not a *necessary* condition in all cases. Suppose, for example, that we modify Example 3 of §20 by taking the height of the triangle in the graph of $y = f_n(x)$ to be $2\sqrt{n}$ instead of $2n$. The convergence is still nonuniform, and $\lim_{n\to\infty} f_n(x) = 0$ when $0 \leqq x \leqq 1$. But now

$$\int_0^1 f_n(x)\,dx = \frac{1}{\sqrt{n}}$$

and so (20.4–1) is true in this case.

EXERCISES

1. If $f(x) = \sum_{n=1}^{\infty} \frac{\cos nx}{n^2}$, show that $\int_0^{\pi/2} f(x)\,dx = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)^3}$. Justify your reasoning.

2. If $f(x) = \frac{\sin 3x}{1\cdot 2} + \frac{\sin 5x}{3\cdot 4} + \frac{\sin 7x}{5\cdot 6} + \cdots$, find a series for $\int_0^{\pi/2} f(x)\,dx$.

3. If $f_n(x) = nxe^{-nx^2}$ and $f(x) = \lim_{n\to\infty} f_n(x)$, show that the sequence converges nonuniformly on the interval $0 \leqq x \leqq 1$, and that

$$\int_0^1 f(x)\,dx \neq \lim_{n\to\infty} \int_0^1 f_n(x)\,dx.$$

4. If $f(x) = \lim_{n\to\infty} f_n(x)$, where $f_n(x) = \frac{2nx}{1+n^5x^4}$, find

$$\int_0^1 f(x)\,dx \quad \text{and} \quad \lim_{n\to\infty} \int_0^1 f_n(x)\,dx.$$

What do you conclude about uniform convergence of the sequence on the interval $0 \leqq x \leqq 1$?

5. From the series expansions of e^t, $\sin t$, and $\cos t$ in powers of t (see §19.1), verify that $\int_a^b e^t\,dt = e^b - e^a$ and that $\int_0^x \sin t\,dt = 1 - \cos x$, $\int_0^x \cos t\,dt = \sin x$. Prove whatever you need about uniform convergence for applying Theorem IV.

20.5 / DIFFERENTIATION OF SEQUENCES AND SERIES

We now come back to questions B, D, B′, and D′ of §20. We shall answer D and D′, and in so doing shall answer B and B′. That is, we shall give conditions which will not only assure us that the limit function is differentiable, but will tell us how to express the derivative itself as a limit function. Here again uniform convergence is the key to the situation. But now we must look for uniform convergence of the sequence $\{f_n'(x)\}$.

THEOREM V. *Let the functions $f_n(x)$ be defined and have continuous derivatives on the interval $a \leqq x \leqq b$. Let the sequence $\{f_n'(x)\}$ be uniformly convergent on the interval, and let the sequence $\{f_n(x)\}$ itself be convergent, with limit function $f(x)$. Then f is differentiable, and*

$$f'(x) = \lim_{n\to\infty} f_n'(x). \tag{20.5–1}$$

For infinite series this is equivalent to the statement that if

$$f(x) = \sum_{n=1}^{\infty} u_n(x)$$

is convergent and if the series of derivatives

$$\sum_{n=1}^{\infty} u_n'(x)$$

is uniformly convergent, then

$$f'(x) = \sum_{n=1}^{\infty} u_n'(x). \tag{20.5–2}$$

Here it is assumed that each term $u_n(x)$ has a continuous derivative.

Proof. Let us denote the limit of the sequence $\{f_n'(x)\}$ by $g(x)$; we do not yet know that $g(x) = f'(x)$. Since $f_n'(x)$ converges uniformly to $g(x)$, Theorem IV permits us to write

$$\int_a^x g(t)\,dt = \lim_{n\to\infty} \int_a^x f_n'(t)\,dt, \tag{20.5–3}$$

for the convergence is also uniform on any subinterval $a \leqq t \leqq x$, where x is any point of $[a, b]$. Now

$$\int_a^x f_n'(t)\,dt = f_n(x) - f_n(a).$$

Since $f_n(x) \to f(x)$ as $n \to \infty$, (20.5–3) becomes

$$\int_a^x g(t)\,dt = f(x) - f(a),$$

or

$$f(x) = \int_a^x g(t)\,dt + f(a). \tag{20.5–4}$$

Now g is continuous, by Theorem III. We conclude from (20.5–4) and Theorem VII, §1.52, that f is differentiable, with $f'(x) = g(x)$. This result is just another way of stating (20.5–1). The transition from (20.5–1) to (20.5–2) is accomplished in the same way as was done for (20.4–1) and (20.4–2) at the end of the proof of Theorem IV.

Example. Use Theorem V to aid in obtaining the series expansions, valid if $|x| < 1$,

$$\frac{1}{(1-x)^2} = 1 + 2x + 3x^2 + 4x^3 + \cdots + (n+1)x^n + \cdots, \tag{20.5–5}$$

$$\frac{2}{(1-x)^3} = 2 + 3 \cdot 2x + 4 \cdot 3x^2 + \cdots + (n+2)(n+1)x^n + \cdots. \tag{20.5–6}$$

We start from the geometric series

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots + x^n + \cdots, \tag{20.5–7}$$

which we know to be convergent if $|x| < 1$. The series (20.5–5) results from (20.5–7) according to the rule of (20.5–2), and (20.5–6) is derived from (20.5–5) in a similar manner. In order to justify this procedure by Theorem V we must prove that each of the series (20.5–5), (20.5–6) is uniformly convergent if $-r \leqq x \leqq r$ for any r such that $0 < r < 1$. If x is any fixed point such that $-1 < x < 1$, we can choose r so that $|x| < r < 1$. Then x will be in an interval of uniform convergence, and the foregoing formulas will be justified.

Suppose then that

$$0 < r < 1.$$

If $-r \leqq x \leqq r$ we have

$$|(n+1)x^n| \leqq (n+1)r^n.$$

Now consider the series

$$1 + 2r + 3r^2 + \cdots + (n+1)r^n + \cdots. \tag{20.5–8}$$

This is a series of positive constants. It is convergent, for the ratio of successive terms is

$$\frac{(n+2)r^{n+1}}{(n+1)r^n} = \frac{n+2}{n+1}\,r,$$

and this ratio converges to r as $n \to \infty$. Since $r < 1$, the series (20.5–8) is convergent, by Theorem VII, §19.22. The uniform convergence of the series (20.5–5) is now a consequence of the M-test (Theorem II, §20.2); the M-series in this case is (20.5–8).

The proof that the series (20.5–6) converges uniformly if $|x| \leqq r$, $r < 1$, is entirely similar, and the details are left to the student.

EXERCISES

1. Which of the following functions $f(x)$ has the property that $f'(x)$ can be calculated for each x on the specified interval by differentiating the series for $f(x)$ term by term?

(a) $f(x) = \sum_{n=1}^{\infty} \frac{\sin nx}{n^2}, \quad 0 \leqq x \leqq 2\pi.$

(b) $f(x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}, \quad 0 \leqq x \leqq 1.$

(c) $f(x) = \sum_{n=2}^{\infty} \frac{x^n}{n^2 (\log n)^2}, \quad -1 \leqq x \leqq 1.$

(d) $f(x) = \sum_{n=0}^{\infty} \frac{x^{2n}}{2n+1}, \quad -1 < x < 1.$

(e) $f(x) = \sum_{n=1}^{\infty} \left(\frac{1}{x - n\pi} + \frac{1}{n\pi} \right), \quad 0 < x < \pi.$

(f) $f(x) = 1 - \frac{x^2}{2^2} + \frac{x^4}{2^2 \cdot 4^2} - \cdots + (-1)^n \frac{x^{2n}}{2^2 \cdot 4^2 \cdots (2n)^2} + \cdots, \quad$ all x.

2. Prove that the series (20.5–6) converges uniformly if $|x| \leqq r$, where $0 < r < 1$.

3. Let $f_n(x) = (2/\pi)x \tan^{-1} nx$, $f(x) = \lim_{n\to\infty} f_n(x)$. Show that $f(x) = |x|$, so that it is clear that f is not differentiable at $x = 0$. Show that $\lim_{n\to\infty} f_n'(x)$ exists for each x, including $x = 0$. What do you conclude about the uniformity of convergence of the sequence of derivatives?

4. Let $f(x) = \lim_{n\to\infty} f_n(x)$, $f_n(x) = (1/n)^{-n^2x^2}$. Show that $\lim_{n\to\infty} f_n'(x) = f'(x)$ for every x, but that the convergence of the sequence of derivatives is not uniform in any interval containing the origin. The original sequence is convergent uniformly on the entire axis.

21 / POWER SERIES

21 / GENERAL REMARKS

The purpose of this chapter is to give a systematic exposition of some of the most important things about power series.

The representation of functions by power series is one of the most useful of mathematical techniques in a wide variety of situations. Sometimes we start from a function that is defined for us in some manner not employing series, and seek to expand the function in a power series. At other times we may form a power series, or have one presented to us, and then undertake to use this function in some way. In either of these situations we need to know something of what properties a function has if it is defined by a power series.

The general form of a power series is

$$a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots + a_n(x - x_0)^n + \cdots; \tag{21–1}$$

this is called a power series in $x - x_0$. Here x_0 is fixed and x is variable. In the special case where $x_0 = 0$ the series takes the form

$$a_0 + a_1x + a_2x^2 + \cdots + a_nx^n + \cdots. \tag{21–2}$$

It turns out that in studying power series it is sufficient to consider (21–2), since the general case (21–1) can be reduced to (21–2) by a translation of the origin along the x-axis. For this reason all the general theory of power series will be developed for series of powers of x, of the form (21–2).

For convenience we shall often refer to the series (21–2) as the series $\Sigma\, a_nx^n$, omitting the limits $n = 0$ and ∞ for greater ease in printing.

21.1 / THE INTERVAL OF CONVERGENCE

The first important fact about a power series is expressed in the following theorem:

THEOREM I. *Suppose a power series* $\Sigma\, a_nx^n$ *is convergent for* $x = x_0$, *where* $x_0 \neq 0$. *Then it is absolutely convergent when* $|x| < |x_0|$. *The same conclusion holds under the weaker assumption that there is some positive constant* A *such that* $|a_nx_0^n| \leqq A$ *for all values of* n.

Proof. If $\Sigma\, a_nx_0^n$ is convergent, $a_nx_0^n \to 0$, and therefore the terms $a_nx_0^n$ are bounded. Under the hypothesis in the theorem we have

$$a_nx^n = a_nx_0^n\left(\frac{x}{x_0}\right)^n, \quad |a_nx^n| \leqq A\left(\frac{|x|}{|x_0|}\right)^n.$$

The geometrical series

$$\sum A\left(\frac{|x|}{|x_0|}\right)^n$$

is convergent if $|x| < |x_0|$; therefore the series $\Sigma |a_n x^n|$ is convergent, by the comparison test (Theorem II, §19.2). This proves the theorem.

Conceivably a particular power series may be convergent for all values of x. This is the case with the series for e^x, for instance (see (19.1–6)). It is also conceivable that a series may not be convergent for any values of x except the one value $x = 0$. Such is the case with the series

$$1!x + 2!x^2 + \cdots + n!x^n + \cdots,$$

for, with $u_n = n!x^n$, if $|x| \neq 0$,

$$\left|\frac{u_{n+1}}{u_n}\right| = (n+1)|x| \to +\infty$$

as $n \to \infty$, and so the series is divergent, by Theorem XIII, §19.4.

Let us now consider the case in which the series $\Sigma a_n x^n$ is convergent for at least one nonzero value of x, but is also divergent for some value of x. Let us denote by R the least upper bound of the positive numbers x for which $\Sigma a_n x^n$ is convergent. Denote this set of positive numbers x by S. The set S is not empty, for by Theorem I, if the series is convergent for the nonzero value x_0, it is convergent for all positive x such that $x < |x_0|$. The set S must be bounded, because of our assumption that the series is not convergent for all values of x. For if S were not bounded, then for any x we could find an x_0 in S such that $|x| < x_0$, and this would imply that $\Sigma a_n x^n$ is convergent. Since S is not empty and has an upper bound, it has a least upper bound (Theorem II, §2.7). This justifies the introduction of the number R.

We now assert the following: The series $\Sigma a_n x^n$ is absolutely convergent if $|x| < R$ and is divergent if $|x| > R$. The proof is a simple consequence of Theorem I. Suppose $|x| < R$. Then choose x_0 so that x_0 is in the class S and $|x| < x_0 \leqq R$; this is possible, since R is the least upper bound of S. But then the series $\Sigma a_n x^n$ converges absolutely, by Theorem I. On the other hand, suppose that $|x| > R$. Then the series $\Sigma a_n x^n$ cannot converge, for if it did, Theorem I would assure us that $\Sigma a_n y^n$ converges if $R < y < |x|$, so that y would be in S, contrary to the fact that R is the least upper bound of S.

We sum up the foregoing conclusions in theorem form.

THEOREM II. *For a power series $\Sigma a_n x^n$ there are three possibilities:*

1. *It is absolutely convergent for all values of x.*
2. *It diverges for every $x \neq 0$.*

3. *There is a positive number R such that the series converges absolutely if $|x| < R$ and diverges if $|x| > R$.*

In case (3) we call R the *radius of convergence* of the series. The interval $-R < x < R$ is called the *interval of convergence.* The series may or may not converge at the end points $x = \pm R$. It may converge at both, at just one, or at neither. These possibilities were illustrated in Chapter 19.

In case (1) we say that the radius of convergence is infinite, and that the interval of convergence is the entire x-axis. It is a convenient symbolism to write $R = \infty$ in this case. In case (2) we write $R = 0$; here there is no interval of convergence.

THEOREM III. *Suppose that the series $\Sigma\, a_n x^n$ has a positive or infinite radius of convergence R. Let $0 < r < R$ (if $R = \infty$, r may be any positive number). Then the series converges uniformly on the closed interval $-r \leqq x \leqq r$.*

Proof. The series $\Sigma\, |a_n|\, r^n$ is convergent, since the power series converges absolutely at $x = r$, by Theorem II. If $-r \leqq x \leqq r$ we have $|a_n x^n| \leqq |a_n|\, r^n$. The uniform convergence is therefore a consequence of the Weierstrass M-test (Theorem II, §20.2), with $M_n = |a_n|\, r^n$.

In the cases most commonly arising in practice, the radius of convergence of a power series may be found by using d'Alembert's ratio test (Theorem XIII, §19.4), as illustrated by Examples 1 and 2 in §19.4. If the limit

$$\lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| = L \tag{21.1–1}$$

exists, and if we write $u_n = a_n x^n$, then

$$\lim_{n\to\infty} \left| \frac{u_{n+1}}{u_n} \right| = \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| |x| = L|x|;$$

the series converges if $L|x| < 1$, and diverges if $L|x| > 1$. From this we conclude that $R = 1/L$ if $L \neq 0$. Also, $R = \infty$ if $L = 0$, and $R = 0$ if $L = \infty$. As a formal statement we have

THEOREM IV. *The radius of convergence of the series $\Sigma\, a_n x^n$ is given by*

$$R = \lim_{n\to\infty} \left| \frac{a_n}{a_{n+1}} \right|, \tag{21.1–2}$$

provided that the limit exists or is $+\infty$.

There are power series for which the limit (21.1–2) does not exist; there is then a means of determining R by examining the sequence $|a_n|^{1/n}$. This problem is considered in §21.5.

Theorem III has important consequences, the chief immediate ones of which have to do with the function defined by a power series.

THEOREM V. *Let a function f be defined by*

$$f(x) = \sum_{n=0}^{\infty} a_n x^n, \tag{21.1–3}$$

where it is assumed that the radius of convergence of the power series is not zero. The function f so defined is continuous in the open interval of convergence of the series. Moreover, if a and b are points of this interval,

$$\int_a^b f(x)\,dx = \sum_{n=0}^{\infty} a_n \frac{b^{n+1} - a^{n+1}}{n+1}; \tag{21.1–4}$$

that is, the integral of the function is equal to the series obtained by integrating the original power series term by term.

Proof. The continuity of f is a direct consequence of Theorem III, §20.3, for if x_0 is any point of the interval of convergence, we may choose $r > 0$ so that $|x_0| < r < R$, and the series converges uniformly on the interval $-r \leqq x \leqq r$, which includes the point x_0. The assertion (21.1–4) is a direct application of (20.4–2) in Theorem IV, §20.4, for the power series converges uniformly on the closed interval $[a, b]$, each point of which is interior to the interval of convergence.

Theorem V leaves something to be desired in one respect. It does not give us any information about what happens at an end of the interval of convergence. Suppose, for example, that the series (21.1–3) happens to be convergent at $x = R$, where R is the radius of convergence. Will the function f be continuous at $x = R$? That is, will

$$f(x) \to \sum_{n=0}^{\infty} a_n R^n$$

as $x \to R$ from the left? The answer to this question is in the affirmative, but the proof is not covered in Theorem V. We shall take up this matter in §21.4. A similar question arises with regard to (21.1–4). Is it ever possible to put $b = R$ or $a = -R$? This also is considered in §21.4.

There are many important practical applications of Theorem V, particularly of (21.1–4).

Example 1. Derive the series expansion

$$\sin^{-1} x = x + \frac{1}{2}\frac{x^3}{3} + \frac{1 \cdot 3}{2 \cdot 4}\frac{x^5}{5} + \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6}\frac{x^7}{7} + \cdots. \tag{21.1–5}$$

We start from the fact that

$$\sin^{-1} x = \int_0^x \frac{dt}{\sqrt{1 - t^2}}. \tag{21.1–6}$$

Formula (21.1–6) is valid if $|x| \leqq 1$; the integral is improper if $x = \pm 1$, since the integrand becomes infinite at $t = \pm 1$. In the binomial series (19.5–5) we replace x by $-t^2$ and set $m = -\frac{1}{2}$. The result is

$$(1-t^2)^{-1/2} = 1 - \tfrac{1}{2}(-t^2) + \frac{(-\frac{1}{2})(-\frac{3}{2})}{2!}(-t^2)^2 + \frac{(-\frac{1}{2})(-\frac{3}{2})(-\frac{5}{2})}{3!}(-t^2)^3 + \cdots$$

$$= 1 + \tfrac{1}{2}t^2 + \frac{1\cdot 3}{2\cdot 4}t^4 + \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}t^6 + \cdots;$$

this result is valid if $|t| < 1$, and the series has radius of convergence $R = 1$. Hence we may integrate the series from $t = 0$ to $t = x$ if $|x| < 1$. Using (21.1–6), we obtain the result (21.1–5). This argument, based on Theorem V, shows that (21.1–5) is valid if $|x| < 1$. As a matter of fact, one can easily verify that the series converges even when $x = \pm 1$, by using Raabe's test (Theorem XIV, §19.4); it is then a consequence of the discussion in §21.4 that (21.1–5) is valid if $-1 \leqq x \leqq 1$.

Example 2. Find an expansion in powers of x of the function

$$f(x) = \int_0^1 \frac{1 - e^{-tx}}{t}\, dt,$$

and use the result to calculate $f(\frac{1}{2})$ approximately.

From the series (19.1–6) for the exponential function we have

$$e^{-tx} = 1 - tx + \frac{t^2x^2}{2!} - \frac{t^3x^3}{3!} + \cdots.$$

Therefore

$$\frac{1 - e^{-tx}}{t} = x - \frac{tx^2}{2!} + \frac{t^2x^3}{3!} - \cdots + (-1)^{n-1}\frac{t^{n-1}x^n}{n!} + \cdots.$$

This series representation is valid for all values of x and t; the radius of convergence of the power series in t is $R = \infty$. Integrating, we have

$$f(x) = x - \frac{x^2}{2\cdot 2!} + \frac{x^3}{3\cdot 3!} - \cdots + (-1)^{n-1}\frac{x^n}{n\cdot n!} + \cdots.$$

In particular,

$$f\left(\frac{1}{2}\right) = \int_0^1 \frac{1 - e^{-t/2}}{t}\, dt = \frac{1}{2} - \frac{1}{2\cdot 2!}\left(\frac{1}{2}\right)^2 + \frac{1}{3\cdot 3!}\left(\frac{1}{2}\right)^3 - \cdots = 1.13 \text{ (approximately)}.$$

EXERCISES

1. Find the radius of convergence of each of the following series.

(a) $\sum \frac{(2n)!}{(n!)^2}x^n$. (c) $\sum \frac{(3n)!}{(n!)^2}x^n$.

(b) $\sum \frac{(n!)^3}{(3n)!}x^n$. (d) $\sum \frac{(n!)^{7/2}}{(4n)!}x^n$.

(e) $\sum \frac{(n+p)!}{n!(n+q)!}x^n$, p and q positive integers.

(f) $\sum \frac{[1\cdot 3\cdot\cdots\cdot(2n-1)]^2}{2^{2n}(2n)!}x^n$.

2. Find the radius of convergence of the series $\sum \frac{(pn)!}{(n!)^q}x^n$, where p is a positive integer and $q>0$. Consider all possibilities.

3. Find the radius of convergence of the following series:

$$\sum_{n=1}^{\infty} \frac{[q(q-2)\cdot\cdots\cdot(q-2n+2)][(q+1)(q+3)\cdot\cdots\cdot(q+2n-1)]}{(2n)!}x^n,$$

where q is not a positive even integer.

4. Let the radii of convergence of $\Sigma a_n x^n$ and $\Sigma b_n x^n$ be R_1 and R_2, respectively. Suppose $|a_n| \leqq |b_n|$ for all sufficiently large values of n. Prove that $R_1 \geqq R_2$.

5. Let $\{a_n\}$ be a bounded sequence, and let R be the radius of convergence of the series $\Sigma a_n x^n$. Prove that $R \geqq 1$.

6. If the radius of convergence of $\Sigma a_n x^n$ is R, prove that the radius of convergence of $\Sigma a_n x^{2n}$ is $R^{1/2}$.

7. Find an expansion as a power series in x for each of the following functions. Indicate the radius of convergence in each case.

(a) $\int_0^x \frac{\sin t}{t}\,dt$. (c) $\int_0^x e^{-t^2}\,dt$.

(b) $\int_0^x \cos u^2\,du$. (d) $\int_0^x \frac{t^p}{\sqrt{1-t^2}}\,dt$, p a positive integer.

8. Deduce the expansion

$$\sinh^{-1} x = x + \sum_{n=1}^{\infty} (-1)^n \frac{1\cdot 3\cdot\cdots\cdot(2n-1)}{2\cdot 4\cdot\cdots\cdot 2n}\frac{x^{2n+1}}{2n+1}.$$

State the justification of your procedure, and indicate for what values of x you are proving the validity of the expansion.

9. If $f(x)=\int_0^x \frac{\log(1+u)}{u}\,du$, find a series for $f(x)$, and calculate the approximate value of $f(\frac{1}{10})$.

10. Find the approximate value of $\int_0^1 \frac{1-\cos x}{x^2}\,dx$.

11. Find a series for the function $f(x)=\int_0^x \frac{\log(1+t)}{1+t}\,dt$, and calculate the approximate value of $f(\frac{1}{10})$. See Example 2, §19.6.

12. Find the approximate value of $\int_0^{1/2} \frac{e^{-x^2}}{\sqrt{1-x^2}}\,dx$.

13. Use series (21.1–6) to write numerical series for $\pi/2$ and $\pi/6$, respectively.

21.2 / DIFFERENTIATION OF POWER SERIES

The principal fact to be established in this section is the following: If a function is defined by a power series which has a positive or infinite radius of convergence, then the function has derivatives of all orders at each point of the

open interval of convergence, and these derivatives are represented by the series which are obtained by differentiation of the original series term by term. Thus, if

$$f(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + \cdots \tag{21.2–1}$$

is convergent when $|x| < R$, then

$$f'(x) = a_1 + 2a_2x + 3a_3x^2 + 4a_4x^3 + \cdots \tag{21.2–2}$$

$$f''(x) = 2a_2 + 3 \cdot 2a_3x + 4 \cdot 3a_4x^2 + \cdots \tag{21.2–3}$$

$$f'''(x) = 3!a_3 + 4 \cdot 3 \cdot 2a_4x + \cdots, \tag{21.2–4}$$

and so on, all these series likewise being convergent if $|x| < R$. To get at these facts we begin with a consideration of the series (21.2–1) and (21.2–2).

THEOREM VI. *In the case of any power series, the series in* (21.2–1) *and the series in* (21.2–2) *have the same radius of convergence.*

Proof. Let R and R' denote the radii of convergence of the series (21.2–1) and (21.2–2), respectively. Suppose $|x| < R$, and choose x_0 so that $|x| < |x_0| < R$. Then the series (21.2–1) is convergent with $x = x_0$, and consequently $a_nx_0^n \to 0$. We may therefore choose a number $A > 0$ such that $|a_nx_0^n| \leqq A$ for all n. Then

$$na_nx^{n-1} = \frac{n}{x_0}\, a_nx_0^n \left(\frac{x}{x_0}\right)^{n-1},$$

$$|na_nx^{n-1}| \leqq \frac{A}{|x_0|}\, nr^{n-1}, \tag{21.2–5}$$

where

$$r = \frac{|x|}{|x_0|} < 1.$$

The series

$$\sum \frac{A}{|x_0|}\, nr^{n-1}$$

is convergent, for the limit of the ratio of the term of index $n + 1$ to the term of index n is

$$\lim_{n\to\infty} \frac{n+1}{n}\, r = r < 1.$$

Consequently, by (21.2–5), the series

$$\sum na_nx^{n-1}$$

is convergent. This is precisely the series (21.2–2), and so we have proved that this series converges if $|x| < R$. It follows that the radius of convergence of (21.2–2) is *not less* than that of (21.2–1), i.e. $R' \geqq R$. If $R = \infty$ this means that $R' = \infty$ also.

To complete the proof we show that $R' > R$ is impossible, so that $R' = R$. For suppose that $R' > R$ and choose x so that $R < |x| < R'$. Then, for this x, the series (21.2–2) is absolutely convergent and the series (21.2–1) is divergent. Now

$$|a_n x^n| = |n a_n x^{n-1}| \left|\frac{x}{n}\right| < |n a_n x^{n-1}|$$

as soon as $n > |x|$. This comparison shows that the series (21.2–1) must be convergent, which is a contradiction. The proof of Theorem VI is now complete.

THEOREM VII. *Let $f(x)$ be defined by the power series* (21.2–1), *and assume that the radius of convergence is not zero. Then f is differentiable at each point inside the interval of convergence, and $f'(x)$ is represented by the series* (21.2–2). *Application of this conclusion to f' in place of f shows that $f''(x)$ is represented by* (21.2–3), *and so on.*

The proof is an immediate consequence of Theorem V, §20.5, and Theorem III, §21.1, in view of Theorem VI of the present section.

It is possible that the series for $f'(x)$ will diverge at an end of the interval of convergence, even though the series for $f(x)$ may be convergent at that point. An example is furnished by the series for $\log(1+x)$ at $x = 1$ [see (19.1–1)]; this is convergent, but the series for the derivative, $(1+x)^{-1}$, is divergent at $x = 1$.

Theorem VII shows that there is a great difference between functions defined by power series and functions defined by series of less special type. Another very important class of series is the class of trigonometric series, of which

$$\frac{\sin x}{1^2} + \frac{\sin 2x}{2^2} + \cdots + \frac{\sin nx}{n^2} + \cdots$$

is an example. We saw in Example 2, §20.1, that this series is uniformly convergent for all values of x. If we differentiate the series term by term we get

$$\frac{\cos x}{1} + \frac{\cos 2x}{2} + \frac{\cos 3x}{3} + \cdots + \frac{\cos nx}{n} + \cdots.$$

This series is convergent for some values of x, but not for all values; for instance, it is divergent when $x = 0$. Another termwise differentiation gives the series

$$-\sin x - \sin 2x - \cdots - \sin nx - \cdots,$$

which is convergent if $x = 0$, but is divergent except when x is an integral multiple of π.

Theorem VII permits us to show the relation between the general theory of power series and the expansion of a function in a power series by means of Taylor's series. In the Taylor's series expansion of a function in powers of x the coefficient of x^n is $\dfrac{f^{(n)}(0)}{n!}$ (see (19.1–4)).

THEOREM VIII. *If a function $f(x)$ is defined by a power series* (21.2–1) *with positive or infinite radius of convergence, the coefficients are related to the function by the formulas*

$$a_n = \frac{f^{(n)}(0)}{n!}. \tag{21.2–6}$$

This means that the power series is the Taylor's series of the function.

The formulas (21.2–6) are established by setting $x = 0$ in the successive series for $f(x)$, $f'(x)$, $f''(x)$, etc. (see (21.2–1)–(21.2–4)). It is clear by induction that the leading term in the series for $f^{(n)}(x)$ is $n!a_n$, and (21.2–6) is a direct consequence.

THEOREM IX. *Suppose that two power series are convergent and have the same sum for all values of x in some interval $|x| < r$:*

$$\sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{\infty} b_n x^n, \quad -r < x < r.$$

Then $a_n = b_n$ for all n.

This theorem is called the *uniqueness theorem* for power series. It is a corollary of Theorem VIII. For, let $f(x)$ be the common sum of the two series. Then by (21.2–6) we see that a_n and b_n are both equal to $\frac{f^{(n)}(0)}{n!}$, and therefore equal to each other.

Example 1. Consider the functions $J_0(x)$, $J_1(x)$ defined as follows:

$$J_0(x) = 1 - \frac{x^2}{(1!)^2 2^2} + \frac{x^4}{(2!)^2 2^4} - \cdots + (-1)^n \frac{x^{2n}}{(n!)^2 2^{2n}} + \cdots, \tag{21.2–7}$$

$$J_1(x) = \frac{x}{2}\left[1 - \frac{x^2}{1!2!2^2} + \cdots + (-1)^n \frac{x^{2n}}{n!(n+1)!2^{2n}} + \cdots\right]. \tag{21.2–8}$$

Show that they are defined for all values of x, and that

$$J_0'(x) = -J_1(x). \tag{21.2–9}$$

The function $J_0(x)$ is called the *Bessel function of order zero of first kind*; $J_1(x)$ *is called the Bessel function of order one of first kind.* These and other varieties of Bessel functions are of great importance because of the way they arise in many kinds of physical problems.

Both series are convergent for all values of x, as is readily verified by the ordinary ratio test (Theorem XIII, §19.4). To verify (21.2–9) we write the series for $J_0(x)$ and $J_1(x)$ in the forms

$$J_0(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{n!n!2^{2n}}, \quad J_1(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{n!(n+1)!2^{2n+1}}.$$

Then, calculating $J_0'(x)$ in the manner justified by Theorem VII, we have

$$J_0'(x) = \sum_{n=1}^{\infty} (-1)^n \frac{2nx^{2n-1}}{n!n!2^{2n}} = \sum_{n=1}^{\infty} (-1)^n \frac{x^{2n-1}}{(n-1)!n!2^{2n-1}}.$$

In the series for $J_0(x)$ the term with $n = 0$ is a constant, so that the series for $J_0'(x)$ will begin with the term for which $n = 1$. If we now write $n + 1$ instead of n in the last summation, the new index n will go from 0 to ∞, and we shall have

$$J_0'(x) = \sum_{n=0}^{\infty} (-1)^{n+1} \frac{x^{2n+1}}{n!(n+1)!2^{2n+1}}.$$

Since $(-1)^{n+1} = -(-1)^n$, we that (21.2–9) is true, by a comparison of the summation expressions for $J_0'(x)$ and $J_1(x)$.

Example 2. Determine what can be said about solutions of the differential equation

$$(1 - x^2)\frac{d^2y}{dx^2} + 6y = 0 \tag{21.2–10}$$

of the form $y = f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n + \cdots$,
i.e., solutions which can be expanded in powers of x, convergent for some interval about $x = 0$,

Assuming tentatively that there is such a series solution, we differentiate it, obtaining

$$\frac{dy}{dx} = a_1 + 2a_2x + 3a_3x^2 + \cdots + na_nx^{n-1} + \cdots,$$

$$\frac{d^2y}{dx^2} = 2a_2 + 3 \cdot 2a_3x + \cdots + n(n-1)\, a_nx^{n-2} + \cdots.$$

$$x^2\frac{d^2y}{dx^2} = 2a_2x^2 + 3 \cdot 2a_3x^3 + \cdots + n(n-1)\, a_nx^n + \cdots.$$

Then

$$(1 - x^2)y'' + 6y = \sum_{n=2}^{\infty} n(n-1)\, a_nx^{n-2} - \sum_{n=2}^{\infty} n(n-1)\, a_nx^n + \sum_{n=0}^{\infty} 6a_nx^n$$

$$= \sum_{n=0}^{\infty} [(n+2)(n+1)\, a_{n+2} - n(n-1)\, a_n + 6a_n]\, x^n$$

(observe that the term $n(n-1)\, a_n$ is zero for $n = 0, 1$). If now (21.2–10) is to be satisfied, we infer by Theorem IX that we must have

$$(n+2)(n+1)\, a_{n+2} - n(n-1)\, a_n + 6a_n = 0, \quad n = 0, 1, 2, \ldots,$$

Since $n(n-1) - 6 = (n-3)(n+2)$, this last relation is equivalent to

$$a_{n+2} = \frac{(n-3)(n+2)}{(n+2)(n+1)}\, a_n = \frac{n-3}{n+1}\, a_n.$$

From this relation we may determine $a_2, a_4, a_6, \ldots$ successively in terms of an arbitrary a_0; likewise $a_3, a_5, a_7, \ldots$ are determined in terms of an arbitrary a_1. We have

$$a_3 = \frac{1-3}{1+1}a_1 = -a_1, \quad a_5 = \frac{3-3}{3+1}a_3 = 0,$$

whence $a_7 = a_9 = a_{11} = \cdots = 0$. For the even subscripts,

$$\begin{aligned} a_2 &= -\tfrac{3}{1}a_0 \\ a_4 &= -\tfrac{1}{3}a_2 \\ a_6 &= +\tfrac{1}{5}a_4 \\ &\vdots \qquad\quad \vdots \\ a_{2n+2} &= \frac{2n-3}{2n+1}a_{2n}. \end{aligned}$$

Clearly none of these coefficients is zero if $a_0 \neq 0$. Now

$$a_{2n+2}a_{2n}\cdots a_4a_2 = \frac{(2n-3)(2n-5)\cdots 1\cdot(-1)\cdot(-3)}{(2n+1)(2n-1)\cdots 5\cdot 3\cdot 1}a_{2n}\cdots a_2a_0;$$

when like factors are cancelled from either side of this relation we find

$$a_{2n+2} = \frac{3}{(2n+1)(2n-1)}a_0. \tag{21.2–11}$$

What we have done thus far shows that *if* there is a solution of (21.2–10) in the form of a series of powers of x, the solution can be written

$$y = a_1(x-x^3) + a_0\left[1 + \frac{3}{1\cdot(-1)}x^2 + \frac{3}{3\cdot 1}x^4 + \cdots + \frac{3}{(2n-1)(2n-3)}x^{2n} + \cdots\right].$$

Moreover, the work shows that this really is a solution within the interval of convergence of the series, provided there is an interval of convergence. Now the infinite series

$$\sum_{n=0}^{\infty}\frac{3}{(2n-1)(2n-3)}x^{2n} \tag{21.2–12}$$

is convergent when $|x| < 1$, as may readily be verified. Thus we have found two linearly independent solutions of the differential equation (21.2–10): the polynomial $x - x^3$ and the infinite series (21.2–12). The coefficients a_0 and a_1 are arbitrary.

EXERCISES

1. The Bessel function of order m of first kind (m a nonnegative integer) is defined as

$$J_m(x) = \sum_{n=0}^{\infty}(-1)^n\frac{1}{(n+m)!\,n!}\left(\frac{x}{2}\right)^{2n+m}$$

Show that **(a)** $J_0(x) = x^{-1}\dfrac{d}{dx}(xJ_1(x))$, **(b)** $J_1(x) = x^{-2}\dfrac{d}{dx}(x^2J_2(x))$, **(c)** $J_2(x) =$

$-x\dfrac{d}{dx}(x^{-1}J_1(x))$. **(d)** State and prove the rule corresponding to **(a)** and **(b)** for Bessel functions of higher orders $m-1$ and m. **(e)** Of what general rule are **(c)** and (21.2–9) special cases? Prove the correctness of your answer.

2. Using only what you know about power series, and nothing of what you know about e^x, find the function $f(x)$, defined by a power series in x, such that $f'(x)=f(x)$ and $f(0)=1$. What is the radius of convergence of the series?

3. Find the power series in x, denoted by $f(x)$, such that $f''(x)+f(x)=0$ and $f(0)=0$, $f'(0)=1$. What is the radius of convergence of the series?

4. **(a)** Find $f(x)$, a power series in x, such that $(1+x)f'(x)=mf(x)$, where m is a constant and not one of the integers 0, 1, 2, . . ., **(b)** Determine the radius of convergence of the series. **(c)** Let $g(x)=\dfrac{f(x)}{(1+x)^m}$, and show from the requirement in **(a)** that $g'(x)=0$, so that $g(x)$ is constant. What can you now conclude about $f(x)$ and $(1+x)^m$ for values of x inside the interval of convergence of the series?

5. Suppose $f(x)$ is defined by a power series in x with positive radius of convergence. Let k be the smallest positive integer such that $f^{(k)}(0)\neq 0$. (We suppose $f(x)$ not a constant, so that there really is such an integer k.) **(a)** Show that f has neither a relative maximum nor a relative minimum at $x=0$ if k is odd. **(b)** Assuming that k is even, show that f has a relative minimum at $x=0$ if $f^{(k)}(0)>0$ and a relative maximum at $x=0$ if $f^{(k)}(0)<0$.

6. **(a)** If the function $f(x)$ represented by a power series in x is *even*, i.e., if $f(x)=f(-x)$, show that all the coefficients of odd powers of x in the series are zero. **(b)** If the function is *odd*, i.e., if $f(-x)=-f(x)$, show that all the coefficients of even powers of x are zero.

7. Suppose $f(x)$ and $g(x)$ are power series in x, each with a positive radius of convergence. Suppose that $f(0)=f'(0)=\cdots=f^{(m-1)}(0)=0$, $f^{(m)}(0)\neq 0$ and that $g(0)=g'(0)=\cdots=g^{(n-1)}(0)=0$, $g^{(n)}(0)\neq 0$. **(a)** Show that neither $f(x)$ nor $g(x)$ can vanish if $0<|x|<h$, provided h is sufficiently small. **(b)** Show that $\lim\limits_{x\to 0}\dfrac{f(x)}{g(x)}=\dfrac{f^{(n)}(0)}{g^{(n)}(0)}$ if $m=n$, and that the limit is 0 if $m>n$. Also show that $\lim\limits_{x\to 0}\left|\dfrac{f(x)}{g(x)}\right|=\infty$ if $m<n$.

8. Find a power series in x for $\dfrac{d}{dx}\left(\dfrac{e^x-1}{x}\right)$, and deduce that

$$1=\sum_{n=1}^{\infty}\frac{n}{(n+1)!}.$$

9. Show that $1=\dfrac{3}{2\cdot 1!}-\dfrac{5}{2^2\cdot 2!}+\dfrac{7}{2^3\cdot 3!}-\cdots$ by considering $\dfrac{d^2}{dx^2}(e^{-x^2})$.

10. Show that $4=\sum\limits_{n=1}^{\infty}(-2)^{n+1}\dfrac{n+2}{n!}$ by considering $\dfrac{d}{dx}(x^2e^{-x})$.

11. Find an expression for the function $f(x)=\sum\limits_{n=0}^{\infty}\dfrac{x^n}{(n+1)^2n!}$.

SUGGESTION: Calculate successively $xf(x)$, $\dfrac{d}{dx}(xf(x))$, $x\dfrac{d}{dx}(xf(x))$ in series form, and identify the last series. Then obtain $f(x)$ by integration.

12. **(a)** Obtain a power-series solution $y=f(x)$ of the differential equation $xy''+$

$(1-x)y' + qy = 0$, where q is an arbitrary constant. Find the radius of convergence of the series. **(b)** If q is a nonnegative integer, show that $f(x)$ is a polynomial of degree q. In this latter case, if $f(0) = q!$, $f(x)$ is denoted by $L_q(x)$ and called the Laguerre polynomial of degree q.

13. **(a)** Show that the differential equation $y'' - 2xy' + 2my = 0$, where m is a constant, has two independent power-series solutions, $f_1(x) = a_0 + a_2x^2 + a_4x^4 + \cdots$, $f_2(x) = a_1x + a_3x^3 + \cdots$. Find the radii of convergence of these series. **(b)** If m is a nonnegative integer, show that one of these solutions is a polynomial of degree m. If the coefficient of x^m in the polynomial is 2^m, it is called the Hermite polynomial of degree m, and denoted by $H_m(x)$.

21.3 / DIVISION OF POWER SERIES

It is quite often advantageous to obtain the power-series expansion of a function by representing it as the quotient of two power series and then dividing one of these series by the other. We begin with an illustrative example.

Example 1. Find several terms in the expansion of $\tan x$ in powers of x. We make use of the known Taylor's series for $\sin x$ and $\cos x$:

$$\tan x = \frac{\sin x}{\cos x} = \frac{x - \dfrac{x^3}{3!} + \dfrac{x^5}{5!} - \cdots}{1 - \dfrac{x^2}{2!} + \dfrac{x^4}{4!} - \cdots}.$$

Then we perform the long division as indicated:

$$\begin{array}{r|l} & x + \frac{1}{3}x^3 + \frac{2}{15}x^5 + \cdots \\ \hline 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 - \cdots & x - \frac{1}{6}x^3 + \frac{1}{120}x^5 - \cdots \\ & x - \frac{1}{2}x^3 + \frac{1}{24}x^5 - \cdots \\ \hline & \frac{1}{3}x^3 - \frac{1}{30}x^5 + \cdots \\ & \frac{1}{3}x^3 - \frac{1}{6}x^5 + \cdots \\ \hline & \frac{2}{15}x^5 - \cdots \\ & \frac{2}{15}x^5 - \cdots \\ \hline \end{array}$$

In this way we are led to the expansion

$$\tan x = x + \tfrac{1}{3}x^3 + \tfrac{2}{15}x^5 + \cdots. \qquad (21.3\text{–}1)$$

Although we have not indicated the general term of this series, it is clear that we may compute as many terms as we please according to this systematic procedure indicated in the long-division process.

None of our previous theorems furnishes any proof of the correctness of the result (21.3–1). Note also that since we have not obtained a general formula for the coefficients in (21.3–1), we are at present unable to determine the radius of convergence of the series.

The thing of the foremost practical importance is that the method of Example 1 really works. By long division we can find the quotient of two power

series as another power series. *The essential condition which must be fulfilled is that the power series in the denominator must begin with a nonzero constant term.* Either of the two infinite series in the quotient may in particular cases terminate (i.e., instead of an infinite series we may have a polynomial). Experience shows that the long-division process is often the best method for obtaining a power-series representation. With a rational function, for instance, the long-division method is much more practical than the method of computing the coefficients in the Taylor's series by differentiation.

THEOREM X. *Consider a function defined as the quotient of two power series:*

$$f(x) = \frac{a_0 + a_1x + a_2x^2 + \cdots + a_nx^n + \cdots}{b_0 + b_1x + b_2x^2 + \cdots + b_nx^n + \cdots}, \tag{21.3–2}$$

where $b_0 \neq 0$, and where both of the series are convergent in some interval $|x| < r$. Then for sufficiently small values of x the function f can be represented as a power series

$$f(x) = c_0 + c_1x + c_2x^2 + \cdots + c_nx^n + \cdots \tag{21.3–3}$$

whose coefficients may be found by the process of long division, or, what is equivalent, by solving the relations

$$\begin{aligned} b_0c_0 &= a_0 \\ b_0c_1 + b_1c_0 &= a_1 \\ &\vdots \\ b_0c_n + b_1c_{n-1} + \cdots + b_nc_0 &= a_n \\ &\vdots \end{aligned} \tag{21.3–4}$$

successively for $c_0, c_1, c_2, \ldots$.

It is impractical to give a complete proof of this theorem with the knowledge presently at our disposal, but we shall go as far as is conveniently possible. In the first place, the assumption $b_0 \neq 0$ means that the function defined by the series $\Sigma\, b_nx^n$ is not zero at $x = 0$. It is therefore different from zero throughout some interval about the origin, since the function is continuous, by Theorem V, §21.1. If we now *assume* that a power series expansion of the form (21.3–3) is valid, we shall have the product relation

$$\left(\sum_{n=0}^{\infty} b_nx^n\right)\left(\sum_{n=0}^{\infty} c_nx^n\right) = \sum_{n=0}^{\infty} a_nx^n \tag{21.3–5}$$

holding throughout some interval in which all three series are absolutely convergent. The rule for multiplication of absolutely convergent series (see §19.6) gives the result

$$\left(\sum_{n=0}^{\infty} b_nx^n\right)\left(\sum_{n=0}^{\infty} c_nx^n\right) = \sum_{n=0}^{\infty} (b_0c_n + b_1c_{n-1} + \cdots + b_nc_0)\, x^n. \tag{21.3–6}$$

By the uniqueness theorem for power series we then conclude from (21.3–5) and (21.3–6) that the general relation (21.3–4) holds. The coefficients c_n determined successively in this way are the same as would be found by the long-division process, as the student may verify for himself. Note that to solve for c_n when $c_0, \ldots, c_{n-1}$ are known, it is essential to know that $b_0 \neq 0$.

The proof is now complete except for justifying the assumption that $f(x)$ can be expanded in a series (21.3–3). It is this part of the proof which we shall not give. It is an easy consequence of certain standard theorems in the theory of functions of a complex variable, but the development of this latter theory is beyond the scope of this text.

Example 2. The long-division method yields the series expansion

$$\frac{2+x}{1+x+x^2} = 2 - x - x^2 + 2x^3 - x^4 - x^5 + 2x^6 - x^7 - x^8 + \cdots. \qquad (21.3\text{–}7)$$

Detailed verification is left to the student. It may be shown by Theorem XVI, §19.4, that the series (21.3–7) has radius of convergence $R = 1$.

It is important to observe that the division method of obtaining the expansions (21.3–1) and (21.3–7) is much easier than using the formula (21.2–6) for computing the coefficients in the expansion.

Let us consider the situation when $b_0 = 0$ in the quotient (21.3–2). The general case may be represented by assuming that b_k is the first of the b's which is not 0, and that a_l is the first of the a's which is not 0. Then, assuming $x \neq 0$,

$$f(x) = \frac{a_l x^l + a_{l+1}x^{l+1} + \cdots}{b_k x^k + b_{k+1}x^{k+1} + \cdots},$$

$$f(x) = x^{l-k}\left(\frac{a_l + a_{l+1}x + \cdots}{b_k + b_{k+1}x + \cdots}\right). \qquad (21.3\text{–}8)$$

This shows that, as $x \to 0$, $f(x)$ behaves like $x^{l-k}\left(\frac{a_l}{b_k}\right)$. In particular, $f(x)$ approaches a finite limit if $l \geqq k$, but not otherwise. Let us agree to define $f(0) = \lim_{x\to 0} f(x)$ if the limit exists. The quotient in the parenthesis in (21.3–8) can be expressed as a single power series by the method of long division, since $b_k \neq 0$.

Example 3. Find a power series in x for $\dfrac{\sin x}{\sin 2x}$.

For $x \neq 0$ we have

$$\frac{\sin x}{\sin 2x} = \frac{x - \dfrac{x^3}{3!} + \dfrac{x^5}{5!} - \cdots}{2x - \dfrac{(2x)^3}{3!} + \dfrac{(2x)^5}{5!} - \cdots}.$$

We remove the common factor x from the two series, and deal by long division

with what is left. In this way we find, for $x \neq 0$,

$$\frac{\sin x}{\sin 2x} = \frac{1 - \frac{x^2}{6} + \frac{x^4}{120} - \cdots}{2 - \frac{4}{3}x^2 + \frac{4}{15}x^4 - \cdots} = \frac{1}{2} + \frac{1}{4}x^2 + \frac{5}{48}x^4 + \cdots.$$

EXERCISES

1. Find several terms in the power series expansions of the following quotients:

(a) $\dfrac{2e^x}{1+e^x}$. **(d)** $\dfrac{1-\cos x}{\sin(x^2)}$.

(b) $\dfrac{1}{\cos x}$. **(e)** $\dfrac{-x}{\log(1-x)}$.

(c) $\dfrac{1}{\cos x - \sin x}$. **(f)** $\dfrac{x\cos x - \sin x}{x \sin x}\left(= \operatorname{ctn} x - \dfrac{1}{x}\right)$.

2. Show that $\sum_{n=0}^{\infty} \dfrac{a_n x^n}{1-x} = \sum_{n=0}^{\infty} (a_0 + a_1 + \cdots + a_n)x^n$,

and use the result to find the function represented by the following series:

(a) $\sum_{n=0}^{\infty} \left(1 + \dfrac{1}{1!} + \cdots + \dfrac{1}{n!}\right)x^n$,

(b) $\sum_{n=1}^{\infty} \left(0 + 1 - \frac{1}{2} + \frac{1}{3} - \cdots + (-1)^{n+1}\dfrac{1}{n}\right)x^n$,

(c) $\sum_{n=0}^{\infty} (0 + 1 + \cdots + n)x^n$.

3. Solve the first $(n+1)$ equations in Theorem X by determinants, and express c_n as $b_0^{-(n+1)}$ times a determinant. For the special case in which $a_0 = 1$, $a_1 = a_2 = \cdots = 0$, show that

$$c_n = \frac{(-1)^n}{b_0^{n+1}} \begin{vmatrix} b_1 & b_0 & 0 & 0 & \dots & 0 \\ b_2 & b_1 & b_0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ b_n & b_{n-1} & \dots & \dots & \dots & b_1 \end{vmatrix}$$

4. **(a)** Suppose that $(\Sigma_{n=0}^{\infty} b_n x^n)(\Sigma_{n=0}^{\infty} c_n x^n) = \Sigma_{n=0}^{\infty} a_n x^n$. Show that, as a consequence, $(\Sigma_{n=0}^{\infty}(-1)^n b_n x^n)\,(\Sigma_{n=0}^{\infty}(-1)^n c_n x^n) = \Sigma_{n=0}^{\infty}(-1)^n a_n x^n$. Take all questions of convergence for granted.

(b) Prove, either directly or as a consequence of **(a)**, that if $b_0 \neq 0$ and

$$\frac{\sum_{n=0}^{\infty} a_{2n}x^{2n}}{\sum_{n=0}^{\infty} b_{2n}x^{2n}} = \sum_{n=0}^{\infty} c_{2n}x^{2n},$$

then

$$\frac{\sum_{n=0}^{\infty} (-1)^n a_{2n}x^{2n}}{\sum_{n=0}^{\infty} (-1)^n b_{2n}x^{2n}} = \sum_{n=0}^{\infty} (-1)^n c_{2n}x^{2n}.$$

(c) Use (b) to prove that if $\tanh x = \Sigma_{n=0}^{\infty} A_n x^{2n+1}$, then

$$\tan x = \sum_{n=0}^{\infty} (-1)^n A_n x^{2n+1}.$$

(d) Use (b) to prove that if $x \operatorname{ctnh} x = \Sigma_{n=0}^{\infty} A_n x^{2n}$, then

$$x \operatorname{ctn} x = \sum_{n=0}^{\infty} (-1)^n A_n x^{2n}.$$

5. Define B_n as $n!$ times the coefficient of x^n in the power-series expansion of $x/(e^x - 1)$:

$$\frac{x}{e^x - 1} = \sum_{n=0}^{\infty} \frac{B_n}{n!} x^n.$$

The numbers B_n are called Bernoulli's numbers.

(a) Show that $B_0 = 1$, $B_1 = -\frac{1}{2}$, $B_2 = \frac{1}{6}$, $B_3 = 0$.

(b) Writing

$$\frac{x}{e^x - 1} + \frac{x}{2} = 1 + \sum_{n=2}^{\infty} \frac{B_n}{n!} x^n,$$

show that the function on the left is equal to the even function $\frac{x}{2} \operatorname{ctnh} \frac{x}{2}$. Deduce from this that $B_3 = B_5 = B_7 = \cdots = 0$ (see Exercise 6, §21.2). Then show that $x \operatorname{ctnh} x = \sum_{n=0}^{\infty} \frac{B_{2n}}{(2n)!} (2x)^{2n}$. The radius of convergence of this series is not readily determined by our present knowledge, but it may be shown to be π.

6. (a) Use Exercises 4(d) and 5(b) to show that

$$x \operatorname{ctn} x = \sum_{n=0}^{\infty} (-1)^n \frac{B_{2n}}{(2n)!} (2x)^{2n}.$$

(b) Use the identity $\tan x = \operatorname{ctn} x - 2 \operatorname{ctn} 2x$ to obtain the expansion

$$\tan x = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{B_{2n}}{(2n)!} 2^{2n}(2^{2n} - 1) x^{2n-1}.$$

(c) Use the identity $\operatorname{ctn} x + \tan \frac{x}{2} = \frac{1}{\sin x}$ to show that

$$\frac{x}{\sin x} = \sum_{n=0}^{\infty} (-1)^{n+1} \frac{B_{2n}}{(2n)!} (2^{2n} - 2) x^{2n}.$$

21.4 / ABEL'S THEOREM

Suppose that a function is defined by a power series:

$$f(x) = \sum_{n=0}^{\infty} a_n x^n, \tag{21.4–1}$$

and suppose that the radius of convergence R of the series is positive and finite. The series may or may not converge at $x = \pm R$. Let us suppose that the series does converge at $x = R$. In §21.1 we raised the question as to whether, in this

circumstance, $f(x) \to f(R)$ as $x \to R$ (with $x < R$). We shall now answer this question, by means of a theorem due to Abel.

THEOREM XI. *If the series* (21.4–1) *converges at* $x = R$, *then it converges uniformly on the closed interval* $0 \leqq x \leqq R$. *As a consequence, the function* f *defined by the series is continuous on* $0 \leqq x \leqq R$. *A like conclusion holds for* $-R \leqq x \leqq 0$ *if the series converges at* $x = -R$.

The proof is made with the aid of an inequality which we state as a lemma of independent interest.

LEMMA. *Suppose that*

$$m \leqq s_k \leqq M, \quad k = 0, 1, \ldots, p, \tag{21.4–2}$$

where

$$s_k = u_0 + u_1 + \cdots + u_k,$$

and that

$$v_0 \geqq v_1 \geqq \cdots \geqq v_p \geqq 0. \tag{21.4–3}$$

Then

$$mv_0 \leqq u_0v_0 + u_1v_1 + \cdots + u_pv_p \leqq Mv_0. \tag{21.4–4}$$

Proof. According to (19.7–1) we have

$$u_0v_0 + \cdots + u_pv_p = s_0(v_0 - v_1) + \cdots + s_{p-1}(v_{p-1} - v_p) + s_pv_p.$$

Now, because of (21.4–2) and (21.4–3) we can write

$$u_0v_0 + \cdots + u_pv_p \leqq M(v_0 - v_1) + \cdots + M(v_{p-1} - v_p) + Mv_p = Mv_0.$$

The other half of inequality (21.4–4) is deduced in a similar manner.

Proof of theorem XI. Suppose ϵ is any positive number. According to Theorem I, §20.1, our proof will be accomplished if we show that there is some integer N such that

$$|a_m x^m + a_{m+1} x^{m+1} + \cdots + a_{m+p} x^{m+p}| < \epsilon \tag{21.4–5}$$

if $N \leqq m$, $0 < p$, and $0 \leqq x \leqq R$. Let us set

$$u_k = a_{m+k}R^{m+k}, \quad v_k = \left(\frac{x}{R}\right)^{m+k}.$$

Then (21.4–5) is equivalent to

$$-\epsilon < u_0v_0 + u_1v_1 + \cdots + u_pv_p < \epsilon. \tag{21.4–6}$$

The conditions (21.4–3) are fulfilled by the v's. In addition, $v_0 \leqq 1$. Since the

series is convergent when $x = R$, we can choose N so that

$$-\epsilon < a_m R^m + a_{m+1}R^{m+1} + \cdots + a_{m+k}R^{m+k} < \epsilon$$

if $N \leqq m$ and $0 \leqq k$ (this is just the Cauchy condition for convergence). This means, in our present notation, that (21.4–2) is satisfied with $m = -\epsilon$, $M = \epsilon$. Applying the lemma, and noting that $-\epsilon \leqq -\epsilon v_0$, $\epsilon v_0 \leqq \epsilon$, we see that the conclusion (21.4–4) of the lemma yields (21.4–6). This completes the proof as regards $0 \leqq x \leqq R$. The case of convergence at $x = -R$ is reduced to the first case by considering $g(x) = f(-x)$ at $x = R$.

Example 1. If the binomial series (19.5–5) converges at $x = 1$, its sum is 2^m. This assertion may be justified as follows: We proved the validity of the binomial series expansion for $(1+x)^m$ when $|x| < 1$. Therefore, by Theorem XI, if the series converges at $x = 1$, its sum there is

$$\lim_{x \to 1} (1+x)^m = 2^m.$$

This always happens if $m > 0$, by what was established in Example 4, §19.4. It may be shown that the series converges at $x = 1$ if $-1 < m$, but diverges if $m \leqq -1$ (see Exercise 4, §19.5).

Next we show how Theorem XI permits us to extend the result of Theorem V, §21.1, with respect to the integration of a power series. If the power series

$$f(x) = \sum_{n=0}^{\infty} a_n x^n \tag{21.4–7}$$

is convergent when $|x| < R$, then

$$\int_0^R f(x)\,dx = \sum_{n=0}^{\infty} \frac{a_n}{n+1} R^{n+1}, \tag{21.4–8}$$

provided the series on the right in (21.4–8) *is convergent*, irrespective of whether or not the series in (21.4–7) is convergent at $x = R$. Of course, if (21.4–7) is not convergent at $x = R$, the integral in (21.4–8) may be improper at the upper limit.

The proof of the foregoing assertion is simple. If $0 < b < R$, we have

$$\int_0^b f(x)\,dx = \sum_{n=0}^{\infty} \frac{a_n}{n+1} b^{n+1}$$

by (21.1–4). Then, provided the series in (21.4–8) is convergent, we have

$$\lim_{b \to R} \int_0^b f(x)\,dx = \sum_{n=0}^{\infty} \frac{a_n}{n+1} R^{n+1},$$

by Theorem XI. Since it is also true that

$$\lim_{b \to R} \int_0^b f(x)\,dx = \int_0^R f(x)\,dx,$$

(21.4–8) is proved.

The remarks at the end of Example 1, §21.1, illustrate the application of (21.4–8).

Example 2. Show that

$$\int_0^1 \frac{\log(1-t)}{t}\,dt = -\left(\frac{1}{1^2}+\frac{1}{2^2}+\frac{1}{3^2}+\cdots+\frac{1}{n^2}+\cdots\right). \tag{21.4–9}$$

We start from the series (see (19.1–2))

$$\log(1-t) = -t-\tfrac{1}{2}t^2-\tfrac{1}{3}t^3-\cdots-\frac{1}{n}t^n-\cdots.$$

Dividing by t, we have

$$\frac{\log(1-t)}{t} = -1-\tfrac{1}{2}t-\tfrac{1}{3}t^2-\cdots-\frac{1}{n}t^{n-1}-\cdots.$$

Observe that the series diverges at $t = 1$. But if we integrate from 0 to x we have

$$\int_0^x \frac{\log(1-t)}{t}\,dt = -x-\frac{1}{2}x^2-\frac{1}{3^2}x^3-\cdots-\frac{1}{n^2}x^n-\cdots.$$

This series converges when $x = 1$, and so we have (21.4–9) as a special case of (21.4–8). The integral is improper at $t = 1$, but not at $t = 0$, since the integrand approaches the finite limit -1 as $t \to 0$ (this may be seen by use of L'Hospital's rule).

EXERCISES

1. (**a**) Express $\int_0^x \tan^{-1} t\,dt$ as a power series in x, and discuss the range of validity of the series, with particular attention to the end points of the interval of convergence. (**b**) Use the result in (**a**) to show that.

$$\frac{\pi}{4}-\log\sqrt{2} = 1-\tfrac{1}{2}-\tfrac{1}{3}+\tfrac{1}{4}+\tfrac{1}{5}--++\cdots.$$

2. Let $a_n = \left(\dfrac{1\cdot 3\cdots 2n-1}{2\cdot 4\cdots 2n}\right)^2$. Show that $\displaystyle\sum_{n=1}^{\infty}\frac{a_n}{2n-1} = \frac{\pi-2}{\pi}$ by considering $\int_0^{\pi/2}(1-x^2\sin^2 t)^{1/2}\,dt$ as a power series in x.

3. (**a**) Justify the formula

$$\int_0^1 \frac{t^{p-1}}{1+t^q}\,dt = \frac{1}{p}-\frac{1}{p+q}+\frac{1}{p+2q}-\cdots,$$

where p and q are positive integers. (**b**) Calculate the value of the integral to two decimal places, if $p = 10$, $q = 40$.

4. Let $f(x) = \displaystyle\sum_{n=1}^{\infty} n\frac{(pn)!}{(n!)^p}x^{n-1}$. For what positive integral values of p is it true that $\displaystyle\int_0^{p^{-p}} f(x)\,dx = \sum_{n=1}^{\infty}\frac{(pn)!}{(n!)^p}p^{-np}$?

21.5 / INFERIOR AND SUPERIOR LIMITS

The subject matter of this section is not properly a part of the theory of power series, but it is relevant to the problem of finding the radius of convergence of a power series. Inferior and superior limits have many other important uses in analysis.

Consider an arbitrary sequence $\{x_n\}$, without any assumption as to whether or not it is convergent.

Definition. *A sequence $\{x_n\}$ is said to* cluster, *or* accumulate, *at a point ξ if every open interval centered at ξ contains x_n for infinitely many values of n. These points x_n need not be distinct from ξ.*

Example 1. Let $x_n = \frac{1}{2}$ if $n = 1, 4, 7, 10, \ldots$, $x_n = 1$ if $n = 2, 5, 8, 11, \ldots$, $x_n = 2$ if $n = 3, 6, 9, 12, \ldots$. This sequence clusters at the three points $\frac{1}{2}$, 1, 2.

Example 2. Let $x_n = 1 + \sin(n\pi/2)$. This sequence clusters at 0, 1, and 2.

Example 3. Let $x_n = (-1)^n (1 + (1/n))$. This sequence clusters at -1 and 1.

If we compare the definition of a cluster point of $\{x_n\}$ with the definition of the limit of a convergent sequence, *we see at once that if $\{x_n\}$ is convergent, with limit ξ, then ξ is the sole cluster point of $\{x_n\}$.*

If a sequence is bounded, it must have at least one cluster point. This is a corollary of Theorem V, §16.31, or it may be proved by an argument much like the one used in proving the Bolzano–Weierstrass theorem (§16.3).

If a sequence is bounded above and has one or more cluster points, there is a cluster point farthest to the right, namely the least upper bound of all the cluster points. This cluster point farthest to the right is called the *limit superior* of the sequence; it is denoted by

$$\overline{\lim_{n\to\infty}}\, x_n \quad \text{or} \quad \limsup_{n\to\infty} x_n.$$

Likewise, if the sequence is bounded below and has one or more cluster points, the cluster point farthest to the left is called the *limit inferior*, and denoted by

$$\underline{\lim_{n\to\infty}}\, x_n \quad \text{or} \quad \liminf_{n\to\infty} x_n.$$

We often drop the $n \to \infty$ part of the notation, as a typographical convenience. Clearly we always have, in the case of a bounded sequence,

$$\underline{\lim}\, x_n \leqq \overline{\lim}\, x_n. \tag{21.5–1}$$

The sequence is convergent if and only if $\underline{\lim}\, x_n = \overline{\lim}\, x_n$, in which case the limit of the sequence coincides with the inferior and superior limit.

If the sequence has no upper bound, we say that

$$\overline{\lim_{n\to\infty}}\, x_n = +\infty;$$

if there is no lower bound, we say that

$$\underline{\lim}_{n\to\infty} x_n = -\infty.$$

If $x_n \to +\infty$, we say that $\underline{\lim}\, x_n = \overline{\lim}\, x_n = +\infty$, and if $x_n \to -\infty$, we say that $\underline{\lim}\, x_n = \overline{\lim}\, x_n = -\infty$.

One of the main reasons for introducing the new concepts of limit superior and limit inferior at just this point in the text is that we can use them in the following theorem.

THEOREM XII. *The radius of convergence R of a power series $\Sigma\, a_n x^n$ is given by*

$$\frac{1}{R} = \overline{\lim_{n\to\infty}} |a_n|^{1/n}. \tag{21.5–2}$$

The understanding is that $R = +\infty$ if the limit superior is 0, and $R = 0$ if the limit superior is $+\infty$.

Proof. We appeal to Cauchy's root test (Theorem XVI, §19.4). Let $u_n = a_n x^n$. Then

$$\overline{\lim}\, |u_n|^{1/n} = \overline{\lim}\, |a_n x^n|^{1/n} = |x|\, \overline{\lim}\, |a_n|^{1/n} = \frac{|x|}{R}, \tag{21.5–3}$$

where R is defined by (21.5–2). Here we have used the fact that, for any sequence $\{y_n\}$,

$$\overline{\lim}\, cy_n = c\, \overline{\lim}\, y_n \quad \text{if} \quad c > 0.$$

It follows from (21.5–3) and Cauchy's root test that the series converges if $|x| < R$ and diverges if $|x| > R$. This proves Theorem XII. We have tacitly assumed R to be finite and positive. The cases $R = 0$ or $R = +\infty$ require special attention but present no difficulties. They are left to the student.

Example 4. Find the radius of convergence of the series in (21.3–7). Here $a_n = 2$ if $n = 0, 3, 6, \ldots$ and $a_n = -1$ for other values of n. Thus $|a_n|^{1/n}$ is either $2^{1/n}$ or $1^{1/n}$, depending on the value of n. We see that $|a_n|^{1/n}$ is convergent, with limit 1, so that $R = 1$, by (21.5–2). Note that R cannot be found by Theorem IV in this case, for the successive values of $\left|\frac{a_n}{a_{n+1}}\right|$, for $n = 0, 1, 2, \ldots$, are $2, 1, \frac{1}{2}, 2, 1, \frac{1}{2}, \ldots$, and the sequence of ratios has no limit.

It is not difficult to prove that, if $\{a_n\}$ is any sequence of positive numbers,

$$\underline{\lim}\, \frac{a_{n+1}}{a_n} \leqq \underline{\lim}\, a_n^{1/n} \tag{21.5–4}$$

and

$$\overline{\lim}\, a_n^{1/n} \leqq \overline{\lim}\, \frac{a_{n+1}}{a_n}. \tag{21.5–5}$$

It is a consequence of these inequalities that if $\lim_{n\to\infty} a_{n+1}/a_n$ exists, then $\lim_{n\to\infty} a_n^{1/n}$ also exists, and the two limits are equal. This is sometimes useful.

Example 5. Show that

$$\lim_{n\to\infty} \frac{n}{(n!)^{1/n}} = e. \tag{21.5–6}$$

We take $a_n = \dfrac{n^n}{n!}$. Then

$$\frac{a_{n+1}}{a_n} = \frac{(n+1)^{n+1}}{(n+1)!}\frac{n!}{n^n} = \left(\frac{n+1}{n}\right)^n = \left(1+\frac{1}{n}\right)^n \to e$$

as we know from (1.62–5). Therefore also

$$a_n^{1/n} = \frac{n}{(n!)^{1/n}} \to e.$$

This result is not so easy to show without using this method. The result (21.5–6) may also be established by using Stirling's formula (see §22.8).

EXERCISES

1. Find the radius of convergence of $\Sigma\, a_n x^n$ in each of the following cases:

(a) $a_n = c^{\sqrt{n}}$, $c > 0$.

(b) $a_n = [1+(1/n)]^{n^2}$.

(c) $a_n = 2^n$ if n is even. $a_n = 3^n$ if n is odd.

(d) $a_n = n!/n^n$.

(e) $a_n = (\sqrt{n})^{-n}$.

(f) $a_n = n^{-\sqrt{n}}$.

2. If $\Sigma\, a_n x^n$ converges for all values of x, prove that $|a_n|^{1/n} \to 0$.

3. Prove Theorem VI, §21.2, by use of Theorem XII.

4. Find the radius of convergence of each of the following series:

(a) $\Sigma\, x^{n^2}/2^n$.

(b) $\Sigma\, n! x^{n!}$.

(c) $\Sigma\, n^n x^{n^2}$.

(d) $\Sigma\, r^{n^2} x^n$, $0 < r < 1$.

(e) $\Sigma\, r^n x^{n^2}$, $0 < r$.

(f) $\Sigma\, n! x^{n^2}$.

5. If $\Sigma\, a_n x^n$ has a finite positive radius of convergence, prove that the radius of convergence of $\Sigma\, a_n x^{n^2}$ is $R = 1$.

6. Use the method of Example 5 to find $\lim a_n^{1/n}$ in each of the following cases:

(a) $a_n = p$, a positive constant.

(b) $a_n = n$.

(c) $a_n = \dfrac{1\cdot 3\cdots(2n-1)}{2\cdot 4\cdots 2n}$.

7. If $\{x_n\}$ is bounded above, show that $\overline{\lim}\, x_n = \xi$ if and only if for each $\epsilon > 0$ it is true that $x_n < \xi + \epsilon$ for all sufficiently large values of n, and $\xi - \epsilon < x_n$ for infinitely many values of n. Devise and prove a similar statement about the limit inferior.

8. Use the formulation in Exercise 7 to prove (21.5–4) and (21.5–5).

SUGGESTION: Use the relation

$$\frac{a_{N+p}}{a_N} = \frac{a_{N+1}}{a_N} \cdot \frac{a_{N+2}}{a_{N+1}} \cdots \frac{a_{N+p}}{a_{N+p-1}}.$$

Choose N suitably, depending on ϵ, and then let $p \to \infty$ after getting an appropriate inequality.

21.6 / REAL ANALYTIC FUNCTIONS

The use of power series to represent functions is so important that it is convenient to have an adjective for functions which can be so represented. This adjective is the word "analytic." There is, of course, a general meaning of this word, not solely mathematical, but we are now using the term "analytic" in a highly technical sense.

Definition. *Let f be a real-valued function of the real variable x, defined on the open interval $a < x < b$, and suppose that f possesses derivatives of all orders at each point of the interval. We say that f is* analytic *on the interval if, for each point x_0 in the interval, $f(x)$ is represented by the Taylor's series*

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!} (x - x_0)^n \tag{21.6–1}$$

in some subinterval $x_0 - h < x < x_0 + h$, where the size of the positive number h may vary with the point x_0.

Example 1. We see by (19.1–7) that the function e^x is analytic on the whole x-axis. Likewise one may show that $\sin x$ and $\cos x$ are analytic on the whole x-axis.

Example 2. The function $\log x$ is analytic on the whole positive x-axis. To see this we proceed as follows: Suppose $x_0 > 0$. Then

$$x = x_0 + x - x_0 = x_0 \left(1 + \frac{x - x_0}{x_0}\right),$$

$$\log x = \log x_0 + \log\left(1 + \frac{x - x_0}{x_0}\right).$$

Now use the series expansion (19.1–1), with $\dfrac{x - x_0}{x_0}$ in place of x. The result is

$$\log x = \log x_0 + \sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{n} \frac{(x - x_0)^n}{x_0^n}.$$

The Taylor's series here is convergent if

$$-1 < \frac{x - x_0}{x_0} \leqq 1, \quad \text{or} \quad 0 < x \leqq 2x_0.$$

Most of the basic elementary functions and the usual combinations of them

are analytic on any open interval on which they are defined. We say that f is analytic at a point x_0 if it is analytic in some open interval centered at x_0. If f is not analytic at x_0, but *is* analytic at all points near x_0, we call x_0 an isolated singular point of f.

Example 3. The function $1/(1-x^2)$ has $x = 1$ and $x = -1$ as isolated singular points. It is analytic everywhere else, as may be shown with the aid of the binomial series.

If two functions f and g are each analytic on an interval $a < x < b$, the sum $f(x) + g(x)$ and the product $g(x)g(x)$ are each analytic on that interval. The quotient $f(x)/g(x)$ is analytic at each point of the interval at which $g(x) \neq 0$. Under appropriate conditions the formation of a composite function from analytic functions yields another analytic function.

Example 4. The function $e^{\sin x}$ is analytic for all values of x, and $\log(\cos x)$ is analytic for values of x such that $\cos x > 0$.

A function defined by a power series is analytic at each point in the interior of the interval of convergence of the power series.

Some of the facts we have been stating may be proved rather easily with what we have learned in Chapter 19 and in the present chapter. The fact that a composite function of analytic functions is analytic is somewhat harder to prove. The analyticity of the quotient of two analytic functions is a special case of the result for composite functions. We omit these proofs. The theory of *real* analytic functions is much clarified and simplified by a study of the theory of analytic functions of a *complex* variable.

It is instructive to know that a function can fail to be analytic at a certain point even though it is everywhere continuous, with a continuous derivative. The function $|x|^{3/2}$ is such a function, but it is not analytic at $x = 0$. This function does not have a second derivative at $x = 0$. Much more surprising is an example of a function which has continuous derivatives of *all* orders for all values of x, yet is not analytic at $x = 0$. The function

$$f(x) = e^{-1/x^2} \quad \text{if} \quad x \neq 0, \quad f(0) = 0$$

is such a function. It is shown in Example 6, §4.5, that f and all its derivatives have the value 0 at $x = 0$. It follows at once that $f(x)$ is not represented by its Taylor's series about the point $x = 0$, for

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n = 0$$

for all values of x, since $f^{(n)}(0) = 0$, whereas $f(x) \neq 0$ if $x \neq 0$.

The following theorem about real analytic functions is interesting and useful:

THEOREM XIII. *Suppose $f(x)$ is defined and has derivatives of all orders when $a < x < b$, and suppose that $f^{(n)}(x) \geqq 0$ for $n = 0, 1, 2, \ldots$ when $a < x < b$. Then f is analytic at each point of the interval, and the Taylor's series about*

the point x_0 converges to $f(x)$ for each x such that $|x - x_0| < h$, where h is the distance from x_0 to the nearer end-point of (a, b).

We shall not give the proof of this theorem; the proof is easily deduced from the result stated in Exercise 3.

EXERCISES

1. What does Theorem XIII imply (a) about e^x? (b) about $(1-x)^{-m}$ if $m > 0$ and $x < 1$? (c) about $-\log(1-x)$ if $x < 1$? (d) about $\sin x + e^{x-a}$ if $x \geqq a$?

2. By induction, or otherwise, prove the formula

$$\phi(x) = \sum_{k=0}^{n} \frac{x^k \phi^{(k)}(0)}{k!} + \frac{x^{n+1}}{n!} \int_0^1 \phi^{(n+1)}(tx)(1-t)^n \, dt,$$

assuming that ϕ and all the derivatives occurring are continuous on the interval from 0 to x, inclusive.

3. Assume that ϕ and all its derivatives are defined and nonnegative in value when $0 \leqq x \leqq r$. Prove by the following steps that $\phi(x)$ can be expanded by Maclaurin's series if $0 \leqq x < r$.

(a) Use the formula in Exercise 2 to show that

$$\int_0^1 \phi^{(n+1)}(tr)(1-t)^n \, dt \leqq \frac{n!}{r^{n+1}} \phi(r).$$

(b) Use the fact that $\phi^{(n+2)}(x)$ is a nonnegative function and the result in (a) to prove that the integral remainder term in the formula of Exercise 2 is not greater than $(x/r)^{n+1}\phi(r)$ if $0 \leqq x < r$. Now complete the proof of the main assertion.

4. Use the result in Exercise 3 to show that $\tan x$ can be expanded by Maclaurin's series if $0 \leqq x < \pi/2$. What can you conclude about the expansion for $x < 0$, seeing that $\tan x$ is an odd function?

5. What can you infer about $\log \dfrac{1+x}{1-x}$ from Exercise 3?

MISCELLANEOUS EXERCISES

1. Prove the binomial coefficient relation

$$\binom{n}{0}^2 + \binom{n}{1}^2 + \cdots + \binom{n}{n}^2 = \binom{2n}{n},$$

using Theorem IX and the fact that $(1+x)^n(1+x)^n = (1+x)^{2n}$.

2. Deduce the formula

$$\frac{\sin^{-1} x}{\sqrt{1-x^2}} = x + \tfrac{2}{3}x^3 + \frac{2 \cdot 4}{3 \cdot 5}x^5 + \frac{2 \cdot 4 \cdot 6}{3 \cdot 5 \cdot 7}x^7 + \cdots,$$

and obtain a series for $\frac{1}{2}(\sin^{-1} x)^2$. Is the latter series valid when $x = 1$?

3. Deduce the formula

$$\frac{(\tan^{-1} x)^2}{2!} = \frac{x^2}{2} - (1 + \tfrac{1}{3})\frac{x^4}{4} + (1 + \tfrac{1}{3} + \tfrac{1}{5})\frac{x^6}{6} - \cdots.$$

4. Deduce the formulas

$$-\frac{\log(1-x)}{1-x} = x + (1+\tfrac{1}{2})x^2 + (1+\tfrac{1}{2}+\tfrac{1}{3})x^3 + \cdots,$$

$$\tfrac{1}{2}[\log(1-x)]^2 = \sum_{n=2}^{\infty} \left(1+\tfrac{1}{2}+\cdots+\frac{1}{n-1}\right)\frac{x^n}{n}.$$

Prove the validity of the latter series when $x = -1$. (Use Euler's constant.)

5. Suppose that $\{x_n\}$ is a sequence with $x_n > 0$. Let

$$\overline{\lim_{n\to\infty}}\, x_n = A, \underline{\lim_{n\to\infty}}\, \frac{1}{x_n} = B.$$

Show that $B = 1/A$ if A is finite and positive. Show also that $B = 0$ if $A = +\infty$, and that $B = +\infty$ if $A = 0$.

6. Apply the results of the preceding exercise to show that (21.5–2) can be replaced by the formula

$$R = \underline{\lim_{n\to\infty}}\, |a_n|^{-1/n}.$$

22 / IMPROPER INTEGRALS

22 / PRELIMINARY REMARKS

In this chapter we shall study improper integrals systematically in somewhat the same way that we studied infinite series in Chapter 19. There are many analogies between the theory of improper integrals and the theory of infinite series. These analogies are seen most clearly in comparing

$$\int_0^\infty f(x)\,dx \quad \text{and} \quad \sum_{n=0}^{\infty} a_n.$$

In the integral we have a variable x, ranging *continuously* from 0 to ∞, while in the series we have a variable n, with the *discrete* range of values $0, 1, 2, \ldots$. The typical function value $f(x)$ is the counterpart of the typical term a_n, and the integration symbol $\int_0^\infty (\ )\,dx$ is the counterpart of the summation symbol $\Sigma_{n=0}^{\infty}$. The counterpart of a partial sum $\Sigma_{k=0}^{n} a_k$ is the "partial integral"

$$\int_0^x f(t)\,dt.$$

Definition. *By an improper integral of* first kind *we mean an integral*

$$\int_a^\infty f(x)\,dx \tag{22–1}$$

in which $f(x)$ is defined when $x \geqq a$ and is integrable (in the sense of §18.1*) over every finite interval $[a, b]$. The integral is defined as the limit*

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

if the limit exists. The integral (22–1) *is then said to be* convergent. *If the limit does not exist, the integral is called* divergent. *In most of the applications and illustrations $f(x)$ is continuous. The integrals*

$$\int_1^\infty \frac{dx}{x^2}, \quad \int_0^\infty \sin t^2\,dt, \quad \int_0^\infty x^n e^{-x}\,dx \quad (n \geqq 0)$$

are all of first kind.

Definition. *By an improper integral of* second kind *we mean an integral*

$$\int_a^b f(x)\,dx \tag{22–2}$$

with finite limits, in which the failure to be an ordinary "proper" integral arises from the behavior of $f(x)$ *either as* $x \to a$ *or as* $x \to b$, but not both. *Thus, if* $f(x)$ *is integrable on* $[a, c]$ *for each* c *such that* $a < c < b$, *but is* not *integrable on* $[a, b]$, *we say that the integral* (22–2) *is improper at* $x = b$. *Sometimes we say that* $f(x)$ *has a* singularity *at* $x = b$. *We then define the integral* (22–2) *as the limit*

$$\lim_{c \to b-} \int_a^c f(x)\, dx$$

if the limit exists. The terms convergent *and* divergent *are applied to the integral according as the limit does or does not exist. Similar definitions are made for integrals of the second kind which are improper at the lower limit of integration.*

Examples. $\displaystyle\int_0^1 \frac{dx}{\sqrt{1-x^2}}$, $\displaystyle\int_0^{\pi/3} \frac{d\theta}{\sqrt{\cos\theta - \frac{1}{2}}}$, $\displaystyle\int_{1/2}^1 \frac{\log x}{(1-x)^{3/2}}\, dx$ are improper at the upper limits, and

$$\int_0^1 \left(\log \frac{1}{x}\right)^3 dx, \quad \int_0^1 \frac{e^{-x}}{\sqrt{x}}\, dx, \quad \int_0^1 \frac{\log x}{1+x}\, dx$$

are improper at the lower limits.

As with infinite series, it is important to be able to test an improper integral for convergence or divergence. There are certain analogies between the tests for series and tests for integrals, which we shall point out as we proceed. In practice, however, we do not need as great a variety of tests for integrals as we do for series.

Just as certain functions may be represented by infinite series whose terms depend on a variable, so certain functions may be represented by improper integrals whose integrands depend on a parameter. As examples, we cite particularly the *gamma function* $\Gamma(x)$, defined by

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, dt, \quad 0 < x,$$

integrals of the form

$$f(s) = \int_0^\infty e^{-st} F(t)\, dt,$$

which are known as *Laplace transforms*, and functions defined by integrals of either of the forms

$$\sqrt{\frac{2}{\pi}} \int_0^\infty f(t) \sin xt\, dt, \quad \sqrt{\frac{2}{\pi}} \int_0^\infty f(t) \cos xt\, dt,$$

which are *Fourier transforms*. Laplace and Fourier transforms are of great importance, both theoretically and practically.

22.1 / POSITIVE INTEGRANDS. INTEGRALS OF THE FIRST KIND

It is convenient to begin by studying improper integrals with positive integrands, for the same reasons that make it useful to study series of positive terms before studying more general types of series.

Let us start with an integral

$$\int_a^\infty f(t)\,dt \tag{22.1–1}$$

of the first kind. This is *convergent* if the integral

$$F(x) = \int_a^x f(t)\,dt \tag{22.1–2}$$

approaches a finite limit as $x \to \infty$; otherwise it is *divergent*. Now let us suppose that $f(t) \geqq 0$ when $t \geqq a$. Usually we shall have $f(t) > 0$, but zero values for $f(t)$ need not be ruled out here. Then it is clear that $F(x)$ does not decrease as x increases, for

$$F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(t)\,dt \geqq 0$$

if $a \leqq x_1 < x_2$. There are now two possibilities: either $F(x)$ is bounded above or it is not. If it is bounded above, there is some constant M such that $F(x) \leqq M$ for all the relevant values of x. If it is not bounded above, for each M there will be some x such that $M < F(x)$, no matter how large M is, and since $F(x)$ never decreases as x increases, this means that $F(x) \to +\infty$ as $x \to \infty$. In this case the integral is divergent. If $F(x)$ is bounded, however, then $F(x)$ approaches a finite limit as $x \to \infty$, and in this case the integral is convergent. The fact that a function $F(x)$ approaches a limit as $x \to \infty$ if it is bounded and nondecreasing as x increases is analogous to a corresponding theorem about sequences (Theorem III, §2.7), and may be proved in an analogous way (see Exercise 9). We state our conclusion about the integral formally:

THEOREM I. *An integral* (22.1–1) *of the first kind with* $f(t) \geqq 0$ *for all* t *is convergent if and only if there is a constant* M *such that*

$$\int_a^x f(t)\,dt \leqq M$$

when $x > a$. *The value of the improper integral is then not greater than* M.

This theorem is the counterpart of Theorem I, §19.2. There is also a counterpart of Theorem II, §19.2; we call it the comparison test for integrals.

THEOREM II. *Let* $\int_a^\infty f(x)\,dx$ *and* $\int_b^\infty g(x)\,dx$ *be two integrals of first kind with nonnegative integrands, and suppose that* $f(x) \leqq g(x)$ *for all values of* x *beyond a certain point* $x = c$. *Then if* $\int_b^\infty g(x)\,dx$ *is convergent, so is* $\int_a^\infty f(x)\,dx$, *and if* $\int_a^\infty f(x)\,dx$ *is divergent, so is* $\int_b^\infty g(x)\,dx$.

For the proof we note that the convergence or divergence of the integrals is not affected if we replace both lower limits by $x = c$. We then have, if $x > c$,

$$\int_c^x f(t)\,dt \leqq \int_c^x g(t)\,dt;$$

the conclusions of the theorem now follow at once from Theorem I.

Example 1. The integral $\int_0^\infty \frac{dx}{\sqrt{1+x^3}}$ is convergent, by comparison with the integral $\int_1^\infty \frac{dx}{x^{3/2}}$, which is convergent, as may be shown directly from the definition; for $(1+x^3)^{-1/2} < x^{-3/2}$ when $x > 0$.

Example 2. The integral $\int_0^\infty \frac{dx}{(1+x^3)^{1/3}}$ is divergent, by comparison with the integral $\int_0^\infty \frac{dx}{1+x}$, which is divergent.

To see the divergence of the second integral, note that

$$\int_0^x \frac{dt}{1+t} = \log(1+x) \to \infty \quad \text{as} \quad x \to \infty.$$

Now

$$\frac{1}{1+x} < \frac{1}{(1+x^3)^{1/3}}$$

when $x > 0$, for this inequality is equivalent to $1 + x^3 < (1+x)^3 = 1 + 3x + 3x^2 + x^3$, which is obviously correct if $x > 0$. Thus the first integral must diverge, by Theorem II.

To avoid troublesome details of working with inequalities in practice, it is often convenient to use the following theorem rather than to use the comparison test directly.

THEOREM III. *Suppose $\int_a^\infty f(x)\,dx$ and $\int_b^\infty g(x)\,dx$ are integrals of the first kind with positive integrands, and suppose that the limit*

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = L \tag{22.1–3}$$

exists (finite) and is not zero. Then either both integrals are convergent, or both are divergent.

This is proved in exactly the same manner as we proved its counterpart for series, Theorem III, §19.2.

Example 3. The integral $\int_1^\infty \frac{x^2\,dx}{2x^4 - x + 1}$ is convergent. To prove this, observe that for large values of x the integrand is about the size of $1/(2x^2)$. More exactly,

taking

$$f(x) = \frac{x^2}{2x^4 - x + 1}, \quad g(x) = \frac{1}{x^2},$$

we see that

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = \tfrac{1}{2}.$$

Now the integral $\int_1^\infty (1/x^2)\,dx$ is convergent, and so the given integral is also, by Theorem III.

It is convenient to state an additional theorem about what conclusion can be drawn if $L = 0$ or $L = +\infty$ in (22.1–3).

THEOREM IV. *For the integrals described in Theorem III suppose that the limit $L = 0$. Then if $\int_b^\infty g(x)\,dx$ is convergent, so is $\int_a^\infty f(x)\,dx$. Or, alternatively, suppose that $L = +\infty$. Then if $\int_b^\infty g(x)\,dx$ is divergent, so is $\int_a^\infty f(x)\,dx$.*

The proof comes directly from Theorem II. For, in the first case we must have $f(x) < g(x)$ beyond a certain point, and in the second case it is clear that $g(x) < f(x)$ beyond a certain point. Nevertheless, it is usually easier to find L than to deal directly with the inequalities.

Example 4. The integral $\int_1^\infty x^\alpha e^{-x}\,dx$ is convergent, no matter what real number α may be. We apply the limit test, using the fact that $\int_1^\infty (1/x^2)\,dx$ is convergent.

$$\frac{x^\alpha e^{-x}}{x^{-2}} = x^{\alpha+2} e^{-x} = \frac{x^{\alpha+2}}{e^x},$$

and

$$\lim_{x \to +\infty} \frac{x^{\alpha+2}}{e^x} = 0,$$

as we see by applying l'Hospital's rule (compare with Example 5, §4.5). The convergence of the given integral now follows by Theorem IV.

In using Theorems III and IV the student needs to have in mind a few simple standard integrals whose convergence or divergence is known. It is easily shown that $\int_a^\infty (1/x^p)\,dx$ (where $a > 0$) is convergent if $p > 1$ and divergent if $p \leqq 1$. In a very large number of practical situations it will be found that the convergence or divergence of an integral $\int_a^\infty f(x)\,dx$ of the first kind with positive integrand can be settled by using Theorem III or Theorem IV with $g(x) = 1/x^p$, choosing an appropriate value of p as determined by trial or inspection.

Example 5. Let $f(x) = \dfrac{3x - 7}{(x+1)^{5/2}}$. For large values of x this function is comparable in value to $3x/x^{5/2} = 3/x^{3/2}$. Thus, applying Theorem III with $g(x) = \dfrac{1}{x^{3/2}}$, we see that $\displaystyle\int_0^\infty \frac{3x - 7}{(x+1)^{5/2}}\,dx$ is convergent.

Example 6. Consider the integral $\int_2^\infty \frac{dx}{(\log x)^p}$, where $p>0$. In this case, we know that $\log x$ increases more slowly than any positive power of x. We apply Theorem IV with $f(x)=(\log x)^{-p}$, $g(x)=x^{-1}$. Then, using l'Hospital's rule,

$$\lim_{x\to\infty}\frac{f(x)}{g(x)}=\lim_{x\to\infty}\frac{x}{(\log x)^p}=\lim_{x\to\infty}\frac{1}{p(\log x)^{p-1}\cdot(1/x)}.$$

This is the same as

$$\lim_{x\to\infty}\frac{x}{p(\log x)^{p-1}}.$$

After a certain number of applications of l'Hospital's rule we find that

$$\lim_{x\to\infty}\frac{f(x)}{g(x)}=+\infty.$$

(One must consider separately the cases in which p is or is not an integer.) Theorem IV then assures us that the given integral is divergent.

The student will note that we have not developed any analogues of the ratio tests of §19.22. In the analogy between series and integrals there is no simple way of formulating a counterpart of a ratio test, because a typical value $f(x)$ of the integrand has no immediate successor, in the sense that a_{n+1} is the successor of a_n.

One other difference between infinite series and improper integrals of the first kind is worth noting. If a series is convergent, its typical term a_n approaches zero as $n\to\infty$. But if an integral $\int_a^\infty f(x)\,dx$ is convergent, it does not necessarily follow that $f(x)\to 0$ as $x\to\infty$. See Exercise 8, and Example 2, §22.3.

EXERCISES

1. Test the following integrals for convergence or divergence, using Theorem II and the known facts about $\int_a^\infty x^{-p}\,dx$ for $a>0$.

(a) $\int_1^\infty \frac{dx}{(1+x)\sqrt{x}}$.

(b) $\int_0^\infty \frac{x\,dx}{(1+x)^2(2+\sqrt{x})}$.

(c) $\int_1^\infty \frac{x+2}{x(x+1)}\,dx$.

(d) $\int_2^\infty \frac{dx}{x\sqrt{1+x^2}}$.

(e) $\int_4^\infty \frac{\sqrt{x+1}}{x+\sqrt{x}}\,dx$.

(f) $\int_2^\infty \frac{x^2+4x+4}{(\sqrt{x}-1)^3\sqrt{x^3-1}}\,dx$.

2. Establish the facts about the values of the exponent p for which the integral $\int_a^\infty \frac{dx}{x(\log x)^p}$ is convergent ($a>1$). Then use either Theorem II or Theorem III to test the following integrals, using the foregoing integral as a standard, with an appropriate value of p in each case.

(a) $\int_1^\infty \frac{dx}{\sqrt{x^2+1}\,[\log(1+x)]^2}$.

(b) $\int_6^\infty \frac{(x+1)\,dx}{(x^2-2)\left(\log\frac{x}{2}-1\right)}$.

3. Find whether each of the following integrals is convergent or divergent:

(a) $\int_0^\infty \frac{x}{(1+x)^3}\,dx$.

(b) $\int_1^\infty \frac{x}{e^x-1}\,dx$.

(c) $\int_0^\infty \frac{x^3}{16+x^4}\,dx$.

(d) $\int_0^\infty e^{-x^2}\,dx$.

(e) $\int_0^\infty x^2e^{-x^2}\,dx$.

(f) $\int_1^\infty \frac{\sqrt{x}}{(1+x)^2}\,dx$.

(g) $\int_0^\infty \frac{\tan^{-1}x}{1+x^2}\,dx$.

(h) $\int_2^\infty \frac{dx}{\sqrt{x}(\log x)^3}$.

(i) $\int_0^\infty e^{-x}x^2(\log x)^3\,dx$.

(j) $\int_1^\infty \frac{\sin\frac{1}{x}}{x}\,dx$.

(k) $\int_1^\infty \frac{\frac{\pi}{2}-\tan^{-1}x}{x}\,dx$.

(l) $\int_0^\infty \left(\frac{\pi}{2}-\tan^{-1}x\right)dx$.

(m) $\int_2^\infty \frac{\frac{\pi}{2}-\tan^{-1}x}{\log x}\,dx$.

(n) $\int_{2/\pi}^\infty \frac{\cos\left(\frac{\pi}{2}-\frac{1}{x}\right)}{1+x}\,dx$.

(o) $\int_4^\infty \frac{dx}{(x-1)\log(x-2)\cdot\log(\log x)}$.

(p) $\int_3^\infty \frac{dx}{\sqrt{1+x^2}\log x[\log(\log x)]^2}$.

4. For what values of α is $\int_1^\infty \frac{x^{\alpha-1}}{1+x}\,dx$ convergent?

5. Show that $\int_0^\infty x^ne^{-x^2}\,dx$ is convergent if $n \geqq 0$. Find the value of the integral if $n = 2m+1$, where m is a nonnegative integer.

6. Find the conditions on m and n which guarantee that $\int_1^\infty e^{-x}x^m(\log x)^n\,dx$ is a convergent integral of the first kind.

7. Let $P(x)$ and $Q(x)$ be polynomials of degrees m and n, respectively, and suppose that r is the largest real root of the equation $Q(x) = 0$. What is the necessary and sufficient condition on m and n to make the integral $\int_c^\infty \frac{P(x)}{Q(x)}\,dx$ converge, if $r < c$? Why is the integrand of constant sign if x is large enough?

8. Suppose $0 < a_n < \frac{1}{2}$, and $b_n > 0$, $n = 0, 1, 2, \ldots$. Define a function $f(x)$ for $x \geqq 0$ as follows: $f(n) = b_n$, $n = 0, 1, 2, \ldots$, $f(x) = 0$ if $n - 1 + a_n \leqq x \leqq n - a_{n+1}$ and $n = 1, 2, \ldots$, $f(x)$ continuous for all $x \geqq 0$, and a linear function on each of the intervals where it has not already been defined. (a) Sketch the graph of the function enough to show its general character. (b) Show that $\int_0^\infty f(x)\,dx$ is convergent if and only if the series $\sum_{n=1}^\infty a_{n+1}b_n$ is convergent. (c) Specialize the sequences $\{a_n\}$ and $\{b_n\}$ so as to obtain a convergent integral and yet have $b_n \to \infty$. This shows that the integral may converge even though $f(x)$ is unbounded.

9. Suppose $F(x)$ is defined when $x \geqq x_0$, and that: (a) $F(x) < M$ for all such x, where M is a constant; (b) $F(x_1) \leqq F(x_2)$ if $x_0 \leqq x_1 < x_2$. Prove that $\lim_{x\to\infty} F(x) = A$, where A is the least upper bound of the values of $F(x)$.

10. Prove the following theorem: If $f(x) > 0$, if $f(x)$ decreases steadily as x increases, and if $\int_a^\infty f(x)\,dx$ is convergent, then $\lim_{x\to\infty} xf(x) = 0$. SUGGESTION: First prove that $\int_{x/2}^x f(t)\,dt \to 0$ as $x \to \infty$. Then use (18.1–3).

22.11 / INTEGRALS OF THE SECOND KIND

Let us consider integrals of the second kind with positive (or nonnegative) integrands. Theorems I–IV have exact analogues whose wordings differ but slightly from the statements of these theorems given in §22.1. Suppose, for example, that we are dealing with integrals improper at the upper limit,

$$\int_a^b f(t)\,dt \quad (b > a) \tag{22.11–1}$$

with $f(t) \geqq 0$. Then, for $a \leqq x < b$,

$$F(x) = \int_a^x f(t)\,dt$$

does not decrease as x increases, and the integral (22.11–1) is convergent if and only if $F(x)$ is bounded; in which case the value of the integral is $\lim_{x\to b} F(x)$. This result enables us to prove a comparison test strictly parallel to Theorem II, from which in turn we deduce limit tests in which the convergence or divergence of two integrals of the same type,

$$\int_a^b f(x)\,dx, \quad \int_a^b g(x)\,dx$$

are related by an examination of the limit

$$\lim_{x\to b-} \frac{f(x)}{g(x)}.$$

Entirely similar results obtain for integrals of the second kind improper at the lower limit of integration.

For integrals of the second kind the basic standard reference integrals are

$$\int_a^b \frac{dx}{(b-x)^p}, \quad \int_a^b \frac{dx}{(x-a)^p}. \tag{22.11–2}$$

These integrals are convergent if $p < 1$, and divergent if $p \geqq 1$. If $p \leqq 0$ they are *proper* integrals, with no singularities of the integrands.

Example 1. The integral $\int_0^1 \frac{dx}{(1-x^3)^{1/3}}$ is convergent. For, let

$$f(x) = \frac{1}{(1-x^3)^{1/3}} = \frac{1}{(1-x)^{1/3}(1+x+x^2)^{1/3}},$$

$$g(x) = \frac{1}{(1-x)^{1/3}}.$$

We have

$$\lim_{x\to 1-} \frac{f(x)}{g(x)} = \lim_{x\to 1-} \frac{1}{(1+x+x^2)^{1/3}} = (\tfrac{1}{3})^{1/3}.$$

Consequently, since $\int_0^1 g(x)\,dx$ is of the type (22.11–2) with $p < 1$, the given integral is convergent (by the counterpart of Theorem III, §22.1).

We often have occasion to recall the fact that $\sin x$ is approximately equal to x when x is small; more precisely,

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

Example 2. For what values of p is $\int_0^{\pi/2}\frac{\sin x}{x^p}\,dx$ convergent?

Here there is a singularity at $x=0$ if $p>1$; if $p=1$ the integrand is bounded and the integral is proper. We write

$$\frac{\sin x}{x^p}=\frac{\sin x}{x}\cdot\frac{1}{x^{p-1}},$$

and take

$$f(x)=\frac{\sin x}{x^p},\quad g(x)=\frac{1}{x^{p-1}},$$

so that

$$\lim_{x\to 0}\frac{f(x)}{g(x)}=1.$$

The integral $\int_0^{\pi/2}\frac{dx}{x^{p-1}}$, and therefore also the integral $\int_0^{\pi/2}\frac{\sin x}{x^p}\,dx$, is convergent if $p-1<1$, or $p<2$. Both integrals are divergent if $p\geqq 2$.

It is worth observing that an integral of second kind can be transformed into an integral of first kind by a simple substitution. For an integral $\int_a^b f(x)\,dx$ with singularity at $x=b$ we can let

$$y=\frac{1}{b-x},\quad x=b-\frac{1}{y},\quad dx=\frac{dy}{y^2}.$$

As $x\to b^-$, $y\to+\infty$, and we get an integral of the form $\int_c^\infty \phi(y)\frac{dy}{y^2}$ with $c=(b-a)^{-1}$.

Example 3. Transform $\int_0^1 \log\left(\frac{1}{1-x}\right)dx$.

We set $y=(1-x)^{-1}$ and obtain

$$\int_0^1 \log\left(\frac{1}{1-x}\right)dx=\int_1^\infty\frac{\log y}{y^2}\,dy.$$

This integral of first kind is convergent, as may be verified by the limit test of Theorem IV, with $g(y)=y^{-3/2}$. The convergence of the original integral could be established directly by a limit test with $g(x)=(1-x)^{-1/2}$. We leave it for the student to verify by l'Hospital's rule that

$$\lim_{x\to 1^-}\frac{\log\left(\frac{1}{1-x}\right)}{(1-x)^{-1/2}}=0.$$

Other changes of variable may also be used.

Example 4. Show that

$$\int_0^1 \left(\log \frac{1}{u}\right)^3 du = \int_0^\infty t^3 e^{-t}\, dt. \tag{22.11-3}$$

Here we set $t = \log(1/u)$, or $u = e^{-t}$, $du = -e^{-t}\, dt$. As $u \to 0$, $t \to +\infty$, and as $u \to 1$, $t \to 0$, whence the result follows as stated. The integral on the right in (22.11-3) has already been proved convergent (Example 4, §22.1).

EXERCISES

1. Determine the convergence or divergence of each of the following integrals by comparison with an appropriate one of the integrals

$$\int_0^1 x^{-p}\, dx, \quad \int_0^b (b-x)^{-q}\, dx,$$

using the known facts as to the convergence or divergence of these latter integrals.

(a) $\displaystyle\int_0^1 \frac{(1+2x)\sqrt{1+x^2}}{1-x^2}\, dx.$ (d) $\displaystyle\int_0^{\pi/2} \frac{\sqrt{x}\, dx}{(x+\sin x)(1+x^2)}.$

(b) $\displaystyle\int_0^1 \frac{x\, dx}{\sqrt{1-x^3}}.$ (e) $\displaystyle\int_0^1 \frac{dx}{\sqrt{x}(x+2x^2)^{1/3}}.$

(c) $\displaystyle\int_1^2 \frac{4(8-x^3)}{(2x-x^2)^2}\, dx.$ (f) $\displaystyle\int_0^1 \left(\frac{1}{x}+x\right)^{1/2} \frac{x^{1/3}}{x-\frac{1}{2}x^2}\, dx.$

2. Test the following integrals for convergence of divergence by methods analogous to those of Theorems III or IV, §22.1.

(a) $\displaystyle\int_1^2 \frac{dx}{\sqrt{(x-1)(3-x)}}.$ (g) $\displaystyle\int_0^{\pi/2} \frac{x^2}{(\sin x)^3}\, dx.$

(b) $\displaystyle\int_0^{10} \frac{dx}{x\sqrt{x^2-1}}.$ (h) $\displaystyle\int_0^{\pi/4} \frac{x^3\, dx}{\sin(x^2)(\tan x)^3}.$

(c) $\displaystyle\int_0^2 \frac{x\, dx}{(16-x^4)^{1/3}}.$ (i) $\displaystyle\int_0^2 \frac{\sqrt{4-\frac{1}{4}x^2}}{\sqrt{x^3-2x^2-4x+8}}\, dx.$

(d) $\displaystyle\int_0^1 \frac{dx}{x^{2/3}(1+x)}.$ (j) $\displaystyle\int_0^1 \frac{-\log x}{\sqrt{x}}\, dx.$

(e) $\displaystyle\int_3^5 \frac{\sqrt{25-x^2}}{x^2-x-6}\, dx.$ (k) $\displaystyle\int_1^2 \frac{\sqrt{x}}{\log x}\, dx.$

(f) $\displaystyle\int_0^1 \frac{x^3 \sin^{-1} x}{\sqrt{1-x^2}}\, dx.$ (l) $\displaystyle\int_0^1 \left(\frac{1}{x}-x\right)^{1/2} \frac{\log(1+x^{1/3})}{\sin x}\, dx.$

3. Give necessary and sufficient conditions on p and q for $\displaystyle\int_0^{\pi/2} \frac{x^p}{(\sin x)^q}\, dx$ to be a convergent improper integral of second kind.

4. For what values of a is the integral $\displaystyle\int_0^1 \frac{x^{a-1}+x^{-a}}{1+x}\, dx$ a proper integral? For what values of a is the integral improper, but convergent?

5. Consider

$$\int_0^1 (1-x^2)^{-1/2}(1-x^3)^{-1/3}\cdots(1-x^n)^{-1/n}\, dx, \quad n \geqq 2.$$

Find all values of n for which the integral is convergent.

6. For what values of x is $\int_0^1 e^{-t}t^{x-1}\,dt$ (a) proper? (b) improper, but convergent?

7. Show that $\int_0^1 \frac{(-\log x)^p}{x^q}\,dx$ is convergent if $p \geqq 0$ and $0 < q < 1$.

8. For what values of p is $\int_1^a (\log x)^{-p}\,dx$ improper, but convergent? Assume $a > 1$.

9. Prove that $\int_0^{\pi/2} \log(\sin x)\,dx$ is convergent. What can you say about $\int_0^{\pi/2} \log(\cos x)\,dx$?

10. If $f(u)$ is continuous, $0 \leqq u \leqq 1$, show that $\int_0^1 \frac{f(u)}{\sqrt{1-u^2}}\,du$ is convergent. Show also that the substitution $u = \sin\theta$ transforms the given integral into a proper integral.

11. By a suitable change of variable transform $\int_0^1 x^{-n}e^{-1/x}\,dx$ into an improper integral of first kind, and show that it is convergent for all values of n.

12. In the integral $\int_0^1 \frac{x^p}{[\log(1+x)]^q}\,dx$, use the power-series expansion of $\log(1+x)$ to show that the integral is proper if $p \geqq q$, improper but convergent if $p < q < p+1$, and divergent if $q \geqq p+1$.

22.12 / INTEGRALS OF MIXED TYPE

Many improper integrals occurring in practice are of mixed type.

Example 1. Consider $\int_0^\infty \frac{dx}{(1+x)\sqrt{x}}$. This is of mixed type, with infinite upper limit, and a singularity of the integrand at the lower limit. There are no other singularities, so we consider the separate integrals

$$\int_0^1 \frac{dx}{(1+x)\sqrt{x}}, \quad \int_1^\infty \frac{dx}{(1+x)\sqrt{x}},$$

of second and first kinds, respectively. The first integral is convergent, as may be shown by using a limit test to compare it with the convergent integral $\int_0^1 \frac{dx}{\sqrt{x}}$. The second integral is also convergent, since the integrand behaves essentially like $x^{-3/2}$ as $x \to \infty$. We then write

$$\int_0^\infty \frac{dx}{(1+x)\sqrt{x}} = \int_0^1 \frac{dx}{(1+x)\sqrt{x}} + \int_1^\infty \frac{dx}{(1+x)\sqrt{x}}.$$

The choice of $x = 1$ as a breaking point is arbitrary. Any other positive value of x could have been used.

When an integral of mixed type is separated into its constituent "pure-type" parts, it is called divergent if any one of the constituents is divergent. If singularities occur within the interval of integration, or at both ends of a finite interval of integration, the integral must be separated into several integrals, each of which is a pure type of either first or second kind.

Integrals with $-\infty$ as a limit of integration may be treated by methods parallel

to those of §22.1, or may be reduced to integrals with $+\infty$ as a limit of integration, by the substitution $x = -u$.

Example 2. Consider $\int_{-\infty}^{\infty} \frac{x\,dx}{e^x + x^4}$. We separate this into

$$(a)\ \int_{-\infty}^{0} \frac{x\,dx}{e^x + x^4} \quad \text{and} \quad (b)\ \int_{0}^{\infty} \frac{x\,dx}{e^x + x^4}.$$

The integral (b) is convergent, since

$$\frac{x}{e^x + x^4} < xe^{-x}$$

if $x > 0$, and $\int_0^\infty xe^{-x}\,dx$ is convergent (Example 4, §22.1). The integral (a) is also convergent; for, as $x \to -\infty$, $e^x \to 0$, and the integrand behaves like x^{-3}.

One may set $x = -u$, and thus get

$$\int_{-\infty}^{0} \frac{x\,dx}{e^x + x^4} = \int_{\infty}^{0} \frac{u\,du}{e^{-u} + u^4} = -\int_{0}^{\infty} \frac{u\,du}{e^{-u} + u^4}.$$

In the transformed integral the integrand behaves like u^{-3} as $u \to +\infty$. The original integral has thus been shown to be the sum of two convergent integrals of first kind.

EXERCISES

1. Examine each of the following integrals as to convergence or divergence, giving a complete analysis of the convergence or divergence of each of the constituent pure types.

(a) $\int_0^\infty \frac{dx}{x\sqrt{1+x^2}}$.

(b) $\int_0^\infty \frac{dx}{\sqrt{x}\sqrt{1+x^2}}$.

(c) $\int_{-\infty}^{\infty} e^{-x^2}\,dx$.

(d) $\int_{-\infty}^{\infty} x^2 e^{-|x|}\,dx$.

(e) $\int_1^\infty \frac{dx}{(x-1)^{1/2}(3-x)^{2/3}}$.

(f) $\int_0^\infty \frac{dx}{x^{1/2}(x-1)^{4/3}}$.

(g) $\int_{-1}^{2} \frac{dx}{x^{1/3}}$.

(h) $\int_0^\infty \frac{3x-7}{(x-1)^3}\,dx$.

(i) $\int_0^\pi \frac{dx}{(\cos x)^{2/3}}$.

(j) $\int_0^\pi \frac{dx}{(\sin x)^{3/2}}$.

2. Proceed as directed in Exercise 1 with each of the following integrals:

(a) $\int_0^\infty \frac{(\log x)^2}{1+x^2}\,dx$.

(b) $\int_0^\infty \frac{\log x}{(1+x^2)^2}\,dx$.

(c) $\int_0^\infty \frac{e^{-x}-1}{x\sqrt{x}}\,dx$.

(d) $\int_0^\infty \frac{(e^{-x}-1)^2}{x^3}\,dx$.

(e) $\int_0^\infty \frac{t^{-2/3}}{1+t}\,dt$.

(f) $\int_{-\infty}^{\infty} \frac{\sinh \frac{\pi x}{2}}{\sinh \pi x}\,dx$.

3. In each of the following integrals the integrand contains a parameter. For each

integral find the range of values of the parameter such that the integral is not divergent.

(a) $\int_0^\infty \frac{e^{-x}-1}{x^p}\,dx$.

(b) $\int_0^\infty \frac{t^{x-1}}{1+t}\,dt$.

(c) $\int_0^\pi \frac{dx}{(\sin x)^p}$.

(d) $\int_0^\infty \frac{x^p}{1+x+x^2}\,dx$.

(e) $\int_1^2 \left(\frac{2-x}{x-1}\right)^{a-1}\frac{dx}{x}$.

(f) $\int_0^1 \left(\log\frac{1}{u}\right)^{x-1} du$.

(g) $\int_0^{\pi/2} (\tan\theta)^{1-2x}\,d\theta$.

(h) $\int_0^\infty \left(\frac{e^{-t}}{t}-\frac{e^{-at}}{1-e^{-t}}\right) dt$.

4. Where must the point (x, y) lie in the xy-plane if the integral $\int_0^\infty \frac{dt}{t^x(1+t^y)}$ is to be convergent?

5. Answer the question of Exercise 4 for the integral $\int_0^\infty \frac{e^{-xu}-e^{-yu}}{u(1+e^{-u})}\,du$.

6. Show that the integral $\int_{-\infty}^\infty \frac{e^{-xu}-e^{-yu}}{1-e^{-u}}\,du$ is convergent if $0<x<y<1$.

22.2 / THE GAMMA FUNCTION

One very interesting and important improper integral is the following, which defines what is known as the *gamma function*:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt. \tag{22.2–1}$$

If $x \geqq 1$ this is an integral of the first kind, and is convergent, by Example 4, §22.1. If $x < 1$, however, the integral is of mixed type, with a singularity of the integrand at $t = 0$, and we have to consider the integral

$$\int_0^1 t^{x-1}e^{-t}\,dt \tag{22.2–2}$$

of second kind. Since $e^{-t}\to 1$ as $t\to 0$, it is clear that the integrand in (22.2–2) behaves like t^{x-1} near $t = 0$. Now

$$\int_0^1 \frac{dt}{t^{1-x}}$$

is convergent if $1-x<1$, i.e. if $0<x$, and divergent if $1-x\geqq 1$, i.e., if $x\leqq 0$. Therefore, by the analogue of Theorem III, §22.1, for integrals of second kind, (22.2–2) is convergent if and only if $x>0$ (it is improper only if $x<1$).

The integral from $t = 1$ to $t = \infty$ is always convergent, as we saw in Example 4, §22.1. The result is, then, that the integral (22.2–1) defining $\Gamma(x)$ is convergent if and only if $x>0$. Putting $x = 1$ we have

$$\int_0^\infty e^{-t}\,dt = \lim_{T\to\infty}\int_0^T e^{-t}\,dt = \lim_{T\to\infty}(-e^{-T}+1) = 1,$$

or

$$\Gamma(1) = 1. \tag{22.2–3}$$

There is a very simple relation between the values of the gamma function at x and $x+1$. This relation is found by carrying out an integration by parts. We start with

$$\Gamma(x+1) = \int_0^\infty t^x e^{-t}\,dt.$$

Setting $u = t^x$, $dv = e^{-t}\,dt$, we have $du = xt^{x-1}\,dt$, $v = -e^{-t}$,

$$\int_0^T t^x e^{-t}\,dt = \left[-t^x e^{-t}\right]_0^T + \int_0^T xt^{x-1}e^{-t}\,dt;$$

letting $T \to \infty$, we see that

$$\int_0^\infty t^x e^{-t}\,dt = -\lim_{T\to\infty} T^x e^{-T} + 0 + x\int_0^\infty t^{x-1}e^{-t}\,dt. \tag{22.2–4}$$

But

$$\lim_{T\to\infty} T^x e^{-T} = 0,$$

as we see by applying l'Hospital's rule n times to $\dfrac{T^x}{e^T}$, where n is the first integer greater than or equal to x. Therefore, by (22.2–4),

$$\Gamma(x+1) = x\Gamma(x). \tag{22.2–5}$$

From this formula and (22.2–3) we have successively,

$$\begin{aligned}
\Gamma(2) &= 1\cdot\Gamma(1) = 1\\
\Gamma(3) &= 2\cdot\Gamma(2) = 2\cdot 1\\
\Gamma(4) &= 3\cdot\Gamma(3) = 3\cdot 2\cdot 1\\
\Gamma(5) &= 4\cdot\Gamma(4) = 4\cdot 3\cdot 2\cdot 1\\
&\vdots \qquad\quad \vdots \qquad\quad \vdots
\end{aligned}$$

In general we can write

$$\Gamma(n+1) = n! \tag{22.2–6}$$

or

$$\Gamma(n) = (n-1)! \tag{22.2–7}$$

In the ordinary elementary sense $n!$ is defined only if n is a positive integer. But since $\Gamma(x)$ has been defined for every positive x, we see by (22.2–7) and (22.2–3) that it is natural to make the agreement that $0! = 1$. This is customarily done.

The gamma function gives us a convenient method of interpolating between the values of the factorials $n!$, and this is one of the primary reasons for the importance of the gamma function. Just now we shall take for granted that $\Gamma(x)$ is a continuous function, though we can prove this later on (§22.5, following

Theorem VIII). In fact, $\Gamma(x)$ has continuous derivatives of all orders, and is analytic. The derivatives are found by differentiating with respect to the parameter x *under the integral sign* in (22.2–1). Recall that

$$t^x = e^{x \log t}, \frac{d}{dx}(t^x) = \log t \cdot e^{x \log t} = t^x \log t.$$

Thus

$$\Gamma'(x) = \int_0^\infty t^{x-1}(\log t)e^{-t}\,dt. \tag{22.2–8}$$

This is the same procedure as that given in (18.5–2) for *proper* integrals dependent on a parameter. For improper integrals further justification is required; the problem is much the same as the problem of justifying the differentiation of a series term by term. We return to this problem systematically in Theorem X, §22.5; for the present let us proceed with our study of the gamma function. We can differentiate a second time, obtaining

$$\Gamma''(x) = \int_0^\infty t^{x-1}(\log t)^2 e^{-t}\,dt. \tag{22.2–9}$$

The integrals for $\Gamma'(x)$ and $\Gamma''(x)$ are convergent integrals of mixed type, with singularities of the integrand at $t = 0$, and the same is true for the integrals giving all the higher derivatives (see Exercise 1).

It is clear from (22.2–9) that $\Gamma''(x) > 0$, and therefore the curve $y = \Gamma(x)$ is concave upward for all $x > 0$. We also see that $\Gamma(x) > 0$, $\Gamma(1) = \Gamma(2) = 1$. From these facts we see that $\Gamma(x)$ has just one minimum value, and that this occurs for a value of x between 1 and 2. To see how $\Gamma(x)$ behaves as $x \to 0$, we observe that if we integrate only from 0 to 1 in (22.2–1), the result is less than $\Gamma(x)$. Furthermore, e^{-t} is a decreasing function, so that $e^{-t} > e^{-1}$ if $0 \leqq t < 1$. Therefore

$$\Gamma(x) > \int_0^1 t^{x-1}e^{-t}\,dt > e^{-1}\int_0^1 t^{x-1}\,dt = \frac{1}{ex}.$$

It follows from this that $\Gamma(x) \to +\infty$ as $x \to 0^+$. From the information which we have now collected it is possible to show the general character of $\Gamma(x)$ on a graph. We leave it for the student to prepare such a graph for himself.

The formula (22.2–1) does not define a function if $x \leqq 0$. Nevertheless we can define $\Gamma(x)$ for certain negative values of x by using formula (22.2–5). If $-1 < x < 0$, then $0 < x + 1$, so that $\Gamma(x + 1)$ has a meaning already defined. We then define $\Gamma(x)$ by requiring

$$\Gamma(x) = \frac{\Gamma(x+1)}{x}. \tag{22.2–10}$$

Thus, for instance

$$\Gamma(-\tfrac{1}{2}) = -2\Gamma(\tfrac{1}{2}).$$

Now suppose that $-2 < x < -1$; then $-1 < x + 1 < 0$, so that $\Gamma(x + 1)$ is already

defined. We then define $\Gamma(x)$ by (22.2–10), e.g.,

$$\Gamma(-\tfrac{3}{2}) = -\tfrac{2}{3}\Gamma(-\tfrac{1}{2}).$$

This process can evidently be continued, so that we obtain a definition of $\Gamma(x)$ for all values of x except $0, -1, -2, -3, \ldots$, and the equation (22.2–10) holds for all other values of x.

It is easy to see that $\Gamma(x) < 0$ when $-1 < x < 0$, and that $\Gamma(x) \to -\infty$ as $x \to 0^-$ or $x \to -1^+$. We leave it for the student to study the situation when $-2 < x < -1$, $-3 < x < -2$, and so on. A rough graph should be constructed. It will be shown later that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ (see (22.41–6)); from this we may calculate $\Gamma(-\frac{1}{2})$, $\Gamma(-\frac{3}{2})$, etc.

EXERCISES

1. Show that $\int_0^\infty t^{x-1}(\log t)^n e^{-t}\,dt$ is convergent for $n = 1, 2, \ldots$ if $0 < x$.

2. Prepare a graph of $y = \Gamma(x)$, showing the general behavior of the gamma function for $x > 0$ and in the intervals $-1 < x < 0$, $-2 < x < -1$, etc.

3. Show that $\Gamma(x) = 2\int_0^\infty u^{2x-1}e^{-u^2}\,du$ if $x > 0$.

4. Calculate the value in terms of $\sqrt{\pi}$ of

(a) $\int_0^\infty x^2 e^{-x^2}\,dx$, **(b)** $\int_0^\infty x^4 e^{-x^2}\,dx$.

5. If $a > 0$, show that $\int_0^\infty x^{n-1}e^{-ax}\,dx = a^{-n}\Gamma(n)$.
What is the implied restriction on n?

6. Calculate in terms of $\sqrt{\pi}$ the values of

(a) $\int_0^\infty x^{-1/2}e^{-2x}\,dx$, **(b)** $\int_0^\infty x^{3/2}e^{-4x}\,dx$.

7. Show by (22.2–5) that, if $n = 1, 2, \ldots$,

$$\Gamma(n + \tfrac{1}{2}) = \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2^n}\sqrt{\pi}.$$

As a consequence show that

$$\sqrt{\pi}\Gamma(2n+1) = 2^{2n}\Gamma(n + \tfrac{1}{2})\Gamma(n+1),$$

and

$$\sqrt{\pi}\Gamma(2n) = 2^{2n-1}\Gamma(n)\Gamma(n + \tfrac{1}{2}).$$

These formulas suggest the conjecture that perhaps

$$\sqrt{\pi}\Gamma(2x) = 2^{2x-1}\Gamma(x)\Gamma(x + \tfrac{1}{2})$$

not merely for $x = n + \frac{1}{2}$ and $x = n$, where n is a positive integer, but for all $x > 0$. The conjecture is correct, as can be proved by later developments (see Exercise 8, §22.7).

8. Show that $\dfrac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots 2n} = \dfrac{\Gamma(n + \frac{1}{2})}{\sqrt{\pi}\Gamma(n+1)}$.

9. Derive the formula $\Gamma(x) = \displaystyle\int_0^1 \left(\log\frac{1}{u}\right)^{x-1} du$ by putting $u = e^{-t}$ in (20.2–1). Then set $u = v^a$, where $a > 0$, and so find the value of

$$\int_0^1 \left(\log\frac{1}{v}\right)^{x-1} v^{a-1}\,dv,$$

where $x > 0$.

10. Utilize the results of Exercise 9 to show that

(a) $\int_0^1 \left(\frac{\log(1/t)}{t}\right)^{1/2} dt = \sqrt{2\pi}$,

(b) $\int_0^1 \left(\frac{t}{\log(1/t)}\right)^{1/2} dt = \sqrt{\frac{2\pi}{3}}$.

22.3 / ABSOLUTE CONVERGENCE

An improper integral of first kind, $\int_a^\infty f(x)\,dx$, is called *absolutely convergent* if the integral $\int_a^\infty |f(x)|\,dx$ is convergent. Exactly the same definition is applied to integrals of second kind, and to integrals of mixed type. The switch from $f(x)$ to $|f(x)|$ corresponds exactly to the switch from $\Sigma\, a_n$ to $\Sigma\, |a_n|$ in defining absolute convergence of the infinite series. If an integral is convergent, but not absolutely convergent, it is called *conditionally convergent.*

The following theorem corresponds to Theorem IX, §19.3:

THEOREM V. *If the integral $\int_a^\infty |f(x)|\,dx$ is convergent, so is $\int_a^\infty f(x)\,dx$.*

In other words, if an integral is absolutely convergent, it is convergent.

Proof of the theorem. First of all we observe that

$$0 \leqq |f(x)| - f(x) \leqq 2|f(x)|. \tag{22.3–1}$$

Both parts of this double inequality may be checked by considering separately the cases when $f(x) \geqq 0$ and $f(x) < 0$. Now let $g(x) = |f(x)| - f(x)$. Since $\int_a^\infty |f(x)|\,dx$ is assumed to be convergent, the integral with $2|f(x)|$ as integrand is also convergent. Then, by (22.3–1) and Theorem II, §22.1, we see that $\int_a^\infty g(x)\,dx$ is convergent. But $f(x) = |f(x)| - g(x)$, and therefore $\int_a^\infty f(x)\,dx$ is convergent, for sums and differences of convergent integrals are convergent, as may be seen directly from the definition of convergence. (What theorem about limits is used at this last step in the argument?)

The theorem and its proof apply to integrals of the second kind; only the limits of integration have to be changed.

To test whether an integral is absolutely convergent, we can apply the methods of §§22.1, 22.11, since the integrand $|f(x)|$ is never negative. If an integral is conditionally convergent, the demonstration of its convergence is usually a more delicate matter. Many of the instances of practical importance can be handled by the following theorem, which is analogous to Dirichlet's test for series (§19.7).

THEOREM VI. *Consider an improper integral of first kind of the form*

$$\int_a^\infty \phi(t) f(t)\,dt, \tag{22.3–2}$$

where the functions ϕ and f satisfy the conditions:

(a) $\phi'(t)$ is continuous, $\phi'(t) \leqq 0$, and $\lim_{t\to\infty} \phi(t) = 0$,

(b) $f(t)$ is continuous, and the integral

$$F(x) = \int_a^x f(t)\,dt \tag{22.3–3}$$

is bounded for all $x \geqq a$. Then the integral (22.3–2) is convergent.

Proof. We note that $F'(x) = f(x)$. Therefore, integrating by parts and noting that $F(a) = 0$, we have

$$\int_a^x \phi(t)f(t)\,dt = \int_a^x \phi(t)F'(t)\,dt = \phi(x)F(x) - \int_a^x \phi'(t)F(t)\,dt. \tag{22.3–4}$$

Let us suppose that M is a bound for $|F(x)|$, that is, $|F(x)| \leqq M$. Then $|\phi(x)F(x)| \leqq |\phi(x)|M$, and so $\phi(x)F(x) \to 0$ as $x \to \infty$, since $\phi(x) \to 0$ by hypothesis. It then follows from (22.3–4) that (22.3–2) is convergent, provided we can show that the integral

$$\int_a^\infty \phi'(t)F(t)\,dt \tag{22.3–5}$$

is convergent. This integral is in fact absolutely convergent. For, since $\phi'(t) \leqq 0$,

$$|\phi'(t)F(t)| = -\phi'(t)|F(t)| \leqq -M\phi'(t).$$

It is then enough to show that

$$\int_a^\infty -M\phi'(t)\,dt \tag{22.3–6}$$

is convergent; then (22.3–5) will be absolutely convergent, by Theorem II, §22.1. Now

$$\int_a^x -M\phi'(t)\,dt = -M\phi(x) + M\phi(a) \to M\phi(a)$$

as $x \to \infty$, by condition (a). Thus (22.3–6) is convergent, and the proof is complete.

Example 1. The integral $\int_0^\infty (1/t)\sin t\,dt$ is convergent. (There is no singularity at $t = 0$; see the remark in Example 2, §22.11.)

Here we take

$$\phi(t) = \frac{1}{t}, \quad f(t) = \sin t.$$

Then

$$F(x) = \int_0^x \sin t\,dt = 1 - \cos x.$$

The conditions of Theorem VI are fulfilled, for $\phi'(t) = -1/t^2$ and $0 \leqq F(x) \leqq 2$. Therefore, the given integral is convergent. It is *not* absolutely convergent, however, as is not difficult to see (for a hint on this see Exercise 6).

Example 2. The integral $\int_0^\infty \sin u^2\, du$ is convergent. We make a change of variable,

$$u = \sqrt{t}, \qquad du = \frac{dt}{2\sqrt{t}}.$$

Then

$$\int_0^x \sin u^2\, du = \tfrac{1}{2}\int_0^{x^2} \frac{\sin t}{\sqrt{t}}\, dt. \tag{22.3–7}$$

Now the integral $\int_0^\infty \frac{\sin t}{\sqrt{t}}\, dt$ is convergent, by an application of Theorem VI very much the same as in Example 1. Hence, letting $x \to \infty$ in (22.3–7), we see that

$$\int_0^\infty \sin u^2\, du = \tfrac{1}{2}\int_0^\infty \frac{\sin t}{\sqrt{t}}\, dt.$$

The integral of Example 2 illustrates the remark made at the end of §22.1, for $\sin u^2$ does not approach zero as $u \to \infty$.

EXERCISES

1. Examine the convergence of each of the following integrals. Where possible, prove that the integral is absolutely convergent. If it is necessary to use Theorem VI, give details of the application of the theorem in the particular case. Proofs that an integral is not absolutely convergent need not be given.

(a) $\int_0^\infty \frac{x \cos x}{a^2 + x^2}\, dx.$

(b) $\int_0^\infty \frac{\cos x}{(1+x^2)(4+x^2)}\, dx.$

(c) $\int_0^\infty \frac{\cos x}{\sqrt{1+x^3}}\, dx.$

(d) $\int_2^\infty \frac{\sin x}{\log x}\, dx.$

(e) $\int_2^\infty \frac{\log(\log x)}{\log x} \cos x\, dx.$

(f) $\int_{-\infty}^\infty \frac{x^2 - x + 2}{x^4 + 10x^2 + 9}\, dx.$

(g) $\int_0^\infty \frac{x(x^2+1)\sin x}{x^4 - x^2 + 1}\, dx.$

(h) $\int_0^\infty \frac{\sin \pi x}{x(x^2-1)}\, dx.$

(i) $\int_{-\infty}^\infty \frac{x \sin x}{1+x^2}\, dx.$

(j) $\int_0^\infty \frac{\sin ax}{1+x^2}\, dx.$

(k) $\int_{-\infty}^\infty \frac{\cos x}{1-x^2}\, dx.$

(l) $\int_{-\infty}^\infty \frac{\cos x}{x^{2/3}}\, dx.$

(m) $\int_0^\infty \frac{\sin x \sin mx}{x}\, dx.$

(n) $\int_0^\infty \frac{e^{-x} \sin x}{x}\, dx.$

(o) $\int_0^\infty \frac{\sin x}{e^{2x} - 1}\, dx.$

(p) $\int_0^\infty \frac{\sin(\sin x)\cos x}{x}\, dx.$

(q) $\int_0^\infty \frac{x+1}{x^{3/2}} \sin x\, dx.$

(r) $\int_0^\infty e^{-x} \cos 2x\, dx.$

2. Show that $\int_0^\infty \frac{\sin ax}{x^p}\,dx$ is convergent if $0<p<2$, and absolutely convergent if $1<p<2$.

3. Show that $\int_0^\infty \frac{1-\cos ax}{x^p}\,dx$ is convergent if $1<p<3$. Is it absolutely convergent for any of these values of p?

4. Show that $\int_0^\infty \frac{\sin x(1-\cos x)}{x^p}\,dx$ is convergent if $0<p<4$. Is it absolutely convergent for any of these values of p?

5. Show that $\int_0^\infty \cos(x^2)\,dx$ and $\int_0^\infty x\cos(x^4)\,dx$ are convergent. Note that the integrand in the second integral is unbounded.

6. Show that

$$\left|\int_{n\pi}^{(n+1)\pi} \frac{\sin x}{x}\,dx\right| > \frac{2}{(n+1)\pi}, \quad n = 0, 1, \ldots.$$

Use this result to prove that $\int_0^\infty \frac{\sin x}{x}\,dx$ is not absolutely convergent.

22.4 / IMPROPER MULTIPLE INTEGRALS. FINITE REGIONS

Consider first the case of an integral

$$\iint_R f(x, y)\,dA, \tag{22.4–1}$$

where R is a closed bounded region, f is continuous in R except at one point (x_0, y_0), and the behavior of f at that point is such that the function is not integrable over R in the sense of §18.6. The cases of greatest practical importance are those in which $f(x, y)$ either becomes infinite or has a factor which becomes infinite as $(x, y)\to(x_0, y_0)$, e.g.,

$$f(x, y) = \frac{1}{r} \quad \text{or} \quad f(x, y) = \frac{x - x_0}{r^2},$$

where

$$r = [(x - x_0)^2 + (y - y_0)^2]^{1/2}. \tag{22.4–2}$$

To define what we mean by the convergence or divergence of the integral (22.4–1) we proceed as follows: Let R' be a region derived from R by discarding a small region ΔR having the point (x_0, y_0) in its interior (R' is the shaded portion of R in Fig. 181). No restriction is placed on the shape of ΔR except that it be a Riemann region in the sense defined in §18.6. Of course, we also assume that R is a Riemann region. Let d be the maximum diameter of ΔR, that is, the distance between two points of ΔR which are as far apart as it is

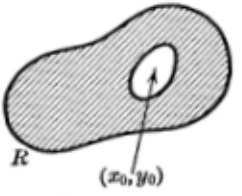


Fig. 181.

possible for them to be when both are in ΔR. We then consider the integral

$$\iint_{R'} f(x, y)\, dA$$

over the region R'. If this integral approaches a limiting value as $d \to 0$, the value of the limit is denoted by

$$\iint_{R} f(x, y)\, dA = \lim_{d \to 0} \iint_{R'} f(x, y)\, dA \tag{22.4–3}$$

and the integral (22.4–1) is said to be *convergent*. If the limit does not exist, the integral is called *divergent*.

We shall confine our attention to the case when the integral of $|f(x, y)|$ is convergent. In this case the integral of $f(x, y)$ is also convergent, as may be proved in much the same way that we proved Theorem V, §22.3. The integral of $f(x, y)$ is then said to be *absolutely convergent*.

The comparison principle is valid in the following form for improper double integrals: *If* $|f(x, y)| \leqq g(x, y)$, *and if* $\iint_R g(x, y)\, dA$ *is convergent, then* $\iint_R f(x, y)\, dA$ *is absolutely convergent.*

If one is trying to determine whether or not the integral of a *positive* function is convergent, it is not necessary to consider regions ΔR of *arbitrary* shape in examining the limit (22.4–3). It is sufficient to confine one's attention to regions of one particular shape, say circles (or, anternatively, squares) with center at (x_0, y_0). Let δ be the radius of such a circle, and let R' be the region which is obtained by deleting from R the interior of the circle in question. Then, if $g(x, y) \geqq 0$, and if the limit of $\iint_{R'} g(x, y)\, dA$ exists as $\delta \to 0$, the integral of g over R will be convergent and will be equal to the limit just mentioned. Proof of this fact is indicated in Exercise 8.

All the foregoing remarks apply without essential change to the case in which (x_0, y_0) is on the boundary of R instead of in the interior of R. The only difference is that we discard from R that part of R which is in a small region having (x_0, y_0) in its interior.

Example 1. Let r be defined by (22.4–2) and consider the integral

$$\iint_{R} \frac{1}{r^m}\, dA, \tag{22.4–4}$$

it being understood that (x_0, y_0) is a point of R and $m > 0$, so that the integral is improper. We shall show that it is convergent if $m < 2$. Since the only singularity is at (x_0, y_0), the typical difficulty is exposed in the case that R is a circle with center at (x_0, y_0), and of radius c. Let us delete a small concentric circle of radius δ, so that R' is the annulus between these circles. In evaluating the integral we may as well assume that (x_0, y_0) is the origin, since the value of the integral is not

affected by the location of the axes. Then, with the use of polar co-ordinates,

$$\iint_{R'} \frac{1}{r^m}\,dA = \int_0^{2\pi} d\theta \int_\delta^c r^{-m} r\,dr = \frac{2\pi}{2-m}[c^{2-m} - \delta^{2-m}].$$

Since $m < 2$ we see that

$$\lim_{\delta \to 0} \iint_{R'} \frac{1}{r^m}\,dA = \frac{2\pi}{2-m} c^{2-m}.$$

Thus the integral (22.4–4) exists if R is a circle with center at (x_0, y_0) and $m < 2$. For regions of other shape, and for (x_0, y_0) located on the boundary of R, the difficulty can easily be resolved in terms of the case we have treated.

Example 2. The integral

$$\iint_R \frac{x - x_0}{r^2}\,dA \tag{22.4–5}$$

is absolutely convergent; for $|x - x_0| \leqq r$, by (22.4–2), and so

$$\left|\frac{x - x_0}{r^2}\right| \leqq \frac{1}{r},$$

whence, by the comparison-test principle and the result of Example 1, the asserted result follows.

Similar considerations apply to improper triple integrals. Improper multiple integrals in which the integrand has just one singular point in the region of integration occur typically in the theory of force fields governed by the inverse-square law, e.g., gravitational or electrostatic fields. From a purely mathematical point of view the study of such fields belongs to what is called *potential theory*. Let R be a bounded closed region in 3-space, and let it be filled with mass of density $\mu(x, y, z)$. If Q is the point (x_0, y_0, z_0), and

$$r = [(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2]^{1/2},$$

the *Newtonian potential* at Q produced by the total mass is

$$\phi(x_0, y_0, z_0) = \iiint_R \frac{\mu}{r}\,dV, \tag{22.4–6}$$

and the *gravitational field* at Q is a vector $\mathbf{F}$ whose first component (in the x-direction) is

$$F_1 = \iiint_R \mu \frac{x - x_0}{r^3}\,dV, \tag{22.4–7}$$

with similar formulas for F_2 and F_3. If Q is a point of R these are improper

integrals; they are absolutely convergent if $\mu(x, y, z)$ is bounded and integrable. One can also discuss the potential and field due to distribution of mass on surfaces, and in particular on plane regions. If Q is a point in the plane region, the double integral defining the potential at Q is absolutely convergent, but the integrals defining the components of the field at Q are divergent except in very special cases.

EXERCISES

1. Let r be the circular region of radius 1 with center at the origin. Determine, for each of the following integrals, if it is convergent or divergent. Polar co-ordinates are denoted by r, θ.

(a) $\iint_R \frac{x^2}{(x^2+y^2)^{7/4}}\,dA$. (d) $\iint_R \log\frac{1}{r}\,dA$.

(b) $\iint_R \frac{x^2y^2}{(x^2+y^2)^3}\,dA$. (e) $\iint_R \frac{1}{r}\log\frac{1}{r}\,dA$.

(c) $\iint_R \frac{\sin\frac{\theta}{2}}{r^{3/2}}\,dA$. (f) $\iint_R \frac{x^2+y^4}{(x^2+y^2)^{3/2}}\,dA$.

2. Let $f(x, y)$ be continuous in R except at (x_0, y_0), and bounded. Let $r^2 = (x-x_0)^2+(y-y_0)^2$. Show that $\iint_R \frac{f(x,y)}{r^m}\,dA$ is absolutely convergent if $m<2$.

3. State and prove a result corresponding to that of Exercise 2, for improper triple integrals.

4. If R is a closed and bounded plane region containing the origin, for what values of p is the integral

$$\iint_R \frac{xy}{r^p}\,dA$$

certainly absolutely convergent?

5. Let R be the unit sphere $x^2+y^2+z^2 \leqq 1$, and let $r^2 = x^2+y^2+z^2$. Find the values of those among the following integrals which are convergent. If a literal constant appears, indicate the restrictions you place on it to insure convergence.

(a) $\iiint_R \frac{1}{r}\,dV$. (d) $\iiint_R \frac{z^2}{r^n}\,dV$.

(b) $\iiint_R r^{-n}\,dV$. (e) $\iiint_R \frac{x^2+y^4+z^6}{r^5}\,dV$.

(c) $\iiint_R \frac{x^2+y^2}{r^{9/2}}\,dV$. (f) $\iiint_R \frac{x^2y^2z^2}{r^{17/2}}\,dV$.

6. (a) Is the integral $\iint_R \frac{dA}{[(x-1)^2+y^2]^{1/2}}$ convergent or divergent, where R is the region $x^2+y^2 \leqq 1$?

(b) What if the exponent $\frac{1}{2}$ is replaced by m?

(c) For what values of p is $\iint_R \frac{x-1}{[(x-1)^2+y^2]^p}\,dA$ certainly convergent?

7. Let R, f, ΔR, d have the meanings used in the discussion of (22.4–3).

(a) Let W be a subregion of R which contains (x_0, y_0) and all the points of R in some neighborhood of (x_0, y_0). Show that $\iint_R f(x, y)\,dA$ is convergent if and only if $\iint_W f(x, y)\,dA$ is convergent, and then

$$\iint_R f(x, y)\,dA = \iint_{R-W} f(x, y)\,dA + \iint_W f(x, y)\,dA,$$

where $R - W$ is the region which results by removing W from R.

SUGGESTION: Consider

$$\iint_{R-\Delta R} f(x, y)\,dA - \iint_{W-\Delta R} f(x, y)\,dA,$$

where ΔR is so small that it is contained in W.

(b) Deduce from the result in (a) that $\lim_{d\to 0} \iint_{\Delta R} f(x, y)\,dA = 0$ if $\iint_R f(x, y)\,dA$ is convergent.

8. Suppose $g(x, y)$ is nonnegative and continuous in R except at (x_0, y_0). Let $\{W_n\}$ be a sequence of subregions of the type of W in Exercise 7(a). Suppose W_1 contains W_2, W_2 contains W_3, etc., and that $d_n \to 0$ as $n \to \infty$, where d_n is the maximum diameter of W_n. Finally, if $R_n = R - W_n$, assume that

$$\lim_{n\to\infty} \iint_{R_n} g(x, y)\,dA \text{ exists, } = I.$$

Prove that $\iint_R g(x, y)\,dA$ is convergent, with value I. SUGGESTION: For a given m, choose ΔR so small that it is contained in W_m. Then choose n so that W_n is contained in ΔR. Now show that

$$\iint_{R-\Delta R} g(x, y)\,dA - I = \iint_{W_m-\Delta R} g(x, y)\,dA + \iint_{R_m} g(x, y)\,dA - I,$$

and that

$$\iint_{W_m-\Delta R} g(x, y)\,dA \leqq \iint_{R_n} g(x, y)\,dA - \iint_{R_m} g(x, y)\,dA.$$

From here it is easy to complete the proof. Write out the whole argument carefully.

9. Prove the validity of the comparison principle stated in the text.

10. Suppose $\iint_R |f(x, y)|\, dA$ is convergent. **(a)** Let $\{W_n\}$ be a sequence of regions of the type described in Exercise 8. By considering the inequality $0 \leqq |f(x, y)| - f(x, y) \leqq 2|f(x, y)|$, show that

$$\lim_{n\to\infty} \iint_{R_n} f(x, y)\, dA \text{ exists, } = I, \text{ say.}$$

(b) If ΔR is so small that it is contained in W_n, show that

$$\left| \iint_{R-\Delta R} f(x, y)\, dA - I \right| \leqq \iint_{W_n} |f(x, y)|\, dA + \left| \iint_{R_n} f(x, y)\, dA - I \right|,$$

and so deduce that $\iint_R f(x, y)\, dA = I$. Note the result of Exercise 7(**b**).

22.41 / IMPROPER MULTIPLE INTEGRALS. INFINITE REGIONS

In this section we are going to consider integrals of the type

$$\iint_R f(x, y)\, dA, \tag{22.41–1}$$

where R is an unbounded region, such as the first quadrant ($x \geqq 0$, $y \geqq 0$), an infinite strip (say between the lines $y = 0$, $y = 1$), or even the entire xy-plane. In any case we assume that the boundary of R is regular enough to cause us no trouble, and we tacitly assume the same about all other regions subsequently mentioned in this section. We also assume that f is integrable, in the ordinary proper sense, over any bounded closed subregion of R. Thus the only problem arises from the fact that R is an infinite region.

We might proceed as follows: Let us take a sequence of concentric circles with centers at O and radii becoming infinite. Let $\{R_n\}$ be the sequence of regions obtained by considering the part of R inside or on the nth circle. Then define

$$\iint_R f(x, y)\, dA = \lim_{n\to\infty} \iint_{R_n} f(x, y)\, dA \tag{22.41–2}$$

provided this limit exists. This seems like a reasonable procedure, as indeed it is, under suitable conditions. But, one may ask, why not use squares instead of circles? Or, why not regions more general than just circles and squares? Will one get the same limit in all cases? The answer is, perhaps not, unless some further restriction is placed on the function f. Suppose, however, that we assume $f(x, y) \geqq 0$. Then it may be shown that, if the limit (22.4–2) exists when we use circles to get R_n, it also exists *and has the same value* when we use squares or any other sequence of regions subject to reasonable restrictions. To understand

the essential principle of what is involved here, let $\{R_n\}$ be a sequence of regions formed by using circles as already described, and let $\{S_n\}$ be a sequence obtained in the same way, using squares (instead of circles) with center at O and length of side becoming infinite. Also, let

$$I_n = \iint_{R_n} f(x, y)\, dA, \qquad J_n = \iint_{S_n} f(x, y)\, dA,$$

and write

$$I = \lim_{n\to\infty} I_n.$$

Since each region R_n contains its predecessor R_{n-1}, and since $f(x, y) \geqq 0$, it is clear that

$$I_1 \leqq I_2 \leqq I_3 \leqq \cdots \leqq I.$$

Likewise

$$J_1 \leqq J_2 \leqq J_3 \leqq \cdots.$$

Now *any* one of the squares is contained in *some* one of the circles, and vice versa. Therefore, given n, there is some N such that

$$J_n \leqq I_N \leqq I,$$

and given m, there is some M such that

$$I_m \leqq J_M.$$

These inequalities show that the sequences $\{I_n\}$ and $\{J_n\}$ both have the same limit.

These arguments show that it is sufficient to define the convergence of the integral (22.41–1) in the manner already indicated if $f(x, y) \geqq 0$.

We define absolute convergence of (22.41–1) in the usual way. It is then easy to see that the definition (22.41–2) is satisfactory for absolutely convergent integrals. This is all we shall deal with. In practice we shall use regions R_n obtained by using squares, circles, or other regions as convenience dictates.

Example 1. If R is the entire xy-plane, and $a > 0$,

$$\iint_R (x^2 + y^2 + a^2)^{-3/2}\, dA = \frac{2\pi}{a}. \tag{22.41–3}$$

Using polar co-ordinates, and letting R' denote an arbitrary circular region $x^2 + y^2 \leqq c^2$, we have

$$\iint_{R'} (x^2 + y^2 + a^2)^{-3/2}\, dA = \int_0^{2\pi} d\theta \int_0^c (r^2 + a^2)^{-3/2} r\, dr$$

$$= 2\pi\left\{-(c^2 + a^2)^{-1/2} + \frac{1}{a}\right\}.$$

Letting $c \to \infty$, we see that there is a limit, so that the integral over all of R is convergent, and given by (22.41–3).

Example 2. If R is the entire first quadrant in the xy-plane,

$$\iint_R e^{-x^2-y^2}\,dA = \frac{\pi}{4}. \tag{22.41–4}$$

To obtain this result, let R' be the part of R in the circle $x^2 + y^2 \leqq c^2$ (see Fig. 182).

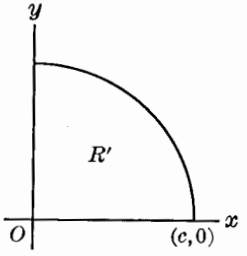


Fig. 182.

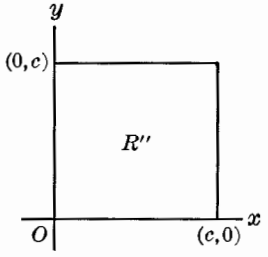


Fig. 183.

Using polar co-ordinates, we have

$$\iint_{R'} e^{-x^2-y^2}\,dA = \int_0^{\pi/2} d\theta \int_0^c e^{-r^2} r\,dr$$

$$= \frac{\pi}{2}\left[-\tfrac{1}{2}e^{-r^2}\right]_0^c = \frac{\pi}{4}(1 - e^{-c^2}).$$

Hence

$$\iint_R e^{-x^2-y^2}\,dA = \lim_{c\to\infty} \iint_{R'} e^{-x^2-y^2}\,dA = \frac{\pi}{4}.$$

Let us consider what happens if we attempt to evaluate the integral in Example 2 by using squares instead of circles in the limiting process. Let R'' be the part of R inside the square with sides along the lines $x = 0$, $x = c$, $y = 0$, $y = c$ (see Fig. 183). Then

$$\iint_{R''} e^{-x^2-y^2}\,dA = \int_0^c dx \int_0^c e^{-x^2-y^2}\,dy$$

$$= \int_0^c e^{-x^2}\,dx \int_0^c e^{-y^2}\,dy = \left(\int_0^c e^{-x^2}\,dx\right)^2.$$

At the last step we have changed y to x under the integral sign, this being permitted since the value of the integral does not depend on the variable of integration.

Now we know from the earlier discussion that the value of the improper double integral is given just as well by the limiting process with squares as by that with circles. Therefore, in view of (22.41–4), we have

$$\frac{\pi}{4} = \lim_{c\to\infty} \left(\int_0^c e^{-x^2}\, dx \right)^2$$

or

$$\int_0^\infty e^{-x^2}\, dx = \frac{\sqrt{\pi}}{2}. \tag{22.41–5}$$

This last result probably appears to the student as an unexpected by-product of our study of improper double integrals. It is a very important result, nevertheless. In fact, it was very largely to get this result that we developed this particular section of the text. Our immediate interest in the integral (22.41–5) is because of its connection with the gamma function. This connection is worked out in the next paragraph, by a change of variable. But the integral (22.41–5) is also of interest in statistics and elsewhere.

In (22.41–5) let us make the substitution

$$x = t^{1/2}, \qquad dx = \tfrac{1}{2}t^{-1/2}\, dt.$$

Then

$$\frac{\sqrt{\pi}}{2} = \tfrac{1}{2}\int_0^\infty t^{-1/2}e^{-t}\, dt = \tfrac{1}{2}\Gamma(\tfrac{1}{2}),$$

by (22.2–1). Thus we have

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}. \tag{22.41–6}$$

EXERCISES

1. Let the xy-plane be given a distribution of static electricity of constant density σ per unit area. Show that the resulting field at the point $(0, 0, a)$ on the positive z-axis is in the z-direction and of magnitude $2\pi\sigma$. Note that the result is independent of the distance a. The field at a point is the force which would be exerted on a unit positive charge there.

2. The electrostatic potential at $(0, 0, a)$ resulting from a charge of constant density σ on a bounded region R in the xy-plane is

$$\iint_R \frac{\sigma\, dA}{(x^2 + y^2 + a^2)^{1/2}}.$$

Would this integral be convergent if R were taken to be the whole xy-plane instead of a bounded region?

3. Suppose that, in Exercise 1, the charge is placed only on the strip of the xy-plane between the lines $x = \pm b$, where $b > 0$. Show that the electrostatic field at $(0, 0, a)$ is $4\tan^{-1}(b/a)$. One method of solving calls for the formula

$$\int \frac{dx}{(A + Bx^2)(C + Dx^2)^{1/2}} = \frac{1}{A}\left(\frac{A}{BC - AD}\right)^{1/2} \tan^{-1}\left[x\left(\frac{BC - AD}{A\{C + Dx^2\}}\right)^{1/2}\right] + \text{Const.},$$

which is not found in all integral tables.

4. Is the integral $\iint_R \frac{dA}{1+x^2+y^2}$ convergent, if R is the first quadrant of the xy-plane?

5. Show that, if R is the first quadrant of the xy-plane, the integral $\iint_R ye^{-y^2(1+x^2)}\,dA$ is convergent. One method is to integrate over the square $0 \leqq x \leqq a$, $0 \leqq y \leqq a$, and study what happens as $a \to \infty$. It is possible to show that the integral over the square is equal to $\frac{1}{2}\tan^{-1} a - \frac{ae^{-a^2}}{2}\int_0^{a^2} \frac{e^{-t^2}\,dt}{a^2+t^2}$. From this one can find the value of the double integral over R.

6. If the integrals $\int_0^\infty f(x)\,dx$ and $\int_0^\infty g(y)\,dy$ are convergent, and if $f(x)$ and $g(y)$ are never negative, show that the double integral $\iint_R f(x)g(y)\,dA$, over the entire first quadrant, is convergent and equal to the product

$$\left(\int_0^\infty f(x)\,dx\right)\left(\int_0^\infty g(y)\,dy\right).$$

This conclusion may *not* be correct if either or both of the functions $f(x)$, $g(y)$ can assume both positive and negative values. An illustration is afforded by $\int_0^\infty (1/x)\sin x\,dx$ and $\int_0^\infty e^{-y}\,dy$. If we attempt to compute the double integral using squares $0 \leqq x \leqq a$, $0 \leqq y \leqq a$, we obtain the product of the two integrals as the limit when $a \to \infty$. But if we integrate over the region R_n defined as we shall indicate, it can be shown by estimation of the integrals over the blocks composing R_n that the double integral over R_n is greater than $\frac{\pi}{\log(\log n)}\log\frac{2n+1}{3} + c_n$, where $c_n \to -\frac{\pi}{2}$ as $n \to \infty$; hence the integral tends to $+\infty$. The blocks composing R_n are $k\pi \leqq x \leqq (k+1)\pi$, $0 \leqq y \leqq \log n$ if $k = 0, 2, \ldots, 2n$ and $k\pi \leqq x \leqq (k+1)\pi$, $0 \leqq y \leqq \log[\log(\log n)]$ if $k = 1, 3, \ldots, 2n-1$.

22.5 / FUNCTIONS DEFINED BY IMPROPER INTEGRALS

There are many important problems in analysis in which one encounters improper integrals depending on a parameter; if such an integral is convergent, it defines a function of the parameter. A typical situation is that of the integral (22.2–1) defining the gamma function. As a general type of problem suppose we have

$$F(x) = \int_c^\infty f(t, x)\,dt, \tag{22.5–1}$$

where the integral is of first kind and convergent for each value of x in some interval. In practice it is important to know whether or not F is continuous; also, under certain conditions it is possible to deal with integration and differentiation of $F(x)$ according to the formulas

$$F'(x) = \int_c^\infty \frac{\partial f}{\partial x}\,dt,$$

$$\int_a^b F(x)\,dx = \int_c^\infty dt \int_a^b f(t, x)\,dx.$$

It is important to know when these formulas are legitimate. The treatment of

these questions is quite analogous to the treatment of the corresponding questions in the case of functions defined by infinite series. The key concept is that of *uniform convergence*, just as it was in Chapter 20.

Definition. *Suppose that the integral* (22.5–1) *is convergent for each x in the interval $a \leqq x \leqq b$. We say that it is* uniformly convergent *on that interval if to each $\epsilon > 0$ corresponds a number s_0, depending (possibly) on ϵ, but not on x, such that*

$$\left| F(x) - \int_c^s f(t, x)\, dt \right| < \epsilon$$

whenever $a \leqq x \leqq b$ and $s_0 \leqq s$.

This definition should be compared with the definition of uniform convergence of an infinite sequence of functions, in §20.1. The $f_n(x)$ of that definition is analogous to $\int_c^s f(t, x)\, dt$ in our present definition, and n is analogous to s. The parallel between Chapter 20 and our present work is so close that we shall in the main merely state the important theorems and omit the proofs. A student who understands Chapter 20 thoroughly will have little additional difficulty in appreciating the present section. He must also be familiar with §18.5, in which similar problems for *proper* integrals depending on a parameter are treated (see particularly Theorems XIII and XIV of §18.5).

THEOREM VII. *Let $f(t, x)$ be continuous in the two variables t, x when $c \leqq t$ and $a \leqq x \leqq b$, and let the integral*

$$F(x) = \int_c^\infty f(t, x)\, dt \tag{22.5–2}$$

be uniformly convergent on the interval $[a, b]$. Then F is continuous on the interval.

This corresponds to Theorem III, §20.3.

There is a convenient practical test for uniform convergence, analogous to the M-test for series (Theorem II, §20.2). It is the following:

THEOREM VIII. *Suppose $g(t) \geqq 0$, and that the integral of first kind*

$$\int_c^\infty g(t)\, dt$$

is convergent. Also suppose $|f(t, x)| \leqq g(t)$ when $a \leqq x \leqq b$, for all values of t beyond some fixed value ($t_0 \leqq t$). Then the integral (22.5–2) *is uniformly convergent on the interval $[a, b]$.*

There are analogous theorems for improper integrals of the second kind, with the obvious modifications in notation.

As an application, we can now prove that $\Gamma(x)$ is continuous when $x > 0$. We write

$$\Gamma(x) = \int_0^1 t^{x-1}e^{-t}\,dt + \int_1^\infty t^{x-1}e^{-t}\,dt. \tag{22.5–3}$$

We prove that $\Gamma(x)$ is continuous on every interval $a \leqq x \leqq b$ with $0 < a < b$. Having fixed a and b, we observe that, if $a \leqq x \leqq b$, and $t \geqq 1$, $0 < t^{x-1}e^{-t} \leqq t^{b-1}e^{-t}$. Since

$$\int_1^\infty t^{b-1}e^{-t}\,dt$$

is convergent, the uniform convergence of the second integral in (22.5–3) follows by Theorem VIII. If $x > 1$, the first integral in (22.5–3) is proper, and is a continuous function of x by Theorem XIII, §18.5. When the integral is improper, however, we use the theorem corresponding to Theorem VIII for integrals of second kind, noting that

$$0 < t^{x-1}e^{-t} \leqq t^{a-1}e^{-t}$$

if $0 < t \leqq 1$ and $0 < a \leqq x \leqq b$, and that

$$\int_0^1 t^{a-1}e^{-t}\,dt$$

is convergent. Thus $\Gamma(x)$, being the sum of two continuous functions, is continuous.

THEOREM IX. *Under the hypothesis of Theorem VII it is true that*

$$\int_a^b F(x)\,dx = \int_c^\infty dt \int_a^b f(t, x)\,dx. \tag{22.5–4}$$

This corresponds to Theorem IV, §20.4.

THEOREM X. *Let the integral*

$$F(x) = \int_c^\infty f(t, x)\,dt$$

be convergent when $a \leqq x \leqq b$. Let the partial derivative $\dfrac{\partial f}{\partial x}$ be continuous in the two variables t, x when $c \leqq t$ and $a \leqq x \leqq b$, and let the integral $\displaystyle\int_c^\infty \frac{\partial f}{\partial x}\,dt$ converge uniformly on $[a, b]$. Then $F(x)$ has a derivative given by

$$F'(x) = \int_c^\infty \frac{\partial f(t, x)}{\partial x}\,dt. \tag{22.5–5}$$

We shall prove this theorem, which corresponds to Theorem V, §20.5. We

set

$$G(u) = \int_c^\infty \frac{\partial f(t, u)}{\partial u}\, dt.$$

Applying Theorem IX, we have

$$\begin{aligned}\int_a^x G(u)\, du &= \int_c^\infty dt \int_a^x \frac{\partial f(t, u)}{\partial u}\, du \\ &= \int_c^\infty [f(t, x) - f(t, a)]\, dt \\ &= F(x) - F(a).\end{aligned}$$

Since G is a continuous function (by Theorem VII), we know from the last formula that

$$F'(x) = G(x)$$

(Theorem VII, §1.52). This result is equivalent to (22.5–5).

With Theorem X we can justify the formula (22.2–8) for $\Gamma'(x)$, for the integrals

$$\int_0^1 t^{x-1}(\log t)e^{-t}\, dt, \quad \int_1^\infty t^{x-1}(\log t)e^{-t}\, dt$$

are uniformly convergent, as may be shown by arguments very similar to those employed in connection with (22.5–3). The procedure extends at once to the higher derivatives, so that, for $n = 1, 2, \ldots$ and $x > 0$

$$\Gamma^{(n)}(x) = \int_0^\infty t^{x-1}(\log t)^n e^{-t}\, dt. \tag{22.5–6}$$

The theorems of this section have applications in justifying the steps in various ingenious methods by which one can sometimes find the value of an improper integral.

Example 1. Show that if $0 < a < b$,

$$\int_0^\infty \frac{e^{-at} - e^{-bt}}{t}\, dt = \log \frac{b}{a}. \tag{22.5–7}$$

The integral is of first kind, for the integrand approaches the finite limit $b - a$ as $t \to 0$. The ingenious device here consists in starting from the formula

$$\frac{1}{x} = \int_0^\infty e^{-xt}\, dt, \tag{22.5–8}$$

which is valid if $x > 0$; this formula is obtained by direct evaluation, using the indefinite integral

$$\int e^{-xt}\, dt = -\frac{1}{x} e^{-xt} + C.$$

The integral (22.5–8) is uniformly convergent on the interval $a \leqq x \leqq b$, by Theorem VIII and the fact that $e^{-xt} \leqq e^{-at}$ when $t \geqq 0$. Hence, by (22.5–4),

$$\int_a^b \frac{dx}{x} = \int_0^\infty dt \int_a^b e^{-xt}\, dx.$$

When we calculate the integral on the left and the inner integral on the right, we obtain (22.5–7).

Example 2. Show that, if $a > 0$,

$$\int_0^\infty \frac{e^{-at} \sin xt}{t}\, dt = \tan^{-1} \frac{x}{a}. \tag{22.5–9}$$

We denote the integral here by $F(x)$. We are justified in calculating $F'(x)$ by Theorem X, getting

$$F'(x) = \int_0^\infty e^{-at} \cos xt\, dt, \tag{22.5–10}$$

for this latter integral is easily seen to converge uniformly for all values of x. Now, by an elementary integration formula

$$\int_0^s e^{-at} \cos x\, dt = \left[\frac{e^{-at}(x \sin xt - a \cos xt)}{a^2 + x^2}\right]_0^s,$$

and hence, as $s \to \infty$ we have, after an easy calculation,

$$\int_0^\infty e^{-at} \cos xt\, dt = \frac{a}{a^2 + x^2}. \tag{22.5–11}$$

Therefore, by integration, taking account of (22.5–10),

$$F(x) = \tan^{-1} \frac{x}{a} + C.$$

To determine the constant of integration C we observe that $F(0) = 0$ by the definition of $F(x)$. Since $\tan^{-1} 0 = 0$ also, we conclude that $C = 0$. This completes the derivation of (22.5–9).

Example 3. Show that, if $x > 0$,

$$\int_0^\infty \frac{\sin xt}{t}\, dt = \frac{\pi}{2}. \tag{22.5–12}$$

This result is deduced from (22.5–9). We consider the integral in the latter formula as a function of the parameter a, writing

$$G(a) = \int_0^\infty \frac{e^{-at} \sin xt}{t}\, dt.$$

Note that $G(a) = \tan^{-1}(x/a)$ if $a > 0$, while

$$G(0) = \int_0^\infty \frac{\sin xt}{t}\, dt.$$

We regard x as fixed. This last integral is convergent, by Theorem VI, §22.3. Now it can be shown that the integral defining $G(a)$ is uniformly convergent when $a \geqq 0$, so that G is continuous for such values of a, by Theorem VII. Then $G(a) \to G(0)$ as $a \to 0^+$. Now

$$\lim_{a \to 0+} G(a) = \lim_{a \to 0+} \tan^{-1} \frac{x}{a} = \frac{\pi}{2}$$

if $x > 0$. Thus (22.5–12) is established. The assertion about uniform convergence is discussed in Exercise 20.

It is easily shown that the integral in (22.5–12) has the value $-\pi/2$ if $x < 0$, for the integral defines an odd function of x. Thus

$$\int_0^\infty \frac{\sin xt}{t}\, dt = \begin{cases} \frac{\pi}{2} \text{ if } x > 0, \\ 0 \text{ if } x = 0, \\ -\frac{\pi}{2} \text{ if } x < 0. \end{cases} \tag{22.5–13}$$

From this it is clear that the function defined by the integral is discontinuous at $x = 0$. It must therefore fail to be uniformly convergent in any closed interval which contains $x = 0$.

EXERCISES

1. (a) Let $F(x) = \int_0^\infty e^{-t^2} \cos xt\, dt$. Assume the applicability of Theorem X, and show that $F'(x) = -\frac{1}{2}xF(x)$. Then find $F(x)$.

(b) By change of variable in the result of (a) show that

$$\int_0^\infty e^{-at^2} \cos xt\, dt = \frac{1}{2}\sqrt{\frac{\pi}{a}}\, e^{-x^2/4a}, \quad a > 0.$$

(c) By suitable use of Theorem VIII, show that the integral

$$\int_0^\infty te^{-t^2} \sin xt\, dt$$

is convergent uniformly with respect to x for all values of x, thus justifying the procedure used in (a).

2. (a) Show that, for all x,

$$\int_0^\infty e^{-t^2-(x^2/t^2)}\, dt = \frac{\sqrt{\pi}}{2} e^{-2|x|},$$

by denoting the integral by $F(x)$ and showing that $F'(x) = -2F(x)$ when $x > 0$. The substitution $u = x/t$ is useful at a certain stage in the work. Explain how you justify the answer when $x < 0$.

(b) Deduce from (a) the result

$$\int_0^\infty e^{-px^2-(q/x^2)}\, dx = \frac{1}{2}\sqrt{\frac{\pi}{p}}\, e^{-2\sqrt{pq}}, \; p > 0, \; q \geqq 0.$$

(c) Prove that the integral $F(x)$ in (a) is convergent uniformly for all values of x, and that the integral $\int_0^\infty (x/t^2)e^{-t^2-(x^2/t^2)}\,dt$ is convergent uniformly for $\delta \leqq x \leqq M$ if $\delta > 0$. Note that the method of calculating $F'(x)$ by differentiating under the integral sign gives a false result at $x = 0$. This is explained by the fact that the foregoing integral is not uniformly convergent in any interval including $x = 0$.

3. (a) Deduce the result

$$\int_0^\infty t^n e^{-xt}\,dt = \frac{n!}{x^{n+1}} \quad (x > 0)$$

by repeated differentiation of both sides of the formula

$$\int_0^\infty e^{-xt}\,dt = \frac{1}{x}.$$

(b) Prove that it is legitimate to differentiate under the integral sign as was done in (a).

4. (a) Deduce the result

$$\int_0^\infty \frac{dt}{(t^2+x)^{n+1}} = \frac{1\cdot 3\cdots(2n-1)}{2\cdot 4\cdots 2n}\,\frac{1}{x^{n+1/2}} \quad (x > 0)$$

by repeated differentiation of both sides of the formula

$$\int_0^\infty \frac{dt}{t^2+x} = \frac{\pi}{2}\,x^{-1/2}.$$

(b) Prove that it is legitimate to differentiate under the integral sign as was done in (a).

5. (a) Let $F_n(x) = \int_0^\infty t^n e^{-xt^2}\,dt$, $x > 0$. Show that $F_n'(x) = -F_{n+2}(x)$. Calculate $F_0(x)$ explicitly from the result (22.41–5), and then use the foregoing relation between F_n' and F_{n+2} to prove by induction that

$$\int_0^\infty t^{2n} e^{-xt^2}\,dt = \sqrt{\pi}\,\frac{1\cdot 3\cdots(2n-1)}{2^{n+1}}\,\frac{1}{x^{n+1/2}}.$$

(b) Prove that the integral defining $F_n(x)$ converges uniformly for $x \geqq \delta$ if $\delta > 0$.

6. Show that $\int_0^\infty e^{-t^2}\sin 2xt\,dt = e^{-x^2}\int_0^x e^{u^2}\,du$.

7. Show that

$$\frac{d^n}{dx^n}\left(\frac{1}{1+x^2}\right) = (-1)^{n/2}\int_0^\infty t^n e^{-t}\cos xt\,dt, \quad n = 0, 2, 4, \ldots$$

$$\frac{d^n}{dx^n}\left(\frac{1}{1+x^2}\right) = (-1)^{(n+1)/2}\int_0^\infty t^n e^{-t}\sin xt\,dt, \quad n = 1, 3, 5, \ldots.$$

8. Prove that $\int_0^\infty \frac{\sin^2 xt}{t^2}\,dt = \frac{\pi}{2}\,x (x > 0)$. Assume that differentiation under the integral sign is legitimate.

9. Prove that, if $F(x) = \int_0^\infty \frac{1-e^{-xt^2}}{t^2}\,dt$, where $x \geqq 0$, then $F'(x)$ can be computed by differentiation under the integral sign when $x > 0$. Deduce the value of $F(x)$.

10. (a) Start with

$$\frac{t}{1+t^2} = \int_0^\infty e^{-tx}\cos x\,dx \quad (t > 0),$$

and obtain

$$\int_0^\infty \frac{e^{-bx} - e^{-ax}}{x} \cos x \, dx = \tfrac{1}{2} \log \frac{1+a^2}{1+b^2},$$

if a and b are positive. Upon what theorem do you depend?

(b) Assume without proof that, as a function of b, the integral in **(a)** is uniformly convergent when $b \geqq 0$. Explain carefully how this implies that, if $a > 0$,

$$\int_0^\infty \frac{1 - e^{-ax}}{x} \cos x \, dx = \tfrac{1}{2} \log(1 + a^2).$$

What theorem do you use?

11. Prove that, if a and b are positive,

$$\int_a^b dx \int_0^\infty \frac{dt}{1+x^2t^2} = \int_0^\infty dt \int_a^b \frac{dx}{1+x^2t^2},$$

and in this way find the value of the integral $\displaystyle\int_0^\infty \frac{\tan^{-1}(bt) - \tan^{-1}(at)}{t} \, dt$.

12. Show that, under certain conditions

$$\int_{-\infty}^\infty \frac{f(bx) - f(ax)}{x^2} \, dx = \int_a^b dt \int_{-\infty}^\infty \frac{f'(tx)}{x} \, dx.$$

Apply this to show that

$$\int_{-\infty}^\infty \frac{\cos bx - \cos ax}{x^2} \, dx = \pi(a - b)$$

if a and b are positive. Take for granted that the "certain conditions" are satisfied.

13. Prove that $\displaystyle\int_0^\infty e^{-x} \frac{1 - \cos xy}{x} \, dx = \tfrac{1}{2} \log(1 + y^2)$.

HINT: $\displaystyle\frac{t}{1+t^2} = \int_0^\infty e^{-x} \sin xt \, dx$.

14. Prove that

$$\int_0^\infty e^{-x^2} \frac{\sin 2xy}{x} \, dx = \sqrt{\pi} \int_0^y e^{-t^2} \, dt.$$

15. From $\displaystyle\int_0^\infty e^{-y^2x^2} \, dx = \frac{\sqrt{\pi}}{2y}$ $(y > 0)$ deduce by integration that

$$\int_0^\infty \frac{e^{-a^2x^2} + e^{-b^2x^2}}{x^2} \, dx = (b - a)\sqrt{\pi}$$

if $0 < a < b$.

16. Show that

$$\int_0^\infty \frac{e^{-t} \sin t}{t} \, dt = \frac{\pi}{4}.$$

17. From (22.5–13) and trigonometric identities, deduce that

$$\int_0^\infty \frac{\sin x \cos mx}{x}\,dx = \begin{cases} \frac{\pi}{2} \text{ if } |m|<1, \\ 0 \text{ if } |m|>1, \\ \frac{\pi}{4} \text{ if } |m|=1. \end{cases}$$

18. Prove that $\int_0^\infty \frac{\cos xy}{1+x^2}\,dx$ converges uniformly with respect to y for all y.

19. Prove that $\int_1^\infty \frac{dx}{x^{1+y}}$ converges uniformly with respect to y if $y \geqq \delta$, where $\delta > 0$.

20. Prove that the integral $G(a) = \int_0^\infty e^{-at}\frac{\sin xt}{t}\,dt$, where x is fixed, is convergent uniformly with respect to a when $a \geqq 0$.

SUGGESTION: Put $u = t^{-1}$, $dv = e^{-at}\sin xt\,dt$, and integrate by parts between $t = T$ and $t = \infty$, where $T > 0$. In this way show that

$$\left|\int_T^\infty e^{-at}\frac{\sin xt}{t}\,dt\right| \leqq \frac{a+|x|}{a^2+x^2}\frac{2}{T}.$$

Explain why this inequality insures uniform convergence. (Consider the maximum of the expression on the right as a varies, assuming $x \neq 0$.)

22.51 / LAPLACE TRANSFORMS

A function $f(s)$ defined by

$$f(s) = \int_0^\infty e^{-st}F(t)\,dt \qquad (22.51\text{–}1)$$

is called the *Laplace transform* of the function $F(t)$. For instance, by (22.5–9) we see that

$$f(s) = \tan^{-1}\left(\frac{1}{s}\right) \quad \text{if} \quad F(t) = \frac{\sin t}{t},$$

and by Exercise 5, §22.2, we see that

$$f(s) = \frac{\Gamma(n)}{s} \quad \text{if} \quad F(t) = t^{n-1}.$$

Laplace transforms are used a good deal in applied mathematics in solving differential equations. We do not have space in this book for an extensive account of the theory and applications of Laplace transforms.

The theorems of §22.5 are useful in the discussion of Laplace transforms. The exponential e^{-st} in (22.51–1) causes the integral to converge very nicely, provided that $F(t)$ does not grow too rapidly as $t \to \infty$. If there is some constant c such that $|F(t)| \leqq e^{ct}$ for all sufficiently large values of t, the integral defining $f(s)$ converges when $s > c$, and f has derivatives of all orders, which can be calculated by differentiating under the integral sign.

One of the most important properties of the Laplace transform is revealed when we integrate by parts in (22.51–1), taking $u = F(t)$, $dv = e^{-st}\,dt$. We first obtain

$$\int_0^b e^{-st}F(t)\,dt = -\frac{1}{s}e^{-st}F(t)\Big|_0^b + \frac{1}{s}\int_0^b e^{-st}F'(t)\,dt.$$

Then, letting $b \to \infty$, we get

$$\int_0^\infty e^{-st}F(t)\,dt = \frac{1}{s}F(0) + \frac{1}{s}\int_0^\infty e^{-st}F'(t)\,dt, \tag{22.51–2}$$

provided certain conditions are satisfied. These conditions must be such as to guarantee the convergence of the integrals, and also such as to guarantee that

$$e^{-st}F(t) \to 0 \quad \text{when} \quad t \to \infty.$$

It is convenient to denote the Laplace transform of a function by using a symbol L in this way: $f(s) = L\{F(t)\}$. Formula (22.51–2) can now be written

$$L\{F'(t)\} = sL\{F(t)\} - F(0). \tag{22.51–3}$$

If the procedure can be repeated with F' in place of F, we obtain

$$L\{F''(t)\} = sL\{F'(t)\} - F'(0).$$

If in this result we substitute for $L\{F'(t)\}$ from (22.51–3), we obtain

$$L\{F''(t)\} = s^2L\{F(t)\} - sF(0) - F'(0). \tag{22.51–4}$$

Formulas for the Laplace transform of higher-order derivatives can be found by carrying on the process.

The foregoing formulas provide the basis for using Laplace transforms to solve differential equations. As a preliminary to further discussion of this question we observe the following formulas:

$$L\{\sin at\} = \int_0^\infty e^{-st}\sin at\,dt = \frac{a}{s^2 + a^2}, \tag{22.51–5}$$

$$L\{\cos at\} = \int_0^\infty e^{-st}\cos at\,dt = \frac{s}{s^2 + a^2}. \tag{22.51–6}$$

These may be proved by direct use of elementary integration formulas [see the derivation of (22.5–11)].

Now we consider a differential equation problem.

Example. Find the function $y = F(t)$ such that

$$\frac{d^2y}{dt^2} + 4y = 3\sin t \tag{22.51–7}$$

and

$$F(0) = 1, \quad F'(0) = 0.$$

We assume that the problem has a solution, and we let $f(s) = L\{F(t)\}$. From

(22.51–7) we see that

$$L\{F''(t)\} + 4L\{F(t)\} = 3L\{\sin t\}.$$

We use (22.51–4) and (22.51–5) to obtain

$$s^2 f(s) - s \cdot 1 - 0 + 4f(s) = \frac{3}{s^2+1}.$$

We solve this equation to find $f(s)$:

$$f(s) = \frac{3}{(s^2+4)(s^2+1)} + \frac{s}{s^2+4}.$$

The first fraction on the right can be expressed in terms of simpler fractions, and we find

$$f(s) = \frac{1}{s^2+1} - \frac{1}{s^2+4} + \frac{s}{s^2+4}.$$

We now observe by (22.51–5) and (22.51–6) that

$$\begin{aligned} f(s) &= L\{\sin t\} - \tfrac{1}{2}L\{\sin 2t\} + L\{\cos 2t\} \\ &= L\{\sin t - \tfrac{1}{2}\sin 2t + \cos 2t\}. \end{aligned}$$

Since $f(s) = L\{F(t)\}$, this suggests that

$$F(t) = \sin t - \tfrac{1}{2}\sin 2t + \cos 2t.$$

A check shows that this function satisfies the differential equation and the initial conditions.

This example illustrates a procedure which can be developed into a systematic technique for solving differential equations of certain types. Of course, the foregoing problem could also have been solved by a variety of standard elementary methods.

EXERCISES

1. Verify that $L\{1\} = s^{-1}$ and $L\{e^{ct}\} = (s-c)^{-1}$.

2. Derive the formulas

$$L\{\sinh at\} = \frac{a}{s^2-a^2}, \quad L\{\cosh at\} = \frac{s}{s^2-a^2}.$$

3. Find a function $y = F(t)$ such that $y'' + 2y' = e^t$ and $F(0) = 1$, $F'(0) = 0$. Use the results in Exercise 1.

4. Find a function $y = F(t)$ such that **(a)** $y'' + 9y = \cos 2t$, $F(0) = -1$, $F'(0) = 1$; **(b)** $y'' - 4y = -3\sinh t$, $F(0) = 0$, $F'(0) = 5$.

5. If $L\{F(t)\} = f(s)$, show that $L\{e^{ct}F(t)\} = f(s-c)$. Use this to find Laplace transforms of **(a)** $t^{n-1}e^{-t}$; **(b)** $e^{-t}\sin t$; **(c)** $e^{2t}\cos 3t$.

22.6 / REPEATED IMPROPER INTEGRALS

We sometimes encounter improper integrals whose integrands are improper integrals, e.g.,

$$\int_0^\infty dy \int_0^\infty e^{-xy} \sin x \, dx, \quad \int_0^1 dy \int_0^\infty y^{-1/4} x^{-1/2} e^{-y(1+x)} \, dx.$$

It is often highly useful, as a technique in solving problems, to be able to invert the order of integration in such integrals. The legitimacy of such an inversion of order is not guaranteed by Theorem IX of §22.5. We do not intend to take up theorems which deal with this problem. Our main intent is to point out that the inversion of the order of integration requires proof. Such proofs are sometimes rather delicate, and a systematic treatment of them would be too long and difficult for this book. We shall nevertheless give some examples of the use of inversions of the order of integration.

Example 1. Assuming the correctness of

$$\int_0^\infty dy \int_0^\infty e^{-xy} \sin x \, dx = \int_0^\infty dx \int_0^\infty e^{-xy} \sin x \, dy, \tag{22.6–1}$$

we can easily show that

$$\int_0^\infty \frac{\sin x}{x} \, dx = \frac{\pi}{2}; \tag{22.6–2}$$

for the right side of (22.6–1) is $\int_0^\infty \frac{\sin x}{x} \, dx$, by (22.5–8). The inner integral on the left is handled in a manner similar to that of deriving (22.5–11). The result is

$$\int_0^\infty e^{-xy} \sin x \, dx = \frac{1}{1+y^2}, \quad y > 0. \tag{22.6–3}$$

Thus (22.6–1) becomes

$$\int_0^\infty \frac{dy}{1+y^2} = \int_0^\infty \frac{\sin x}{x} \, dx.$$

The left side here is $\lim_{y\to\infty} \tan^{-1} y = \pi/2$. Thus (22.6–2) is established. We thus have an alternative to the method of Example 3, §22.5.

Example 2. Show that

$$\int_0^\infty \frac{\sin t}{\sqrt{t}} \, dt = \frac{2}{\sqrt{\pi}} \int_0^\infty \frac{ds}{1+s^4}. \tag{22.6–4}$$

To get started, substitute $x = s\sqrt{t}$ in (22.41–5), taking s as the new variable of integration. The result is

$$\frac{2}{\sqrt{\pi}} \int_0^\infty e^{-s^2 t} \, ds = \frac{1}{\sqrt{t}}.$$

Thus

$$\int_0^\infty \frac{\sin t}{\sqrt{t}}\,dt = \frac{2}{\sqrt{\pi}}\int_0^\infty dt \int_0^\infty e^{-s^2t}\sin t\,ds.$$

If we now assume that it is legitimate to invert the order of integration on the right, we obtain the repeated integral

$$\frac{2}{\sqrt{\pi}}\int_0^\infty ds \int_0^\infty e^{-s^2t}\sin t\,dt.$$

The inner integral here is evaluated by (22.6–3), and we have

$$\frac{2}{\sqrt{\pi}}\int_0^\infty \frac{ds}{1+s^4}; \tag{22.6–5}$$

thus (22.6–4) has been derived. The integral (22.6–5) can be calculated by elementary methods, and is found to have the value $\frac{\pi}{2\sqrt{2}}$ (see Exercise 9). Thus we arrive at the formula

$$\int_0^\infty \frac{\sin t}{\sqrt{t}}\,dt = \left(\frac{\pi}{2}\right)^{1/2}. \tag{22.6–6}$$

EXERCISES

In the following exercises it is to be assumed by the student that reversal of the order of integration in the repeated integrals is legitimate.

1. Let $I = \int_0^\infty e^{-x^2}\,dx$. If $x > 0$, show that $\int_0^\infty xe^{-x^2y^2}\,dy = I$. Thus

$$I^2 = \int_0^\infty e^{-x^2}\,dx \int_0^\infty xe^{-x^2y^2}\,dy.$$

Now reverse the order of integration and deduce the value of I. This is an alternative derivation of (22.41–5).

2. (**a**) Verify that $2\int_0^\infty ye^{-(1+x^2)y^2}\,dy = 1/(1+x^2)$. From this deduce that

$$\int_0^\infty \frac{\cos ax}{1+x^2}\,dx = \frac{\pi}{2}e^{-|a|}.$$

Use the result of Exercise 1(**b**), §22.5.

(**b**) Use the method of (**a**) to show that $\int_0^\infty \frac{x\sin ax}{1+x^2}\,dx = \frac{\pi}{2}e^{-a}$, $a > 0$. In the process you will have to deal with $\int_0^\infty xe^{-x^2y^2}\sin ax\,dx$ by integration by parts, and with $\int_0^\infty y^{-2}e^{-y^2-(a^2/4y^2)}\,dy$ by an appropriate substitution of a new variable of integration.

3. By reversing the order of integration in the repeated integral

$$\int_0^\infty dy \int_0^\infty y^{-1/4}x^{-1/2}e^{-y(1+x)}\,dx,$$

show that

$$\int_0^\infty \frac{dx}{x^{1/2}(1+x)^{3/4}} = \frac{\Gamma(\frac{1}{4})\Gamma(\frac{1}{2})}{\Gamma(\frac{3}{4})}.$$

4. Use the formula $1/(1+x^2) = \int_0^\infty e^{-xt} \sin t\, dt$ to show that

$$\int_0^\infty \frac{x^{a-1}}{1+x^2}\, dx = \Gamma(a) \int_0^\infty \frac{\sin t}{t^a}\, dt$$

if $0 < a < 2$.

5. Show that, if $a \geqq 0$, $\int_0^\infty \frac{e^{-ax}}{1+x^2}\, dx = \int_0^\infty \frac{\sin t}{a+t}\, dt$.

6. Show that $\int_0^\infty \frac{\cos ax}{1+x^2}\, dx = \int_0^\infty \frac{t \sin t}{a^2+t^2}\, dt$, and that, if $a > 0$, these integrals are equal to $\int_0^\infty \frac{x \sin ax}{1+x^2}\, dx$. From this result deduce that the value of the first integral is $\frac{\pi}{2} e^{-|a|}$. Assume the legitimacy of all procedures of reversing order of integration and differentiation under the integral sign.

7. Show that $\int_0^\infty t^{a-1} e^{-x\sqrt{t}}\, dt = 2\frac{\Gamma(2a)}{x^{2a}}$ if $a > 0$. Use this and the result $\int_0^\infty e^{-x\sqrt{t}} \sin x\, dx = \frac{1}{1+t}$ to show that $\int_0^\infty \frac{t^{a-1}}{1+t}\, dt = 2\Gamma(2a) \int_0^\infty \frac{\sin x}{x^{2a}}\, dx$ if $0 < a < 1$.

8. Use the formula $\int_0^\infty t^{a-1} e^{-xt}\, dt = \frac{\Gamma(a)}{x^a}$ $(x > 0,\ a > 0)$ to show that $\int_0^\infty \frac{\cos x}{x^a}\, dx = \frac{1}{\Gamma(a)} \int_0^\infty \frac{t^a}{1+t^2}\, dt$ if $0 < a < 1$.

9. Prove that $\int_0^\infty \frac{dx}{1+x^4} = \frac{\pi\sqrt{2}}{4}$ by expressing the integrand in terms of partial fractions with quadratic denominators and deriving the integration formula

$$\int \frac{dx}{1+x^4} = \frac{\sqrt{2}}{8} \log \frac{x^2+x\sqrt{2}+1}{x^2-x\sqrt{2}+1} + \frac{\sqrt{2}}{4}[\tan^{-1}(x\sqrt{2}-1) + \tan^{-1}(x\sqrt{2}+1)].$$

10. **(a)** Prove that $\int_0^\infty \frac{x^2\, dx}{1+x^4} = \int_0^\infty \frac{dx}{1+x^4}$ by substituting $y = \frac{1}{x}$ in the integral.

(b) Prove that $\int_0^\infty \frac{\cos t}{\sqrt{t}}\, dt = \left(\frac{\pi}{2}\right)^{1/2}$ by the method of illustrative Example 2.

22.7 / THE BETA FUNCTION

The beta function is defined by

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\, dt. \tag{22.7-1}$$

If $x \geqq 1$ and $y \geqq 1$, the integral is proper. If $x > 0$ and $y > 0$, and either $x < 1$ or $y < 1$ (or both), the integral is improper, but convergent.

We shall show that there is an important connection between the beta

function and the gamma function. In fact, if x and y are positive.

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}. \tag{22.7–2}$$

We shall verify this relation without trying to explain how it might be discovered.

To begin with, let us substitute

$$t = \frac{u}{1+u}, \qquad dt = \frac{du}{(1+u)^2}$$

in (22.7–1), thus getting

$$\begin{aligned} B(x, y) &= \int_0^\infty \left(\frac{u}{1+u}\right)^{x-1}\left(\frac{1}{1+u}\right)^{y-1}\frac{du}{(1+u)^2} \\ &= \int_0^\infty \frac{u^{x-1}}{(1+u)^{x+y}}\,du. \end{aligned} \tag{22.7–3}$$

Next, in the formula (22.2–1) let us substitute $t = vu$, $dt = v\,du$, $(v > 0)$. Then

$$\Gamma(x) = \int_0^\infty (vu)^{x-1}e^{-vu}v\,du = \int_0^\infty v^x u^{x-1}e^{-vu}\,du.$$

Multiply this by $v^{y-1}e^{-v}$ and integrate with respect to v:

$$\Gamma(x)\int_0^\infty v^{y-1}e^{-v}\,dv = \int_0^\infty dv \int_0^\infty v^{x+y-1}u^{x-1}e^{-v(1+u)}\,du.$$

The left side here is $\Gamma(x)\Gamma(y)$. We assume that we can reverse the order of integration on the right. Then we have

$$\Gamma(x)\Gamma(y) = \int_0^\infty u^{x-1}\,du \int_0^\infty v^{x+y-1}e^{-v(1+u)}\,dv.$$

In the inner integral set

$$v = \frac{w}{1+u}, \quad dv = \frac{dw}{1+u}.$$

Then

$$\begin{aligned} \int_0^\infty v^{x+y-1}e^{-v(1+u)}\,dv &= \frac{1}{(1+u)^{x+y}}\int_0^\infty w^{x+y-1}e^{-w}\,dw \\ &= \frac{\Gamma(x+y)}{(1+u)^{x+y}}, \end{aligned}$$

and so

$$\Gamma(x)\Gamma(y) = \Gamma(x+y)\int_0^\infty \frac{u^{x-1}}{(1+u)^{x+y}}\,du = \Gamma(x+y)B(x, y),$$

by (22.7–3). Thus (22.7–2) is established. The justification of the reversal of order of integration is a problem of the type referred to in §22.6. There are various

methods for proving (22.7–2). Some of the methods use the theory of improper double integrals, as sketched in §22.41.

A useful alternative form for the beta function comes from putting

$$t = \sin^2\theta, \quad dt = 2\sin\theta\cos\theta\, d\theta$$

in (22.7–1). The new formula is

$$\mathrm{B}(x, y) = 2\int_0^{\pi/2} \sin^{2x-1}\theta\cos^{2y-1}\theta\, d\theta. \tag{22.7–4}$$

In particular, with $x = n$, $y = \frac{1}{2}$, we have

$$\int_0^{\pi/2} \sin^{2n-1}\theta\, d\theta = \tfrac{1}{2}\mathrm{B}(n, \tfrac{1}{2}) = \frac{1}{2}\cdot\frac{\Gamma(n)\Gamma(\frac{1}{2})}{\Gamma(n+\frac{1}{2})}. \tag{22.7–5}$$

By repeated use of (22.2–5) we can write

$$\begin{aligned}\Gamma(n+\tfrac{1}{2}) &= (n-\tfrac{1}{2})\Gamma(n-\tfrac{1}{2}) = (n-\tfrac{1}{2})(n-\tfrac{3}{2})\Gamma(n-\tfrac{3}{2})\\ &\quad\cdots\cdots\cdots\\ &= (n-\tfrac{1}{2})(n-\tfrac{3}{2})\cdots\tfrac{1}{2}\Gamma(\tfrac{1}{2}),\end{aligned}$$

or

$$\Gamma(n+\tfrac{1}{2}) = \frac{1\cdot 3\cdot 5\cdots(2n-1)}{2^n}\Gamma(\tfrac{1}{2}). \tag{22.7–6}$$

When this is combined with (22.7–5) we have

$$\int_0^{\pi/2} \sin^{2n-1}\theta\, d\theta = \frac{2^{n-1}\Gamma(n)}{1\cdot 3\cdot 5\cdots(2n-1)} = \frac{2\cdot 4\cdots(2n-2)}{1\cdot 3\cdots(2n-1)}. \tag{22.7–7}$$

In a similar manner it may be shown that

$$\int_0^{\pi/2} \sin^{2n}\theta\, d\theta = \frac{1\cdot 3\cdots(2n-1)}{2\cdot 4\cdots(2n)}\cdot\frac{\pi}{2}. \tag{22.7–8}$$

These two formulas are valid for $n = 1, 2, \ldots$. They may be derived in a completely elementary fashion by using the reduction formula

$$\int_0^{\pi/2} \sin^m\theta\, d\theta = \frac{m-1}{m}\int_0^{\pi/2} \sin^{m-2}\theta\, d\theta.$$

Our purpose in deriving (22.7–7) and (22.7–8) here is so that we may prove the formula

$$\frac{\pi}{2} = \lim_{n\to\infty}\left[\frac{2\cdot 4\cdots(2n)}{1\cdot 3\cdots(2n-1)}\right]^2\frac{1}{2n+1}. \tag{22.7–9}$$

This formula is called Wallis's formula, after a 17th century English mathematician. To prove (22.7–9) we observe that, if $0 < \theta < \frac{\pi}{2}$,

$$\sin^{2n+1}\theta < \sin^{2n}\theta < \sin^{2n-1}\theta,$$

and so

$$\int_0^{\pi/2} \sin^{2n+1}\theta\, d\theta < \int_0^{\pi/2} \sin^{2n}\theta\, d\theta < \int_0^{\pi/2} \sin^{2n-1}\theta\, d\theta.$$

Therefore

$$\frac{2\cdot 4\cdots(2n)}{1\cdot 3\cdots(2n+1)} < \frac{1\cdot 3\cdots(2n-1)}{2\cdot 4\cdots(2n)}\cdot\frac{\pi}{2} < \frac{2\cdot 4\cdots(2n-2)}{1\cdot 3\cdots(2n-1)}.$$

These inequalities can be rewritten in the form

$$\frac{2n}{2n+1}\cdot\frac{\pi}{2} < \left[\frac{2\cdot 4\cdots(2n)}{1\cdot 3\cdots(2n-1)}\right]^2 \frac{1}{2n+1} < \frac{\pi}{2},$$

and (22.7–9) follows at once.

EXERCISES

1. Show directly from (22.7–1) that $B(y, x) = B(x, y)$ (make a simple change of the variable of integration).

2. Show that $B(x, 1) = x^{-1}$.

3. Using integration by parts, show directly from (22.7–1) that $yB(x+1, y) = xB(x, y+1)$. Hence, if $m+1$ and $n+1$ are positive integers, show that $B(m+1, n+1) = \dfrac{m!n!}{(m+n+1)!}$ without recourse to (22.7–2).

4. Show directly from (22.7–1) that $B(x+1, y) + B(x, y+1) = B(x, y)$. Combine this with the result in Exercise 3 to show that

$$B(x+1, y) = \frac{x}{x+y}B(x, y).$$

Hence, if n is a positive integer, show that

$$B(n, y) = \frac{(n-1)!}{y(y+1)\cdots(y+n-1)}$$

without recourse to (22.7–2).

5. Show that

$$\int_0^1 \frac{x^{2a}}{\sqrt{1-x^2}}\,dx = \frac{\sqrt{\pi}\,\Gamma(a+\frac{1}{2})}{2\Gamma(a+1)}.$$

What is the restriction on a?

6. Show that

$$\int_0^1 \frac{dx}{\sqrt{1-x^n}} = \frac{\sqrt{\pi}\,\Gamma\left(\frac{1}{n}\right)}{n\Gamma\left(\frac{n+2}{2n}\right)}.$$

7. Express $\int_0^1 x^m(1-x^p)^n\,dx$ in terms of the beta function, and hence in terms of values of the gamma function. As particular cases, calculate

(a) $\int_0^1 (1-\sqrt{x})^2\,dx$, **(b)** $\int_0^1 x^{3/2}(1-\sqrt{x})^{1/2}\,dx$,

(c) the total area enclosed by the curve $x^{2/3} + y^{2/3} = 1$, (d) $\int_0^1 \frac{x^4}{\sqrt{1-x^4}}\,dx$.

8. Show that $B(x, x) = 2\int_0^{1/2} (t - t^2)^{x-1}\,dt$. Make the change of variable $u = 4(t - t^2)$, and deduce the relation $B(x, x) = 2^{1-2x}B(x, \frac{1}{2})$. From this deduce that $\sqrt{\pi}\Gamma(2x) = 2^{2x-1}\Gamma(x)\Gamma(x + \frac{1}{2})$. Compare with Exercise 7, §22.2.

9. The gamma function can be expressed in the form

$$\Gamma(x) = 2\int_0^\infty s^{2x-1}e^{-s^2}\,ds,$$

as we see putting $t = s^2$ in (22.2–1). Show that, as a consequence,

$$\tfrac{1}{4}\Gamma(x)\Gamma(y) = \iint_R e^{-s^2-t^2}s^{2x-1}t^{2y-1}\,dA,$$

where R is the entire first quadrant of the st-plane. Using a method similar to that of Example 2, §22.41, evaluate this improper double integral with the aid of polar coordinates, and show that it has the value $\frac{1}{4}B(x, y)\Gamma(x + y)$. This gives us a new derivation of the formula (22.7–2).

10. Show that $\int_1^\infty \frac{y^{a-1}}{1+y}\,dy = \int_0^1 \frac{x^{-a}}{1+x}\,dx$ if $0 < a < 1$, and hence (see (22.7–3)) that $B(a, 1-a) = \int_0^1 \frac{x^{a-1} + x^{-a}}{1+x}\,dx$.

11. Use the geometric series for $(1+x)^{-1}$ in the last integral in Exercise 10, and integrate the resulting series term by term. Show that the result can be put in the form $\Gamma(a)\Gamma(1-a) = \frac{1}{a} + 2a\sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{n^2 - a^2}$. It can be shown that the series on the right has the value $\pi/(\sin a\pi)$. Thus

$$B(a, 1-a) = \int_0^\infty \frac{t^{a-1}}{1+t}\,dt = \frac{\pi}{\sin a\pi} \text{ if } 0 < a < 1.$$

12. If $0 < a < 1$, show, using (22.7–3), that $B(a, 1-a) = \frac{1}{a}\int_0^\infty \frac{dx}{1+x^{1/a}}$. Hence show that, if $n > 1$, $\int_0^\infty \frac{dx}{1+x^n} = \frac{\pi}{n}\csc\frac{\pi}{n}$. Use the result stated in Exercise 11.

22.8 / STIRLING'S FORMULA

The factorial function features in many problems in mathematics and related fields, especially statistics and physics. A significant fact about $n!$ is that it grows rapidly as n increases, and it is very useful in practice to have a convenient approximate formula for $n!$ when n is large. We shall show that for every positive integer, n, there is a number θ (depending on n) such that $0 < \theta < 1$, and

$$n! = n^n e^{-n}\sqrt{2\pi n}\,e^{\theta/12n}. \tag{22.8–1}$$

This is known as *Stirling's formula*. Its usefulness depends on the fact that $e^{\theta/12n}$ is close to 1 when n is large. Accordingly, the *percentage error* in using

$n^n e^{-n}\sqrt{2\pi n}$ as an approximation for $n!$ approaches zero as $n \to \infty$. We shall elaborate on this point later. Stirling did his work on approximating $n!$ about 1730.

The ingenious derivation of Sterling's formula which we shall give depends on a clever method of obtaining a formula for the logarithm of $n!$ We start from the fact that

$$\log n! = \sum_{p=2}^{n} \log p. \tag{22.8–2}$$

Now consider the graph of $y = \log x$ from $x = p - 1$ to $x = p$. In Fig. 184 we have exaggerated the curvature of the graph to make the region between the graph and the dotted chord easier to see. When p is large the graph of the logarithm from $p - 1$ to p is not very different from the chord and the area between the two, which we shall denote by ϵ_p, is quite small.

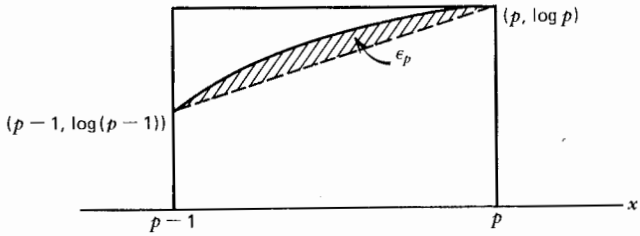


Fig. 184

In the figure, the area of the rectangle is the same as the height, namely $\log p$. We can think of this rectangular area as being made up of the trapezoidal area under the chord plus the triangular area above the chord. These parts are, respectively,

$$\int_{p-1}^{p} \log x\, dx - \epsilon_p \qquad \text{and} \qquad \tfrac{1}{2}[\log p - \log(p-1)].$$

Therefore

$$\log p = \int_{p-1}^{p} \log x\, dx + \tfrac{1}{2}[\log p - \log(p-1)] - \epsilon_p. \tag{22.8–3}$$

Forming the sum from $p = 2$ to $p = n$, we find, using (22.8–2), that

$$\log n! = \int_{1}^{n} \log x\, dx + \tfrac{1}{2}\log n - \sum_{p=2}^{n} \epsilon_p.$$

By carrying out the integration we get

$$\log n! = n \log n - n + 1 + \tfrac{1}{2}\log n - \sum_{p=2}^{n} \epsilon_p. \tag{22.8–4}$$

Our next problem is to deal with the sum of the ϵ_p's. By carrying out the integration in (22.8–3), solving for ϵ_p, and simplifying, we obtain $\epsilon_p =$

$(p-\frac{1}{2})\log p-(p-\frac{1}{2})\log(p-1)-1$, which can be written

$$\epsilon_p=\frac{2p-1}{2}\log\frac{p}{p-1}-1. \tag{22.8–5}$$

Our purpose now is to show that the infinite series $\Sigma_{p=2}^{\infty}\epsilon_p$ is convergent and to obtain an estimate for the remainder of this series after the term ϵ_n. The step we now take depends on a clever use of the infinite series

$$\log\frac{1+t}{1-t}=2(t+\tfrac{1}{3}t^3+\tfrac{1}{5}t^5+\cdots),$$

which, by (19.6–1) we know to be valid if $|t|<1$. We put $t=\dfrac{1}{2p-1}$; clearly $0<t\leqq\frac{1}{3}$ if $p\geqq 2$, and it is easily shown that $\dfrac{1+t}{1-t}=\dfrac{p}{p-1}$. Substituting in the foregoing series gives

$$\log\frac{p}{p-1}=2\left[\frac{1}{2p-1}+\frac{1}{3(2p-1)^3}+\frac{1}{5(2p-1)^5}+\cdots\right].$$

It now follows from (22.8–5) that we have, for $p\geqq 2$,

$$\epsilon_p=\frac{1}{3(2p-1)^2}+\frac{1}{5(2p-1)^4}+\frac{1}{7(2p-1)^6}+\cdots \tag{22.8–6}$$

We compare this expression for ϵ_p with the quantity δ_p defined as follows:

$$\delta_p=\frac{1}{3(2p-1)^2}+\frac{1}{3(2p-1)^4}+\frac{1}{3(2p-1)^6}+\cdots \tag{22.8–7}$$

This is a geometric series whose sum we can calculate, because

$$r^2+r^4+r^6+\cdots=\frac{r^2}{1-r^2}\ \text{ when }\ |r|<1.$$

Letting $r^2=\dfrac{1}{(2p-1)^2}$ in (22.8–7) we see that

$$\delta_p=\frac{1}{3(2p-1)^2}\,\frac{1}{1-[1/(2p-1)^2]}. \tag{22.8–8}$$

Simplifying this, we find that

$$\delta_p=\frac{1}{12p(p-1)}.$$

In order to find a simple expression for $\sum_2^n\delta_p$ we observe that

$$\frac{1}{p(p-1)}=\frac{1}{p-1}-\frac{1}{p},$$

and that

$$\sum_{p=2}^{n} \delta_p = \frac{1}{12}\left[\left(1-\frac{1}{2}\right)+\left(\frac{1}{2}-\frac{1}{3}\right)+\cdots+\left(\frac{1}{n-1}-\frac{1}{n}\right)\right] = \frac{1}{12}\left(1-\frac{1}{n}\right).$$

The cancellation of terms makes finding the sum easy. We have here what is called a telescoping sum. We see from the foregoing that

$$\sum_{p=2}^{\infty} \delta_p = \lim_{n\to\infty} \sum_{p=2}^{n} \delta_p = \frac{1}{12},$$

and that

$$\sum_{p=n+1}^{\infty} \delta_p = \sum_{p=2}^{\infty} \delta_p - \sum_{p=2}^{n} \delta_p = \frac{1}{12n}.$$

But then, because it is clear from (22.8–6) and (22.8–7) that $0 < \epsilon_p < \delta_p$, we conclude that

$$\sum_{p=n+1}^{\infty} \epsilon_p = \frac{\theta}{12n} \tag{22.8–9}$$

where θ is the ratio of $\sum_{p=n+1}^{\infty} \epsilon_p$ to $\sum_{p=n+1}^{\infty} \delta_p$. The important thing for us is that $0 < \theta < 1$. The value of θ depends on n, of course, and at times it is desirable to emphasize the dependency by a subscript: $\theta = \theta_n$.

Now let

$$\alpha = \sum_{p=2}^{\infty} \epsilon_p, \quad \text{so that} \quad \sum_{p=2}^{n} \epsilon_p = \alpha - \frac{\theta_n}{12n}. \tag{22.8–10}$$

We substitute this last result in (22.8–4) and go to exponentials on both sides of the equation, thus obtaining

$$n! = n^n e^{-n} \sqrt{n}\, e^{1-\alpha} e^{\theta_n/12n}. \tag{22.8–11}$$

It remains to find the value of the constant term $e^{1-\alpha}$, which we shall denote by C. We can achieve our purpose by using Wallis's formula (22.7–9). We observe that $2 \cdot 4 \cdot \cdots \cdot 2n = 2^n n!$ and that

$$1 \cdot 3 \cdot 5 \cdots (2n-1) = \frac{(2n)!}{2 \cdot 4 \cdot \cdots \cdot (2n)} = \frac{(2n)!}{2^n n!},$$

so that Wallis's formula can be written

$$\lim_{n\to\infty} \frac{2^{4n}(n!)^4}{[(2n)!]^2(2n+1)} = \frac{\pi}{2}. \tag{22.8–12}$$

From (22.8–11), with $e^{1-\alpha} = C$,

$$(n!)^4 = C^4 n^{4n} e^{-4n} n^2 e^{\theta_n/3n}, \quad \text{and}$$
$$(2n)! = C 2^{2n} n^{2n} e^{-2n} \sqrt{2n}\; e^{\theta_{2n}/24n}.$$

Substituting these into (22.8–12) gives a rather formidable-looking expression

which can, after much cancellation, be reduced to

$$\lim_{n\to\infty} \frac{C^2 n e^{\theta_n/3n}}{2(2n+1)e^{\theta_{2n}/12n}} = \frac{\pi}{2}.$$

But the limit on the left side is, clearly, $C^2/4$. Therefore $C^2/4 = \pi/2$, or $C = \sqrt{2\pi}$. Substituting this value for C into our above formula for $(n!)^4$, and taking the fourth root of both sides, we get (22.8–1). In our derivation of Stirling's formula we have made use of the hypothesis that $n \geqq 2$, but it is a simple matter to verify (22.8–1) directly in the special case where $n = 1$.

It follows from Stirling's formula that

$$\lim_{n\to\infty} \frac{n!}{n^n e^{-n}\sqrt{2\pi n}} = 1. \tag{22.8–13}$$

The fact that the above ratio approaches 1 as $n \to \infty$ is expressed in words by saying that $n!$ is *asymptotically equal* to $n^n e^{-n}\sqrt{2\pi n}$. It is important to notice that even though the ratio of $n!$ to $n^n e^{-n}\sqrt{2\pi n}$ approaches 1 as n approaches infinity, this does not imply that the difference between $n!$ and $n^n e^{-n}\sqrt{2\pi n}$ approaches zero—in fact, it approaches infinity. So asymptotic approximations can be somewhat different from the kinds of approximations we have been used to.

We can easily obtain some notion of the behavior of the fraction on the left in (22.8–13), as a function of n, by observing from (22.8–1) that

$$\frac{n!}{n^n e^{-n}\sqrt{2\pi n}} = e^{\theta/12n} \tag{22.8–14}$$

and estimating the value of $e^{\theta/12n}$. Clearly it is larger than 1, because e^x is an increasing function which equals 1 when $x = 0$. Using the expansion of e^x in powers of x we see that

$$e^{1/12n} = 1 + \sum_{p=1}^{\infty} \frac{1}{p!}\left(\frac{1}{12n}\right)^p < 1 + \sum_{p=1}^{\infty}\left(\frac{1}{12n}\right)^p.$$

The last series here is a geometric series whose sum is

$$\frac{1}{1-(1/12n)} = \frac{12n}{12n-1} = 1 + \frac{1}{12n-1}.$$

Thus we see that

$$1 < \frac{n!}{n^n e^{-n}\sqrt{2\pi n}} < 1 + \frac{1}{12n-1}, \tag{22.8–15}$$

and this shows that the ratio in question is rapidly squeezed downward closer and closer to 1 as n increases. The fourth column of the following table (which can be easily checked with the aid of a pocket calculator) shows how close the ratio is to 1, even for small values of n. In view of this, the fifth column is rather

surprising in showing a startling rate of increase in the *difference* between the numerator and the denominator.

n	$n!$	$n^n e^{-n}\sqrt{2\pi n}$	$\dfrac{n!}{n^n e^{-n}\sqrt{2\pi n}}$	$n! - n^n e^{-n}\sqrt{2\pi n}$
5	120	118	1.01695	2
10	3,628,800	3,598,696	1.00837	30,104
15	1,307,674,368,000	1,300,430,725,000	1.00557	7,243,642,000

It is an interesting and not difficult exercise to improve the first (left-hand side) of the two inequalities in (22.8–15) by showing that

$$1+\frac{1}{12n+1}<\frac{n!}{n^n e^{-n}\sqrt{2\pi n}}. \tag{22.8–16}$$

We use a technique analogous to that used in going from (22.8–6) to (22.8–7) (a geometric series) and then estimating sums of the δ_p's by use of telescoping sums. This time we want something *smaller* than ϵ_p. We let

$$r_p=\frac{1}{3(2p-1)^2}+\frac{1}{3^2(2p-1)^4}+\frac{1}{3^3(2p-1)^6}+\cdots.$$

It is clear that $\epsilon_p > r_p$ because

$$\frac{1}{5}>\frac{1}{3^2},\ \frac{1}{7}>\frac{1}{3^3},\ldots,\frac{1}{2n+1}>\frac{1}{3^n},\ldots$$

as may be proved easily by mathematical induction. Also, by the formula for the sum of a geometrical series,

$$r_p=\frac{1}{3(2p-1)^2}\frac{1}{1-\dfrac{1}{3(2p-1)^2}}=\frac{1}{3(2p-1)^2-1}.$$

Next, we want to get a lower bound for the value of $\Sigma_{p=n+1}^{\infty} r_p$ by use of a telescoping series. Things are not as simple here as they were with the δ_p's. We write

$$r_p=\frac{1}{12(p^2-p+\frac{1}{6})}$$

and observe that, to get a telescoping series, we want an inequality of the form

$$r_p \geqq \frac{1}{12(p-1+a)(p+a)}=\frac{1}{12}\left(\frac{1}{p-1+a}-\frac{1}{p+a}\right)$$

for some positive number a, so that when we sum the expressions on the right

we will get a telescoping series. What we need then, is to have

$$p^2 - p + \tfrac{1}{6} \leqq (p - 1 + a)(p + a) = p^2 - p + ap + ap - a + a^2,$$

or

$$\tfrac{1}{6} \leqq 2ap - a + a^2 = a(2p - 1 + a) \qquad \text{when } p \geqq 2.$$

This inequality will hold when $p > 2$ if it holds when $p = 2$, so we want $\frac{1}{6} \leq a(a + 3)$. Various choices are possible for a; $a = \frac{1}{18}$ and $a = \frac{1}{12}$ are both acceptable. For simplicity of arithmetic we shall choose $a = \frac{1}{12}$. We then have

$$\sum_{p=n+1}^{\infty} \epsilon_p > \sum_{p=n+1}^{\infty} r_p > \frac{1}{12} \sum_{p=n+1}^{\infty} \left(\frac{1}{p - 1 + \frac{1}{12}} - \frac{1}{p + \frac{1}{12}} \right) = \frac{1}{12n + 1}.$$

Since, by (22.8–9),

$$\frac{\theta}{12n} > \frac{1}{12n + 1},$$

we get

$$\frac{\theta}{e^{12n}} > e^{\frac{1}{12n+1}}.$$

But now, because $e^x = 1 + x + \dfrac{x^2}{2!} + \cdots > 1 + x$ when $x > 0$, we see that

$$e^{\theta/12n} > 1 + \frac{1}{12n + 1}.$$

In view of (22.8–14) this gives us (22.8–16). The reasoning which we used made use of the assumption that $n \geqq 2$, but it can be easily verified that both (22.8–15) and (22.8–16) continue to hold when $n = 1$. In summary, we have shown that for every positive integer n,

$$1 + \frac{1}{12n + 1} < \frac{n!}{n^n e^{-n} \sqrt{2\pi n}} < 1 + \frac{1}{12n - 1}.$$

MISCELLANEOUS EXERCISES

1. Prove the following:

(a) $\displaystyle\int_0^{\infty} \frac{x}{(1+x)^3}\,dx = \frac{1}{2}\int_0^{\infty} \frac{dx}{(1+x)^2} = \frac{1}{2}.$

(b) $\displaystyle\int_1^{\infty} \frac{\sqrt{x}}{(1+x)^2}\,dx = \frac{1}{2} + \frac{\pi}{4}.$

(c) $\displaystyle\int_a^{\infty} \frac{dx}{x^4\sqrt{a^2+x^2}} = \frac{2-\sqrt{2}}{3a^4} \quad (a > 0).$

(d) $\displaystyle\int_0^{\infty} \frac{dx}{(a^2+x^2)(b^2+x^2)} = \frac{\pi}{2ab(a+b)}.$

(e) $\displaystyle\int_0^{\infty} \frac{x^2\,dx}{(a^2+x^2)(b^2+x^2)} = \frac{\pi}{2(a+b)}.$

(f) $\displaystyle\int_0^{\infty} \frac{x^2\,dx}{(x^2-a^2)^2 + b^2x^2} = \frac{\pi}{2b} \quad (b > 0).$

(g) $\displaystyle\int_0^{\infty} \frac{x^4\,dx}{[(x^2-a^2)^2 + b^2x^2]^2} = \frac{\pi}{4b^3} \quad (b > 0).$

2. Examine the following integrals as to convergence or divergence:

(a) $\int_1^\infty \frac{\log[1+(1/x)]}{x}\,dx$. **(b)** $\int_1^\infty \log(1+e^{-x})\,dx$.

3. Suppose $f(x) \geqq 0$, $a_n = \int_{x_n}^{x_{n+1}} f(x)\,dx$, where $c = x_0 < x_1 < x_2 < \cdots$, and $x_n \to \infty$. Prove that $\int_c^\infty f(x)\,dx$ is convergent if and only if $\Sigma\, a_n$ is convergent.

4. Let $F(x) = \int_0^{\pi/2} \log(1 - x^2\cos^2\theta)\,d\theta$, $0 \leqq x^2 \leqq 1$.

(a) Find $F'(x)$ when $0 \leqq x < 1$, by differentiation under the integral sign (this is justified by Theorem XIV, §18.5); evaluate the resulting integral by use of standard tables. Then integrate and find $F(x) = \pi \log \frac{1+\sqrt{1-x^2}}{2}$, at least if $0 \leqq x^2 < 1$.

(b) Prove that F is continuous at $x = 1$, by use of Theorem VII, and hence deduce from **(a)** that $\int_0^{\pi/2} \log\sin\theta\,d\theta = (\pi/2)\log\frac{1}{2}$.

5. Let $I = \int_0^\pi \log\sin\theta\,d\theta$. Show that

$$I = 2\int_0^{\pi/2} \log\sin\theta\,d\theta = 2\int_0^{\pi/2} \log\cos\theta\,d\theta.$$

Then use the formula $\sin\theta = 2\sin(\theta/2)\cos(\theta/2)$ to deduce that $I = -\pi\log 2$.

6. Show that $\int_0^\infty xe^{-ax}\cos bx\,dx = \frac{a^2-b^2}{(a^2+b^2)^2}$ $(a>0)$.

7. In the integral $F(x) = \int_0^\infty e^{-t^2-(x^2/t^2)}\,dt$, assume that $x > 0$, and make the change of variable $u = t - (x/t)$. The resulting integral will be of the form $\int_{-\infty}^\infty \phi(x,u)\,du$. Express $\int_{-\infty}^0 \phi(x,u)\,du$ in the form $\int_0^\infty \phi(x,-u)\,du$ and in this way deduce the value of $F(x)$.

8. Suppose $\int_0^\infty f(x)\,dx$ is convergent. Suppose also that $\phi(x) > 0$, that $\phi'(x)$ is continuous, and $\phi'(x) \leqq 0$. Show that $\int_a^\infty \phi(x)f(x)\,dx$ is convergent.

9. Use the result of Exercise 8 to show that $\int_1^\infty \frac{\cos x^2}{\tan^{-1}x}\,dx$ is convergent.

10. Let $F(x) = \frac{2}{\pi}\int_0^\infty \frac{x\,dt}{x^2+t^2}$. Find the value of $F(x)$ for each x. Is F continuous at $x = 0$?

11. Let $F(y) = \int_0^\infty y^3 e^{-xy^2}\,dx$. Show that $F(y) = y$ for all values of y. Verify that $F'(y) = \int_0^\infty \left[\frac{\partial}{\partial y}(y^3e^{-xy^2})\right]dx$ if $y \neq 0$, but that this is false if $y = 0$. What do you conclude from this about the uniform convergence of the integral $\int_0^\infty y^2(3-2y^2x)e^{-xy^2}\,dx$?

12. Is the equation

$$\int_0^a dy\int_0^\infty (2y - 2xy^3)e^{-xy^2}\,dx = \int_0^\infty dx\int_0^a (2y - 2xy^3)e^{-xy^2}\,dy$$

true or false when $a \neq 0$?

13. Let $f(x,y) = \sin(x^2+y^2)$. Let R be the square region $0 \leqq x \leqq a$, $0 \leqq y \leqq a$, and let T be the portion of the circular region $x^2+y^2 \leqq a^2$ which lies in the first quadrant. Show that, if $I = \iint_R f(x,y)\,dA$ and $J = \iint_T f(x,y)\,dA$, then $I \to \pi/4$ as $a \to \infty$, but J does not approach any limit.

14. **(a)** Let $F(x) = \int_0^\infty \frac{\sin xt}{t(1+t^2)}\,dt$. Show that $F''(x) - F(x) = -\pi/2$ if $x > 0$. Take for granted that $F''(x)$ can be calculated by differentiating under the integral. Solve the

differential equation and evaluate $F(0)$, $F'(0)$, and thus show that, if $x \geqq 0$,

$$\int_0^\infty \frac{\sin xt}{t(1+t^2)}\,dt = \frac{\pi}{2}(1-e^{-x}),$$

$$\int_0^\infty \frac{\cos xt}{1+t^2}\,dt = \frac{\pi}{2}e^{-x}.$$

(b) Show how to deduce the value of $\int_0^\infty \frac{t \sin xt}{1+t^2}\,dt$ from the results in **(a)**. State what must be proved to be legitimate for the success of your method, but do not prove it.

15. Suppose $f(x)$ is continuous when $x \geqq 0$, that $\lim_{x\to 0+} \frac{f(x)-f(0)}{x}$ exists, and that $\int_1^\infty \frac{f(u)}{u}\,du$ is convergent. From these assumptions it may be proved that

$$\int_0^\infty \frac{f(bx)-f(ax)}{x}\,dx = f(0)\log\frac{a}{b}$$

if a and b are positive. Write out a proof with the aid of the following suggestions:
(a) Show that, if $y > 0$,

$$\int_0^y \frac{f(ax)-f(0)}{x}\,dx = \int_0^{ay} \frac{f(u)-f(0)}{u}\,du,$$

explaining why the integrals are not improper at the lower limit.
(b) Explain why

$$\lim_{y\to\infty} \int_{ay}^{by} \frac{f(u)}{u}\,du = 0.$$

(c) Complete the proof.

16. Show that the functions $\cos x$ and $(\pi/2) - \tan^{-1} x$ satisfy the conditions on $f(x)$ in Exercise 15, and so find the values of the integrals

$$\int_0^\infty \frac{\cos bx - \cos ax}{x}\,dx, \quad \int_0^\infty \frac{\tan^{-1} bx - \tan^{-1} ax}{x}\,dx.$$

17. Suppose that $f(x)$ is continuous when $x \geqq 0$, and that $f'(x)$ is continuous when $x > 0$. Suppose also that $\lim_{x\to\infty} f(x)$ exists, and denote the limit by $f(\infty)$. From these assumptions it may be proved that, if a and b are positive,

$$\int_0^\infty \frac{f(bx)-f(ax)}{x}\,dx = (f(\infty)-f(0))\log(b/a).$$

Write out a proof with the aid of the following suggestions:
(a) Apply Theorem IX to $\int_a^b dt \int_0^\infty f'(tx)\,dx$.
(b) Show that $\int_1^\infty f'(tx)\,dx$ is convergent uniformly for $t \geqq a$, and the same for $\int_0^1 f'(tx)\,dx$ in case this integral is improper at $x = 0$.

18. Apply the result of Exercise 17 to each of the functions e^{-x}, $\tan^{-1} x$.

19. Show that, if $0 < a < 1$, $\int_{-\infty}^\infty \frac{e^{at}}{1+e^t}\,dt = \Gamma(a)\Gamma(1-a)$.

20. **(a)** Let $F(\theta) = \int_0^\infty e^{-t\cos\theta}\, t^{a-1} \cos(t \sin\theta)\,dt$,

$G(\theta) = \int_0^\infty e^{-t\cos\theta}\, t^{a-1} \sin(t \sin\theta)\,dt$,

where $-\pi/2 < \theta < \pi/2$ and $a > 0$. Show that $F'(\theta) = -aG(\theta)$ and $G'(\theta) = aF(\theta)$, whence $F''(\theta) + a^2F(\theta) = 0$. From this deduce that $F(\theta) = \Gamma(a)\cos a\theta$ and $G(\theta) = \Gamma(a)\sin a\theta$.

(b) Show how to obtain the formulas

$$\int_0^\infty \frac{\cos t}{t^{1-a}}\,dt = \Gamma(a)\cos\frac{\pi a}{2}, \quad \int_0^\infty \frac{\sin t}{t^{1-a}}\,dt = \Gamma(a)\sin\frac{\pi a}{2}$$

from the results in **(a)**. State without proof the facts about uniform convergence which would be sufficient to justify the procedure.

(c) The first formula in **(b)** is valid if $0 < a < 1$. The second is actually valid if $-1 < a < 1$, $a \neq 0$, $\Gamma(a)$ for $a < 0$ being defined as in §22.2 (see (22.2–10)). Deduce that $\int_0^\infty \frac{\sin t}{t^a}\,dt = \Gamma(1-a)\cos\frac{\pi a}{2}$ if $0 < a < 2$, $a \neq 1$.

21. Some improper integrals can be evaluated by expanding the integrand in a series and integrating term by term. For improper integrals the justification of this procedure is not covered by Theorem IV, §20.4. In the following, assume the legitimacy of the procedure, and deduce the results as listed.

(a) $\displaystyle\int_0^1 \frac{\log x}{1+x}\,dx = -\left(\frac{1}{1^2} - \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + \cdots\right)$.

(b) $\displaystyle\int_0^1 \frac{1}{x}\log\frac{1+x}{1-x}\,dx = 2\left(\frac{1}{1^2} + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots\right)$.

(c) $\displaystyle\int_0^1 \frac{\log x}{1-x^2}\,dx = -\left(\frac{1}{1^2} + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots\right)$.

(d) $\displaystyle\int_0^\infty \log\frac{e^x+1}{e^x-1}\,dx = 2\left(\frac{1}{1^2} + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots\right)$.

(e) $\displaystyle\int_0^\infty \frac{x^2}{e^x-1}\,dx = 2\left(\frac{1}{1^3} + \frac{1}{2^3} + \frac{1}{3^3} + \frac{1}{4^3} + \cdots\right)$.

22. Deduce the formula $\int_0^\infty e^{-ax^2}\cos 2bx\,dx = \frac{1}{2}\sqrt{\pi/a}\,e^{-b^2/a}$ (assuming $(a > 0)$ by expressing the cosine as a power series and integrating term by term.

23. In picking a value for the number a such that

$$r_p \geqq \frac{1}{12(p-1+a)(p+a)}$$

for all $p \geqq 2$ (see §22.8), it was observed that many choices are possible. For simplicity we took $a = \frac{1}{12}$. Considering only positive values of a, what choice makes the upper and lower bounds for

$$\frac{e^n}{n^n e^{-n}\sqrt{2\pi n}}$$

as close together as possible?

When $n = 10$, what percentage reduction in the difference between these two bounds is achieved by using the best positive value of a rather than $\frac{1}{12}$?

Answers to Selected Exercises

CHAPTER 1

§1.1 Pages 9–12

1. (a) 8; (c) n; (e) $-\frac{1}{4}$; (h) 2.

2. (a) 0; (b) 1; (d) $\frac{\pi}{2}$.

3. (a) Yes; (c) Yes, $\lim_{x\to 0} f(x) = 0$; no, f not continuous at $x = 0$. (d) yes; (e) no.

4. (a) 2; (b) 1; (c) 12. **9.** (a) 0; (b) 1; (c) -1.
10. (a) Approaches 0; (b) becomes infinitely large and positive; (c) does not exist.
13. (a) and (c). **14.** (c) 0.
15. (b) $f(x) = -1 - x$ if $0 < x < 1$; $f(x) = 1 + x$ if $1 < x$.
16. (a) $f(-1) = 2$; (c) the limit does not exist.
17. (a) $f(\frac{3}{2}) = 1$; (b) yes; (d) no.
18. (a) 0, 1, -1, 0, -1; (b) no.
19. Take δ to be the smaller of the numbers 1, $\epsilon/5$.
21. Take δ to be the smaller of the numbers 1, $\epsilon/7$.
24. No; $\lim_{x\to 0+} f(x) = 1$, $\lim_{x\to 0-} f(x) = 3$.

§1.11 Pages 18–20

9. 200π. **10.** $x = ct^3$. **11.** 0. **14.** (c) No. **15.** (c) 0.
18. (a) $n \geqq 2$; (b) $n \geqq 4$; (c) $n \geqq 5$.
21. f is continuous but not differentiable at $x = 1$, $x = 2$. $f'(\frac{1}{2}) = 1/\sqrt{2}$.
24. In first case yes, with $f'(0) = 0$; in second case f is not differentiable at $x = 0$.

§1.12 Pages 24–26

1. (b) Max. 0 at $x = -5$, min. -226 at $x = 06$.
2. Max. 7000, min. $200(35 - 27\sqrt{5})$. **3.** 10. **5.** $\frac{32}{27}$.
6. Abs. min. 48 $\sqrt[3]{4}$; no abs. max. **7.** Abs. min. 125.
8. (a) 2, -2; (b) $\frac{3}{2}$, -3. **9.** $\frac{4c}{3}$. **11.** (c) $2r < h$.
12. (b) Max. 4, min. π. HINT: Use symmetry about line $x = \frac{1}{2}$, and study the graph of $F(x)$, where $f'(x) = \dfrac{F(x)}{x^2(1-x)^2}$.
13. (d) 25. **14.** (c) $c_1 > c_2$ and $c_2 a < b\sqrt{c_1^2 - c_2^2}$. **15.** Max. $\frac{7}{2}$.
16. (a) Quickest to walk all way around from A to B.
(b) It is never quickest to mix rowing and walking. Quickest to row directly across if $v/u < \dfrac{\pi}{2}$, and to walk all the way if $v/u > \dfrac{\pi}{2}$.
17. $\dfrac{a^2 - b^2}{2ab}$.

§1.2 Pages 30–32

3. $2 - \frac{1}{3}\sqrt{21}$. **4.** $C \geqq 1$.

§1.3 Pages 34–35

1. (a) $dy = -2$; (b) $dx = 2$. **5.** $d\psi = \dfrac{r'^2 - rr''}{r^2 + r'^2}\, d\theta$.

§1.4 Page 38

3. If $F(x) = 0$ when $0 \leqq x \leqq 1$, $F(x) = x - 1$ when $1 \leqq x \leqq 2$, and $F(x) = 2x - 3$ when $2 \leqq x \leqq 3$, then $F'(x) = [x]$ except at $x = 1, 2, 3$, but $F'(x)$ is not defined at these points.

§1.5 Pages 43–44

1. (a) 10, −2; (b) 7, 1; (c) 4.
2. (a) 1.09458; (b) $s = 0.99564$, $S = 1.21786$.
6. (b) $A_3(n) = \frac{1}{4}n^2(n+1)^2$. **9.** Low 1.3339, high 1.4052.

§1.51 Page 46

2. $\frac{1}{4}\pi a$. **3.** $\frac{1}{3}\pi R^2$.

§1.52 Page 49

2. (a) $1/\sqrt{3}$. **3.** (b) $2/\pi$; (c) 0. **4.** (a) 1; (b) 0.
5. (a) $x = 0, 1$; (b) 0. **6.** 0.

§1.53 Pages 51–53

3. (a) $\log 4$; (b) $\log 2$; (c) $\log \frac{3}{4}$; (d) $\log \frac{4}{3}$; (e) $\frac{1}{6}\log 10$.
5. (a) $\pi/3$; (b) $\frac{7}{12}\pi$; (c) $5\pi/12\sqrt{3}$; (d) $\pi/2$. **8.** (d) $(\pi/2) - \frac{1}{2}\tan^{-1} 2$.
15. $\frac{1}{8}\pi(b-a)^2$. **18.** $\pi^2/4$.

§1.61 Pages 57–58

6. Take M the larger of the numbers $\frac{1}{3}$, $(7+\epsilon)/(3\epsilon)$.
11. (a) $f'(0) = +\infty$; (b) $f'_+(0) = +\infty$, $f'_-(0) = -\infty$.

§1.62 Pages 65–66

1. (b), (e), and (f) convergent. **2.** $N > 1/(3\epsilon)$ will do. **3.** $A = 6$.
4. Take N greater than 10 and $10^{11}/(10!\epsilon)$. **7.** (b) 2^r. **8.** $\frac{4}{3}$.
9. (a) 0; (b) $\frac{1}{2}$. **11.** $f(x) = 0$ if $x \neq 1$, $f(1) = \frac{1}{2}$. **14.** (a) $\frac{1}{2}$. **20.** (b) e^2.

Miscellaneous Exercises Pages 70–71

3. $\frac{1}{2}(a+b)$. **5.** 2. **7.** (b) $(1 - \cos \pi\alpha)(\pi\alpha)$. **9.** $a + b$.
10. Continuous, but not differentiable.
12. $F'(0) = 2$. **13.** Max. 8, min. −4.

CHAPTER 2

§2.7 Page 82

5. The limit is 2. **6.** The limit is $\frac{1}{2}(1 + \sqrt{1+4c})$.

Miscellaneous Exercises Pages 83–84

4. $N = 3$. **6.** The l.u.b. is 2.
7. The l.u.b. is $\frac{3}{2}$; the g.l.b. is -1. **8. (b)** 1.

CHAPTER 3

§3 Page 86

1. No.
3. The answer is "yes" in both cases. In the second case, consider $f(x) = -x$ if $x < 0$, $f(x) = 1$ if $x \geqq 0$, and $g(x) = 1$ if $x < 0$, $g(x) = x$ if $x \geqq 0$.
4. The answers are "no" and "yes", respectively.
5. No.

§3.1 Page 88

1. Yes. **2.** Yes. **3. (b), (c), (d).** **4. (a)** 14; **(b)** 6; **(c)** 13.

§3.2 Page 90

2. $M = 1$, $m = 0$; M not attained.
3. $M = \pi/2$, $m = -\pi/2$; neither attained.
4. $M = 1$, $m = 0$; m not attained.

§3.3 Pages 91–92

5. Consider the function $F(x)$ defined by $F(x)\left(\int_a^b f(t)\,dt\right) = \int_a^x f(t)\,dt$.

Miscellaneous Exercises Pages 92–94

1. (ii). **(a)** g undefined at 1, -1; **(b)** continuous elsewhere; **(c)** may be defined at 1 by $g(1) = \frac{1}{2}$, and is then continuous there.
(iv). **(a)** F undefined at 4, 9, and for $x < 0$; **(b)** continuous elsewhere; **(c)** continuous at 4 if one defines $F(4) = -32$.
(vi). **(a)** H undefined at 2, and for $x < \frac{5}{3}$; **(b)** continuous elsewhere; **(c)** no.
3. Let $x = y = 0$ and use $f(0+0) = f(0) + f(0)$ to conclude that $f(0) = 0$. Observe that $\lim_{x \to x_0} f(x) = f(x_0)$ is equivalent to $\lim_{h \to 0} f(x_0 + h) = f(x_0)$.
Explain why $\lim_{h \to 0} f(h) = 0$. Now use $f(x_0 + h) = f(x_0) + f(h)$ to complete the proof.
7. There must be at least one positive root and at least one negative root.
11. **(a)** Define L as the set of all real numbers x such that $f(u) < 0$ if $u < x$. Define R as the set of all real numbers not in L; that is, x is in R if and only if there is some u with $u < x$ and $f(u) \geqq 0$. It needs to be verified that L and R define a cut and that, because of the continuity of f at the cut number c, $f(c)$ can be neither positive nor negative.
16. **(a)** Only at $x = \frac{1}{2}$. **17. (c)** Yes.
19. HINT: Choose a point u such that $a < u < b$ and $f(u) = (a + b)/2$. Why is this possible? Then consider the intervals $[a, u]$ and $[u, b]$.

CHAPTER 4

§4.3 Pages 103–105

1. $x^4 = 81 + 108(x-3) + 54(x-3)^2 + 12(x-3)^3 + (x-3)^4$.
3. (a) $\sin^2 x = x^2 - \frac{1}{3}x^4 \cos 2X$.
   (e) $\dfrac{\log(1-x)}{1-x} = -x + x^2\left[\dfrac{2\log(1-X)-3}{2(1-X)^3}\right]$.
4. $\dfrac{\cos X}{5!}\left(x - \dfrac{\pi}{2}\right)^5$. 9. $x = \dfrac{\pi}{2(1+\lambda)}$.

§4.5 Pages 112–114

1. (a) $+\infty$; (b) $\frac{1}{2}$; (e) e^2; (f) 1.
2. (b) 0; (e) 2; (f) 0; (g) $+\infty$; (h) 1.
3. (b) 1; (d) e^{-2}; (f) 1.
4. (a) 0; (c) $+\infty$; (e) $\frac{2}{3}$.
7. (b) 0; (c) 0; (d) $\dfrac{e}{2}$.

Miscellaneous Exercises Pages 114–115

6. (a) 0; (b) e^{-1}.
7. The limit is a_n in the general case. 8. 1. 11. $\theta \to 1/(n+2)$.

CHAPTER 5

§5.1 Pages 121–122

1. S is open. $B(S)$ is the set described in Exercise 2.
2. S is closed. It has no interior points.
3. S is closed. $B(S)$ is composed of the line segment $y = 1$, $-1 \leqq x \leqq 1$, and the parabolic arc $y = x^2$, $-1 \leqq x \leqq 1$.
4. S is neither open nor closed. $B(S)$ is composed of the part of the curve $xy = 1$ in the first quadrant, and the nonnegative portion of each co-ordinate axis.
5. S has no interior points. It is not closed. $B(S)$ consists of S and the line segment $x = 0$, $-1 \leqq y \leqq 1$.
6. S is not open. $B(S)$ consists of the semicircular arc $y = \sqrt{4-x^2}$, the segment $y = 0$, $-2 \leqq x \leqq 2$, the segment $x = 1/n$, $0 < y \leqq 1$, and the segment $x = 0, 0 < y \leqq 1$. The points of this last segment are in S. S is a region, but $C(S)$ is not.
7. This set is open, and therefore a region. Its boundary consists of the y-axis and the curve $y = \sin(1/x)$.
8. This set is open. Its boundary consists of the half-lines $x = n\pi$, $y \geqq 0$, and the segments $y = 0$, $2n\pi \leqq x \leqq (2n+1)\pi$, $n = 0, \pm 1, \pm 2, \ldots$.

§5.2 Pages 124–125

2. No. 8. $\delta = 2\sqrt{\epsilon}$. 10. $\delta = \sqrt{\epsilon/2}$ will do. 12. Yes.

§5.3 Page 127

2. No. 4. (a) Yes; (b) no. 5. No discontinuities.
7. Define $f(x, x) = 0$. 8. (a) No; (b) continuous elsewhere.
9. No. 10. Yes. Define $f(0, 0, 0) = 0$.

CHAPTER 6

§6.1 Page 134

2. $\dfrac{\partial u}{\partial x} = \dfrac{yu + 4xv}{2(u^2 + v^2)}, \dfrac{\partial v}{\partial x} = \dfrac{4xu - yv}{2(u^2 + v^2)}$.

3. $\dfrac{\partial z}{\partial x} = 2, \dfrac{\partial z}{\partial y} = -1$. **4.** $\dfrac{\partial z}{\partial x} = \dfrac{yz - x^2}{z^2 - xy}$.

§6.2 Page 138

1. (a) $x + z = 1$; (c) $y = z$. **2.** $5x + 7y - 21z + 9 = 0$.
4. The angle is $\cos^{-1}(1/\sqrt{3})$.

§6.3 Pages 143–144

1. $(3, 4, 5)$. **3.** $\frac{1}{3}$, at $(\frac{1}{2}, \frac{1}{3})$.
4. Min. at $(6, 6)$; no max. **5.** $4A^2(a^2 + b^2 + c^2)^{-1}$.
6. Max. $\frac{5}{4}$; min. -1. **7.** Max. $\frac{1}{2}$; min. -1. **10.** $\frac{3}{8}\sqrt{3}$.
11. Max. $2e^{-1}$ at $(0, \pm 1)$.
12. Max. 16 at $(0, 2)$ and $(2, 0)$; min. 0 at $(0, 0)$.
13. Max. 45 at $(3, 3)$; min. -30 at $(2, 1)$.
14. 18. **15.** $\frac{1}{864}$.
18. (a) Along x-axis and at $(0, 1)$; (b) $x = 0, 0 < y < 1$;
(c) min. 1 at $x = 0, 0 \leqq y \leqq 1$.
19. 1 at $(0, 1)$. **20.** 1 at $(0, 0, 0)$.

§6.4 Page 153

4. $dz = \pi^2$. **5.** 0.02.

§6.5 Pages 159–162

2. $\dfrac{\partial G}{\partial y} = -2y\dfrac{\partial F}{\partial u} + 2x\dfrac{\partial F}{\partial v}$ (partial answer).

4. $\dfrac{\partial G}{\partial u} = \dfrac{\partial F}{\partial x} + \dfrac{\partial F}{\partial y} + \dfrac{\partial F}{\partial z}, \dfrac{\partial G}{\partial w} = \dfrac{\partial F}{\partial z}$ (partial answer).

11. $\left(\dfrac{\partial w}{\partial x}\right)^2 - \left(\dfrac{\partial w}{\partial y}\right)^2$. **22.** (a) 3.

§6.52 Pages 166–168

5. $-4c^2\dfrac{\partial^2 u}{\partial \xi\, \partial \eta}$. **9.** (a) $F(r) = Ar^{-1} + B$, A and B constant.

§6.8 Pages 186–188

1. $abc/27$. **4.** $A^2/(3\,abc)$. **5.** $(\sqrt{a} + \sqrt{b} + \sqrt{c})^2$. **7.** a^2b^2.
8. $(a^{3/2} + b^{3/2} + c^{3/2})^{-2}$. **10.** (c) Max. $\frac{8}{3}$, min. 1; (d) 9.
11. $8A^3/(27\,abc)$. **13.** $bc/(1 + b^2)^{1/2}$. **14.** $|a - b|$.
19. Max. 54; min. $153 - 45\sqrt{5}$. **20.** $(\Sigma\, a_i^2)^{-1}$. **21.** $\Sigma\, a_i^2$.
22. $\frac{4}{3}$. **23.** $(13 - 4\sqrt{3})^{1/2} = 2\sqrt{3} - 1$.

§6.9 Pages 193–194

1. Semi-axes 1, $\sqrt{2}/2$. 2. Max. at $(2/\sqrt{5}, -1/\sqrt{5})$.
3. Max. 10. 4. Max. 2; min. -1.
5. (a) $\lambda_1 = \sqrt{2}+1$, $\lambda_2 = 0$, $\lambda_3 = 1-\sqrt{2}$; (b) $\lambda_1 = 18$, $\lambda_2 = \lambda_3 = 9$.
6. (a) $\lambda_1 = 1$, $\lambda_2 = \lambda_3 = -\frac{1}{2}$; (b) 1; (c) $x = y = z = \pm\dfrac{\sqrt{3}}{3}$.
7. (a) and (b) hyperboloids of one sheet; (c) ellipsoid.
8. Max. 16; min. 4. 9. $\frac{27}{10}$.

Miscellaneous Exercises Pages 194–195

1. $a = \frac{4}{15}$, $b = \frac{4}{5}$.
3. (b) $(0, a)$, (a, a), $(a, 0)$, $(\frac{2}{3}a, \frac{2}{3}a)$; (c) rel. max. at $(\frac{2}{3}a, \frac{2}{3}a)$, others neither max. nor min.; (d) no.
4. $24\sqrt{3}$. 5. $(8, 4, 12)$. 6. Max. 25; min. $\frac{675}{43}$.
7. $\dfrac{y}{x} = 1 + 1/\sqrt{5}$, $\dfrac{x}{z} = \sqrt{5}/2$. 8. Max. 4; min. 3.
9. $(a^q + b^q + c^q)^{1/q}$, where $q = 1/(1-p)$.
11. Max. at $(0, 1)$; min. at $(\frac{1}{3}, 0)$.
12. (a) $5\sqrt{5}$; (b) $\sqrt{5}$; (c) no in (a), yes in (b).

CHAPTER 7

§7.1 Page 199

1. Yes.

§7.4 Pages 206–207

1. (a) $1/\sqrt{3}$; (b) $1/\sqrt{3}$; (c) $\frac{5}{9}$ (approx.). 7. (a) $1/\sqrt{3}$; (b) $\frac{2}{3}$; (c) $\frac{1}{2}$.

§7.5 Pages 210–211

1. $\sin h \sin k = hk - \frac{1}{6}h(h^2 + 3k^2)\cos\theta h \sin\theta k - \frac{1}{6}k(3h^2 + k^2)\sin\theta h \cos\theta k$.
3. $F(3+h, 3+k) = \dfrac{1}{2!}(18h^2 - 18hk + 18k^2) + \dfrac{1}{3!}(6h^3 + 6k^3)$.
4. $\log 2 + \frac{1}{2}[(x-1) + y] - \frac{1}{8}[(x-1)^2 + 2(x-1)y - y^2] + \cdots$.
5. $2(x-1) + y + \frac{1}{2}[2(x-1)^2 + y^2] + \cdots$.
7. $1 + (2xy - y^2) + (2x^2y^2 - 2xy^3 + \frac{1}{2}y^4) + \cdots$.
10. (a) $x - y$; (b) $\frac{1}{4}\pi + \frac{1}{10}(x-3) - \frac{4}{5}(y - \frac{1}{2})$.
11. (a) $h^4 - h^2k^2 + k^4$; (b) $(\pi^2/96)(h+k)^4 - \frac{1}{2}h^2k^2$.

§7.6 Pages 220–221

1. (b) Saddle at $(0, 0)$; min. at $(0, 1)$ and $(0, -2)$. (d) Max. at $(0, 0)$; four saddle points. (e) Saddle at $(0, 0)$, $(0, 3)$, and $(4, 0)$; max. at $(\frac{4}{3}, 1)$. (f) Max. at $(a/2, a/3)$; degenerate if $x = 0$ or $y = 0$. (g) Saddle at $(1, 1)$, $(0, 1)$, $(1, 0)$; max. at $(\frac{2}{3}, \frac{2}{3})$. (j) Min. at $(6, 8)$. (k) Min. at $(6, 8)$; saddle at $(-\frac{6}{5}, -\frac{8}{5})$.
3. 5 critical points.
4. Shortest distance $= \sqrt{2}$. Saddle point at $(x, -x)$, where $x^3 + 2x - 1 = 0$.

5. Min. $= a$ if $0 < a \leqq 1$; min. $= \sqrt{2a-1}$ if $a > 1$.
6. Min. at $(-2, 0, 1)$; max. at $(\frac{16}{7}, 0, -\frac{8}{7})$.
7. Max. at $(1, 1, 1)$; min. at $(-1, -1, -1)$. **9.** Min. at $x = y = z = \frac{1}{4}$.

Miscellaneous Exercises Page 221

4. Does not have tangent plane at origin.

CHAPTER 8

§8.2 Pages 228–230

1. Yes. **2.** $f(1+h, k) \sim -2h - k$. **3.** Yes. No. **4.** Yes.
7. **(b)** No. **(c)** A spherical surface and its center. **11.** No.
12. An infinite number of isolated points, among them $(0, \pi/2)$, $(\pi, 5\pi/2)$, $(-\pi, \pi/2)$.
14. **(a)** $f(x) = x\sqrt{x-1}$. **(c)** $(0, 0)$ and $(1, 0)$.
15. **(b)** $f(x) = x^2 + x^{5/2}$. **16.** $a^2 \neq 3b^2$. **17.** Yes. **20.** No.

§8.3 Pages 234–236

1. Yes. **2. (a)** $u_0 v_0 w_0 \neq 0$. **3.** $u_0 c_0 \neq 0$, $x_0^2 \neq y_0^2$, $x_0 y_0 \neq 0$.
4. **(b)** $x_0 \neq y_0$, $y_0 \neq z_0$, $z_0 \neq x_0$. **5.** $y_0 \neq z_0$.
6. **(a)** $x_0^2 + y_0^2 \neq 0$. **(b)** $z_0 \neq 0$, $y_0 \neq 2x_0$.

CHAPTER 9

§9.2 Pages 251–252

1. The circle $u^2 + v^2 \leqq 1$.
3. The triangle bounded by $u = v$, $u = -v$, $u = 1$.
5. Regions in the first and third quadrants, bounded by $y = x + 3$, $y = x - 3$, $xy = 4$, $xy = 16$.
6. $u = x + y$, $v = y/(x + y)$. The u-curves are parallel lines of slope -1. The v-curves are straight lines through the origin. The x-curves are hyperbolas with asymptotes $u = 0$, $v = 1$. The y-curves are hyperbolas with asymptotes $u = 0$, $v = 0$.
8. **(a)** $(1 - x - y)^{-3}$. **(b)** $x = u(1 + u + v)^{-1}$, $y = v(1 + u + v)^{-1}$. **(c)** The u-curves are straight lines through $(0, 1)$, the v-curves are straight lines through $(1, 0)$. **(e)** A quadrangle with vertices at $(\frac{2}{3}, 1)$, $(1, 1)$, $(1, 2)$, $(\frac{1}{2}, \frac{3}{2})$.
9. **(a)** Ellipses if $u > 0$ or $u < -1$; hyperbolas if $-1 < u < 0$. **(c)** Along the lines $x = 0$ and $y = 0$.

§9.5 Pages 262–263

1. **(b)** Yes; all points where $u = 0$ are singular. **(c)** The regions where $y^2 > x^2$.
4. $u = [\frac{1}{2}(\sqrt{x^2 + ry^2} + x)]^{1/2}$, $v = [\frac{1}{2}(\sqrt{x^2 + 4y^2} - x)]^{1/2}$.
6. **(a)** The u-curves are hyperbolas. The v-curves are ellipses. **(b)** Singular points correspond to the foci $x = \pm 1$, $y = 0$.
8. **(a)** $r^2 \sin \phi$. **(b)** A sphere. A nappe of a cone. A half plane.
9. The singular points correspond to the circle $x^2 + y^2 = 1$, $z = 0$. The u-surfaces are tori. The v-surfaces are spheres.

§9.6 Pages 266–267

1. No. **2.** $ad - bc = 0$.
4. (a) $w = u^2 - 2v$; (b) $w = -(1 + uv)/(u + v)$.

Miscellaneous Exercises Page 267

1. (c) $-r^{-6}$. **2.** $r^3 \sin^2\phi \sin\psi$.

CHAPTER 10

§10.11 Page 274

1. (a) Neither; (b) perpendicular; (c) neither; (d) collinear.
2. (a) 7; (b) 6; (c) 3; (d) 5. **3.** $\frac{4}{5}$. **7.** $2\sin(\theta/2)$.

§10.12 Pages 279–280

3. $\mathbf{v}_1 = \mathbf{i}$, $\mathbf{v}_2 = \mathbf{j}$, $\mathbf{v}_3 = \mathbf{k}$.
6. (b) $\mathbf{V} = 4\mathbf{A} - \mathbf{C}$.

§10.2 Page 283

1. (a) $\mathbf{i} - 2\mathbf{j} + \mathbf{k}$; (c) $-11\mathbf{i} - 5\mathbf{j} - \mathbf{k}$. **2.** $\sqrt{110}$. **3.** (b) $\frac{1}{2}\sqrt{69}$.
4. (a) -79, 79. **5.** (e) 2; (f) both zero. **6.** $(\mathbf{A} \times \mathbf{B}) \cdot (\mathbf{C} \times \mathbf{D}) = 0$.

§10.3 Pages 285–286

5. (a) $(1, 26, 3)$; $(-16, -31, 29)$.
6. (a) $-\mathbf{i}' + \mathbf{j}' - \sqrt{3}\mathbf{k}'$; (b) $-2\mathbf{i} - \sqrt{2}(\mathbf{j} - \mathbf{k})$.

§10.51 Page 295

1. (a) $\frac{1}{2}(x'^2 - 6x'y' + y'^2 - 2z'^2)$; (b) $F(x, y, z) = 2y$;
(c) $-x' + \sqrt{2}z'$, $y'\dfrac{1}{\sqrt{2}}(x' - y') + z'$; (d) $-\left(\dfrac{x}{\sqrt{2}} + z\right)\mathbf{i} + \dfrac{3y - z}{2\sqrt{2}}\mathbf{j} + \left(x - \dfrac{y + z}{2\sqrt{2}}\right)\mathbf{k}$.
2. (b) $\frac{1}{49}(-23x^2 + 41y^2 + 31z^2 - 48xy + 72xz + 24yz)$;
(c) $\frac{1}{7}(3y' + 2z')\mathbf{i}' - \frac{1}{7}(3x' + 6z')\mathbf{j}' - \frac{1}{7}(2x' - 6y')\mathbf{k}'$.

§10.6 Pages 298–300

1. (a) $(4x - 3y - 4z)\mathbf{i} + (2y - 3x)\mathbf{j} + (12z - 4x)\mathbf{k}$;
(d) $r^{-5}[-3xz\mathbf{i} - 3yz\mathbf{j} + (r^2 - 3z^2)\mathbf{k}]$, where $r^2 = x^2 + y^2 + z^2$.
2. $-t : 11t : -t$, where $t = (123)^{-1/2}$. **3.** $-\mathbf{i} - 2\mathbf{j} - 2\mathbf{k}$; rate $= 3$.
4. (a) 0; (b) $14\sqrt{2}$. **7.** (e) $nr^{n-2}\mathbf{R}$; (f) $-2e^{-r^2}\mathbf{R}$. **8.** (c) $r^{-1} + C$.

§10.7 Page 305

4. (b) $Ar^{-1} + B$. **5.** (b) Cr^{-3}. **6.** (b) $\nabla \cdot \mathbf{V} = 0$.

§10.8 Pages 307–308

3. (a) 0; (c) $-\mathbf{i} - \mathbf{j} - \mathbf{k}$; (e) 0.

Miscellaneous Exercises Page 308

2. (a) $nr^{n-2}\mathbf{R}\cdot\mathbf{A}$ and $nr^{n-2}\mathbf{R}\times\mathbf{A}$; (b) 0.

CHAPTER 11

§11–§11.10 Pages 331–334

7. The locus $\|\mathbf{x}\|_1 = 1$ is the periphery of the square with corners at $(\pm 1, 0)$, $(0, \pm 1)$. The locus $\|\mathbf{x}\|_m = 1$ is the periphery of the square with corners at $(\pm 1, \pm 1)$ (four combinations of signs).

CHAPTER 12

§12.1–§12.8 Pages 372–375

9. (b) $d\mathbf{F}(t, h) = \mathbf{f}'(\mathbf{u} + t\mathbf{w})[\mathbf{w}]h$.
11. g is not continuous at $(0, 0)$.
13. (b) $D_{\mathbf{u}}f(\mathbf{a}) + D_{\mathbf{v}}f(\mathbf{a}) = \|\mathbf{u} + \mathbf{v}\|\, D_{\mathbf{w}}f(\mathbf{a})$, where $\mathbf{w} = (\mathbf{u} + \mathbf{v})/\|\mathbf{u} + \mathbf{v}\|$.

CHAPTER 13

§13.3 Pages 389–390

1. (b) $\frac{1}{3}a^3$; (d) $\frac{1}{2}\pi Ha^2$; (e) $\frac{4}{3}\pi abc$; (f) $\frac{1}{12}\pi a^2 h$; (g) $\frac{8}{15}abc$.
3. $\frac{2}{3}a^3$.
4. (a) $(\frac{2}{3}a, \frac{1}{3}b)$; (b) $(\frac{1}{3}(a + b), \frac{1}{3}c)$; (c) $((4a/3\pi), 0)$; (e) $(\frac{3}{5}B, \frac{3}{8}H)$; (g) $(a/2, \frac{2}{5}b)$.
7. πa^3.
8. (a) $\left(\dfrac{187 + 4\sqrt{2}}{310}, \dfrac{29 + 116\sqrt{2}}{310}\right)$; (c) $(\frac{5}{14}, \frac{32}{35})$.

§13.4 Pages 394–395

1. (a) $\frac{2}{3}a^3$; (c) $\frac{1}{2}\pi a^2 h$; (e) $\frac{16}{9}a^3(3\pi - 4)$; (f) $\frac{32}{9}a^3$;
   (g) $\dfrac{a^2 h}{12}\left[1 + \dfrac{3\sqrt{2}}{2}\log(1 + \sqrt{2})\right]$.
2. (c) $\mathbf{y} = \dfrac{5a}{6}$; (e) $\mathbf{x} = \dfrac{a}{2}\dfrac{3\sqrt{3} + 8\pi}{3\sqrt{3} + 2\pi}$, $\mathbf{y} = \dfrac{11a}{2(3\sqrt{3} + 2\pi)}$;
   (f) $\mathbf{x} = \dfrac{a\pi\sqrt{2}}{8}$.

3 One part is $\frac{1}{9}a^3(12\sqrt{3}\pi - 20)$. 5. $\pi/32$.

§13.5 Pages 399–401

1. (a) $M = \frac{1}{3}ka^2 b$, $\bar{x} = \frac{3}{4}a$, $\bar{y} = \frac{3}{8}b$; (c) $M = \frac{1}{3}ka^3$, $\bar{x} = \frac{3}{16}\pi a$, $\bar{y} = \frac{3}{8}a$;
   (e) $M = \frac{1}{6}k\pi a^3$, $\bar{x} = \bar{y} = \frac{3}{2}a/\pi$; (g) $M = \frac{4}{15}kab^2$, $\bar{x} = \frac{5}{16}a$, $\bar{y} = \frac{3}{7}b$;
   (i) $M = \frac{16}{9}ka^3$, $\bar{x} = \frac{6}{5}a$, $\bar{y} = \frac{9}{20}a$.

3. (b) $I_x = \frac{1}{4}Ma^2$; (d) $I_x = \frac{1}{6}Mb^2$, $I_y = \frac{1}{2}Ma^2$;
(f) $I_x = \frac{4}{5}Ma^2$, $I_y = \frac{12}{7}Ma^2$;
(h) $I_x = \frac{Mb^2}{4}\left(1 - \frac{2\sin^3\beta\cos\beta}{3\beta - 3\sin\beta\cos\beta}\right)$,
$I_y = \frac{Mb^2}{4}\left(1 + \frac{2\sin^3\beta\cos\beta}{\beta - \sin\beta\cos\beta}\right)$.

4. (b) $\frac{\sigma a^4}{8} = \frac{Ma^2}{2\pi}$; (d) $-\frac{Ma^2}{4}$; (f) 0.

7. (a) $\frac{\sigma a^3 b^3}{6(a^2+b^2)}$; (c) $\frac{656\sigma a^4}{35}$.

8. (a) $(1+\sqrt{2})y = x$, $y = -(1+\sqrt{2})x$; (c) $y = \pm x$; (e) $y = 0$, $x = 0$.

9. $\frac{1}{4}\pi a^6$ about $y = -x$, $\frac{1}{12}\pi a^6$ about $y = x$.

§13.51 Page 404

4. $2\sigma b \log(1+\sqrt{2})$. **5.** $4\sigma b$.

6. $u = 2\sigma\left(b\tan^{-1}\frac{b}{\sqrt{b^2+2a^2}} + \frac{a}{2}\log\frac{\sqrt{b^2+2a^2}+a}{\sqrt{b^2+2a^2}-a}\right) - \frac{\pi\sigma b}{2}$.

§13.7 Pages 411–412

2. $(\frac{1}{4}, \frac{1}{4}, \frac{3}{4})$. **3.** (a) $(\frac{1}{4}a, \frac{1}{4}b, \frac{1}{4}c)$; (b) $\frac{1}{10}M(a^2+c^2)$.

4. $\bar{x} = \frac{3}{16}\pi a$, $\bar{y} = \frac{3}{8}a$. **5.** (b) $2\sqrt{2} - \sqrt{3} - 1$.

7. (a) $(\frac{3}{8}a, \frac{3}{8}b, \frac{3}{8}c)$; (b) $\frac{1}{5}M(b^2+c^2)$.

8. (a) $M = 8a^3$, avg. density $= 1$; (b) $I_z = \frac{38}{45}Ma^2$; (c) $\frac{19}{15}$.

§13.8 Page 413

2. $\frac{2}{5}Ma^2$. **3.** (a) $\frac{1}{2}Ma^2$; $\frac{1}{12}M(3a^2+h^2)$.

§13.9 Page 416

4. (b) $1 + 2\sqrt{2}$; (c) $\frac{4}{3}\pi\mu a$.

CHAPTER 14

§14.2 Pages 420–421

2. Avg. speed $= \frac{1}{2}\sqrt{2+\omega^2} + \frac{1}{\omega}\log\frac{\omega + \sqrt{2+\omega^2}}{\sqrt{2}}$.

3. $4\pi a$. **9.** $a\sqrt{2}\sinh t_1$.

§14.32 Pages 426–428

5. (a) $\mathbf{T} = \mathbf{i}$, $\mathbf{N} = \mathbf{j}$, $\mathbf{B} = \mathbf{k}$, $\kappa = 2$, $\tau = 3$; (b) $\frac{2}{27}\sqrt{5}$;
(c) $\kappa = \frac{216\sqrt{334}}{(1621)^{3/2}a}$, $\tau = -\frac{9}{167a}$;
(d) $\mathbf{T} = \mathbf{i}$, $\mathbf{N} = -\frac{1}{2}\sqrt{2}(\mathbf{j}+\mathbf{k})$, $\mathbf{B} = \frac{1}{2}\sqrt{2}(\mathbf{j}-\mathbf{k})$, $\kappa = \frac{1}{4}\sqrt{2}$, $\tau = 0$;
(e) $\kappa = (1+2e^{2t})^{1/2}(1+e^{2t})^{-3/2}$, $\tau = 2e^t(1+2e^{2t})^{-1}$;
(g) $\mathbf{T} = \frac{3}{5}\mathbf{i} + \frac{4}{5}\mathbf{k}$, $\mathbf{B} = -\frac{4}{5}\mathbf{i} + \frac{3}{5}\mathbf{k}$, $\mathbf{N} = \mathbf{j}$, $\kappa = \tau = \frac{6}{25}$.

8. (a) $\mathbf{T} = -\mathbf{i}$, $\mathbf{N} = \frac{1}{\sqrt{5}}(2\mathbf{j} - \mathbf{k})$, $\mathbf{B} = -\frac{1}{\sqrt{5}}(\mathbf{j} + 2\mathbf{k})$, $\kappa = \frac{\sqrt{5}}{2a}$, $\tau = 0$;

(b) $\mathbf{T} = -\frac{\sqrt{3}}{3}(\sqrt{2}\mathbf{j} - \mathbf{k})$, $\mathbf{N} = -\frac{1}{\sqrt{39}}(6\mathbf{i} + \mathbf{j} + \sqrt{2}\mathbf{k})$,

$\mathbf{B} = \frac{1}{\sqrt{13}}(\mathbf{i} - 2\mathbf{j} - 2\sqrt{2}\mathbf{k})$, $\kappa = \frac{\sqrt{39}}{9a}$, $\tau = -\frac{3\sqrt{2}}{13a}$;

(c) $\mathbf{T} = \frac{1}{\sqrt{2}}(\mathbf{i} + \mathbf{k})$, $\mathbf{N} = -\mathbf{j}$, $\mathbf{B} = \frac{1}{\sqrt{2}}(\mathbf{i} - \mathbf{k})$, $\kappa = \frac{1}{2a}$, $\tau = -\frac{3}{8a}$.

§14.4 Pages 432–433

3. (a) $(x^2/a^2) + (y^2/b^2) = z^2$; (c) $(x^2/a^2) + (y^2/b^2) - (z^2/c^2) = 1$;
(e) $(x^2/a^2) - (y^2/b^2) = z$;
(f) The part of the cone $(x^2/a^2) = (y^2/b^2) + (z^2/c^2)$ on which $x \geqq a$.
6. Cylinder, $x^2 + y^2 = a^2$. 7. Plane, $x + y = z$.

§14.5 Page 436

1. Circles in which planes through z-axis intersect the torus; $ds = b\,d\phi$.
3. The u-curves are circles $y^2 + z^2 = a^2$ in planes perpendicular to the x-axis; the v-curves are straight lines parallel to the x-axis. $ds^2 = du^2 + a^2\,dv^2$.
5. (a) $ds^2 = (1 + 4k^2u^2)\,du^2 + u^2\,dv^2$;
(c) $ds^2 = (1 + r^2\sin^2 2\theta)\,dr^2 + r^3 \sin 4\theta\,dr\,d\theta + r^2(1 + r^2\cos^2 2\theta)\,d\theta^2$;
(e) $ds^2 = (1 + 4u^2 + 9u^4)\,du^2 + 2(1 + 4uv + 9u^2v^2)\,du\,dv + (1 + 4v^2 + 9v^4)\,dv^2$.
6. (b) $\theta \tan \omega = \log \tan(\frac{1}{2}\lambda + \frac{1}{4}\pi)$.

§14.6 Pages 443–444

2. $8\sqrt{2}a^2$. 3. $2\pi a^2\sqrt{2}$. 4. $\frac{2}{3}\pi[(1 + a^2)^{3/2} - 1]$.
5. $\frac{2}{3}\pi ab(2\sqrt{2} - 1)$. 6. (a) $\frac{2}{3}\sqrt{2}(a + b)\sqrt{ab}$; (b) $\frac{1}{3}\sqrt{2}$.
7. $\pi[\sqrt{2} + \log(1 + \sqrt{2})]$. 9. $8a^2(\pi - 2)$. 11. $\frac{14}{9}\pi a^2$.
15. $2a^2$.

CHAPTER 15

§15.12 Pages 449–451

1. (a) $\frac{4}{3}$; (b) $\frac{16}{3}$; (c) $-\frac{32}{3}$; (d) $\frac{256}{3}$; (e) $\frac{8}{15}$; (f) -2; (g) $\frac{4}{5}$; (h) $-\frac{8}{3} - \pi$.
2. (a) $\frac{4}{15}$; (b) 16; (c) $\frac{1}{4}\pi a^2$; (d) $-\frac{1}{4}\pi a^4$; (e) 5π; (f) 4π; (g) 2π; (h) $\frac{4}{3}ab^2$.
3. (a) $\frac{3}{4}\pi$; (b) $\frac{1}{3}\pi$; (c) $-\frac{1}{2}\pi$. 4. $\frac{1}{2}\pi(a^2 + 1)$. 6. $\frac{1}{4}$. 7. $\frac{164}{3} + 2\pi$.
8. $\frac{2}{3} - \frac{\pi}{2}$. 9. (a) $-\frac{16}{9}$; (b) $\frac{32}{3\sqrt{3}}$; (c) $\frac{32}{9\sqrt{3}}$.
14. $2\pi \int_C xy\,dy$.

§15.13 Pages 454–455

1. $2a^2$. 2. $2\sqrt{2}a$. 3. $2w + \frac{8}{3}$. 4. (a) $2a$; (b) none.
5. 50 ft. lb.

§15.3 Pages 462–463

1. (a) 56; (b) 0; (c) 4π; (d) 0; (e) $-\frac{1}{4}\pi a^4$; (f) 0.

§15.32 Pages 468–469

1. $53\frac{1}{3}$. **2.** $(a^2+b^2)^{-1}$. **3.** $\frac{1}{30}$. **4.** $\bar{x}=5\frac{1}{2}$. **5.** $\frac{1}{2}\pi ab$.
6. 3. **7.** $\frac{1}{4}(e-e^{-1})$. **8.** $\frac{1}{2}(1-\log 2)$. **9.** $\frac{21}{8}$. **11.** $\frac{1}{32}$.
12. 1. **14.** $2\log 2-\frac{1}{2}$.

§15.41 Pages 477–478

1. (b) and (e) not exact. **2.** (b) $-\frac{8}{5}$; $-\frac{148}{9}$. **3.** (b) $2\log\frac{5}{3}$.
4. $w=e^x$; $e^x(x\sin y-\sin y+y\cos y)$.
5. (b) Along $y=0$, $x>-1$; (c) along $y=0$, $x^2<1$.

§15.51 Pages 483–484

1. $z=\dfrac{a}{2}$. **2.** (a) 9π; (b) 8π; (c) $\frac{20}{3}\pi$.
3. $\sqrt{2}$. **4.** 0. **5.** $\frac{1}{2}\pi^2 a$. **6.** 2.
7. $A=2a^2(\pi-2)$, $A\bar{x}=\frac{8}{3}a^3$, $A\bar{y}=\frac{2}{3}a^3(3\pi-8)$, $Az=\pi a^3$.
8. $16a^4\sigma\dfrac{3\pi-7}{9}$. **10.** $\dfrac{e}{2}\left(1-\dfrac{1}{\sqrt{5}}\right)$, where e is the total charge.
15. $-\dfrac{4\pi}{3}$, $\dfrac{8\pi a^3}{3c^3}$, $\dfrac{2\pi}{3}$ in the corresponding cases.
16. $\dfrac{4\pi\sigma a^2}{\sqrt{a^2+c^2+|a-c|}}$.

§15.6 Pages 490–492

4. $3a^4$. **6.** $4\pi b^5$. **8.** $8\pi b^2$.

§15.62 Pages 497–499

2. (a) 900; (b) $(5,5,-2)$. **3.** $\frac{4}{15}\pi(a^2+b^2)abc$.
4. $\frac{315}{32}\pi$. **5.** $\frac{1}{4}$. **6.** $\frac{1}{24}$ and $\frac{1}{2}$. **7.** $\frac{21}{8}\pi$.

§15.7 Pages 504–505

1. $-2\sqrt{2}\pi b^2$. **2.** $-\pi$. **3.** $-\pi a^2$. **4.** (d) $-4\pi a^2$. **5.** 0.
6. (a) $\frac{2}{3}(3\pi-8)a^2$; (b) πa^2; (c) $\frac{8}{3}a^2$.

§15.8 Pages 509–510

1. (b) a^2b^2c; (d) 1. **2.** $\dfrac{a}{a^2+b^2+c^2}-1$.
3. $x^2y+y\log z$. **4.** $\dfrac{x^2+y^2+z^2}{z}-3$. **5.** $(xy-z)e^{-xy}$.
6. (b) $x^2+y^2<a^2$, $z=0$; (c) $x^2+y^2=0$, $z\neq 0$.

Miscellaneous Exercises Pages 510–511

2. $e^x \cos y - 3 \sin^{-1} x + 7 \tan y$. **4.** Discontinuous along $y = 0$, $x^2 > 1$.
5. $\sqrt{3}$. **7.** $8/(3\pi)$.
9. $4\pi\sigma(b/a)\{(a+b)E(k)+(a-b)K(k)\}$. **10.** $\pi^2/16$.

CHAPTER 16

§16.2 Pages 516–517

1. (a) Open; (b) closed; (c) neither; (d) neither. **2.** 0 and 1.
3. (a) No; (b) yes, 0 and 1.
6. (a) 1. Not closed. (b) 0. Closed. (c) 1 and -1. Not closed.

§16.41 Pages 521–522

1. Yes. **2.** (c) $x = 0$, $-1 \leqq y \leqq 1$.
3. Neither. All points of the square $0 \leqq x \leqq 1$, $0 \leqq y \leqq 1$.

CHAPTER 18

§18.1 Page 539

2. (b) 1. **3.** $I = 2$, $J = 4$.

§18.21 Pages 547–548

1. $\int_a^b F(x)f'(x)\,dx$.

§18.4 Pages 549–550

4. $\frac{1}{3}$.
7. (b) $F(x) = -\frac{1}{2}x^2 + \frac{1}{2}$ if $x < 0$, $F(x) = \frac{1}{2} + \frac{1}{2}x^2$ if $x > 0$.
8. (a) $F(x) = -x - 1$ if $x < 0$, $F(x) = -1 + x$ if $x > 0$.

§18.5 Page 554

1. $F'(u) = \int_0^4 \frac{-2ux^2}{1-u^2x^2}\,dx$. **2.** $(1-a^2)^{-1/2}$.

3. $e^{-x^4} - 2x \int_0^x t^2 e^{-x^2t^2}\,dt$.

4. $(x^2 - \sin^2 x)^n \cos x - (x^2 - x^4)^n 2x + \int_{x^2}^{\sin x} 2xn(x^2 - t^2)^{n-1}\,dt$.

7. $\frac{\partial^2 u}{\partial x^2} = \frac{1}{x^2} f\left(x, \frac{1}{x}\right) + \int_{1/x}^{y} f_1(x, t)\,dt$.

$\frac{\partial^2 u}{\partial y^2} = \frac{1}{y^2} f\left(\frac{1}{y}, y\right) + \int_{1/y}^{x} f_2(s, y)\,ds$.

§18.6 Pages 556–557

1. Yes. **2.** Yes. **3.** No.
4. No. Use the fact that f is uniformly continuous.

§18.9 Page 565

1. 3 in each case. 2. $-\frac{1}{2}, \frac{7}{2}$, 3 respectively.
3. 5, −15, 35 respectively. 4. (a) $2e^{-1}-1$; (b) $\frac{1}{2}(e^{-2}-1)$.
5. (a) $-\frac{1}{2}\log 2$; (b) $\frac{7}{288}\pi^2$.
6. (a) 0, 1, 1, 3 respectively; (b) $4\sqrt{7}-\sqrt{2}-3$.
7. (a) $w(0)=0$, $w(x)=1+\frac{1}{2}x$ if $0<x<5$, $w(5)=\frac{9}{2}$, $w(x)=2+\frac{1}{2}x$ if $5<x<10$, $w(10)=8$; (b) 40 and $\frac{275}{3}$ respectively.
8. (a) Second moment about y-axis is $\int_a^b x^2\,dA(x)$; (b) $\bar{y}\int_a^b dA(x)=\frac{1}{2}\int_a^b \phi(x)\,dA(x)$.
9. $\bar{x}\int_a^b dV(x)=\int_a^b x\,dV(x)$.

CHAPTER 19

§19 Pages 568–569

1. (b) $|x|<1, \frac{8x}{1-x^2}$; (d) $|x|>1, \frac{10}{x-1}$; (f) $x>0, \frac{1+x}{x}$;
(h) $\frac{1}{e}<x<e, \frac{\log x}{1-\log x}$.
4. (a), (b), (d) are divergent; (c) is convergent. 5. Divergent.

§19.2 Pages 576–577

1. (a) and (b) divergent; (c) and (d) convergent.
5. (b), (c), and (f) divergent; (a), (d), (e), and (g) convergent.

§19.21 Page 579

1. (a), (d), and (g) divergent; (b), (c), (e), (f), and (h) convergent.
3. $p>1$. 5. $p>1$ and any q. Also $p=1$ and $q<-1$.

§19.22 Page 581

1. (c) and (e) divergent; (a), (b), (d), (f), and (g) convergent.
3. (a) and (d).

§19.3 Page 585

3. (a) and (d).

§19.32 Pages 589–590

2. (b) and (d) divergent; (a) and (c) convergent.

§19.4 Pages 595–596

1. (b) Convergent if $|x|\leqq 1$, divergent if $|x|>1$; (d) convergent if $|x|<\frac{5}{2}$, divergent if $|x|\geqq\frac{5}{2}$; (f) convergent if $|x|<2$, divergent if $|x|>2$, indecisive if $|x|=2$; (h) convergent if $|x|<\frac{1}{2}$, divergent if $|x|>\frac{1}{2}$, indecisive if $|x|=\frac{1}{2}$.
2. (b) Convergent if $|x|<1$, divergent if $|x|\geqq 1$; (d) indecisive; (f) convergent if $|x|<\sqrt{\frac{3}{2}}$, divergent if $|x|\geqq\sqrt{\frac{3}{2}}$; (h) convergent if $|x|<\frac{2}{3}$, divergent if $|x|\geqq\frac{2}{3}$.

3. (b) Convergent; (d) divergent; (f) convergent if $p > \frac{2}{3}$, divergent if $p < \frac{2}{3}$, indecisive if $p = \frac{2}{3}$.
4. (a) Divergent; (c) convergent; (e) convergent if $p > 2$, divergent if $p \leqq 2$; (f) divergent if $p = \frac{2}{3}$ (see answer to 3(f)); (g) divergent.
8. (b) The first two series are convergent; the third is divergent.

Miscellaneous Exercises Pages 607–609

1. Convergent if and only if $-1 \leqq x < 1$.
6. (a) Convergent, others divergent.
7. (b) Convergent if and only if $-2 \leqq x < 2$. 8. $x > \frac{1}{2}$.
9. (a) No. Yes. (c) All values. (d) Convergent.
10. (a) Convergent if and only if $|x| \leqq 2$. (b) Convergent if and only if $|x| \leqq 1$. (c) Convergent if and only if $-e^{-1} \leqq x < e^{-1}$.

CHAPTER 20

§20.1 Pages 617–618

1. (a) No. (b) $0 < a < b < \pi$. (e) $ab > 0$.
4. (a) $1 + x^2$. 0. Yes. (c) No. 5. (b) No. No. Yes.

§20.2 Page 620

1. Uniform convergence for the ranges of x as indicated: (a) $|x| \leqq c$ if $0 < c < 1$. (b) All x. (c) All x. (d) $x \geqq c$ if $c > 0$. (e) All x on any finite interval.

§20.3 Page 621

1. (a) Not uniformly convergent on any interval having 0 inside or at one end.
3. The function is continuous at 0.

§20.4 Pages 623–624

4. $\int_0^1 f(x)\,dx = 0$; $\int_0^1 f_n(x)\,dx \to \pi/2$.

§20.5 Page 626

1. All but (a) and (b).

CHAPTER 21

§21.1 Pages 631–632

1. (a) $\frac{1}{4}$. (b) 27. (c) 0. (d) ∞. (e) ∞. (f) 4.
2. ∞ if $p < q$; 0 if $p > q$; p^{-p} if $p = q$.
3. 1. 9. 0.0976. 10. 0.4864. 11. 0.0045. 12. 0.482.

§21.2 Pages 637–639

11. $\dfrac{1}{x}\displaystyle\int_0^x \frac{e^t - 1}{t}\,dt$.

12. $L_q(x) = q!\left[1 + \sum_{n=1}^{q} (-1)^n \frac{q(q-1)\cdots(q-n+1)}{(n!)^2} x^n\right]$.

13. $H_m(x) = (2x)^m - \frac{m(m-1)}{1!}(2x)^{m-2} + \frac{m(m-1)(m-2)(m-3)}{2!}(2x)^{m-4} - \cdots$
(breaking off with the term of exponent 0 or 1 according as m is even or odd).

§21.3 Pages 642–643

1. (a) $1 + \frac{1}{2}x - \frac{1}{24}x^3 + \frac{1}{240}x^5 - \cdots$.
(c) $1 + x + \frac{3}{2}x^2 + \frac{11}{6}x^3 + \cdots$.
(e) $1 - \frac{1}{2}x - \frac{1}{12}x^2 - \frac{1}{24}x^3 - \cdots$.
(f) $-\frac{1}{3}x - \frac{1}{45}x^3 - \frac{2}{945}x^5 - \cdots$.

2. (a) $\frac{e^x}{1-x}$. (b) $\frac{\log(1+x)}{1-x}$. (c) $\frac{x}{(1-x)^3}$.

§21.4 Page 646

3. (b) 0.09. 4. $p > 3$.

§21.5 Pages 649–650

1. (a) 1. (b) e^{-1}. (c) $\frac{1}{3}$. (d) e. (e) ∞. (f) 1.
4. (a) 1. (b) 1. (c) 1. (d) ∞. (e) 1. (f) 1.

CHAPTER 22

§22.1 Pages 659–660

1. (a), (b), (d), convergent; (c), (e), (f) divergent.
2. (a) Convergent. (b) Divergent.
3. (a), (b), (d), (e), (f), (g), (i), (j), (k), (n), (p) convergent; others divergent.
5. $\frac{1}{2}m!$. 7. $m + 1 < n$.

§22.11 Pages 663–664

1. (a) Divergent. (c) Divergent. (e) Convergent.
2. (a), (b), (c), (d), (f), (j), convergent; others divergent.
4. Improper but convergent if $0 < a < 1$.
5. $n = 2, 3$. 6. (b) $0 < x < 1$. 8. $0 < p < 1$.

§22.12 Pages 665–666

1. (b), (c), (d), (e), (g), (i) convergent; others divergent.
2. (d) Divergent; others convergent.
3. (a) $1 < p < 2$. (b) $0 < x < 1$. (c) $p < 1$. (d) $-1 < p < 1$.
(e) $0 < a < 2$. (f) $0 < x$. (g) $0 < x < 1$. (h) $0 < a$.
4. To the left of $x = 1$ and above $x + y = 1$.

§22.2 Pages 669–670

4. (a) $\frac{1}{4}\sqrt{\pi}$. (b) $\frac{3}{8}\sqrt{\pi}$. 6. (a) $\sqrt{\pi/2}$. (b) $\frac{3}{128}\sqrt{\pi}$.

§22.3 Pages 672–673

1. (b) Abs. convergent. (d) Convergent. (f) Abs. convergent.
(h) Abs. convergent. (j) Abs. convergent. (l) Convergent.
(n) Abs. convergent. (p) Convergent. (r) Abs. convergent.
3. Yes, whenever $1 < p < 3$. **4.** Yes, whenever $1 < p < 4$.

§22.4 Pages 676–678

1. (b) Divergent; others convergent. **4.** $p < 4$.

5. (a) 2π. (b) $\dfrac{4\pi}{3-n}$, $(n < 3)$. (c) Convergent.
(d) $\dfrac{4\pi}{3(5-n)}$, $(n < 5)$. (e) Divergent. (f) $\dfrac{8\pi}{105}$.

6. (a) Convergent. (c) $p < \frac{3}{2}$.

§22.41 Pages 681–682

2. No. **5.** The value is $\pi/4$.

§22.5 Pages 687–690

9. $\sqrt{\pi x}$. **11.** $\frac{1}{2}\pi \log(b/a)$.

§22.51 Page 692

3. $\frac{1}{2} + \frac{1}{6}e^{-2t} + \frac{1}{3}e^{t}$.
4. (a) $-\frac{6}{5}\cos 3t + \frac{1}{3}\sin 3t + \frac{1}{5}\cos 2t$. (b) $2\sinh 2t + \sinh t$.
5. (a) $\Gamma(n)(s+1)^{-n}$. (b) $(s^2+2s+2)^{-1}$. (c) $3(s^2-4s+13)^{-1}$.

§22.7 Pages 698–699

5. $-\frac{1}{2} < a$.
7. (a) $\frac{1}{6}$. (b) $\dfrac{2^{11}4!5!}{11!}$. (c) $\dfrac{3\pi}{8}$. (d) $\dfrac{\sqrt{\pi}\Gamma(\frac{1}{4})}{12\Gamma(\frac{3}{4})}$.

Miscellaneous Exercises Pages 705–708

2. Both convergent.
23. By choosing $a = .0545632$, the distance between the upper and lower bounds for $\dfrac{n!}{n^n e^{-n}\sqrt{2\pi n}}$ is reduced about 17% below what it was when a was taken to be $\frac{1}{12}$. (This is the reduction when $n = 10$).

Index

Abel, Niels H., 605
Abel's summability identity, 605
Abel's test, 607
Abel's theorem, 644
Absolute maximum, minimum, 22, 23, 128
Absolute convergence
    of improper integrals, 670, 674
    of infinite series, 582
Absolute value, 3, 74
Acceleration vector, 425
Accumulation point, 515, 520
Affine function, 318
Alembert, d', see d'Alembert
Alternating series, 587–588
Ampère, 101
Analytic function, 650
Approximating sum, 39
Arc, 418
Archimedean law, 78
Arc length, 418–419
Areal density, 377
Area of a surface, 437ff., 478–479
Area under a curve, 39, 536
Asymptotic approximation, 703
Average value of a function, 46, 383
Axiom of continuity, 77
Axis of reals, 80
Azimuth, 413

Ball (open, closed), 321
Basis, 276
Bernoulli numbers, 643
Bessel functions, 635, 637
Beta function, 695–696
Binomial series, 597–599
Binormal, 425
Bolzano, Bernard, 517
Bolzano–Weierstrass theorem, 517, 520–521
Borel, Émile, 523
Boundary point, 119
Bounded function, 86, 125–126, 529
Bounded linear transformation, 324
Bounded sequence, 61
Bounded set, 86, 125, 321, 517
Bound of a linear transformation, 324
Bracket function (definition), 2–3

Calculus, 1–2, 55
Cartesian product, 363
Cauchy, Augustin-Louis, 55
Cauchy's convergence condition, 522, 617
Cauchy's form of remainder, 102, 104, 105
Cauchy's generalized law of the mean, 95
Cauchy–Schwarz inequality, 333
Cauchy's inequality, 118, 275
Cauchy's root test, 594, 595
Center of curvature, 35, 425
Center of gravity (or mass), 377, 409
Centroid, 389
Chain rule, 15, 16, 155–156, 202–203, 338–339
Characteristic value, 370
Characteristic vectors, 370
Class $C^{(1)}$, 354
Closed interval, 20, 515
Closed set, 119, 321, 515, 516
Cluster point, 647
Colatitude, 413
Column matrix, 317
Comparison test
    for improper integrals, 656
    for infinite series, 574
Complement (of a set), 118, 514
Composite function 15, 154
Composition of two functions (notation for), 241, 322
Conditional convergence, 583, 670
Connected set, 205, 533
Constraint, 178
Continuity
    and boundedness, 86, 126, 529
    and extreme values, 89, 126, 529
    and intermediate values, 90, 534
    and sequential limits, 527–528
    definition of, 7, 9, 85, 125, 321–322, 527–528
    of sums, products, and quotients, 85, 125, 532
Continuous differentiability, 354
Continuous variable, 54
Convergence
    of an improper integral, 654, 656, 674, 678–679
    of an infinite series, 567
    of a sequence, 59, 321, 521
Convex set, 205
Co-ordinate-free, 269, 290
Coulomb's law, 401, 409
Cramer's rule, 238
Critical point, 215
Cross product, 280, 288

Curl, 306
Curvature
  center of, 35, 425
  of a plane curve, 34, 35
  of a space curve, 424
  radius of, 425
Curve
  definition, 417
  rectifiable, 419
  sectionally smooth, 418
  simple closed, 418
Curvilinear co-ordinates, 250, 257, 258ff.
Cut number, 77
Cylindrical co-ordinates, 412

d'Alembert, J. le R., 590
Darboux, J. G., 543
Darboux's theorems, 93, 543, 556
Decimal fractions, 79
Dedekind, 78
Definite integral, 38–39, 537
Denumerability, 512
Derivative
  definitions, 12, 338
  notations for, 12, 130–132
  of a function from $\mathbb{R}^n$ to $\mathbb{R}^m$, 338
  of an integral, 47–48, 549
  one-sided, 13
  partial, 130
  second and higher order partial, 130–132
Dieudonné, Jean, 348
Difference quotient, 12
Differentiable function, 13, 145, 146, 150, 337
Differentiability
  and continuity, 15, 19, 147, 338
  and tangent plane, 149–150
  definition of, 13, 145, 146, 150, 197, 336–337
  necessary and sufficient conditions for, 341
  of a function from $\mathbb{R}^n$ to $\mathbb{R}^m$, 337
  of a real function of a real variable, 32, 145
  of a real function of n real variables, 150, 197, 336–337
  of a real function of two real variables, 146
  of a scalar product of vector functions, 367
  sufficient conditions for, 197, 199
Differential form, 469, 505
Differential geometry, 417
Dimensionality
  of a vector space, 311–312
  of $\mathbb{R}^n$, 276
Directional derivative, 297, 343
Dirichlet, P. G. Lejeune, 605
Dirichlet's test, 604–606
Discriminant of a quadratic form, 190, 193
Disjoint sets, 120
Distance
  in a vector space, 319
  in $\mathbb{R}^n$, 275
Distance function, 320
Divergence
  of an improper integral, 654, 656
  of an infinite series, 567
  of a vector field, 301, 491, 492
Divergence theorem, 484, 488
Domain of a function, 2, 243
Dot product, 270, 274
Double integral, 380, 554ff.
Duhamel's principle, 419, 447, 545–546

e (the base of natural logarithms), 64, 66
Electrostatic field, 302, 402
Ellipse of inertia, 400
Ellipsoid of inertia, 412
Elliptic integrals, 420–421
Empty (or void) set, 80, 206, 512
Euclidean norm, 319
Euclidean spaces $\mathbb{R}^2$, $\mathbb{R}^3$, $\mathbb{R}^n$, 268–269, 274–275
Eudoxus, 79
Euler, Leonard, 170
Euler's constant, 579, 608
Euler's theorem, 170, 172
Even function, 638
Evolute, 35
Exact differentials, 469–470, 505–506
Existence problem, 65
Extremal problem with constraints, 177ff., 182ff.
Extremum, 20

Factorial, 667
Field, 73
Finite-dimensionality, 311
Finiteness of a set, 512
First-class variables, 155
First moment, 399
Fluid kinematics, 162–164
Fourier transforms, 655
Function
  affine, 318
  bounded, 86, 125–126, 529
  composite, 15, 154
  continuous, 7, 9, 85, 125, 321–322, 527–528
  continuously differentiable, 354
  definition of, 2, 243
  differentiable, 13, 145, 146, 150, 337

elementary, 16, 37
even, 638
extreme values of 88ff., 126, 138ff., 177ff.
homogeneous, 168–169
implicitly defined, 132, 222–223, 361ff.
integrable, 537, 555
multiple-valued, 2
odd, 638
of class $C^{(1)}$, 354
of several variables, 116ff.
real analytic, 650
single-valued, 2
value of, 3, 243
Functional dependence, 266
Functional notation, 3

Gamma function, 655, 666ff.
Gauss, Karl Friedrich, 593
Gauss's test, 593, 595
Gauss's theorem, 457
Generalized inverse, 369
Generalized solution, 368
General solution, 36, 37
Geometric mean, 187–188
Geometric series, 566
Gibbs, Josiah W., 309
Global theorem, 361
Gradient, 296, 297, 342–343
Gradient field, 506
Gram, J. P., 277
Gramian, 277
Gram–Schmidt process, 277
Gravitation, 294
Gravitational field, 294, 409–410, 675
Greatest lower bound, 81
Green's identities, 492
Green's theorem, 457, 463–464
Gundelfinger's rule, 219–220

Harmonic series, 574, 578
Heine–Borel theorem, 523, 525, 531
Heine, Edward, 523
Helix, 428
Hermite polynomial, 639
Hesse, Otto, 354
Hessian, 353
Hölder's inequality, 188
Homogeneous function, 169
Hospital's rule, 107

Identity transformation, 239–241
Image, 248
Implicit function concept, 132, 222–223
Implicit function theorems, 225, 228, 232, 364, 365
Improper integral, 559, 577, 654, 673–674, 678–679
Inequalities, 74
Infinite-dimensionality, 312
Infinite series
absolutely convergent, 582
alternating, 587–588
conditionally convergent, 583
convergent, 567
definition, 567
divergent, 567
geometric, 566
harmonic, 574, 578
multiplication of, 600–602
Infinity, 55
Inner product, 270
Integers, 72, 73, 76
Integrability
of absolute value of a function, 539
of a continuous function, 539
of a function, 537
of a monotonic function, 540
Integral
as a limit of sums, 67, 543
improper, 559, 577, 654, 673–674, 678–679
iterated, 386, 406, 557
of a derivative, 49, 550
Integral test, 578
Integration, theory of, 38ff., 535ff.
Interior point, 119, 321
Intermediate-value theorem, 90, 534
Intersection of sets, 120, 320
Interval of convergence, 629
In the large, 250, 428
In the small, 250
Invariant, 287, 288
Inverse
of a linear operator, 327
of a transformation, 240, 242–243
of differentiation, 35, 48
Inverse function theorem, 242, 356–357
Inverse function theory, 237ff.
Inversion theorem, 356–357
Invertible operator, 327–328, 330
Irrotational field, 306
Isomorphic vector spaces, 312
Iterated integral, 386, 406, 557

Jacobi, Carl, 175
Jacobian
determinant, 175
identical vanishing of, 264, 266

Jacobian—*contd*
matrix, 341
sign of, 251, 466
Jordan curve, 418

$\mathscr{L}(\mathbb{R}^n)$, 327
$\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, 313
Lagrange, J. L., 101, 182
Lagrange's form of remainder, 100, 104, 105, 106
Lagrange's method, 182ff.
Lagrange's multiplier, 183
Laguerre polynomial, 639
Lamina, 395
Laplace, Pierre Simon de, 303
Laplace's equation, 303
Laplace transform, 655, 690
Laplacian, 303, 492, 494
Law of the mean
extensions of, 95, 99, 102
formulations of, 26, 58, 204, 206
for vector functions, 350–351
Least squares solution, 368
Least upper bound, 80
Leibniz, 2, 55
Leibniz's rule, 552
Level curve, 128
Level surface, 129
Limit
defining a definite integral, 67
from right or left, 5
of a function, 4, 54, 55, 122
of a sequence, 59
of sums, products, quotients, 7, 53, 68
Limit inferior, 647
Limiting processes, 4
Limit point, 515
Limit superior, 647
Linear algebra, 309
Linear dependence, independence, 276
Linear functional, 316
Linear operator, 312, 327
Linear space, 311
Linear transformation
definition of, 312
inverse of, 312, 327
matrix representation of, 314–315
norm of, 325–326
Line integral, 446–447
Localization, 223, 250, 264, 361, 428
Locus, 223
Lower bound, 81
Lower sum, 40, 536
Loxodrome, 436

Maclaurin's series, 652
Mapping, 247ff.
Mathematical induction, 75, 77
Matrix product, 313–314
Matrix representation, 315
Maxima and minima, 20ff., 138ff., 177ff., 182ff.
Mean-value theorem
for derivatives, 26
for integrals, 45, 105, 383
Mesh fineness, 542
Method of steepest descent, 344ff., 370–371
Metric, 320
Minkowski's inequality, 188
Moebius band, 503–504
Moment
first, 399
of inertia, 397
second, 398
Monotonic function, 540
Monotonic sequence, 60, 61

Neighborhood, 20, 118, 120, 321, 514, 520
Nested intervals, 83
Newton, 2, 55
Newtonian potential, 675
Newton–Raphson method, 347
Newton's law of gravitation, 409
Newton's method, 347, 349, 371
Nondegenerate critical points, 215, 218
Nonsingular operator, 327–328
Norm, 271, 275, 318, 325–326
Normal derivative, 456
Normal to a surface, 136–137
Normal vector, see principal normal
Normalized vectors, 273
Numbers
complex, 72
irrational, 72
natural, 72, 75–77
pure imaginary, 72
rational, 72, 78
real, 72

Oblique co-ordinates, 257
Odd function, 638
One-to-one correspondence, 248, 250, 512
One-sided surface, 504
Open interval, 20
Open set, 118, 321, 514, 520
Ordered field, 74
Orientable surface, 503–504
Orientation
of a curve, 446
of a surface, 500, 503–504

Orthogonality, 137
Orthogonal vectors, 273, 275
Orthonormal triad (or set), 273, 275
Osculating plane, 425
Osgood, W. F., 546
Outer content, 555
Outer normal, 484

Pappus, 440, 460
Parallelogram law, 271
Parametric representation
  of a curve, 417
  of a surface, 429
Parametric surface, 429
Partial derivatives, 130ff.
Partial differentiation
  changes in order of, 199ff.
  technique of, 130ff., 132ff., 154ff., 164ff.
Partial sum, 567
Particular solution, 36
Partition, 536
Partition weighting, 545
Pi (π), computation of, 573
Point function, 291, 293, 445
Point set, 118, 512, 514, 520 (see also Set)
Polar co-ordinates, 255–256, 390–392
Polygonal arc, 205
Polynomial of degree n, 10–11
Potential, 403, 410, 675
Power series, 627
Preimage, 248
Principal axes of inertia, 400, 412
Principal normal, 424
Product of inertia, 399, 409, 411
Product of two transformations, 253
Pythagoras, 79

Quadratic forms, 189, 193, 218

$\mathbb{R}^n$, see Euclidean space $\mathbb{R}^n$
Raabe's test, 591–592, 595, 596
Radical sign (meaning), 14
Radius of convergence, 629, 648
Radius of curvature, 425
Radius of gyration, 397
Range of a function, 2, 243
Rational fractions, 72
Rational function, 11
Ratio test, 580, 590
Real number scale, 80
Real number system, 72, 78
Rearrangement of a series, 586–587
Rectifiable curve, 419
Region
  convex, 508
  definition of, 120
  regular, 451
  Riemann, 556
  simply connected, 475
  x-simple or y-simple, 458
  xy-simple, 485
  xyz-simple, 487
  xyz-standard, 487
Relative extreme
  necessary conditions for, 139
  sufficient conditions for, 211–213, 215, 217
Relative maximum or minimum, 20, 138
Remainder, see Taylor's formula with remainder
Residual, 368
Resultant of two transformations, 253
Riemann, Bernhard, 535
Riemann integral, 535
Riemann region, 556
Rigid motion, 283
Rolle, Michel, 27
Rolle's theorem, 27
Rotation of axes, 189–192, 255, 284
Row matrix, 317

Saddle point, 139, 215
Scalar, 269
Scalar field, 295
Scalar point function, 291
Scalar product, 270
Schlömilch's form of remainder, 104
Schmidt, E., 277
Schwarz, H. A., 201
Schwarz inequality, 332–333
Schwarz's theorem, 201, 202
Second-class variables, 155
Second moment, 398
Sectional smoothness, 418
Separated sets, 533
Sequence
  bounded above, below, 61
  convergent, 62, 521
  decreasing, 61
  definition of, 58–59
  defining e, 64, 66
  nondecreasing, 60
  strictly increasing, 60
Serret–Frenet formulas, 426–427
Set
  boundary of, 119
  bounded, 86, 125, 321, 517
  closed, 119, 321, 515, 516

Set—*contd*
complement of, 118, 514
concept of, 80, 118, 512, 514
connected, 205, 533
convex, 205
denumerable, 512–513
finite, 512
infinite, 512
interior of, 119
notation, 320
open, 118, 321, 514
Simple closed curve, 418
Simple-connectedness, 475
Simple surface element, 429
Singular point, 257, 259
Smooth curve, 379, 418
Solenoidal field, 302
Solid angle, 489
Solution set, 363
Space of n-dimensions, 243
Sphere (in a vector space), 321
Spherical co-ordinates, 261, 413–414
Standard basis for $\mathbb{R}^n$, 276
Steepest descent, 344–347, 370–371
Stieltjes integral, 560ff.
Stieltjes, T. J., 560
Stirling's formula, 649, 699
Stokes's theorem, 499, 504
Surface
area of a, 437–440, 442–443, 478–480
closed, 430
concept of a, 428–430
parametric, 429
smooth, 430
Surface element, 429
Surface integral, 480

Tangent plane, 135–136, 149
Tangent vector, 421
Taylor, Brook, 101
Taylor's formula
for functions of several variables, 207–210
with derivative remainder, 99, 102
with integral remainder, 99
Taylor's series
for functions of one variable, 570, 634–635, 651
for functions of several variables, 208
Tensor, 436
Tensor analysis, 158
Toroidal co-ordinates, 263
Torsion, 426
Torus, 428, 434
Transformation, 237, 243, 255, 312, 335
Translation (of space), 271
Transpose (of a matrix), 315, 332
Triangle inequality, 275, 318, 319, 320
Triple integral, 405

u-curve, 250, 435
Uniform continuity, 324, 529–531
Uniform convergence, 610, 617, 620, 621, 624, 683
Union of sets, 120, 320
Unit sphere (in a vector space), 319
Unit vector, 319
Upper bound, 80
Upper sum, 40, 536

Variable,
dependent, 3
first or second class, 155
independent, 2
limits of integration, 46ff.
of integration, 46
v-curve, 250, 435
Vector field, 294
Vector point function, 293
Vector product, 280
Vectors
algebra of, 270, 310
concept of, 269, 309–310
Vector space
abstract, 311
$\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, 313
$\mathbb{R}^2$, $\mathbb{R}^3$, 268–269
$\mathbb{R}^n$, 274
Vector velocity, 423
Void (or empty) set, 80, 206, 512
Volume under a surface, 378

Wallis's formula, 697
Weierstrass, Karl, 517
Weierstrass M-test, 619
Work, 453

x-simple (or y-simple) region, 458
$[x]$, the largest integer function, 2–3
xy-simple region, 485
xyz-simple region, 487
xyz-standard region, 487